中国医学伦理学
Chinese Medical Ethics
ISSN 1001-8565,CN 61-1203/R

《中国医学伦理学》网络首发论文

题目：卫生领域人工智能的伦理与治理：多模态大模型指南

作者：王玥，宋雅鑫，王艺霏，于莲，王晶

网络首发日期：2024-03-06

引用格式：王玥，宋雅鑫，王艺霏，于莲，王晶．卫生领域人工智能的伦理与治理：多模态大模型指南[J/OL]．中国医学伦理学.
https://link.cnki.net/urlid/61.1203.R.20240304.1833.002

卫生领域人工智能的伦理与治理：多模态大模型指南*

王玥[1]，宋雅鑫[1]，王艺霏[1]，译；于莲[2]，王晶[3]，审校

（1 西安交通大学法学院，陕西 西安 710049；2 西安交通大学公共卫生学院，陕西 西安 710061；3 首都医科大学附属北京中医医院，北京 100010）

* Ethics and governance of artificial intelligence for health. Guidance on large multi-modal models. Geneva: World Health Organization; 2024. Licence: CC BY-NC-SA 3.0 IGO.

本译文并非源自世界卫生组织（WHO），世卫组织对译文的内容或准确性概不负责。英文原版应作为具有约束力的作准文本。

原文版本号： ISBN 978-92-4-008475-9（电子版） ISBN 978-92-4-008476-6（印刷版）

摘要

人工智能（Artificial Intelligence，AI）是指集成到系统和工具中的算法从数据中学习的能力，这样它们就能执行自动化的任务，而无需人工对每个步骤进行明确的编程。生成式人工智能是算法在可用于生成新内容（如文本、图像或视频）的数据集上进行训练的一种人工智能技术。本指南针对其中一种类型的生成式人工智能，即多模态大模型（Large Multi-modal Model，简称“LMM”）。这种模型可以接受一种或多种类型的数据输入，并产生不局限于输入算法的数据类型的多种输出。据预测，多模态大模型将广泛应用于医疗保健、科学研究、公共卫生和药物开发等领域。多模态大模型也被称为“通用基础模型”（General-purpose Foundation Model），尽管尚未证实多模态大模型能否完成各种任务和目的。

多模态大模型的普及速度超过了历史上任何消费者应用。它们之所以引人注目，是因为其促进了人机交互，可以模仿人类交流，并对查询或数据输入作出类似人类且看似权威的回应。随着消费者的快速采用和接受，并考虑到其颠覆核心社会服务和经济部门的潜力，许多大型科技企业、初创企业和政府都在投资并竞相引导生成式人工智能的发展。

2021 年，世界卫生组织（WHO，以下简称“世卫组织”）发布了《卫生领域人工智能的伦理与治理》的综合指南[1]。世卫组织咨询了 20 位人工智能领域的顶尖专家，他们确定了在卫生领域使用人工智能的潜在益处和潜在风险，并发布了以协商方式达成一致的六项原则，供正在使用人工智能的政府、开发者和提供者在制定政策和实践时考虑。这些原则应指导包括政府、公共机构、研究者、企业和实施者在内的广泛利益相关者在卫生领域开发和部署人工智能。这六项原则分别是（1）保护人类的自主性；（2）增进进人类福祉、安全和公共利益；（3）确保透明、可以解释和可以理解；（4）培养责任感和实行问责制；（5）确保包容性和公平；（6）推广反应迅速且可持续的人工智能（图 1）。

图1：世卫组织就卫生领域人工智能的伦理原则达成共识

世卫组织发布本指南的目的是协助成员国规划与卫生领域多模态大模型有关的益处和

挑战，并为适当开发、提供和使用多模态大模型提供政策和实践方面的指导。本指南提供了与指导原则相一致的企业内部、政府和国际合作的治理建议。本指南的基础是考虑到人类使用卫生领域生成式人工智能独特方式的指导原则和治理建议。

多模态大模型的应用、挑战和风险

多模态大模型在卫生领域的潜在应用与其他形式的人工智能类似，然而多模态大模型的接入和使用方式是全新的，它既有新的益处，也有社会、卫生系统和终端用户尚未做好应对准备的新风险。表 1 总结了多模态大模型的主要应用及其潜在益处和风险。

表 1　卫生领域多模态大模型的各种用途的潜在益处和风险

使用场景	潜在或可能的益处	潜在的风险
诊断和临床保健	● 协助管理复杂病例和审查常规诊断 ● 减轻医疗服务提供者的通信工作量（“键盘解放”） ● 从各种非结构化的健康数据中提供新的见解和报告	● 不准确、不完整或错误的回答 ● 质量不合格的训练数据 ● 偏见（训练数据和回答的偏见） ● 自动化偏见 ● 技能退化（医护者） ● 知情同意（患者）
指导患者使用	● 生成信息以提高对医疗状况的理解（作为患者或护理者） ● 虚拟的医疗助理 ● 临床试验注册	● 不准确、不完整或虚假陈述 ● 操纵 ● 隐私 ● 减少临床医生与患者之间的互动 ● 认知上的不公正 ● 在卫生系统外提供医疗服务的风险
文书和行政任务	● 协助处理临床护理所需的文书和文件工作 ● 协助语言翻译 ● 完成电子病历 ● 在患者就诊后起草临床笔记	● 不准确和错误 ● 根据提示做出不一致的回答
医疗和护理教育	● 适合每个学生需要的动态课本 ● 模拟对话，促进沟通并练习在不同情况下与不同患者交流 ● 在回答问题时进行连环推理	● 助长自动化偏见 ● 错误或虚假信息有损医学教育质量 ● 学习数字技能的新负担
科学研究与药物开发	● 从科学数据和研究中生成见解 ● 生成用于科学论文、手稿提交或同行评议的文本 ● 分析和总结研究数据 ● 校对 ● 新药研发	● 无法让算法对内容负责 ● 高收入国家视角下的算法编码偏见 ● 生成不存在的信息和/或参考文献 ● 破坏科学研究的关键原则，如同行评议 ● 加剧科学知识获取的差异

与使用多模态大模型相关的系统性风险包括下列可能影响医疗卫生系统的风险（表 2）。

表 2　在卫生领域使用多模态大模型对医疗卫生系统造成的风险

风险类型	描述
高估多模态大模型的效益	可能存在一种“技术解决主义”的倾向，或者过高估计多模态大模型的益处，而忽视或淡化其使用过程中的挑战，包括其安全性、有效性和实用性方面。
可及性和可负担性	由于“数字鸿沟”和接入多模态大模型的订阅费用等原因，可能无法公平地接入多模态大模型。
全系统偏见	使用不断扩大的数据集可能会增加多模态大模型编码中的偏见，这些偏见可能会自动贯穿整个医疗卫生系统。
对劳动力的影响	在一些国家，使用多模态大模型可能会导致失业，医疗工作者需要接受再培训并适应多模态大模型的使用。数据标注和过滤可能导致低工资和无法治疗的心理压力。
医疗卫生系统对不合适的多模态大模型的依赖	如果不对多模态大模型进行维护或（在低收入和中等收入国家）只为在高收入国家使用而更新多模态大模型，那么对多模态大模型的依赖会使卫生系统变得脆弱。此外，缺乏对隐私和保密性的保护，可能会削弱人们对卫生系统的信任。
网络安全风险	恶意攻击或黑客行为可能会破坏在卫生领域使用多模态大模型的安全性和信任。

使用多模态大模型还可能会带来更广泛的监管和系统性风险。一个值得关注的问题是（一些数据保护机构正在研究），多模态大模型是否符合包括国际人权义务、国家数据保护法规在内的现有的法律或监管制度。因为多模态大模型训练数据的收集方式、对已收集数据（或由终端用户输入的数据）的管理和处理、多模态大模型开发者的透明度和责任分配问题以及多模态大模型出现“幻觉”的可能性，算法可能不适用于现行法律。多模态大模型还可能违反消费者保护法。

随着不断增长的多模态大模型的使用，开发多模态大模型需要更加海量的算力、数据、人力和财政资源，一项与之相关的更广泛的社会风险（包括在卫生领域中使用此类算法）是多模态大模型普遍由大型科技企业开发和部署这一事实。相对于较小企业和政府而言，这可能会加强这些科技巨头在开发和使用人工智能方面的主导地位，包括引导公共和私营部门的人工智能研究重点。对大型科技企业潜在主导地位的其他担忧还在于企业对伦理和透明度的承诺不足。企业之间以及企业与政府之间的新的自愿承诺，可以在短期内降低一些风险，但不能替代最终可能实施的政府监督。

另一个社会风险是多模态大模型的碳足迹和水足迹。与其他形式的人工智能一样，多模态大模型需要大量能源，并产生不断增加的水足迹。虽然多模态大模型和其他形式的人工智能可以带来重要的社会效益，但不断增长的碳排放可能会成为气候变化的主要因素，而不断增加的耗水量则会对水资源紧张的社区产生进一步的负面影响。与多模态大模型的出现相关

的另一个社会风险是，多模态大模型尽管提供的是似是而非的回应却逐渐被视为知识来源，这最终可能会削弱人类知识的权威，包括在医疗保健、科学和医学研究领域。

卫生保健与药品领域多模态大模型的伦理和管理

多模态大模型可被视为一个或多个行为者关于编程和产品开发方面作出的一系列（或一连串）决策的产物（图 2）。在人工智能价值链的每个阶段做出的决定都可能对下游参与开发、部署和使用多模态大模型的主体产生直接或间接的影响。政府可以通过在国家、地区和全球范围内颁布和执行法律和政策来影响和规范这些决策。

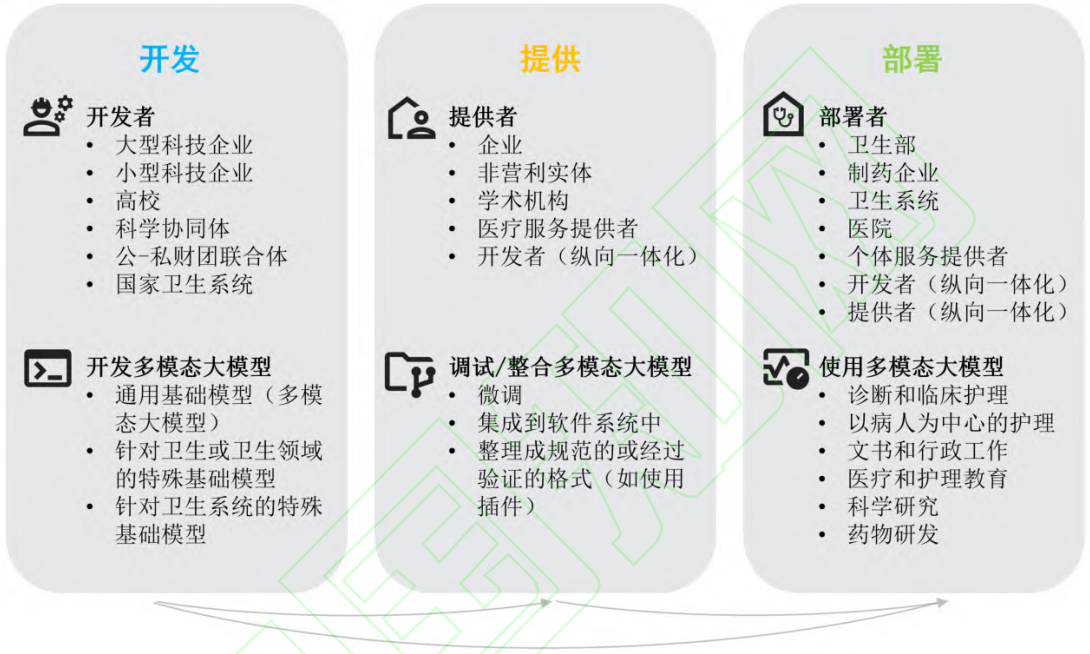

图2：开发、提供和部署多模态大模型的价值链

人工智能价值链通常始于一家大型科技企业，在本指南中称为“开发者”。开发者也可以是高校、较小的科技企业、国家卫生系统、公-私联合体或其他拥有资源和能力使用若干投入的实体。这些投入组成了“人工智能基础设施”（政府在立法和监管中用来描述多模态大模型的术语），如用来开发通用基础模型的数据、算力和人工智能专业技能。这些模型可直接用于执行各种通常意想不到的任务，包括与医疗保健相关的任务。有几种通用基础模型是专门为卫生保健与药品领域的使用而训练的。

第三方（“提供者”）可以通过主动编程接口，将通用型基础模型用于特定目的或用途。这包括：（i）对新的多模态大模型进行微调，这可能需要对基础模型进行额外的训练；（ii）将多模态大模型集成到应用程序或更大的软件系统中，为用户提供服务；或（iii）集成被称为“插件”的组件，以正规或规范格式引导、过滤和组织多模态大模型，生成“可消化”的结果。[1]

[1] 世卫组织卫生领域人工智能的伦理和治理专家 Leong Tze-Yun 的来文。

此后，提供者可向客户（或“部署者”）出售基于多模态大模型的产品或服务，如卫生部门、医疗卫生系统、医院、制药企业甚至个人，如医疗服务提供者。购买或获得许可使用产品或应用程序的客户可以直接将其用于患者、医疗服务提供者、卫生系统的其他实体、非专业人士或自身业务。价值链可以是“纵向一体化”的，因此，收集数据并训练通用基础模型的企业（或其他实体，如国家卫生系统）可以针对特定用途修改多模态大模型，并直接向用户提供应用程序。

治理是通过现行法律和政策，新制定或修订的法律、准则、内部行为规则和开发者程序，以及国际协定和框架，来体现伦理原则和人权义务的一种手段。

构建多模态大模型治理框架的一种方法是将其纳入人工智能价值链的三个阶段：（i）设计和开发通用基础模型或多模态大模型；（ii）提供基于通用基础模型的服务、应用程序或产品；以及（iii）部署医疗服务或应用程序。在本指南中，从三个方面对每个阶段进行审查：

1. 在价值链的每个阶段应该应对（如上所述的）哪些风险？哪些行为者最适合应对这些风险？
2. 为了应对风险，相关行为者可以做些什么？必须坚持哪些伦理原则？
3. 政府的作用是什么，包括相关法律、政策和法规？

某些风险可以在人工智能价值链的各个阶段加以解决，而某些行为者可能在降低各种风险和维护伦理价值观方面发挥更重要的作用。虽然在开发者、提供者和部署者之间的责任归属问题上可能会存在分歧和紧张关系，但在一些明确的领域中，行为者各自都处于最有利的应对位置，或者是唯一有能力应对潜在或实际风险的实体。

通用基础模型（多模态大模型）的设计与开发

在设计和开发通用基础模型的过程中，责任在于开发者。政府有责任制定法律和标准，要求采取或禁止某些做法。本指南第 4 章提供了一些建议，以帮助在开发多模态大模型过程中应对风险并实现效益最大化。

通用基础模型（多模态大模型）的提供

在提供服务或应用的过程中，政府有责任界定对开发者和提供者的要求与义务，以应对与医疗环境中使用的基于人工智能的系统相关的特定风险。本指南第 5 章提供了一些建议，以便在使用多模态大模型为医疗保健提供服务和应用时，应对风险并实现效益最大化。

通用基础模型（多模态大模型）的部署

即使在开发和提供多模态大模型的过程中适用了相关的法律、政策和伦理实践，在使用过程中也会出现风险，部分原因是多模态大模型的不可预测性及其提供的响应，用户可能以开发者和提供者都没有预料到的方式应用通用基础模型，并且多模态大模型的输出可能随着时间的推移而改变。本指南第 6 章就使用多模态大模型和应用过程中应解决的风险和挑战提出了建议。

通用基础模型（多模态大模型）的责任

随着多模态大模型在卫生保健与药品领域的广泛使用，出现错误、误用并最终对个人造成伤害在所难免。因此，问责制可以确保受到多模态大模型伤害的用户得到充分赔偿或其他形式的补救，以减轻受到伤害的用户的举证责任，确保他们得到充分和公平的赔偿。

政府可以通过引入因果关系推定来做到这一点。政府也可以考虑引入严格责任标准，以处理因部署多模态大模型造成的伤害。虽然严格的问责制可以确保对受到伤害的人进行赔偿，但同时也可能阻碍对日益复杂的多模态大模型的使用。政府也可以考虑设立无过错、无责任的赔偿基金。

通用基础模型（多模态大模型）的国际治理

各国政府必须共同努力，建立新的体制结构和规则，确保国际治理跟上技术全球化的步伐。政府还应确保加强联合国系统内的合作与协作，以应对在卫生领域和社会、经济领域内更广泛部署人工智能应用的机遇和挑战。

为了确保各国政府对其在开发和部署基于人工智能的系统方面的投资和参与负责，并确保各国政府出台维护伦理原则、人权和国际法的适当法规，进行国际治理十分必要。国际治理还能确保企业开发和部署的多模态大模型符合适当的国际安全和效率标准，并遵守伦理原则和人权义务。各国政府还应避免出台对企业或政府本身具有竞争优势或劣势的法规。

为了赋予国际治理意义，这些规则必须由所有国家共同制定，而不仅仅是由高收入国家（以及与高收入国家政府合作的科技企业）制定。正如联合国秘书长在 2019 年所提出的，人工智能的国际治理可能需要所有利益相关者通过网络化多边主义进行合作，这将使联合国大家庭、国际金融机构、区域组织、贸易集团和包括民间团体、城市、企业、地方当局和青年在内的其他方面更加密切、有效和包容地合作。

卫生领域人工智能的伦理与治理：多模态大模型

1 简介

本指南涉及多模态大模型在卫生领域相关应用中的新兴用途。[2]它包括在卫生保健与药品领域使用多模态大模型的潜在益处和风险，以及最能确保遵守伦理、人权和安全准则和义务的多模态大模型治理方法。本指南以世卫组织 2021 年 6 月发布的指南《卫生领域人工智能的伦理与治理》（Ethics and governance of artificial intelligence for health）为基础[1]。《卫生领域人工智能的伦理与治理》探讨了卫生领域人工智能的伦理挑战和风险，为了确保给所有在卫生领域利用人工智能的国家带来公共利益而确定了六项原则，并为了最大限度地实现该技术的前景而提出了加强卫生领域人工智能治理的建议。

人工智能指的是集成到系统和工具中的算法从数据中学习以执行自动任务的能力，每个步骤都无需人工明确编程。生成式人工智能是一种人工智能技术，其中机器学习模型用于在数据集上训练算法以生成新的输出，如文本、图像、视频和音乐。生成式人工智能模型在训练数据的过程中学习模式和结构，从而根据所学模式预测并生成新数据。生成式人工智能模型可以通过人类的反馈进行强化学习，实现改进，即训练人员对生成式人工智能模型提供的回应进行排序，以训练算法给出人类认为价值最大的回应。生成式人工智能可应用于设计、内容生成、模拟和科学发现等各个领域。

大语言模型（Large Language Model）是一种特殊的生成式人工智能，它接收文本类型的输入并提供同样类型的文本的回应，因此备受关注。大语言模型是大型单模态模型的典范，也是集成这些模型的聊天机器人早期版本的运行基础。虽然大语言模型参与了对话，但模型本身并不清楚自己在生成什么。它们只是根据前面的单词、学习到的模式或单词组合从而对下一个单词进行预测[2]。

本指南探讨了多模态大模型（包括大语言模型）日益广泛的用途，这些模型在卫生保健与药品领域的应用是通过高度多样化的数据集进行训练的，这些数据集不仅包括文本，还包括生物传感器、基因组、表观基因组、蛋白质组、成像、临床、社会和环境数据[3]。因此，多模态大模型可以接受多种类型的输入，并产生不局限于输入数据类型的输出。多模态大模型可广泛应用于医疗保健和药物开发中。

多模态大模型与以往的人工智能和机器学习不同。虽然人工智能已被广泛集成到许多消费者应用中，但大多数算法的输出既不要求也不邀请客户或用户参与，除了集成到社交媒体平台中的人工智能初级形态，这些平台通过策划用户生成的内容来吸引眼球[4]。多模态大模型与其他类型的人工智能的另一个区别在于其多功能性。以前和现有的人工智能模型，包括用于医疗用途的模型，都是针对特定任务而设计的，因此缺乏灵活性。它们只能执行训练集及其标签[5]中定义的任务，如果不使用不同的数据集进行再训练，就无法适应或执行其他功能。因此，尽管美国食品药品监督管理局已经批准了 500 多个用于临床医学的人工智能模型[5]，但大多数模型仅被批准用于一到两个范围较窄的任务。与此相反，多模态大模型在不同的数据集上经过训练，可用于多种任务，包括一些没有经过明确训练的任务[5]。

多模态大模型通常有一个便于人机交互的界面和格式，可以模仿人与人之间的交流，从而引导用户给算法注入类似人类的品质。因此，与其他形式的人工智能不同，多模态大模型的使用方式及其生成和提供的回应内容看似“与人类一样”，这也是多模态大模型被公众空前采用的原因之一。此外，由于它们提供的回应似乎具有权威性，即使多模态大模型无法保证回应的正确性，无法将伦理规范或道德推理融入其生成的回应中，许多用户也仍然不加批判地将其视为正确的。多模态大模型已被用于教育、金融、通信和计算机科学等众多领域，

[2] 在本指南中，“大型多模态模型”（LMM）和“通用基础模型”这两个术语可以互换使用，后一个术语尤其在讨论治理问题时使用。不过，大型多模态模型是否能完成一般用途的各种任务，目前还不得而知。

而本指南说明了多模态大模型在卫生保健与药品领域中使用（或设想中使用）的不同方式。

多模态大模型可被视为一个或多个行为者在编程和产品开发方面的一系列（或一连串）决策的产物。在人工智能价值链的每个阶段做出的决策，都可能对下游参与多模态大模型开发、部署和使用产生直接或间接的影响。这些决策可能受到在国家、地区和全球范围内颁布和执行法律政策的政府的影响和监管。

人工智能价值链通常始于开发模型的大型科技企业。开发者也可以是高校、较小的科技企业、国家卫生系统、公-私联合体或其他拥有资源和能力使用若干投入的实体。这些投入组成了“人工智能基础设施”，如用来开发通用基础模型的数据、算力和人工智能专业技能。这些模型可直接用于执行各种通常意想不到的任务，包括与医疗保健相关的任务。有几种通用基础模型是专门为卫生保健与药品领域的使用而训练的。

第三方（“提供者”）可以通过主动编程接口，将通用型基础模型用于特定目的或用途。这包括：（i）对新的多模态大模型进行微调，这可能需要对基础模型进行额外的训练；（ii）将多模态大模型集成到应用程序或更大的软件系统中，为用户提供服务；或（iii）集成被称为“插件”的组件，以正规或规范格式引导、过滤和组织多模态大模型，生成“可消化”的结果。[3]

此后，提供者可向客户（或“部署者”）出售基于多模态大模型的产品或服务，如卫生部门、医疗卫生系统、医院、制药企业甚至个人，如医疗服务提供者。购买或获得许可使用产品或应用程序的客户可以直接将其用于患者、医疗服务提供者、卫生系统的其他实体、非专业人士或自身业务。价值链可以是“纵向一体化”的，因此，收集数据并训练通用基础模型的企业（或其他实体，如国家卫生系统）可以针对特定用途修改多模态大模型，并直接向用户提供应用程序。

世卫组织认识到，人工智能可为医疗卫生系统带来巨大惠益，包括改善公共卫生和实现全民健康覆盖。然而，正如世卫组织《卫生领域人工智能的伦理与治理》指南[1]所述，人工智能会带来重大风险，既可能损害公共卫生，也可能危及个人尊严、隐私和人权。尽管多模态大模型相对较新，但其被接受和传播速度已促使世卫组织提供该指南，以确保它们有可能在全球范围内获得成功和可持续的使用。世卫组织认识到，在发布该指南时，人们对人工智能的潜在益处和风险、设计和使用人工智能应适用的伦理原则以及治理和监管的方法存在许多相互较量的观点。由于该指南是在多模态大模型首次应用于卫生领域后不久、在更强大的模型陆续发布之前发布的，世卫组织将更新该指南，以适应技术的快速发展、社会对其使用的处理方式以及在卫生保健与药品领域之外使用多模态大模型对医疗健康造成的影响。

1.1 通用基础模型（多模态大模型）的重要性

虽然多模态大模型相对较新且未经测试，但已在包括医疗和药品在内的各个领域对社会产生了巨大影响。ChatGPT 是一种大语言模型，由一家美国科技企业连续发布了多个版本。据估计，在 2023 年 1 月，即推出仅 2 个月后，该模型的月活跃用户数就达到 1 亿人。这使其一时之间成为历史上增长最快的消费者应用程序[6]。

目前，许多企业都在开发多模态大模型或将多模态大模型集成到消费者应用中，如互联网搜索引擎。大型科技企业也正在迅速将多模态大模型集成到大多数应用软件中或开发新的应用软件[7] [8]。在数百万美元私人投资的支持下，初创企业也正在竞相开发多模态大模型[9]。由于开源平台的可用性，其开发的多模态大模型比巨头企业开发的多模态大模型更快、更便宜[10]。

多模态大模型的出现促进了技术领域的新投资和新产品的不断推出，但是一些企业也承认他们并不完全清楚多模态大模型为何会生成某些回应[11]。尽管根据人类反馈进行了强化学

[3] 世卫组织卫生领域人工智能的伦理和治理专家 Leong Tze-Yun 的来文。

习，但多模态大模型生成的内容依然不总是具有可预测性和可控性，可能会在参与“对话”时生成让用户感到不舒服[12]，或者生成错误但极易令人信服的内容[13]。即便如此，对多模态大模型的支持大多不仅仅是对其功能的热衷，还包括在未经同行评议的出版物中对多模态大模型性能的无条件的、不加批判的主张[14]。

用于训练多模态大模型的数据集尚未公开[15]，但多模态大模型已被迅速采用，因此很难或不可能知道这些数据是否有偏见，是否合法获取并符合数据保护规则和原则，以及能够进行任务或查询的执行是否反映了它已就相同或类似的问题接受过训练、已获得解决问题的能力。其他有关用于训练多模态大模型数据的问题，如是否符合数据保护法，将在下文讨论。

个人和政府都没有为发布多模态大模型做好准备。个人没有接受过有效使用多模态大模型的培训就可能不会明白，即使多模态大模型聊天机器人给人留下了准确可靠的印象，其回应也并不总是准确或可靠的。一项研究发现，大语言模型 GPT-3 虽然“与人类相比……能生成更容易理解的准确信息”，但也能生成“更有说服力的虚假信息”，而且人类无法区分出多模态大模型生成的内容和人类生成的内容[16]。

各国政府也基本上没有做好准备。为治理人工智能的使用而制定的法律法规可能无法应对与多模态大模型相关的挑战或机遇。欧盟已就颁布一项全欧盟范围适用的《人工智能法案》达成协议，但考虑到多模态大模型[17]，不得不在起草的最后阶段修改其立法框架。其他国家的政府也在迅速制定新的法律或法规[18]，或颁布临时禁令（其中一些已被撤销）[19]。预计未来几个月内，各企业将陆续推出功能和性能更强大的多模态大模型，这可能会带来新的益处，但也会带来新的监管挑战。在这种动态环境中，本指南以包括伦理指南在内的以往指南为基础，为在卫生保健与药品领域使用多模态大模型提出了意见和建议。

1.2 世卫组织关于卫生领域人工智能的伦理与治理的指南

世卫组织关于卫生领域人工智能的伦理和治理的第一版指南[1]审查了机器学习的各种方法和卫生领域人工智能的各种应用，但没有具体审查生成式人工智能或多模态大模型。在制定该指南期间以及在 2021 年发布该指南时，没有证据表明生成式人工智能和多模态大模型将很快得到广泛应用，并应用于临床护理、医疗研究和公共卫生领域。

然而，该指南提出的基本伦理挑战、核心伦理原则和建议（见方框 1）对于评估和有效、安全地使用多模态大模型仍然具有现实意义，尽管在这一新技术方面已经出现并将继续出现更多的治理空白和挑战。这些挑战、原则和建议也是本指南中专家组对多模态大模型采取方法的基础。

方框 1. 世卫组织关于使用卫生领域人工智能的共识伦理原则概述

- 保护人类的自主性：人类应继续掌控卫生保健系统和医疗决策。医疗服务提供者掌握安全有效地使用人工智能系统所需的信息。人们了解人工智能系统在卫生领域发挥的作用。以适当的数据保护法律框架、有效的知情同意来保护数据隐私和保密性。
- 增进人类福祉、安全和公共利益：人工智能的设计者要满足监管部门对明确或暗示的安全、准确和有效地使用的要求。应提供实践中的质量控制措施和随着时间的推移在人工智能使用方面的质量改进措施。如果人工智能会造成精神或身体伤害，而这种伤害是可以通过使用其他做法或方法来避免的，那么就不应该使用人工智能。
- 确保透明、可以解释和可以理解：人工智能技术应该让开发者、医疗专业人员、患者、用户和监管机构都能理解或明白。在设计或部署人工智能之前公布或记录足够

的信息，有助于就如何设计人工智能以及如何使用或不使用人工智能进行有意义的公众咨询和辩论。人工智能可根据被解释者的能力进行解释。

- 培养责任感和实行问责制，以确保人工智能在适当的条件下由经过适当培训的人员使用。患者和临床医生对人工智能的开发和部署进行评估。通过设立人工监督点，在算法的上下游应用监督原则。建立适当机制，对受到人工智能决策的不利影响的个人和群体进行慰问和补救。
- 确保包容性和公平：人工智能的设计和共享是为了鼓励尽可能广泛、适当、公平地使用和获取，不分年龄、性别、性别认同、收入、种族、民族、性取向、能力或其他特征。人工智能不仅可用于高收入国家，也可用于低收入和中等收入国家。人工智能不会产生不利于可识别群体的偏见。人工智能可最大限度地减少不可避免的权力差距。对人工智能进行监测和评估，以确定其对特定人群造成的过度影响。
- 推广反应迅速且可持续的人工智能：人工智能技术是与更广泛地促进医疗卫生系统、环境和工作场所的可持续性发展相符合的。

I. 通用基础模型（多模态大模型）的应用、挑战和风险

2 卫生领域通用基础模型（多模态大模型）的应用和挑战

卫生领域人工智能的应用包括诊断、临床护理、研究、药物开发、医疗管理、公共卫生和监测。多模态大模型的许多应用并不是人工智能的新用途；然而，临床医生、患者、非专业人士以及医疗保健专业人员接入和使用多模态大模型的方式各不相同。本章将讨论多模态大模型在卫生领域的潜在应用以及与使用多模态大模型相关的实际和预期挑战与风险。许多应用和用途仍未得到证实，最终可能无法带来如同宣传效果的益处。

2.1 诊断和临床保健

人工智能已被用于诊断和临床保健，如在放射学和医学成像、结核病和肿瘤学等领域的辅助诊断。人们希望临床医生能在会诊期间利用人工智能整合患者记录、识别高危患者、帮助做出高难度的治疗诊断，并发现临床错误[1]。多模态大模型可以将基于人工智能的系统扩展到整个诊断和临床保健过程中，包括线上会诊和面对面会诊，个别专家预计多模态大模型“对医生的重要性将超过以前的听诊器”[20]。一些多模态大模型已通过美国医师执照考试；然而，通过死记硬背医学知识来完成医学测试并不等于能够提供安全有效的临床服务[21]，多模态大模型未能通过以前未在网上公布材料而儿童可以轻松解决的测试[22]。一项关于大语言模型临床知识的研究认为，“从用于回答医疗问题的大语言模型过渡到医疗服务提供者、管理者和消费者都能使用的工具，需要进行大量的额外研究，以确保该技术的安全性、可靠性、有效性和隐私”[23]。

诊断被认为是一个特别有前景的领域，因为多模态大模型可用于识别复杂病例中的罕见诊断或“异常表现”[24]。医生已经在使用互联网搜索引擎、在线资源和鉴别诊断生成器。多模态大模型将成为又一诊断工具。多模态大模型还可用于常规诊断，为医生提供额外的意见，以确保不忽视明显的诊断。所有这些都可以快速完成，因为多模态大模型可以比医生更快地扫描患者的全部病历[24]。

不过，目前在试点项目中用于支持临床医生的几种流行的多模态大模型并没有接受过电子病历或医疗及其他相关健康数据方面的专门训练，尽管它们的数据集确实包括这些信息。例如，在美国的几个卫生保健系统中，一家科技企业提供的多模态大模型正在进行试点测试，读取患者的信息并起草医生的回应，以缩短医务人员为患者解答的时间。这种做法旨在减少医疗保健专业人员每天处理成千上万条信息的倦怠感，使他们能够专注于临床工作（“键盘解放”）[25]。因此，当收到患者信息时，多模态大模型会根据患者信息和电子病历版本显示回应草稿。然而，人工智能只用于回应需要大量的编辑工作的那些患者提问[25]。不过，美国的一项研究发现，在回答在线论坛上提出的问题时，由 ChatGPT 驱动的聊天机器人比合格的医生表现得更好。在所选的 195 个问题中，近 80%的情况下，独立评估人员更喜欢聊天机器人的回答，而不是医生的回答[26]。聊天机器人还可以帮助回答标准化的“非正式医疗咨询”问题，并在患者初次就诊时提供信息和回应，或总结实验室检验结果[27]。

企业和高校也在开发使用医疗和健康数据或电子病历训练的多模态大模型，其中包括基于小型数据集的多模态大模型。例如，有一个多模态大模型是在大约 3 万份病例报告的数据集上进行训练的，目的是学习医疗状况和症状之间的关系，以辅助诊断[28]。另一个多模态大模型是在一个包含 10 万多张胸部 X 光片的数据集上进行训练的，目的是识别异常情况，并最终提供见解或识别病情[29]。一些已经公开评估过的多模态大模型是在数百万份电子病历和其他来源的专业和一般医学知识基础上用算法训练出来的。这种方法提高了算法处理不同形

式的书面医疗信息并给出回应（“医疗问题解答”）的能力[30]。

几家最大的科技企业正在将其通用多模态大模型调整为可辅助临床诊断和护理的多模态大模型。一家科技企业正在开发 Med-PaLM2，该系统旨在从医学文本中回答问题和总结见解，目前，它正在往合成图像（如 X 光片和乳房 X 光片）、撰写报告并回答后续问题的方向不断发展，以方便临床医生进行更多查询，这一功能可减轻医疗保健专业人员与计算机之间的“同行分歧”[31]。

人类的长期愿景是开发“通用医学人工智能”，使医疗保健专业人员能够与多模态大模型灵活对话，根据定制的、临床医生驱动的查询生成回应。因此，用户可以用日常语言描述所需的内容，让通用医学人工智能模型适应新任务，而无需重新训练多模态大模型或训练多模态大模型接受不同类型的非结构化数据来生成回应[5]。

在诊断和临床保健中使用多模态大模型的风险

多模态大模型在临床治疗中大有可为的同时，也存在着重大的风险，其中一些风险早在多模态大模型出现之前就已存在。在诊断和临床保健中使用多模态大模型有以下五大风险：

- *不准确、不完整、有偏见或错误的回应*：关于多模态大模型的一个担忧是，聊天机器人容易根据多模态大模型“发明”的数据或信息（如参考文献）做出不正确或完全错误的回答[32]，以及因重复训练数据中编码的缺陷而做出有偏见的回答[33]。多模态大模型还可能造成场景偏差（contextual bias），在偏差出现的情况下，对人工智能使用场景的预设会导致其给出针对其他场景的建议[1]。例如，低收入和中等收入国家训练数据和观点的代表性不足。因此，如果要求多模态大模型为低收入国家卫生部门提供指导，总结一种疾病的治疗模式，它可能会照搬只适合高收入国家的方法[34]。此外，多模态大模型可能提供不完整的答案，或未考虑到使用环境中的变化情况的答案，或者根本没有答案。

  虚假反应，俗称“幻觉”，与多模态大模型生成的事实准确反应无法区分，因为即使是通过人类反馈进行强化学习的多模态大模型，其训练目的也不是生成事实，而是产生看似事实的信息。一项研究发现，大语言模型在根据一组简单的事实进行总结时，有3%到 27%的情况下会产生幻觉[35]。目前，多模态大模型还依赖于与其有效沟通的人类“提示工程”[36]。因此，即使多模态大模型经过专门的医疗数据和健康信息训练，也不一定能做出正确的反应。对于某些基于多模态大模型的诊断，可能没有确证测试或其他方法来验证其准确性[24]。在医学和其他公共卫生决策领域，即使多模态大模型在大多数情况下都是正确的，其准确性可能不足以和其开发成本或在卫生保健系统中安全有效的实施成本相匹配。

- *数据质量和数据偏见*：多模态大模型生成有偏见或不准确回应的一个原因是数据质量差。目前可供公众使用的许多多模态大模型都是在互联网等大型数据集上训练出来的，这些数据集可能充斥着错误信息和偏见。大多数医疗和健康数据也存在偏见，无论是种族、民族、血统、性别、性别认同还是年龄。由于大多数数据都是从高收入地区收集的，因此根据健康数据训练的多模态大模型通常会对这些偏见进行编码。例如，基因数据往往是针对欧洲后代收集的[1]。多模态大模型也经常根据电子病历进行训练，而电子病历中充满了错误和不准确的信息[24]，或者依赖于从体检中获得的信息，而这些信息可能是不准确的，从而影响多模态大模型的输出[25]。数据质量和偏见问题会影响包括多模态大模型在内的所有人工智能模型[1]。

  一家科技企业在其多模态大模型（GPT-4）的系统卡上声明“我们发现，与早期的语言模型一样，GPT-4-early 和 GPT-4-launch 有许多的局限性，例如产生有偏见和不可靠的内容”[37]。多模态大模型可能会受到训练算法的数据的截止日期的限制，尽管现在已经有一些多模态大模型可以从互联网上获取最新信息。例如，ChatGPT-4 的训练数据

截至 2021 年 9 月[38]，但它现在可以搜索或浏览互联网获取最新信息[39]。不过，这可能会生成更多虚假或不准确的信息，因为原本的“截止日期”可以防止引入新发布的虚假资料[39]。在医学领域，最新和高度准确的信息对于达到医疗标准和了解某些疾病都是至关重要的。

- *自动化偏见*：与其他形式的人工智能一样，多模态大模型可能会助长专家和医疗保健专业人员（以及患者，见下文）的自动化偏见，这加剧了人们对多模态大模型生成错误、不准确或有偏见的回应的担忧。在自动化偏见中，临床医生可能会忽略本可以由人工发现的错误[1]。还有人担心，医生和医疗保健专业人员可能会利用多模态大模型做出存在与伦理或道德考虑相悖的决定[20]。尽管最近的实验表明，ChatGPT 等多模态大模型作为道德顾问的作用可能很不一致，即使用户知道自己得到的是聊天机器人的建议，但其道德判断依然可能会被影响[40]。在医生无法做出困难的判断或决定的时候，使用多模态大模型进行道德判断可能会导致“道德失能”[20]。
- *技能退化*：在医疗实践中越来越多地使用人工智能，会带来一种长期风险，即临床医生作为医疗专业人士的能力会下降或削弱，因为他们越来越多地将日常责任和职责转移给计算机。能力的丧失可能导致医生无法自信地推翻或质疑算法的决定，或者在网络故障或安全漏洞的情况下，医生将无法完成某些医疗任务和程序[1]。
- *知情同意*：在日益增长的多模态大模型运用（亲身地，尤其是线上的运用）中，应当让患者知道，人工智能技术可能会协助患者给出回应，或可能最终负责模仿临床医生生成反馈。然而，如果多模态大模型和其他形式的人工智能被纳入常规医疗实践，患者或其护理人员即使不满或不愿意完全或部分地依赖人工智能技术，也可能无法撤回使用该技术的同意，尤其是在其他选择（非基于人工智能的选择）不容易获得的情况下，或在将责任移交给计算机的临床医生不使用 AI 就无法提供医疗服务的情况下。

2.2 以患者为中心的应用

人工智能正开始改变患者管理自身病情的方式。患者已经在很大程度上承担起了自我护理的责任，包括服药、改善营养和饮食、参加体育锻炼、护理伤口或注射。据预测，人工智能工具将通过使用聊天机器人、健康监测和风险预测工具以及专为残障人士设计的系统等方式，提高自我护理水平[1]。

多模态大模型可能会加快患者和非专业人士将人工智能用于医疗目的的趋势。二十年来，人们一直使用互联网搜索来获取医疗信息。因此，在向患者和非专业人士提供信息方面，大语言模型可以发挥核心作用，包括将其整合到互联网搜索中。由大语言模型驱动的聊天机器人可以在搜索信息方面取代搜索引擎[41]，包括有关自我诊断和就医之前的搜索。

多模态大模型驱动的聊天机器人可利用日益多样化的数据形式，成为高度个性化、广泛关注的线上健康助手。一项研究显示“线上健康助手可以利用个人资料……促进行为改变、回答与健康相关的问题、区分症状或适时与医疗服务提供者沟通”[3]。特定的多模态大模型驱动的聊天机器人可以提供心理健康等方面的治疗[2]。

以患者为中心的多模态大模型的第三种应用是识别临床试验或此类试验的招募[28]。基于人工智能的程序已经可以帮助患者匹配临床试验研究人员[42]，多模态大模型则可以根据患者的相关医疗数据以同样的方式进行匹配[28]。使用人工智能既能降低招募成本，又能提高速度和效率，同时还能让个人有更多机会寻求难以通过其他渠道识别和获取的试验和治疗[42]。

方框 2. 儿童使用多模态大模型的伦理考虑

虽然最近发布了在人工智能和机器学习（ACCEPT-AI）中安全、合乎伦理地使用儿科数据的广泛指南[43]，但必须特别考虑到儿童使用多模态大模型的潜在影响。

开放式多模态大模型的广泛使用为不同年龄段的用户提供了机会。然而，关于儿童接触或使用多模态大模型的信息却很有限。虽然在更广泛的教育背景下，人们讨论了使用多模态大模型的潜在机会和弊端[44]，但目前还不清楚使用多模态大模型会如何影响儿童的心理或身体健康。必须对儿童使用多模态大模型的情况进行长期监测，以了解其益处和潜在危害。

各国关于儿科同意、允准和对父母合法参与的规定的法律和政策各不相同。因此，缺乏连贯、统一、全球性和对儿童的专门监管和监督，可能会造成不明的、无人监控的伤害，尤其是使用多模态大模型造成的伤害。

具体来说，目前还不清楚多模态大模型能够多么准确地概括儿科健康状况。研究表明，成人数据集对儿科人群的通用性可能有限。因此，儿科数据应在测试和训练数据集中独立出来[45]。

开发者应在训练数据中提供包括年龄在内的人口统计信息，必须鼓励开发者提供目标人群的明确描述，如年龄范围，以便与多模态大模型进行适当、安全的接触。在法律允许的情况下，应通过纳入年轻用户的适当参与和反馈来改进多模态大模型。

风险与挑战

个人随意使用多模态大模型可能意味着巨大的风险，如下文所列：

- *不准确、不完整或虚假陈述*：与临床医生和医护专业人员使用多模态大模型的情况一样，患者和非专业人士使用多模态大模型也有可能出现虚假、有偏见、不完整或不准确的陈述，包括声称提供医疗信息的人工智能程序的陈述。没有医学专业知识的人或是儿童使用的风险会更大，因为他们没有质疑多模态大模型提供的回应的依据，也没有其他信息来源（方框 2）。尽管人们使用互联网搜索获取医疗信息已有几十年的历史，但多模态大模型仅以其他多模态大模型（具有同样的风险）作为参考并进行快速比较，从而提供看似“正确”的答案。
- *操纵*：许多由多模态大模型驱动的聊天机器人应用都有独特的聊天机器人对话方式，这种对话方式有望变得更有说服力、更容易上瘾[46]，而且聊天机器人最终可能会根据每个用户的情况调整对话模式 [41]。聊天机器人可以提供对问题的解答或参与对话，以说服个人采取违背自身利益或福祉的行动 [12]。一些专家呼吁采取紧急行动，控制聊天机器人可能带来的负面影响，并指出聊天机器人可能进行“情感操纵”[47] [48]。一个广为人知的案例是，一名比利时焦虑症患者在与聊天机器人进行了 6 周的深入对话后自杀身亡[49]。
- *隐私*：患者和非专业人士使用多模态大模型时可能不注意隐私，也可能不注重他们共享的个人信息和健康信息的保密性。出于其他目的使用多模态大模型的用户往往会分享敏感信息，如企业专有信息[50]。这些被共享的数据不一定会很快消失，因为企业可能会利用用户在多模态大模型上共享的数据来改进其人工智能模型 [50]，尽管利用共享数据进行再次训练可能没有法律依据。当然最终可能会从企业服务器上删除这些数据[51]。多模态大模型与多模态大模型的其他用户共享信息是一个与此相关的问题，无论是因为其他用户明确要求多模态大模型披露此类信息[52]，还是多模态大模型错误地披露了其他人的聊天记录（即使不是聊天内容）[53]。因此，如果一个人的可识别医疗信息被输入多模态大模型，就有可能被披露给第三方[54]。
- *减少临床医生、非专业人员和患者之间的互动*：患者或其护理人员使用多模态

大模型可能会从根本上改变医患关系。在过去二十年里，患者在互联网上搜索的增加已经改变了这种关系，因为患者可以利用他们找到的信息来质疑医疗服务提供者或向其寻求更多信息。虽然多模态大模型可以改善这种对话，但患者或护理人员可能会决定完全依赖多模态大模型进行病情预测和治疗，从而减少或消除对专业医疗判断和支持的恰当依赖。与此相关的一个担忧是，如果人工智能技术减少了医疗服务提供者与患者之间的接触，就会减少临床医生诊疗的机会，并可能破坏一般的支持性护理，例如人与人之间的互动，而此时人往往是最脆弱的[1]。一般来说，人们担心人工智能会使临床护理“去人性化”。

- *认知上的不公正*：用多模态大模型的判断取代医疗服务提供者的判断的另一个潜在后果是给患者带来认知上的不公正。“认知上的不公正”是指“针对一些人作为知识主体（例如卫生系统中的病人）的能力状况做出错误对待”[55]。认知上的不公正的一种形式是解释学不正义，它发生在共同理解和知识（所谓的“集体解释资源”）存在差距的情况下，这种差距使一些人在生活经验、社会经验、卫生领域以及他们对自己身体或精神状况的理解方面处于不利地位[55]。多模态大模型即使接受过大量数据的训练，也会对其所能识别和应对的事物以及词汇量之外的概念和观念区别对待。在临床环境中，如果患者的经历没有得到多模态大模型的认可或承认，就可能导致医疗服务提供者无法提供适当的护理，从而对患者造成伤害。对于数据集中已经被忽视和代表性不足的弱势群体[55]，如残障人士，这种情况更加可能发生（方框 3）。
- *在卫生系统之外提供医疗服务*：卫生领域人工智能不再仅限于卫生保健系统内部或在家庭护理中使用。因为卫生领域人工智能可以很容易地被非卫生系统实体接入和使用，或者仅仅由企业引进，例如那些提供供公众使用的多模态大模型的企业。这就提出了这样的问题：此类技术是应该作为需要更严格的监管审查的临床应用来进行监管，还是作为需要较少的监管审查的“健康应用”来予以监管。可以说，这类技术目前处于这两者之间的灰色地带。

如果患者在没有监管保障的情况下使用“轻度”监管的多模态大模型，包括使用多模态大模型提供医疗建议或进行自我诊断，可能会带来风险。令人担忧的是，患者可能会收到虚假或误导性建议（见上文），如果个人未与医疗服务机构取得联系，患者安全可能会受到影响，如缺乏对有自杀倾向的个人使用人工智能聊天机器人的支持性护理，即使聊天机器人不具有“操纵性”。即使信息是正确的，没有受过医疗培训的人在使用这些信息进行自我诊断时也可能会误读或误用。由于包括多模态大模型在内的此类应用正在继续扩散，而应用未必会标明是医疗卫生应用，因此医疗卫生的整体质量可能会受到影响。这可能会进一步加剧获得优质医疗服务的不平等，尤其是缺乏其他选择的人可能会求助于此类应用程序[1]。

方框 3. 与多模态大模型有关的伦理考虑及其对残障人士的影响

过去，残障人士被排除在工作场所、教育系统和适当的医疗支持之外[56]，因此也被排除在用于训练人工智能系统的数据集之外。这些系统可能会歧视面部不对称、手势、交流方式、行为和行动模式不寻常的人。受影响最严重的群体是有残障、认知或感官障碍或自闭症障碍的人[57]。

这种偏见和排斥可能体现在生成式人工智能上。例如，在患者的描述或简历中，多模态大模型可能会对与“残障”相关的关键词或短语引入负面含义或情绪[58]。聊天机器人可能会因为不同的行为或行动模式而将残障人士识别为“没有生命”、“非人类”或“情感

平淡”的对象。对于有语言障碍的人来说，语音识别系统可能不够准确进而会导致误解。

解决和克服与残障有关的偏见需要在整个人工智能发展过程中采取干预措施：将残障人士纳入人工智能系统的开发和设计中；进行审计，评估数据集中的残障偏见和人工智能系统的性能；确保旨在保护和促进残障人士权利的立法考虑到与人工智能技术有关的挑战，同时也确保法律和政策规制人工智能，以应对残障人士在越来越多地使用基于人工智能的系统时可能面临的挑战和障碍。针对人工智能的立法可涵盖“残障专用”这一分类，包括特定的范围和条件如何受到人工智能系统的影响。

2.3 文书和行政任务

多模态大模型开始用于协助医疗保健专业人员处理行医过程中的文书、行政和财务工作。在许多情况下，医生和其他医疗保健专业人员都需要处理越来越多的文书工作，比如在电子病历中记录患者信息和数据，为私人、保险或公共卫生保健系统开具账单以及其他行政任务。虽然完成电子病历等许多此类职责的初衷是为了“解放”医疗保健专业人员，但大多数义务现已成为导致医生和医疗保健专业人员倦怠的主要原因[59]。一项研究显示，文书工作占医生工作时间的四分之一到二分之一，占护士工作时间的五分之一[59]。

多模态大模型被认为是将最宝贵的时间返还给医疗保健专业人员的一种手段，既可以减少职业倦怠，也可以分配更多时间为每位患者提供护理服务，或照顾更多的患者。一位使用多模态大模型编码软件记录患者就诊情况的医生表示，“人工智能让我作为一名医生能够100%地为患者服务”，而且该软件每天可节省多达 2 小时的时间[60]。

多模态大模型目前和预期的用途包括：

- 通过简化医学术语和使沟通更加“对患者友好”，以帮助翻译或改善临床医生与患者的沟通[34]；
- 填补电子病历中缺失的信息[61]；
- 利用其他形式的人工智能，在每次患者就诊（线上或亲自就诊）后起草临床记录[62]。
- 多模态大模型的使用还将包括预先自动撰写处方、预约、账单代码、安排检查时间、保险企业预授权、手术笔记和出院小结[5]。随着更复杂的多模态大模型的开发，它们还可用于更复杂的文书记录，例如放射科医生“自动起草放射报告，既描述异常情况，又描述相关的正常结果，同时考虑到患者的病史……这些模型可以通过将文本报告与交互式可视化相结合的方式，如突显每个阶段所描述的区域，为临床医生提供进一步的帮助”[5]。

风险与挑战

与其他使用多模态大模型的情况一样，不准确、错误（如转录、翻译或提炼）或“幻觉”可能会导致严重的错误。因此，大多数文书和行政职能不能完全自动化。即使监督和审核需要花费医疗保健专业人员的时间，多模态大模型仍可能减轻现有的负担。另一个问题是多模态大模型可能不一致；对提示或问题稍作改动，就会生成完全不同的电子病历，不过随着时间的推移，不一致的情况预计会减少[63]。

2.4 医疗和护理教育

预计多模态大模型还将在医疗和护理教育中发挥作用。它们可用于创建“动态课本”，与一般课本相比，动态课本可根据学生的具体需求和问题量身进行定制[63]。集成到聊天机器

人中的多模态大模型可提供模拟对话，以改善临床医生与患者之间的沟通并解决问题，包括练习医学问诊、诊断推理和解释治疗方案。聊天机器人还可以量身定制，为学生创造残障患者或罕见病症患者等各种虚拟患者。多模态大模型还可以为医学生提供指导，它通过包括生理和生物过程在内的“思维链”进行推理，从而回答学生提出的问题[63]。

风险与挑战

虽然使用人工智能可以改善或提高医疗保健专业人员的培训和技能，但它也可能带来风险，即专业人员放弃自己（或人类同行）的判断，转而听从计算机的判断。如果多模态大模型提供不正确的信息或回应（或编造回应），可能会影响医学教育的质量。

另一个令人担忧的问题是，将多模态大模型用于教育或简化行政和文书职能，可能会给尚未掌握数字知识的医疗保健专业人员带来额外的负担，他们必须在日常工作中学习使用人工智能辅助技术的新能力[1]。预计多模态大模型的新功能需要医疗保健专业人员不断进行再培训和调整[1]。开发者最终可能会推出有自然语言或视觉效果的人工智能辅助技术，让外行人也能轻松地使用交流界面。

2.5 科学、医学研究及药物开发

人工智能已被用于科学、临床研究与药物开发。利用人工智能，可以对电子病历进行分析，以确定临床实践模式并开发新的临床实践模型。机器学习还可用于基因组医学，例如提高对疾病的认识和确定新的生物标志物[1]。人工智能几乎可被应用于药物开发周期的每一个阶段，包括简化化合物筛选、预测蛋白质的三维形状（“蛋白质折叠问题”）[1]、预测正在进行临床前开发的化合物的毒性和有效性以及改善临床试验期间的人员招募、注册和监测过程。[4]

多模态大模型扩展了人工智能支持科学、医学研究与药物开发的方式。多模态大模型可用于科学研究的各个方面。它们可以生成科研论文中的文本，用于提交稿件或撰写同行评议[34]。它们可用于总结学术论文摘要之类的文本或生成摘要。多模态大模型还可用于分析和总结数据，以便在临床和科学研究中获得新的见解。它们还可用于编辑文本，提高论文和基金申请书等书面文件的语法准确性、可读性和简洁性。此外，多模态大模型还可用于从其所训练的数据中获得洞见[34]。有一种多模态大模型是在数百万篇学术文章中训练出来的，据称可以通过对科学研究进行分析，从而回答问题、提取信息或生成相关文本[64]。多模态大模型也用于药物研发，特别是用于新药设计，以开发具有特定性质的新化合物[65]。

风险与挑战

领先的医学和科学期刊已经对多模态大模型的出现、潜力及其对科学研究的影响做出了回应。例如，一家学术出版企业制定了两条规则：（1）不接受多模态大模型作为研究论文的署名作者；（2）使用多模态大模型的研究人员应在方法部分和致谢中记录其使用情况[66]。世界医学编辑学会将作者限定为人类[67]。

在科学研究中使用多模态大模型的普遍担忧包括：

- *缺乏问责制*：科学或医学研究论文的作者需要承担责任，而人工智能工具无法承担这种责任[66]。缺乏问责制是一家主流学术出版商和世界医学编辑学会决定不接受多模态大模型作为可署名作者的依据。
- *高收入国家偏见*：用于训练多模态大模型的科学和医学研究大多在高收入国家进行。因此，任何多模态大模型的回应都可能偏向于高收入国家的观点[34]。这可能会加

[4] 世卫组织即将出版的《人工智能在药物开发中的伦理与治理》（Ethics and governance of artificial intelligence in drug development）。

强或加剧对于来自低收入和中等收入国家研究成果的忽视和不引用的趋势[68]，尤其在出版物由非拉丁文字撰写的情况下。

- *幻觉和/或错误信息*：多模态大模型可能会总结或引用不存在的学术文章或其他信息而产生“幻觉”[69]。
- *破坏信任*：使用多模态大模型开展诸如同行评议等活动可能会破坏对这一过程的信任[69]。
- *多模态大模型和多模态大模型产生的知识的可及性*：与科学和医学研究中使用的其他工具、技术和信息一样，多模态大模型也将“被置于付费专区”，从而加剧数字和知识鸿沟，并影响到那些寻求参与科学和医学研究但缺乏支持和资金的科学家[34]。

虽然人工智能（包括多模态大模型）有利于药物开发，但在这一领域使用人工智能也令人担忧，一份即将出版的世卫组织出版物将对此进行研究。

3 通用基础模型（多模态大模型）对卫生系统与社会和伦理问题的风险

虽然与多模态大模型相关的许多风险和问题会影响个人用户（如医疗保健专业人员、患者、研究人员或护理人员），但它们也可能构成系统性风险。与卫生领域多模态大模型和其他基于人工智能的技术相关的新出现或预期风险包括：（i）可能影响国家卫生系统的风险，（ii）监管和治理风险，（iii）国际社会关切。

3.1 卫生系统

医疗卫生系统以六个组成部分为基础：提供医疗服务、医疗人力资源、医疗信息系统、获得基本药物、财务以及领导与治理[70]。这六大基石都可能受到多模态大模型直接或间接地影响。与使用可能影响医疗卫生系统的多模态大模型相关的风险描述如下：

高估多模态大模型的效益而低估风险

有些人倾向于夸大和高估人工智能的作用，这可能会导致尚未经过严格的安全性和有效性评估的产品和服务被采用[1]。这很大程度上是由于“技术解决主义”的持久吸引力，即在人工智能和多模态大模型等技术被证明有用、安全和有效之前，这些技术就被当成是消除更深层次的社会、结构、经济和体制障碍的“灵丹妙药”[1]。

多模态大模型是新颖却未经测试的，如上所述，它产生的不是事实，而是类似于事实的信息，可能是不准确的。消费者、政界和公众都对多模态大模型颇感兴趣，但这可能会导致政策制定者、医疗服务提供者和患者高估了它们的益处，忽视了多模态大模型可能带来的挑战和问题。对于政策制定者来说，在多模态大模型尚未开发和使用之前，可能无法获得确定多模态大模型使用范围的必要证据。使用多模态大模型的选择不应优先于已经在使用的基于人工智能的技术，也不应优先于可能资金不足和使用不足但已被证明具有治疗或公共卫生益处的非人工智能或非数字的解决方案。不平衡的医疗保健政策和误导性投资，可能会转移对已被证明有效的干预措施的注意力和资源，并加剧卫生部门对于压缩公共医疗支出的压力，尤其不利于资源有限的低收入和中等收入国家[1]。

可及性和可负担性

很多因素都会影响为医疗服务提供者和患者带来益处的多模态大模型的公平接入。其中之一就是数字鸿沟，它限制了某些国家、地区或部分人口使用数字工具。数字鸿沟还导致了其他差异，其中许多因素都会影响人工智能的使用，而人工智能本身也会强化和加剧差距。另一个可能影响人们使用多模态大模型的因素是，不同于互联网，许多多模态大模型只有在付费或订阅后才能使用，因为开发和运营多模态大模型可能很昂贵。据估计，ChatGPT 每天的运营成本为 70 万美元[71]。一些企业正在对新版本的多模态大模型收取订阅费[72]，这将不仅使得低收入和中等收入国家无法负担，而且会使得高收入国家资源匮乏环境中的个人、卫生系统或地方政府也负担不起[54]。相反，所有国家的穷人都只能使用“具有成本效益的解决方案”的多模态大模型，而只有较富裕的人才能获得“真正的”医疗保健专业人员的服务。第三个因素是，目前大多数多模态大模型只使用英语。因此，虽然它们可以接收输入并以其他语言提供输出，但它们更有可能产生虚假信息或错误信息[73]。

全系统偏见

如上所述，用于训练人工智能模型的数据集是有偏见的，因为许多数据集没有包括女童和妇女、少数族裔、老年人、农村地区和弱势群体。一般来说，人工智能偏向于数据最多的人群，因此，在不平等的社会中，人工智能可能会使少数群体处于不利地位[1]。多模态大模

型的一个特殊问题在于，偏见可能会随着模型规模的扩大而增加[74]，即使正在开发所谓的较小多模态大模型，但是用于训练连续模型的数据不断增加就会导致偏见的增加。偏见可能会引发整个医疗卫生系统的歧视，影响人们获得包括医疗服务和高质量护理在内的基本服务[75]。但同时，多模态大模型很可能包含一些可以“反击”各种形式的偏见和刻板印象的数据。研究人员发现，提示模型不要依赖刻板印象就会对算法的响应产生巨大的积极影响[74]。

对劳动力和就业的影响

据一家投资银行估计，多模态大模型最终将导致至少 3 亿个工作岗位丧失（或“退化”）[76]。经济合作与发展组织的一份报告指出，在其成员国中，受人工智能驱动的自动化影响最大的职业是高技能工作，部分原因是在使用多模态大模型后，特别是“金融、医疗和法律活动中的职业……可能会突然发现自己面临人工智能自动化的风险”[77]。然而，对于许多国家来说，卫生领域不是一个行业，而是政府的一项核心职能，医疗保健专业人员可能不会被技术取代。此外，许多国家仍然存在医疗保健专业人员短缺的问题[1]，包括在 COVID-19 大流行之后[78]。世卫组织预计，到 2030 年，医疗保健专业人员的短缺将达到 1000 万名[79]，短缺主要集中在低收入和中等收入国家。因此，被验证为安全有效的多模态大模型可用于缩小提供医疗服务所需的劳动力与现有劳动力之间的差距。

另一个值得关注的问题是，多模态大模型投入市场对于当前和未来医疗保健专业人员数量的影响。一家大型科技企业估计，多达 80%的工作将受到人工智能到来的影响[80]。咨询企业埃森哲（Accenture）预计，人类 40%的工作时间可能会受到多模态大模型的影响，并乐观地指出，“对人类创造力和生产力的积极影响将是巨大的”[81]。然而，如上所述，引入多模态大模型可能会给许多医疗保健专业人员带来重大挑战，他们需要接受培训并适应多模态大模型。卫生系统必须考虑到它给医疗服务提供者带来的挑战以及给患者和护理人员带来的风险。

第三个令人担忧的问题是，负责审查内容、标注训练数据以及从数据集中删除包含辱骂、暴力或精神困扰类内容的人员所付出的精神和心理代价。负责过滤此类内容的人员通常在低收入和中等收入国家工作，工资很低，可能会因审查此类内容而遭受心理困扰，却无法获得咨询或其他形式的医疗服务[73]。

医疗卫生系统对不适配多模态大模型的依赖

虽然多模态大模型可以解决医疗保健专业人员持续短缺的问题、扩大医疗卫生系统的覆盖范围，但也可能导致这些系统可能会过分依赖多模态大模型，特别是依赖工业界开发的多模态大模型技术。因此，如果医疗和公共卫生领域的多模态大模型得不到维护、被减少或仅在为高收入环境中使用而设计和更新，那么依赖这些多模态大模型的医疗卫生系统将不得不进行调整，并有可能需要在没有多模态大模型的情况下提供医疗卫生服务。此时，若医疗保健专业人员已经“去技能化”、将某些职责外包给人工智能，或者患者期望使用人工智能，这可能会使得情况十分艰难。相关的风险在于，如果多模态大模型不能保护患者的隐私和保密性，过度依赖多模态大模型可能会破坏个人和社会对卫生保健系统的信任，因为人们将不再有信心在不危及隐私的情况下获得医疗卫生服务。

网络安全风险

随着卫生保健系统越来越依赖人工智能，这些技术可能成为恶意攻击和黑客行为的目标，某些系统可能会被关闭，训练数据可能会被操纵以改变其性能和回应，数据也可能会被“绑架”以索取赎金[1]。如上所述，一个独特的危及安全的风险是，敏感数据被输入到不受披露、授权使用等方法保护的多模态大模型中。多模态大模型本身也可能受到网络安全风险的影响，例如“提示词攻击”攻击。这种攻击是指第三方向多模态大模型输入数据，导致多模态大模型不以开发者所预想的方式运行[82]。例如，“提示词攻击”可以指示旨在回答数据库问题的多模态大模型从数据库中删除或更改信息。目前还没有解决这一漏洞的方法。提示词攻

击目前被安全研究人员用来说明多模态大模型所面临的挑战，但它们也可能被恶意行为者用来窃取数据或欺诈用户[83]。

3.2 法律与监管合规

虽然可以颁布新的法律来规范人工智能的使用，但某些现行法律法规，特别是数据保护方面的法律和国际人权义务也适用于多模态大模型的开发、提供和部署。目前已开发并供公众使用的某些多模态大模型可能会违反若干主要的数据保护法律，如欧盟的《通用数据保护条例》[84]。该条例涵盖了各种权利，如免受自动决策影响的权利。这些权利、保护和要求必须指导人工智能的发展[85]。

在欧盟成员国[83]和加拿大[86]等地，一些此类违规行为引发了对多模态大模型的调查。违规行为包括（i）多模态大模型在未征得个人同意的情况下（也没有收集这些数据的“合法权益”）从互联网上收集和使用个人数据[87]；（ii）未曾告知公众正在使用他们的数据，且不能给予用户纠正错误、删除数据（“被遗忘权”）或拒绝使用这些数据的权利的多模态大模型[87]；（iii）在使用用户提供给聊天机器人或其他消费者界面的敏感数据时，不够完全透明的多模态大模型（尽管法律规定用户必须能够删除聊天记录数据）[83]；（iv）没有适当的“年龄门槛”系统来过滤 13 岁以下用户和未获得父母同意的 13 至 18 岁用户的多模态大模型[88]；（v）不能防止个人信息泄露的多模态大模型[87]；（vi）部分由于幻觉而发布不准确个人信息的多模态大模型[89]。其他可能违反《通用数据保护条例》的行为还包括“解释权”要求，即利用个人数据进行自动化处理的实体需要解释该系统（如多模态大模型）是如何做出决定的。[5]如上所述，尽管一些企业正在研究满足“可解释性”要求的方法，但目前还无法解释多模态大模型是如何做出决定的[90]。

许多违法违规行为都较为重大，这与多模态大模型的训练方式、使用方式以及数据控制者的管理方式有关。多模态大模型有可能永远无法遵守《通用数据保护条例》或其他数据保护法律[91]。2023 年，一份向某欧盟成员国数据保护机构提交的投诉称，一家企业的大语言模型及其开发和现在的运行方式系统性地违反了《通用数据保护条例》[92]。

许多违反数据保护法的行为也可能违反消费者保护法[93]。更广泛地说，如果这些问题不能得到解决，它们也直接违反了世卫组织关于卫生领域人工智能的指导原则，包括保护人类的自主性的原则以及确保透明、可以解释和可以理解的原则。

企业无法遵守现行法律可能是导致一些企业对即将出台的人工智能法规热烈关切的原因。针对欧盟计划出台的《人工智能法案》，一家大企业的负责人表示，该企业可能无法在欧洲提供其主要的多模态大模型产品，因为它可能无法遵守相关法规[94]。这一最后通牒可能会导致隐私权和其他保护措施受到侵蚀，也可能导致医疗服务的提供取决于国家是否愿意放弃某些人权。

3.3 社会的关切与风险

与其他人工智能技术一样，多模态大模型预计将产生更广泛的社会影响，超出卫生系统的范围，这不是一项法律或政策所能解决的。这些影响包括：多模态大模型可能会加强一小部分在多模态大模型商业化方面处于领先地位的科技企业（及其高管）的权力和权威。多模态大模型还可能对环境和气候产生负面影响，因为在训练和使用多模态大模型的过程中会消耗碳和水。在人类能够确保人工智能不会用信息、证据和建议取代人类的认识权威之前，这种给出不准确、错误、有偏见的信息且缺乏道德或场景推理的技术很快就会“以一种文明无法安全吸收的速度与数十亿人的生活纠缠在一起”[4]，包括在卫生保健与药品领域。此外，人们还严重担心多模态大模型可能会增强由技术推动的性别暴力，包括网络欺凌、仇恨言论

[5] 详见《通用数据保护条例》第 15(1)(h)条及其序言（Recital）第 71 条。

和未经同意使用图像和视频，如“深度伪造”。后一种风险在本指南中没有涉及，但值得世卫组织更广泛地关注，因为它对人工智能使用所针对的人群，尤其是未成年与成年女性的健康和福祉具有严重的负面影响[95]。

与大型科技企业有关的挑战

随着多模态大模型参数和规模的扩大，多模态大模型的出现加强了开发和部署人工智能的少数大型科技企业的主导地位和中心地位[96]。很少有企业和政府拥有足够的人力和财力、专业知识、数据和算力来开发日益复杂的多模态大模型[96]。多模态大模型的算力和投资都在增加，而且随着对人工智能需求的增长，招聘“人工智能人才”的费用也十分昂贵[97] [98]。拥有功能最强大微芯片的多模态大模型需要多台计算机连续工作数周甚至数月且数千个芯片协同工作，才能完成训练[99]。

随着多模态大模型的训练、部署和维护成本不断上升，少数企业有可能会“产业俘获”许多产品和服务（包括卫生领域）的潜在组成部分，从而排挤高校（学术界）、初创企业甚至政府[100]。在人工智能研究方面，已经有令人信服的证据表明，最大的企业正在排挤高校和政府。

人工智能专业的博士毕业生选择的工作岗位就是一个例子。目前，选择在企业工作的人数“前所未有”。2004 年，只有约 20%的毕业生进入企业工作，而到 2020 年，将近 70%的毕业生进入企业工作[101]。在美国和其他国家，专门从事人工智能研究的教师从高校被聘用到企业工作的数量自 2006 年以来增加了八倍[101]。在算力和大型数据集的使用方面，产业界也已成为政府和学术界的主导。2021 年，工业模型是学术模型的 29 倍[101]。此外，原始支出，尤其是高收入国家政府的原始支出，远远落后于产业界。一项研究指出，“2021 年，美国非国防政府机构为人工智能拨款 15 亿美元。同年，欧盟委员会计划投入 10 亿欧元（12 亿美元）。相比之下，到 2021 年，全球产业界在人工智能上的投入将超过 3400 亿美元，远远超过公共投资”[101]。

“人工智能投入”的主导地位意味着大型科技企业现在也主导着产出和成果。最大人工智能模型的行业份额从 2010 年的 11%增加到 2021 年的 96%，而一个或多个行业合著的研究报告的数量在 2000 年至 2020 年期间增加了 16%[101]。

大型科技企业的主导地位不仅决定了人工智能的应用和用途，也越来越多地决定了早期研究的优先事项[101]。产业的主导地位和政府投资的缺乏也意味着，符合公众利益的重要人工智能技术的替代方案，包括医疗和药品，可能会变得越来越少。这与制药部门的情况不同，例如，在制药部门，政府、非营利组织和慈善机构对研发进行了大量投资，尤其是在药物开发的早期关键阶段，以及某些治疗方法的后期开发阶段[102]。因此，企业将会越来越多地监督、支撑包括医疗卫生系统在内的经济和社会部门的运行，这引发了人们对公民和政府对自身生活掌控能力的担忧[101]。

在缺乏替代品和监管的情况下（即使 2023 年颁布法律，也可能需要数年时间才能全面实施），大型科技企业如何做出内部决策以及如何与社会和政府建立联系变得越来越重要。企业可能会通过例如前沿模型论坛[103]或与高收入国家政府合作以解决各种问题，包括与美国政府做出的几项自愿承诺[104]和即将与欧盟做出的承诺[105]。

另一个令人担忧的问题是，企业可能不会坚持伦理方面的社会责任。例如，一些大型科技企业成立伦理团队的目的是确保人工智能模型的设计和开发遵守内部伦理原则[106]，通过引入“摩擦”从而要求企业放缓或停止某些开发活动，但大型科技企业往往要么将伦理团队搁置一边，要么将其裁撤。裁撤负责人工智能伦理相关问题的整个团队，意味着伦理原则没有“与产品设计紧密联系在一起”[107]，而是被搁置了。

一些大型科技企业通过前沿模型论坛，承诺确保“负责任和安全地开发前沿人工智能模型”，包括多模态大模型，“确定负责任地开发和部署前沿模型的最佳实践”，并“与政策

制定者、学术界、民间团体和企业合作，分享有关信任和安全风险的知识”[103]。在与美国政府的自愿承诺中，科技企业保证避免有害的偏见和歧视，并保护隐私[104]。不过，自愿承诺或合作伙伴关系是否足以取代对伦理规范的有力承诺，目前还不清楚。例如，一家企业的伦理团队曾建议停止发布新的多模态大模型，但后来他们修改了文件，对之前记录的风险进行了淡化[106]。

大型科技企业既没有开发医疗产品和服务的历史，也不够专业。因此，它们可能对医疗卫生保健系统、医疗服务提供者和患者的需求不敏感，可能无法解决隐私或质量保证等传统医疗卫生企业和公共医疗机构所熟悉的问题。随着时间的推移，它们的敏感性可能会提高，正如其他提供了几十年医疗产品和服务的企业一样。

许多正在开发多模态大模型的企业对政府、监管机构或可能使用其模型的企业都不公开透明，这些企业可能（i）需要证据、数据、性能和其他信息，以评估多模态大模型的风险和受益[97] [106]，（ii）需要模型中的参数数量，以衡量模型有多强大的指标[8]。使用此类模型开发产品和服务的企业也不披露他们评估伦理挑战和风险的方法、所采取的保障措施、多模态大模型对这些保障措施的反应以及应限制或停止使用技术的情况。研究者根据 100 项指标对 10 家领先的大语言模型开发者的基础模型透明度指数进行了评估，发现“没有一家主要的基础模型开发者提供接近于充分的透明度，这表明人工智能行业从根本上缺乏透明度”[108]。美国联邦政府与几家大型科技企业达成的自愿协议包括两项透明度承诺。这些企业承诺：（i）与行业、政府、民间团体和学术界共享风险管理信息；（ii）公开报告其人工智能系统的能力、局限性以及适配与不适配的领域[104]。虽然这些承诺可能比现状有所改进，但它们都是自愿且可以自由解释的，如果没有具体的监管要求，可能无法实现充分披露。

由于内部商业压力或外部竞争[106]，企业在充分了解多模态大模型的功能之前[109]，就急于将新的多模态大模型尽快推向市场，而无论适当的测试、保障措施以及伦理风险和关切是否已经确定和解决[106] [110]。一位企业高管说，“在这个时候担心那些以后可以解决的问题，绝对是一个致命的错误”[106]。企业寻求先发优势，因为多模态大模型在某些领域（如互联网搜索）的市场份额能带来收入。据一家企业称，搜索引擎每增加 1%的市场份额，就相当于增加 20 亿美元的收入[107]。一家大型科技企业的高管表示，该企业的多模态大模型“并不完美”，但因为“市场需要”会予以发布[8]。在没有充分识别、验证、说明和降低风险的情况下发布多模态大模型的企业会积累“道德债务”，其最终后果不是由企业承担，而是由那些最容易受到此类技术负面影响的人承担[109]。前沿模型论坛的成员致力于“推进人工智能安全研究”和“确定最佳实践”[103]，且其对美国政府的自愿承诺包括在人工智能系统发布前对其进行内部和外部测试[104]。

商业压力不仅可能导致企业急于将多模态大模型尽快推向市场，还可能导致企业为了优先考虑可以创收的服务，而取消或放弃对公众健康有重大益处的产品和服务。2023 年，一家大型科技企业“砍掉”了开发 ESMFold 的团队，ESMFold 一种蛋白质语言模型，可以从单个序列预测完整的原子级蛋白质结构，还生成了一个包含 6 亿多个蛋白质结构的数据库。人们担心该企业可能不愿意“承担维持数据库运行的成本，以及允许科学家在新蛋白质序列上运行 ESM 算法的另一项服务”[111]。

多模态大模型碳足迹和水足迹

多模态大模型规模不断扩大的另一个后果是其对环境的影响。多模态大模型需要大量数据，而训练数据需要消耗大量能源[112]。在一家大型企业，训练一个新的多模态大模型需要耗费大约 3.4 千兆瓦时（2 个月），相当于 300 个美国家庭的年能耗[112]。虽然有些多模态大模型是在使用可再生能源或无碳能源的数据中心训练的，但大多数人工智能模型是在使用化石燃料供电的电网中进行训练的[112]。随着越来越多的企业引入多模态大模型，电力消耗将继续增加，最终可能对气候变化产生重大影响。

世卫组织认为，气候变化是一项紧迫的全球健康挑战，现在和未来几十年都需要优先采取行动。在 2030 年至 2050 年期间，气候变化预计将导致因营养不良、疟疾、腹泻和热应激死亡的人数每年增加约 25 万。到 2030 年，直接损害健康的成本估计为每年 20 亿至 40 亿美元。在医疗基础设施薄弱的地区，主要是低收入和中等收入国家，如果没有做好准备和应对援助，其应对能力将最差[1]。

多模态大模型对水的使用也有很大影响。一家大型科技企业的早期多模态大模型训练消耗了 70 万升淡水，而其他数据中心的用水量可能更大[113]。尽管许多开发者越来越意识到他们的碳足迹，但许多人并没有意识到他们的水足迹[114]。一个多模态大模型的简短对话（20—50 个问题和回答）所需的水量相当于一瓶 500 毫升的水。训练多模态大模型所消耗的所有水，包括人工智能服务器的制造、运输和芯片制造，其总体水足迹可能要大得多[114]。数据中心会对当地供水造成压力。例如，一家企业的数据中心使用了美国俄勒冈州一个城市全部用水量的 25%以上[114]。另一个大型科技企业正计划在一个严重干旱的国家建造数据中心，这迫使当地居民不得不饮用含盐的水[115]。跟踪水足迹很困难，因为虽然人们对碳足迹有了更多的认识、测量和理解，但企业对水足迹却没有同样的理解，或者不对其进行测量[114]。

取代人类认识权威的“危险”算法

与多模态大模型的出现相关的一个更普遍的社会风险是，多模态大模型尽管提供的是似是而非的回应却逐渐被视为知识来源，这最终可能会削弱人类知识的权威，包括在医疗保健、科学和医学研究。事实上，多模态大模型并不产生知识，也不理解自己在“说什么”，在回答问题时也没有任何道德或场景推理。

如果这种情况持续下去，社会可能还没有为计算机生成推理的后果做好准备。早期的人工智能通过社交媒体算法提供信息，传播错误信息，对心理健康造成负面影响，并加剧了两极分化和分裂[4]。即使科技企业一再警告多模态大模型的危险，它们仍继续在没有保障或监管的情况下直接向社会发布多模态大模型，其方式不仅可能取代人类对知识生产的控制，而且可能削弱人类在医疗、医药和其他领域安全使用知识的能力，而这些领域正是社会所依赖的系统。这种伤害尤其会影响到资源匮乏环境中的人群和社区，因为他们的数据不太可能被用于训练人工智能系统，这降低了系统回应的准确性。然而，这些群体可能会听从人工智能系统的建议，尤其是在没有专业医疗人员或医疗服务提供者对多模态大模型生成的错误或不准确的回应进行背景分析或纠正的情况下。

多模态大模型向公共领域和知识库释放越来越多的不完善信息或虚假信息，最终可能导致“模型崩塌”，即根据不准确或虚假信息训练的多模态大模型也会污染互联网等公共信息来源[116] [117]。为了避免出现这种情况，同时最大限度地发挥多模态大模型在医疗和其他重要社会领域的作用，政府、民间团体和私营部门必须引导这些技术为共同利益服务。

II. 卫生领域通用基础模型（多模态大模型）的伦理和治理

世卫组织专家组确定的伦理原则（见上文）为利益攸关方提供了指导，使其了解在开发、部署和评估卫生领域多模态大模型时应遵循的基本伦理要求。这些原则应成为政府、公共部门机构、研究人员、企业和实施者治理多模态大模型应用的基础。

治理包括政府和其他决策者（包括国际卫生机构）制定指导和规则的职能，以实现有利于全民健康覆盖的国家和全球卫生政策。治理也是一个政治过程，涉及平衡各种相互竞争的影响和需求[1]。现行的法律和政策不可能满足有效治理多模态大模型的需要，因为许多法律和政策是在早期版本的多模态大模型发布之前制定的。与人工智能的整体治理一样，对多模态大模型的治理涵盖适用现行和新的法律法规、“软法”（如伦理原则）、人权义务、实践准则以及企业、行业协会和标准制定机构的内部程序。

目前，多模态大模型的部署速度超过了我们充分了解其能力和缺点的速度。为消除对多模态大模型的担忧，早期的建议是禁止或暂停开发多模态大模型[118]。虽然有些国家确实限制甚至禁止某些多模态大模型的使用，但大多数国家的政府现在都力求通过适当的治理，确保多模态大模型的使用能够产生有益于社会的结果。领先的人工智能企业也呼吁谨慎、审慎地开发多模态大模型和其他形式的人工智能。然而，政府和企业都无法避免竞争。一些国家的政府正在为争夺技术优势进行“军备竞赛”，而即使是呼吁监管的人工智能企业也难逃商业压力[119]。乐观人士认为，人工智能的许多挑战和风险都可以通过设计来解决，包括不断扩大数据集和设计更强大的算法，但批评者指出，多模态大模型的局限性是系统性的，增加训练数据和模型参数的规模不仅不能克服缺点，反而会放大这些缺点[59]。

多模态大模型的治理必须跟上其快速发展和日益增加的用途，不应给予寻求技术优势的政府和寻求商业利益的企业特权。以下初步建议将伦理原则和人权义务置于适当治理的核心，包括企业可以引入的程序和做法以及政府可以颁布的法律和政策。

多模态大模型可被视为一个或多个参与者在编程和产品开发方面一系列（或一连串）决策的产物。在人工智能价值链的每个阶段做出的决定都可能对下游参与开发、部署和使用多模态大模型的各方产生直接或间接的影响。政府可以通过在国家、地区和全球范围内颁布和执行法律和政策来影响和规范这些决策。人工智能价值链首先要整合若干投入，组成“人工智能基础设施”，如数据、算力和人工智能专业技能，以开发通用基础模型。这些模型可直接用于执行各种通常意想不到的任务（包括与医疗保健相关的任务）。有几种通用基础模型是专门为卫生保健与药品领域的使用而训练的。

应在价值链的每个阶段，从数据收集到医疗应用的部署，对卫生领域多模态大模型进行适当治理。因此，人工智能价值链的三个关键阶段是：

- 通用基础模型的设计与开发（设计与开发阶段）；
- 通用基础模型服务、应用程序或产品的定义（提供阶段）；
- 医疗服务应用程序或服务的部署（部署阶段）。

在人工智能价值链的每个阶段，都要提出下列问题：

1. 哪个行为者（开发者、提供者和/或部署者）最适合应对相关风险？应该应对哪些人工智能价值链中的风险？
2. 相关行为者如何应对这些风险？他们必须坚持哪些伦理原则？
3. 政府在应对风险方面应该扮演什么角色？政府可以引入或应用哪些法律、政策或投资来要求人工智能价值链中的行为者坚持特定的伦理原则？

在设计和开发阶段，重点是开发者可以采用哪些做法来维护伦理承诺和规范，政府可以进行政策制定和投资。在提供阶段，重点是政府可以采取哪些措施来评估和规范多模态大模型在医疗和药品方面的使用。在部署阶段，政府和价值链中的所有参与者都要采取措施，确保识别并避免对用户造成任何潜在或实际的伤害。

4 通用基础模型（多模态大模型）的设计与开发

通用基础模型通常需要在大量数据基础上进行训练，需要巨大的算力。多模态大模型的开发还需要专门的具有科学和工程专业知识的人力资源。世卫组织《卫生领域人工智能的伦理和治理》[1]建议，医学人工智能的开发者“应投资于改善其产品的设计、监督、可靠性和自我监管的措施”。

虽然下文的大多数研究结果和建议可适用于所有通用基础模型，但本指南的目的是针对可能或正在用于卫生保健与药品领域（由用户直接使用或通过应用程序或服务使用）的此类模型。以下建议也旨在指导设计和使用经过专门训练的多模态大模型，这些多模态大模型可直接由用户使用，也可通过应用或服务使用。

4.1 通用基础模型（多模态大模型）开发过程中需要应对的风险

通用基础模型的设计和开发可能会带来严重的风险，如果不加以纠正，可能会对社会产生广泛的影响，或对多模态大模型的用户造成特殊的负面后果。消除或降低这些风险是开发者的责任，因为只有开发者才能够（或可以）在设计和开发过程中做出某些决定，而这些决定是可能会使用算法的提供者和部署者无法控制的（而且提供者、部署者或用户也无法通过正确使用技术来降低这些风险）[120]。例如，这些决定涉及用于训练多模态大模型的数据[121]。确保数据保护义务、质量以及减少偏差的义务的实现也不在应用程序下游开发者的控制范围之内[121]，为确保多模态大模型不会发出“人工智能引发的毒性”而必须采取的措施也是如此[122]。如果不追究多模态大模型开发者对此类设计缺陷的责任，就会像一份报告所指出的那样，使拥有最多资源的企业逃避“解决问题的责任……当他们急于主导一种新的应用人工智能形式时，他们的方法可能会不经意地掺杂其中”[122]。

通用基础模型的开发者至少应通过政府法律法规等途径应对八种风险：

- 偏见（与设计和训练数据有关）；
- 隐私（训练数据和其他输入数据）；
- 劳工问题（外包数据过滤，删除攻击性内容）；
- 碳足迹和水足迹；
- 虚假信息、仇恨言论或错误信息；
- 安全性与网络安全；
- 维护人类的认识权威；
- 对多模态大模型的独家控制。

4.2 开发者可以采取的应对通用基础模型（多模态大模型）风险的措施

开发者可以采取许多措施或做法来应对这些风险，无论是作为对伦理原则或政策的承诺，还是为了满足政府的要求。

人工智能专业技能（科学和工程人员）：开发者可以确保其科学和编程人员能够识别和规避风险。世卫组织的伦理指南[1]就科学和工程人员的培训以及设计过程的包容性提出了若干建议。特别是世卫组织专家组建议开发者考虑“对‘高风险’人工智能（包括医学人工智能）开发者的许可或认证要求”。

开发出应用于或可用于医疗、科学研究或药品领域的多模态大模型的企业和其他实体，应考虑进行认证或培训，使自己符合医疗行业的要求，并提高对其产品和服务的信任度[1]。任何由开发者或专业学会引入和执行的标准，都应与政府监管机构合作或由政府监管机构制定，并应符合世卫组织增进人类福祉、安全和公共利益的伦理原则。暂未计划但可以预见其

多模态大模型可能用于卫生领域的开发者最好应确保拥有内部专业知识，以预先准备和应对此类用途。

数据：尽管人力资源和算力对开发多模态大模型至关重要，数据可能才是最关键的基础设施要求。用于训练多模态大模型的数据质量和类型决定其是否符合核心伦理原则和法律要求[123]。虽然人工智能开发者在定性调查中一致认为数据质量“很重要”，需要投入大量时间，但与数据相关的工作往往价值不高，这可能对医疗和药品等“高风险”领域的人工智能产生重大负面影响 [123]。如果数据没有达到适当的质量，就可能违反世卫组织的若干指导原则，包括增进人类福祉、安全和公共利益，以及如果数据导致偏见则应确保包容性和公平的原则。

由于将数据用于医疗保健可能需要严格遵守知情同意的法律规定，因此开发者在训练医疗多模态大模型时，可能不得不依赖较小的数据集 [59]。较小的数据集可能对于确保数据的质量而言是可取的，其多样化的数据可以避免偏差 [59]，也可以反映多模态大模型服务对象的构成和实际情况。然而，较小的数据集可能会增加重新识别个人身份的风险，从而使他们现在或将来受到伤害。依赖较小的数据集可能会带来更多益处，包括减少模型的碳足迹和水足迹[112]，以及使较小的实体有可能参与或开发需要较少数据、算力、人力和财政资源的多模态大模型类模型 [59]。无论数据集的规模如何，开发者都应进行“数据保护影响评估”，按照《通用数据保护条例》的要求，这将要求开发者在处理数据之前，评估数据处理操作对个人权利和自由的风险，以及对个人数据保护的影响 [1]。从低收入和中等收入国家收集数据可能构成“数据殖民主义”，即数据被用于商业或非商业目的，却不对同意、隐私或自主权给予应有的尊重 [1]。

评估范围可超出隐私风险，包括数据质量，如数据是否公正准确。检查或审核此类数据集的人工智能研究人员指出，虽然为人工智能创建数据集很简单，但审核却很困难、耗时且昂贵。一位研究人员指出“做脏活累活工作要难得多”[124]。

开发者可以采取其他措施来提高数据质量并遵守数据保护法。与早期多模态大模型的开发方式不同，无论模型大小如何，开发者都应根据按照数据保护规则的最佳实践收集数据进而对多模态大模型进行训练。因此，开发者应避免使用数据经纪人等第三方来源的数据，因为他们的数据可能是旧的、有偏见的、组合不正确的，或有其他可能尚未纠正的缺陷[125]。谨慎收集数据还能确保多模态大模型不违反版权法或数据保护法，因为这些行为可能会带来法律后果，某些多模态大模型可能会因此被贴上非法标签[126]。

如果使用第三方数据提供者的数据，可以对其进行认证，以建立信任并确保其专业性和合法性[127]。开发者用于训练多模态大模型的所有数据，无论是直接收集的还是从第三方收集的，都必须保持更新。如上所述，一些领先的人工智能模型并未使用最新数据进行训练 [38]，这可能会影响模型在卫生保健与药品领域的表现，因为新的证据和信息会对决策产生有意义的影响。数据集应不断更新且准确无误，这样多模态大模型才能适配且与应用场景相关。

确保数据足够透明可能十分困难。推出新多模态大模型的企业的训练数据越来越不透明。一家发布新多模态大模型的领先人工智能企业表示“鉴于竞争格局和大规模 GPT-4 等大规模模型的安全影响，本报告不包含有关架构（包括模型大小）、硬件、训练计算、数据集构建、训练方法或类似内容的更多细节”[128]。

然而，不愿意让数据透明的做法不符合世卫组织关于确保透明、可以解释和可以理解的伦理原则。开发者应该对用于训练模型的数据保持透明，以便下游用户，包括对多模态大模型进行微调的用户、使用多模态大模型开发医疗应用的用户以及直接使用多模态大模型的用户，了解训练数据集的不足或不完整之处。

当开发者使用低收入和中等收入国家的数据工作者来筛查内容中有关辱骂、暴力或攻击性的材料并对数据进行标注，从而提高数据质量时，应向这些工作者支付维持生计的工资，并向他们提供心理健康服务和其他形式的咨询；开发者应采取保障措施，保护工作者远离任

何形式的痛苦。各国政府应更新其劳工标准，将福利扩大到所有数据工作者，促进各企业之间的“公平竞争环境”，并确保劳工标准随着时间的推移得到维持和改善。

伦理设计和价值设计：将伦理和人权标准纳入人工智能技术开发的一种方法是“价值设计”，这是一种将人类尊严、自由、平等和团结等价值作为设计基础，并将其视为非功能性要求的范例[1]。世卫组织最初的人工智能技术设计专家指南中的几项建议，包括“价值设计”，值得在此重申。

该指南建议，人工智能技术的设计和开发不应仅由科学家和工程师完成，“潜在的终端用户和所有直接和间接的利益相关者应从人工智能开发的早期阶段就参与到结构化、包容性和透明的设计工作中，并应有机会提出伦理问题、表达担忧和为可能采用的人工智能应用提供意见”[1]。因此，在基础模型的开发过程中，可能使用模型或从模型中受益的人可以参与初始开发。其中一项建议是引入所谓的“人类监督协会”，这将有助于让患者代表参与旨在直接或通过医疗提供者间接使患者或护理者受益的多模态大模型开发过程。[6]医疗保健专业人员、研究科学家、患者、非专业人员和弱势群体也可参与多模态大模型的设计、数据标注和测试。例如，多模态大模型设计的包容性可以保护人的自主权，因为医疗服务提供者的参与可以防止或减少服务提供者的自动化偏见。包容性设计可以促进世卫组织确保包容性和公平的指导原则，特别是如果设计团队能够根据年龄、能力、种族、民族、性倾向或性别认同给出不同观点。

世卫组织先前的指南还建议：“设计者和其他利益相关者应确保人工智能系统在设计上能够以必要的准确性和可靠性执行定义明确的任务，以提高医疗卫生系统的能力并促进患者的利益。设计者和其他利益相关者还应能够预测和理解潜在的次生结果”[1]。甚至在开始开发多模态大模型之前，开发者就可以进行所谓的“事前剖析（pre-mortem）”[33]，以考虑“假设的故障”，这样开发团队就可以反向设计这些意料之外的故障。这样，开发者就能识别已知和未知风险，并制定替代方案[33]。通用基础模型的几位开发者提出的第二个建议是“红队”[129]，即对模型或系统进行评估，找出可能导致不良行为（如多模态大模型提供有偏见的意见）的真实世界模拟中的薄弱环节，这样开发者就可以修正模型或系统，确保其可靠性和安全性。一家企业宣布，它将在 2023 年 8 月向黑客大会 DEFCON 会议提交最新多模态大模型，以便“专家对其能力进行进一步分析和压力测试”[130]。

世卫组织先前的指南还建议，“设计者用于‘价值设计’的程序应参考并更新共识原则、最佳实践（如保护隐私的技术和技巧）、设计伦理标准以及不断发展的专业规范”[1]。适当的设计可以限制未经授权披露输入多模态大模型的数据，或解决与多模态大模型的训练和使用有关的环境（碳和水）问题（见下文）。它还可以确保用户知道多模态大模型生成的内容是由人工智能系统而非人类生成的，以避免取代人类的认识权威中心地位。这种通知可以提醒用户、社区和社会，虽然多模态大模型可以生成有用的信息，但它不能替代人类的知识生产。

尊重环境因素的设计：如上所述，多模态大模型的一个主要问题是其碳足迹和水足迹。开发者应采取一切可能的措施减少能源消耗，如提高模型的能效，一些大型科技企业正在尝试这种方法。例如，一家企业开发了一种与外部数据库相结合的多模态大模型，该数据库比多模态大模型运行效率高，使用更多变量进行训练，其性能优于能效较低的多模态大模型[112]。另一家企业正在试验一种不基于一个神经网络而是将变量分配给 64 个较小的神经网络的多模态大模型。经过训练，它可以只使用两个神经网络来完成每项任务，因此每次推理只使用一小部分变量[112]。

提高能效的另一种方法是开发较小的多模态大模型，这些多模态大模型在较小的数据集上进行训练，因此在训练或运行时不需要那么多能源。较小的多模态大模型不仅可以减少能源消耗，还可以为较小的企业或实体开发多模态大模型提供机会，并提高输出结果的准确性

[6] 世卫组织卫生领域人工智能伦理与治理专家 David Gruson 的来文。

[59]。小型多模态大模型对于开发“专用多模态大模型”可能特别有用，例如专门用于医疗、科学研究和医学研究的多模态大模型。一些大型科技企业已经推出了几种这样的多模态大模型[59]。

4.3 政府法律、政策和公共部门投资

可以执行或制定一些现有的或潜在的法律或政策，以减少或避免通用基础模型设计和开发过程中的风险。此外，政府可以进行公共部门投资，促进或支持通用基础模型的伦理设计与开发。

规范数据使用的法律和政策：世卫组织支持应用和执行包括数据保护规则在内的标准，这些标准对如何使用和将如何使用数据来训练多模态大模型作出了规定。数据保护法通常以基于权利的方法为基础，包括数据处理监管标准，既保护个人权利，又规定公共和私人数据控制者和处理者的义务，还包括对侵害法定权利行为的制裁和补救措施。数据保护法已在150多个国家通过，这为包括多模态大模型在内的所有人工智能技术的发展奠定了坚实的基础[1]。数据保护法的局限性在于，大多数数据保护法是在人工智能的生成及其他类型和用途出现之前颁布的，数据保护机构可能不愿意过于激进地适用这些法律，因为最初的法律可能并没有这样的意图[120]。

一项应当执行的数据保护要求是，数据的收集和处理必须合法，尤其是用于训练多模态大模型的医疗数据。这通常需要数据主体提供有效的知情同意，同意将其数据用于所述目的。任何后续处理都应有法律依据，因为不能假定后续处理与最初的目的相一致。开发和发布多模态大模型的企业和其他实体已经受到审查，因为它们可能会使用在未获得知情同意的情况下获取的数据。开发越来越大的多模态大模型需要越来越大的数据集，这可能会导致开发者忽视法律要求[83]，这也违反了世卫组织保护人类的自主性的指导原则。因此，世卫组织专家组建议各国政府“为使用健康数据和保护个人权利制定明确的数据保护法律法规，包括获得有意义的知情同意的权利”。

政府监督和规范用于训练多模态大模型的数据收集和使用的其他措施包括中国政府颁布的、于2023年8月生效的生成式人工智能法规。中国国家互联网信息办公室规定了几项义务，包括（i）提供者应采取有效措施，避免在选择训练数据时出现歧视和偏见；（ii）提供者应使用清晰的标签，并评估数据标签的质量；（iii）开发者应采取“有效措施”，实现数据的真实性、准确性、客观性和多样性目标[131]。这些要求预计不会严格适用于只需要采取有效措施确保适当的数据质量的企业。这些措施只适用于向中国公众提供服务的企业[132]。

与数据有关的立法规定可包括以下要求：描述用于训练基础模型的数据来源，以及使用受数据治理约束的数据，以确保这些数据具有适用性和偏见纠正的措施[133]。

政府在设计和开发过程中可以采取的其他措施如下：

- *目标产品简介*：各国政府和国际机构可发布目标产品简介，说明拟用于卫生保健与药品领域的多模态大模型的偏好和特点，特别是如果政府打算购买此类技术用于政府管理的卫生系统。
- *设计和开发标准与要求*：政府可要求开发者确保通用基础模型的设计和开发在其整个生命周期内实现某些成果。其中可包括对模型的可预测性、可解释性、可追溯性、安全性和网络安全的要求[134]。
- *预认证程序*：监管机构可引入法律义务并制定激励措施，要求并鼓励开发者通过包括预认证程序在内的措施，识别并避免伦理风险，如偏见或损害自主权[1]。世卫组织先前的人工智能伦理指南提出建议，“政府监管机构应激励开发者在产品设计和开发过程中识别、监测和解决相关的安全和人权问题，并应将相关准则纳入预认证计划”[1]。

- *审计*：政府可以对基础模型的初始发展阶段进行审计。一项建议提出了三类审计：对多模态大模型提供者的“治理审计”，对多模态大模型的审计，以及对建立在多模态大模型基础上的下游产品和服务的“应用审计”。其中，应用审计不适用于多模态大模型的开发阶段 [121]。可将审计纳入对卫生保健与药品领域多模态大模型的审评要求中（见下文）。为使审计有效，应对审计质量进行评估，以确保审计达到预期目的。
- *环境足迹*：政府可要求通用基础模型的开发者解决其碳足迹和水足迹问题。例如，政府可以要求开发者测量其能源消耗，减少训练过程中的能源消耗 [133]，并达到尚未定义的环境标准 [134]。
- *对多模态大模型生成内容为“机器生成”予以明确告知*：各国政府可要求开发者确保在部署通用基础模型时，向终端用户发出通知和提醒，说明内容是由机器而非人类生成的 [133]。
- 各国政府还可以考虑要求或鼓励开发者在早期阶段对用于卫生保健与药品领域的人工智能算法或系统予以注册。早期注册可以鼓励发表负面结果，防止发表偏见或对结果过于乐观的解释，并促进有利于患者的知识整合。

为公众利益开发多模态大模型的公共基础设施：随着多模态大模型在卫生领域的应用不断扩大，包括算力和公共数据集，可以通过提供非营利或公共基础设施，鼓励开发符合伦理原则的多模态大模型数据集。公共、私营和非营利部门的开发者都可以利用这种基础设施，可以要求用户在获得使用权后遵守道德原则和符合价值观。它还有助于避免开发者对多模态大模型的独家控制，并在无法使用此类基础设施和资源的最大企业与开发者之间“公平竞争”。

受独立监督的政府可以建设基础设施，然后由开发者用来开发医疗多模态大模型。例如，一个由 1000 名学术志愿者、Hugging Face 企业等组成的国际团队在法国政府的资助下，训练了一个名为 BLOOM 的多模态大模型，其中包含 1,750 亿个参数，训练这些参数的计算时间需要耗费 700 万美元的成本 [112]。

努力创造公平的竞争环境也适用于学术界及其资源劣势。加拿大政府的国家高级研究计算平台（Advanced Research Computing Platform）服务于该国的学术部门；中国政府批准了一个国家计算能力网络系统，使学者和其他人能够访问数据和算力；美国国家人工智能研究资源（National AI Research Resource）工作组“提议创建公共研究云和公共数据集”[101]。欧洲民间团体也呼吁政府在建立所谓的“欧洲大型生成式模型”方面发挥更坚实的作用，政府将为此提供特定的人工智能计算、数据基础设施、科学和研究支持[135]。

4.4 开源通用基础模型（多模态大模型）

开源多模态大模型在整合伦理原则和应对已知风险方面的作用尚不确定。一般来说，通过使用开源软件设计人工智能技术或公开软件源代码，可以提高透明度和参与度 [1]。开源软件的贡献和反馈都是开放的，这使用户能够了解系统如何工作，发现潜在问题，并对软件进行扩展和调整 [1]。开源多模态大模型可为解决医疗多模态大模型的一些问题提供机会。由于开源模型既非专有也非封闭，它们允许较小的企业和实体（如非营利机构）以较低的成本设计多模态大模型[136]。建立在开源模型基础上的多模态大模型可以接受严格审查，因为代码和数据都可供审查。用户社区的参与和监督有助于确保开源模型的长期稳健性 [136]。

但是，如果以前提供模型的大型科技企业选择不再继续提供模型，那么开源多模态大模型可能无法持续下去 [10]。大多数开源多模态大模型的开发都是基于 Meta（前身为 Facebook）有限发布的多模态大模型 [10]。自多模态大模型及其权重被泄露[137]以来，该企业已表明其对开源方法的承诺，并指出开放性“会带来更好的产品、更快的创新和繁荣的市场，使 Meta 受

益，也使许多其他……最终，开放是消除关于人工智能的恐惧的最佳良药”[130]。然而，独立观察人士注意到，虽然 Meta 企业以非商业方式提供其多模态大模型，但其使用条款存在限制条件，因此其提供多模态大模型的方式并不符合开源原则[138] [139]。

开发者很难满足使用开源模型监测其性能和结果的额外要求；然而，这些模型的益处并不能取代监管、避免危害，例如与使用开源模型相关的安全问题[140]。开源模型容易被滥用[141]，其漏洞也可能被用以网络攻击[142]。一组研究人员最近发现，在开源人工智能系统上测试的方法规避了人工智能安全措施和保障措施，也可以绕过所谓封闭系统的保障措施[143]。归根结底，开源模型是基于与其他多模态大模型相同的黑箱技术。

鼓励开源多模态大模型的一种方法是，政府要求利用政府资金或知识产权建立的基础模型能够被广泛接入，与政府要求开放政府资助的研究成果一样。政府还可以鼓励在公共设施中进行开源研发，包括在有公共监督的受控条件下进行下一代模型的研发。公共监督和参与可能比 Meta 泄露的允许任何人“下载并在 MacBook M2 上运行”的模型的新现实更好[144]。

建议：

- 设计应当或可以用于医疗、科学研究或医药领域的多模态大模型的开发者应考虑对程序员进行伦理认证或培训。这将使人工智能开发者符合医疗行业的要求，并增加对其产品和服务的信任。
- 无论数据集的大小如何，开发者在处理这些数据之前都应进行“数据保护影响评估”，这将要求开发者评估数据处理操作侵害个人权利和自由的风险，以及对个人数据保护的影响。
- 开发者应根据数据保护规则的最佳实践，对收集的数据进行多模态大模型训练。
- 用于训练多模态大模型的所有数据集，无论是开发者直接收集的还是通过第三方收集的，都应保持更新并适合系统的使用环境。
- 开发者应对用于训练模型的数据保持透明，以便对多模态大模型进行微调、使用多模态大模型开发医疗应用或直接使用多模态大模型的用户了解训练数据集的任何不足或不完整之处。
- 开发者应向数据工作者支付维持生计的工资，并为他们提供心理健康服务和其他形式的咨询。开发者还应采取保障措施，保护工人免受任何困扰。各国政府应更新劳动标准，将这些福利扩展到所有数据工人，促进各企业之间的“公平竞争环境”，并确保这些劳动标准随着时间的推移得到维护和改进。
- 开发者应确保多模态大模型不仅由科学家和工程师设计，潜在用户以及所有直接和间接的利益相关者，包括医疗提供者、科学研究人员、医疗保健专业人员和患者，应从人工智能开发的早期阶段就参与到结构化、包容性和透明的设计中来，并应有机会提出伦理问题、表达担忧和为可能采用的人工智能应用提供意见。这种意见可通过“人类监督协会”提供。
- 开发者应确保多模态大模型的设计能以必要的准确性和可靠性执行明确界定的任务，以提高卫生系统的能力，促进患者的利益。开发者还应能够预测和了解潜在的次生结果。满足这些要求的技术包括“事前剖析”和“红队”。
- 开发者用于“价值设计”的程序应参考并更新共识、最佳实践（如保护隐私的技术和技巧）、设计伦理标准和不断发展的专业规范，如披露多模态大模型生成的内容是由人工智能系统生成的。
- 开发者应采取一切可能的措施减少能源消耗（如提高模型的能效）。
- 各国政府应为医疗数据的使用制定适用于多模态大模型开发的强有力的、强制执行的数据保护法律法规。这些法律法规必须能够有效保护个人的权利，并为

个人提供保护自身权利所需的工具，包括获得有意义的知情同意的权利。对于为医疗多模态大模型而收集和处理的数据，可能还需要更多的工具。

- 各国政府和世卫组织等国际机构应发布“目标产品简介”，对医疗多模态大模型的偏好和特点进行界定，特别是如果政府预计最终会购买此类工具用于政府运营的卫生系统。
- 政府应要求开发者确保通用基础模型的设计和开发能在产品生命周期内取得某些成果。其中可包括对模型的可预测性、可解释性、可追溯性、安全性和网络安全性的要求。
- 监管机构应引入法律义务并建立激励机制，如预认证程序，要求并鼓励开发者识别并避免伦理风险，包括偏见或破坏自主性。
- 各国政府应对基础模型开发的初始阶段进行审计。
- 政府应要求通用基础模型的开发者解决通用基础模型的碳足迹和水足迹问题。
- 各国政府应要求开发者确保在使用通用基础模型时，通知并提醒用户内容是由机器而非人类生成的。
- 各国政府应考虑要求或制定激励措施，鼓励开发者在早期阶段对用于卫生保健与药品领域人工智能算法或系统予以注册。早期注册可鼓励发表负面结果，防止发表偏见或对结果的过度乐观解释，并可促进纳入对患者有益的知识。
- 政府应投资或提供非营利或公共基础设施，包括算力和公共数据集，供公共、私营和非营利部门的开发者使用，要求用户遵守伦理原则和价值观，以换取访问的权利。
- 各国政府应鼓励开发开源多模态大模型，要求利用政府资金或知识产权建立的基础模型能够被广泛访问，就像各国政府要求对政府资助的研究开放访问一样。各国政府应支持在公共设施中进行开源研发，包括在公共监督的受控条件下进行下一代模型的研发。

5 通用基础模型（多模态大模型）的提供

通用基础模型的用途取决于是用户提示多模态大模型生成与医疗保健相关的输出，还是开发者允许提供者将多模态大模型集成到与医疗保健相关的应用软件、产品或服务中。这两种情况都会带来新的风险，都必须由开发者、提供者或者两方一起共同解决。政府有责任在部署此类技术之前评估和规范其使用。

5.1 使用通用基础模型（多模态大模型）提供医疗服务或应用程序时应注意的风险

对于通用基础模型和应用程序在用于医疗目的的产品中或由用户直接使用时是否都应该得到评估和批准，可能存在分歧。几家最大的科技企业一直在悄悄游说政府官员（例如欧盟）放弃多模态大模型的评估框架，转而将监督重点放在那些可能被政府认为“有风险”的应用上[145]。这既涉及开发和销售包含基础模型的医疗应用程序的提供者，也涉及选择直接或间接通过人工智能系统使用多模态大模型的用户，如提供者或患者。这些企业认为，针对多模态大模型本身的监管会将责任“完全转嫁”给开发者，然而价值链中的其他企业也应承担责任[145]。

让通用基础模型的开发者为多模态大模型的所有使用负责可能不合适，但让提供者、部署者或用户独自承担也不合适，因为他们没有参与模型的开发，可能不了解相关的限制和风险。这将造成多模态大模型的开发者逃避责任，尽管他们拥有巨大的权力、资源、监督和对多模态大模型的理解，而且将在治理人工智能技术用于健康的尝试中打开一个“巨大的漏洞”[145]。

开发人员可能会试图避免将多模态大模型用于医疗目的（或其他用途）。如果开发者不希望将多模态大模型用于医疗（特别是临床医药）目的，它可以通过阻止开发医学应用程序的实体在应用程序编程接口上使用（许可）多模态大模型来阻拦应用；当多模态大模型直接被使用者用于医疗的目的时，可以通过阻止查询或对包含健康或医疗信息的任何响应附加明确警告来阻止应用，并将用户引导至可以提供适当帮助的信息或服务。

如果不采取这些措施，或者如果开发者打算由用户直接或通过提供者间接将其多模态大模型应用于医疗保健，开发者将承担只有它才能履行的具体责任。此外，开发者和提供者都有进一步的解决卫生领域多模态大模型风险的义务。

下文所述的责任在政府授权的法律、政策和法规中得到界定，因为政府必须最终决定是否允许基于人工智能的系统用于卫生领域。如果多模态大模型要用于卫生领域，开发者和提供者也必须履行共同责任。如果没有制定或更新法律来解释这些责任，这些责任可以由政府界定，也可以由双方通过合同约定。

部署前必须解决的主要风险包括全系统偏见、医疗用途的虚假信息或幻觉、输入多模态大模型数据的隐私、操纵和自动化偏见。

5.2 政府可采取的应对此类风险的措施和应坚持的伦理原则

多模态大模型和包含多模态大模型的应用的发展速度要求各国政府迅速制定法规和具体标准，以便将这些人工智能算法用于卫生保健系统、其他科学和医学研究。该方法应涵盖由一个监管机构（如医疗器械或药品监管机构）对拟用于医疗或药品领域的人工智能技术进行评估和审批，当然政府也可为此目的设立一个新的机构。低收入和中等收入国家面临的一个挑战是，它们的监管机构已经资源不足，而且被药品监管压得喘不过气来。

已经有至少一个高收入国家的政府与最大的科技企业达成一致，将对其基础模型进行自

愿公开评估，并披露评估结果，以便向公众和研究人员提供有关模型的信息，并鼓励企业纠正任何错误[146]；然而，自愿的方式可能既不充分，也不持久。

对多模态大模型和应用的评估不应只针对卫生系统中使用的人工智能系统或算法，因为在临床和“健康”应用之间的灰色地带中使用多模态大模型和应用也存在重大风险。鉴于此类技术的迅速扩散，政府至少应在初期识别此类应用，制定共同的标准和法规，并禁止向公众部署不符合标准和法规的应用。

在需要时，开发者和提供者应承担举证责任，证明打算投入使用的卫生领域人工智能符合法律或政策规定的最低要求。鉴于与多模态大模型相关的已知风险和挑战，不应假定具有多模态大模型的人工智能算法和应用是安全有效的，或它们优于已经广泛使用的人工智能或非人工智能方法。

下文介绍了可能适用于在卫生保健与药品领域使用多模态大模型的若干法律、政策和整合的要求。

披露（透明度）要求：适当的监管不仅要求政府有能力和自由裁量权来决定他们可以评估和批准使用什么，还要求有足够的信息来进行这样的评估。披露对于充分监管人工智能技术和确保人工智能价值链中的其他参与者能够安全使用该技术都是必要的。例如，除非开发者披露通用基础模型的性能（如产生幻觉的倾向），否则提供者可能无法获得必要的信息来对模型进行微调或拒绝推销该技术。提供者或开发者的这种形式的披露也可以帮助医疗机构之类的用户，决定不使用可能提供错误信息的多模态大模型，或更仔细地检查输出结果。

公开和透明是世卫组织的指导原则，也是提高基于人工智能的系统的“可解释性”和可理解性的措施，在评估通用基础模型或应用时应要求公开和透明。世卫组织先前的指南[1]建议：“政府监管机构应要求人工智能技术某些方面的透明度，同时考虑到专有权，以改善对安全性和有效性的监督和评估。这可能包括人工智能技术的源代码、数据输入和分析方法”。与多模态大模型相关的新披露形式可能包括其内部测试表现及其碳足迹、水足迹。此外，可能还需要制定“开放权重”标准，让监管者、其他开发者、民间团体和提供者了解算法的训练结果，或多模态大模型在训练过程中获得的知识[147] [148]。

有几种形式的披露可以帮助提供者、使用者或监管者，包括描述基础模型的能力和局限性，根据公共或行业标准基准对模型进行评估，报告内部和外部测试的结果以及模型的优化[134]。一位研究人员将信息披露比作“营养标签”[129]，尤其是可能与多模态大模型或其应用相关的风险信息的披露。

数据保护法：多模态大模型的开发以及开发者管理训练多模态大模型所需的数据的方式可能会违反数据保护法。另一个问题是，可能包括敏感个人信息的、为产生特定输出而被输入多模态大模型或应用程序的数据，会因为意外或被诱导泄露。许多大企业，包括正在开发多模态大模型并将其商业化的科技企业，禁止自己的员工使用此类算法，原因就在于潜在的信息泄露风险[149]。

披露数据违反了开发者保护人类的自主性的义务。如果敏感数据的保存时间超过数据最小化要求所允许的时间，开发者还可能违反数据保护法[85]。有开发者允许用户选择删除他们为改进聊天机器人性能而提供的任何内容[150]。允许使用多模态大模型的政府应确保制定、扩展和执行数据保护规则，以涵盖输入多模态大模型的数据。中国政府对多模态大模型的规定中包含了这样的要求，尽管这种保护只适用于中国的用户[151]。

评估卫生领域通用基础模型和/或应用：人权法与基于风险的对比框架：目前正在制定若干立法框架，以评估和规范人工智能技术。与这些框架有关的一个问题是，人工智能技术是否必须满足人权义务（欧盟认为是“基本权利”），或者是否应采用不同的方法，如在基于风险的框架内评估人工智能技术。欧盟《人工智能法案》采用了基于风险的框架[152]。有人认为，基于风险的框架可以有助于确定对某项技术必须提出的要求或举证责任，举证责任随

着技术的风险程度而增加。

所有卫生领域人工智能系统或工具都应尊重影响个人尊严、自主权或隐私等方面的伦理义务和人权标准。这包括通用基础模型。人权和伦理原则是不容讨价还价的，无论人工智能技术带来的风险或受益如何，都必须坚持[153]。人工智能算法被认为是“低风险”的，但这并不能使其免于审查，开发者或提供者应确保算法尊重人权和道德义务。可以进行人权影响评估，以确定多模态大模型或应用程序是否遵守这些承诺，从而可以安全使用。

世卫组织先前的指南[1]建议，“各国政府应颁布法律和政策，要求政府机构和企业对人工智能技术进行影响评估，评估应涉及伦理、人权、安全和数据保护，并贯穿人工智能系统的整个生命周期”。该指南还指出，“在引入人工智能技术前后，应由独立的第三方对影响评估进行审核，并予以公布”[1]。影响评估的结果应公开披露，同时考虑到专有或敏感信息，并应提供给公众和可能受影响的群体。与审计相同（见上文），影响评估可能需要仔细审查、特别是如果影响评估是由提供工具或服务的第三方进行的，以确保其质量和严谨性。

例如，影响评估可以揭示人工智能技术是否会带来全系统范围的偏见，是否会危及共享个人数据的用户的隐私，或是否会导致用户被操纵。应通过提供者和开发者之间的合作来解决隐私风险问题，开发能保护个人隐私的多模态大模型。美国的一家企业和一家医院系统正在开发这种多模态大模型，但由于数据无法完全去标识化，该项目被认为不太可能成功[154]。影响评估还可以确保通用基础模型或应用程序的使用保持人类在体系中，以避免用户在自动决策过程中，收到虚假信息或错误信息，或医疗服务提供者或患者不加批判地依赖多模态大模型的输出，以至于产生自动化偏见。

相反，政府可能会选择对卫生保健与药品领域多模态大模型采用基于风险的框架。对于那些被认为风险较高的功能，如为严重抑郁症患者提供处方或心理健康建议，或为弱势或边缘人群使用人工智能技术，举证责任会更重。有人担心，如果政府选择基于风险的方法，它将被认为是充分的，或被用来替代基于人权的方法[153]。基于风险的框架可能会将某些看似低风险但最终可能导致伤害的多模态大模型或应用排除在评估之外。

其他问题包括人工智能监管评估是否应适用于基础模型，无论其最终使用情况如何，评估是否应仅适用于规模最大、使用最广泛的基础模型（“系统性基础模型”[7]）；以及此类评估何时应适用于提供者。

本指南不包括关于是否所有基础模型都应接受基于风险和/或基于权利的评估的建议，无论其如何使用。本指南也没有给出对通用基础模型的人工智能监管评估是否应仅适用于最大的（系统性）多模态大模型的建议。专家组确实注意到，对于适用于所有多模态大模型（无论规模大小）的评估，有一个担忧，即它可能会“冻结”最大企业的主导地位，因为标准可能是这样的：只有这些企业能够可行地遵守这些标准，或者这些标准最适合它们的商业模式和目标[155]。这种担忧已经引起了竞争管理机构的注意，它们发现，先行者可能会使用“不公平的竞争方法来巩固其现有实力，或利用这种实力来控制新的人工智能生成市场”[141]。竞争管理机构预计将对多模态大模型的使用进行更严格的审查，尽管它们关注的是开发多模态大模型的企业所使用的方法[156]。

提供者也应接受人工智能监管评估，因为他们对多模态大模型的使用可能会改变其目的和功能，使之与由开发者确定但由提供者控制的目的和功能不同。因此，如果一个通用基础模型经提供者调整后用于医疗或药品领域，且开发者也同意，那么开发者和提供者都应遵守在卫生保健与药品领域使用多模态大模型的要求。如果医疗服务提供者使用的产品或应用程序与基础模型有重大差异，或以超出开发者控制范围的方式改变了基础模型，则医疗服务提供者的监管负担应更重。

医疗器械监管：政府可认定通用基础模型或应用程序符合医疗器械的条件。虽然关于哪

[7] 2023 年 6 月 5 日，欧洲议会 Axel Voss 办公室主任 Kai Zenner 在 AI 向善会议上的介绍。

些多模态大模型符合医疗器械的标准几乎没有指导，但一家监管机构表示，“仅面向一般用途的多模态大模型，如果开发者没有声称其软件可用于医疗目的，则不太可能符合医疗器械的标准”[157]。然而，监管机构也指出：“为特定医疗目的而开发或改编、修改或定向的多模态大模型有可能被认定为医疗器械。此外，如果开发者声称其多模态大模型可用于医疗目的，这也可能意味着产品符合医疗器械的条件”[157]。

根据欧盟和美国现行监管标准，基于多模态大模型的聊天机器人提供医疗建议很可能被定性为医疗器械[158]。世卫组织先前的指南[1]建议“政府监管机构应要求对人工智能系统的性能进行测试，并从随机试验的前瞻性测试中获得可靠证据，而不仅仅是将该系统与实验室中的现有数据集进行比较”。

如果多模态大模型或其应用要被作为医疗器械进行监管，开发者和/或提供者应承担举证责任，提供证据证明该器械具有上市需要的性能，并符合现行或修订后国家法律的要求。这可能包括各种要求，如遵守与偏见和隐私相关的道德义务。欧盟和美国新提出的医疗器械人工智能技术法规可能会纳入与卫生领域人工智能应用相关的伦理原则，包括“可解释性”、控制偏见和透明度。目前包含多模态大模型的聊天机器人不太可能达到这些标准[158]。

用于辅助临床决策的多模态大模型已被用于实验。虽然这些多模态大模型有免责声明，但它们并不妨碍医疗器械法的应用，“医疗器械法规定，此类实验只能在经授权的临床试验中进行，并在适当的控制下保护患者和产生临床相关结果”[158]。各国政府可在监管沙盒中审查此类多模态大模型的受控实验用途，这将允许在实际临床环境中进行测试，并有保障措施和监督，以保护医疗卫生系统免受风险或意外后果的影响。不过，只有在新的医疗产品和服务及其规格受正式监管和数据保护规定约束的国家，才适合使用这种方法[1]。

消费者保护法：各国政府应制定和使用消费者保护法，以确保多模态大模型和应用程序的任何负面影响不会波及用户和患者。例如，消费者保护法可用于防止类似于操纵的行为[159]。在美国，一些政府部门和机构正在实施消费者保护法和其他法规，以防止自动化系统中的歧视和偏见[159]。此类法律可成为政府要求寻求将此类技术商业化的实体解决任何负面后果的依据，并保护患者及其家属免受当前或未来的任何伤害[93]。可以利用消费者保护法或其他法规，要求限制多模态大模型和应用程序使用可能误导最终用户或使其误认为多模态大模型具有类似人类品质的语言。因此，此类法律可以限制或防止多模态大模型或其应用程序使用“我认为”、“我想”或“我建议”等词语。

建议：

- 各国政府应在资源允许的情况下，指定一个现有的或新的监管机构来评估和批准拟用于医疗或药品领域的多模态大模型及其应用软件。
- 多模态大模型及其应用软件的某些方面应该透明，以便监管机构对其安全性和有效性进行监督和评估。这可能包括源代码、数据输入、模型权重和分析方法。政府应考虑的其他披露形式包括多模态大模型或应用在内部测试中的表现及其碳足迹和水足迹。
- 政府应确保数据保护规则适用于用户输入多模态大模型或应用软件的数据。
- 政府的法律、政策和法规应确保卫生领域多模态大模型和应用软件，无论人工智能技术带来的风险或受益如何，都符合可能影响个人尊严、自主权或隐私等方面的道德义务和人权标准。
- 各国政府应颁布法律和政策，要求提供者和开发者在人工智能系统的整个生命周期内对多模态大模型和应用程序进行影响评估，其中应涉及伦理、人权、安全和数据保护。影响评估应在引入人工智能技术之前和之后由独立的第三方进行审

计，并向公众公开。

- 如果产品或应用软件在很大程度上偏离或改变了基础模型，而这种偏离或改变又超出了模型开发者的控制范围，那么提供者的监管负担就应当加重。
- 各国政府应确保，对于作为医疗器械进行监管的多模态大模型或应用软件，开发者和/或提供者应负举证责任，证明器械的性能符合上市要求，并符合国家法律或经修订的法律的要求。
- 各国政府应确保尚未获准使用的多模态大模型或支持临床决策的应用软件不得在授权临床试验环境之外的实验基础上使用。各国政府可通过监管沙盒促进多模态大模型的受控实验使用。沙盒允许在实际临床环境中的实时环境内进行测试，并提供保障和监督，以保护卫生系统免受风险或意外后果的影响。
- 各国政府应利用消费者保护法，确保使用多模态大模型和应用软件的任何负面后果不会影响包括患者在内的用户。例如，可以利用消费者保护法来防止类似于操纵的行为，或者解决多模态大模型或应用软件造成其他负面后果的原因，以保护患者及其家人免受当前或未来的任何伤害。

6 通用基础模型（多模态大模型）的部署

即使多模态大模型或具有多模态大模型的应用的设计符合伦理规范并经过了适当的监管审查，它在商业化时仍然可能存在风险。卫生领域人工智能应用或工具的部署者可以是多模态大模型或应用的开发者或提供者，也可以是卫生部门、医院、医疗保健企业或制药企业。

6.1 部署通用基础模型（多模态大模型）医疗服务或应用时应注意的风险

部署期间的风险可能源于多模态大模型及其提供回答的不可预测性、以开发者和提供者无法预料到的方式使用通用基础模型的可能性，以及多模态大模型生成的回答可能会随着时间的推移而变化。

部署多模态大模型时必须解决的主要风险有：

- 不准确或错误的回答；
- 偏见；
- 多模态大模型输入和输出数据的隐私；
- 多模态大模型的可访问性和可负担性；
- 对劳动力和就业的影响；
- 自动化偏见和技能退化；
- 医疗服务提供者与患者之间的互动质量。

本章介绍包括用户在内的人工智能价值链中的行为者如何减轻或预防风险，以及在人工智能工具部署后的使用过程中政府监管所能发挥的作用，同时对医疗工作者和卫生系统中的其他行为者进行装备和培训，以最大程度地适当使用多模态大模型。

6.2 开发者和提供者在部署阶段中的持续责任

即使在多模态大模型或应用程序获准使用后，开发者和提供者仍有责任和义务，这是因为开发者或提供者部署了多模态大模型，或是因为某些风险只能由开发者或提供者在部署后解决。这种义务可能必须由法规或法律来规定，以保证开发者和提供者分配足够的资源和注意力。

第一，当大规模部署多模态大模型时，政府应引入发布后强制性审计和影响评估，包括由独立的第三方进行数据保护和人权评估[155] [160]。发布后的审计和影响评估应予以公布，并应按用户类型（如年龄、种族或残障）分列结果和影响。

第二，如果多模态大模型在发布后出现不准确、虚假或有害内容，而提供者和开发者均未采取措施进行纠正或避免，政府可追究提供者或开发者的责任。例如，中国政府关于生成式人工智能的法规规定，生成式人工智能不得产生“虚假和有害”的信息[151]，政府可能会强制执行。在欧盟，在产品或服务中纳入多模态大模型可能会给多模态大模型的开发者和提供者带来额外的责任。例如，如果将多模态大模型集成到属于《欧盟数字服务法》等数字服务监管范围的服务中，多模态大模型将间接受到监管审查，由于多模态大模型容易产生幻觉，因此可能需要监管监督[120]。

第三，可能要求开发者和提供者持续提供操作披露，以便政府和用户安全使用多模态大模型。这可能包括充分的技术文档[133] [134]。

6.3 部署者的职责

部署者还负责避免或减轻与使用多模态大模型或其应用程序相关的风险。

首先，部署者应该使用开发者或提供者提供的信息，如训练数据的偏见、使多模态大模

型不适合该环境的场景偏移、其他可避免的错误或部署者已知的潜在风险，来做出禁止在不适当的环境中使用多模态大模型或应用程序的决定。如果部署者收到关于此类风险的明确、充分的警告，却仍将多模态大模型用于不适当的环境，部署者应该对由此产生的一切损害负责。

其次，部署者应该告知他们理应知道的使用多模态大模型可能导致的一切风险，以及对用户可能造成损害的一切错误或失误。这些警告不应采用过小的字体，也不应是容易忽视的内容。在某些情况下，即使法律或法规没有要求，部署者也可能有责任暂停使用或从市场上撤下多模态大模型或应用程序以避免伤害。

第三，部署者可以采取措施提高多模态大模型的可负担性和可接入性。部署者可以确保使用多模态大模型的定价或订阅费符合政府或其他用户的支付能力，并应保证多模态大模型经过了被忽视或被排除在技术利益之外的人使用的语言文字的训练。部署者还应要求提供者和开发者确保当前和未来的多模态大模型有多种语言版本。

6.4 政府计划和做法

将多模态大模型引入卫生保健系统并用于其他与医疗有关的用途，需要医疗保健专业人员作出重大调整。无论是开发者还是提供者，都没有兴趣、资源或专业知识来确保医疗保健专业人员恰当使用多模态大模型，或将其用于其他涉及受过专门培训和/或具有专门知识的个人用途。

与设计通用基础模型一样（见上文），政府可以让医疗保健专业人员和患者共同参与“人类监督协会”，以确保新的多模态大模型和用于临床决策的应用程序得到适当使用，并且不会损害患者的权利。[8]

政府、高校（健康研究机构）或医院等医疗提供者也可确保医疗工作者有效地使用多模态大模型提供临床护理，并适当接受其他用途的培训。医护专业人员和临床医生应接受以下方面的培训：（i）了解多模态大模型如何做出决定以及做出此类决定的局限性，（ii）对恰当使用担忧进行识别，（iii）掌握避免自动化偏见的方法，（iv）接触可能或正在考虑使用多模态大模型的患者并对其进行教育，（v）使用多模态大模型的相关网络安全风险[161]。

在医务工作者进行培训和继续教育中，特别重要的是对于患者、非专业人员和其他第三方何时由多模态大模型提供建议或多模态大模型提供的信息已被提供者用于做出医疗决定或其他医疗功能的告知。在此类告知中，应向患者或非专业人员充分告知使用多模态大模型的相关风险，以维护其知情同意权。

对于医疗工作者的培训也至关重要，以确保当他们专业使用多模态大模型时，他们不会在不知不觉中违反法律，尤其是那些与保护健康数据和信息相关的法律。例如，将“受保护的健康信息”引入多模态大模型聊天机器人的医疗服务提供者可能会违反法律，如美国的《健康保险携带和责任法案》等法律 [150]。例如，随着受欢迎的多模态大模型得到医疗保健专业人员的“信任”，它们可能会披露比自己意识到的更多的患者数据 [154]。

卫生系统中的其他利益攸关方应接受教育，了解多模态大模型在医疗保健中的益处、风险、用途和挑战，以及多模态大模型与其他生成信息或建议的技术有何不同，以及在医疗保健中又是如何用于其他目的的。应提高公众对医疗和其他领域多模态大模型的使用的认识。世卫组织先前的指南[1 建议]：“公众应参与卫生领域人工智能的发展，以了解数据共享和使用的形式，对社会和文化上可接受的人工智能形式发表意见，并充分表达他们的关切和期望。此外，应提高公众对人工智能技术的认识，使他们能够确定哪些人工智能技术是可以接受的”。

向卫生系统提供多模态大模型或应用软件的政府可利用其采购权，在开发者、提供者和

[8] 世卫组织卫生领域人工智能伦理治理专家 David Gruson 的来文。

部署者之间促进某些做法。如果人工智能技术不会取代其他可能更有效、更公平和更负担得起的医疗投资，那么采购关键的多模态大模型或应用软件用于医疗保健系统将消除接入和可负担性的障碍。公共采购可以在数据训练、质量保证、风险评估、缓解和外部审计方面制定透明度要求。如果一个国家既没有相关立法，也没有具备有效监管多模态大模型资源的监管机构，这些要求可能至关重要。

建议：

- 在大规模部署多模态大模型时，各国政府应引入独立第三方进行发布后强制性审计和影响评估，包括数据保护和人权评估。审计和影响评估应予以公布，并应包括按用户类型分类的结果和影响，例如按年龄、种族或残障分类的结果和影响。
- 政府可要求提供者或开发者对多模态大模型发布后未纠正或避免发布不准确、虚假或有毒的内容负责。
- 政府应要求开发者和提供者不断披露操作信息，以确保多模态大模型和应用程序的安全使用。其中可包括充分的技术文件。
- 根据从开发者或提供者处获得的信息，部署者不应在以下情况下使用多模态大模型或应用程序，由于训练数据的偏见、使多模态大模型不适合特定环境的场景偏差，或部署者已知并可避免的其他潜在错误或风险，如多模态大模型发布的不准确、虚假或有害内容。
- 部署者应告知他们应合理知道的使用多模态大模型可能导致的任何风险，以及已对用户造成的伤害的错误；此类警告不应以过小（或容易遗漏）的字体出现。在某些情况下，即使法律法规没有要求，部署者也可能有责任暂停使用或从市场上撤下多模态大模型或应用程序，以避免未来的危害。
- 部署者应确保使用多模态大模型的定价或订阅费符合政府或其他用户的支付能力，并应保证多模态大模型经过了被忽视或被排除在技术利益之外的人使用的语言文字的训练。部署者应要求提供者和开发者确保目前和未来的多模态大模型都是用多种语言开发的。
- 各国政府应为医护专业人员和患者参与“人类监督协会”提供便利，以确保新的多模态大模型和用于临床决策的应用程序得到适当使用，并且不会损害患者的权利。
- 卫生部门和高校（健康研究机构）应培训医疗保健专业人员和临床医生：（i）了解多模态大模型如何做出决定（以及了解这些决定是如何做出的局限性），（ii）识别和了解对合理使用的担忧，（iii）掌握避免自动化偏见的方法，（iv）与可能使用多模态大模型的患者接触并对其进行教育，（v）与使用多模态大模型相关的网络安全风险。
- 政府、医疗服务提供者、医疗研究人员和资助者应让公众参与进来，使他们了解不同形式的数据共享和使用，就社会和文化是否以及如何接受多模态大模型发表意见，并充分表达他们的关切和期望。此外，应提高公众对人工智能技术的认识，使他们能够识别可接受的多模态大模型用途和类型。
- 通过卫生系统提供多模态大模型或应用程序的政府应确保其采购部门促进开发者、提供者和部署者采取某些做法，包括透明度。

7 通用基础模型（多模态大模型）的责任

随着多模态大模型在卫生保健与药品领域得到更广泛的应用，错误、滥用和最终对个人的伤害是不可避免的。必须使用问责制来补偿个人的此类伤害，并在当前方法不充分或过时的情况下建立新的补救机制。

人工智能技术的设计、开发、质保和部署涉及不同的实体，每个实体都扮演着不同的角色。这可能会使责任分配复杂化。开发者可能会要求包括提供者和部署者在内的下游实体对使用多模态大模型造成的一切伤害负责，而下游实体可能会声称先前的行为，如用于训练算法的数据的选择，是损害出现的原因。开发者和提供者也可能声称，一旦医学人工智能技术被监管机构批准使用，他们就不应该再对损害承担责任（预防性监管）[1]。沿着价值链建立责任是立法者和决策者面临的挑战。

民事责任规则的一个关键功能是确保损害受害者可以要求赔偿和补救，无论在参与人工智能技术开发和部署的实体之间分配职责和责任有多么困难。如果受害者发现获得赔偿太难，就毫无正义可言，人工智能价值链中的各方也没有动力在未来避免类似的伤害了。规则还应该确保赔偿足以弥补所遭受的伤害。

欧盟在其拟议的《人工智能责任指令》（AI Liability Directive）中引入了“因果关系推定”[162]，从而简化了受害者的举证责任。因此，如果受害者能够证明一个或多个实体没有遵守与损害相关的义务，并且很可能与人工智能的表现有因果关系，法院可以推定不遵守该义务是损害发生的原因 [162]。因此，责任方有责任反驳这一推定，例如指出另一方是损害的原因。立法的范围不仅限于人工智能系统的原始制造商，还包括人工智能价值链的任何行为者[162]。当人工智能价值链中的所有行为者都承担连带责任时，他们可以证明他们在评估和减轻风险以减少责任方面的有效性。

然而，责任机制仍有可能无法为人工智能驱动的产品和服务造成的伤害提供完全清晰的责任和补救措施，特别是如果个人不知道多模态大模型被用于做出医疗决策。新规则可能会在人工智能驱动的医疗技术造成的伤害的责任上留下空白[163]。由于多模态大模型的高度投机性、人们对其知之甚少且被匆忙推向市场，政府可能希望将用于医疗保健的多模态大模型视为开发者、提供者和部署者要求遵守严格责任标准的产品。让这些行为者对所有错误负责的做法，可能可以确保患者在错误影响到他们时得到赔偿[1]，尽管这取决于患者是否知道使用了多模态大模型。虽然这种持续的责任可能会阻碍越来越复杂的多模态大模型的使用，但它也可能会降低承担不必要风险的意愿，并在其许多风险和潜在危害被完全识别和解决之前，将新的多模态大模型部署到医疗保健或公共卫生环境中[1]。

然而，人工智能的问责制可能不足以分配过错，因为算法正在以开发者、提供者和部署者都无法完全控制的方式发展。此外，可能仍然存在受害者无法追偿的情况和司法管辖范围。例如，在美国，因直接使用多模态大模型寻求建议而受伤的患者可能无法获得损害赔偿，因为人工智能系统本身并不包括在职业责任规则中，而且产品或消费者责任法的例外或限制可能会排除赔偿[163]。在医疗保健的其他领域，有时会在不确定过错或责任的情况下提供赔偿，例如疫苗不良影响造成的医疗伤害。世卫组织先前指南建议确定“无过错、无责任赔偿基金，是否是向因使用人工智能技术而遭受医疗伤害的个人提供赔偿的适当机制，包括如何盘活资源来支付任何索赔”[1]。该建议今天也有效，可以作为确定多模态大模型或多模态大模型应用程序造成的损害赔偿的一种手段。

建议：

- 各国政府应沿着多模态大模型和应用程序的开发、供应和部署的价值链确立责任，以确保损害受害者能够要求赔偿，而不管追究责任的困难以及参与技术开

发和部署的不同实体的责任。

8 通用基础模型（多模态大模型）的国际治理

各国政府应支持集体制定国际规则，以治理卫生领域多模态大模型和其他形式的人工智能，因为此类用途的人工智能正在全球范围内激增。世卫组织的《2020 年—2025 年的数字健康全球战略》是一个例子。这一过程应包括加强联合国系统内合作与协作，以应对在卫生领域中部署人工智能及其在社会和经济领域更广泛应用中所带来的机遇和挑战。除非各国政府共同努力制定适当的、可执行的标准，否则不符合适当的法律、伦理和安全标准的多模态大模型和其他形式的人工智能的数量将会增加，如果不出台法规和其他类型的保护措施，或由于自愿或资源不足而没有得到完全的执行，就有可能造成危害。世卫组织最近与世界各地的监管机构协商发布了一份新出版物，概述了政府和监管机构在制定新的人工智能指南或调整现有指南时可以遵循的关键原则[164]。

国际治理可以避免寻求先发优势而忽视安全和效率标准的企业之间的“逐底竞争”，以及避免为争夺技术霸主地位而在地缘政治竞赛中寻求优势的政府之间的“逐底竞争”。因此，国际治理可以确保所有企业都符合安全和效率的最低标准，也可以避免出台为企业或政府提供竞争优势或劣势的法规。国际治理可以让政府对其投资和参与开发、部署的基于人工智能的系统负责，并确保政府制定尊重伦理原则、人权和国际法的适当法规。缺乏全球可执行的标准也可能对产品采用产生负面影响。

国际治理可以采取多种形式。有建议提出，建立一个由多个国家的政府资助的公共研究机构，像国际合作组织欧洲核子研究组织一样，利用资金和人力资源来开展大型变革性项目，并公开分享项目成果[165] [166]。在另一项提案中，有人建议设立一个实体，负责在高度安全的设施中开发最先进、风险最大的人工智能，从而使其他试图构建此类人工智能的尝试成为非法[167]。目前，这种大规模项目不属于公共资助的公益项目范畴，而属于存在商业竞争关系的各个大型科技企业的职权范围。包括世界各国领导人和技术高管在内的其他领导人呼吁将人工智能与核武器同等对待，建立一个类似于核武器使用条约的全球监管框架 [109]。

无论采取何种形式的国际治理，重要的是不能完全由高收入国家，或者说——主要与或仅与世界上最大的科技企业合作的高收入国家来制定[168]。由高收入国家和科技企业主导并为其制定的标准，无论是针对人工智能的普遍应用，还是针对多模态大模型在卫生保健与药品领域中的具体使用，都将使低收入和中等收入国家的大多数人在制定标准方面没有任何作用或发言权。这将使未来的人工智能技术在可能最终获益最多的国家变得危机四伏或无效。

正如联合国秘书长在 2019 年所提议的[169]，人工智能的国际治理可能需要所有利益攸关方通过网络化多边主义进行合作，这将使联合国大家庭、国际金融机构、区域组织、贸易集团和包括民间团体、城市、企业、地方当局和青年在内的其他方面，更加密切、有效和包容地合作。将伦理和人权置于多模态大模型开发和部署的核心位置可以为实现全民健康覆盖做出重大贡献。

建议：

- 各国政府应支持集体制定人工智能治理的国际规则。无论采取何种治理形式，都不能完全由高收入国家，或者说由主要或仅与世界上最大的科技企业合作的高收入国家来制定，因为这种做法将使低收入和中等收入国家的大多数人在制定人工智能国际治理方面无法发挥作用或失去发言权。

致谢

本世界卫生组织（WHO）指南的制定由 Andreas Reis（卫生研究部卫生伦理与治理的共同领导）和 Sameer Pujari（数字卫生保健与创新司）牵头，由 John Reeder（卫生研究部主任）、Alain Labrique（数字卫生保健与创新司司长）和 Jeremy Farrar（首席科学家）的全面指导。

Rohit Malpani（顾问，法国）为主要撰稿人。世卫组织卫生领域人工智能伦理与治理专家组共同主席 Effy Vayena（瑞士苏黎世联邦理工学院）和 Partha Majumder（印度统计研究所和印度国家生物医学基因组学研究所）为报告起草工作提供了全面指导，并领导了专家组的工作。

世卫组织感谢以下个人为本指南的制定做出了贡献。

世卫组织卫生领域人工智能伦理和治理专家组成员

Najeeb Al Shorbaji，约旦安曼电子健康发展协会；Maria Paz Canales，智利圣地亚哥全球合作伙伴数字组织；Arisa Ema，日本东京大学；Amel Ghouila，美国西雅图比尔及梅琳达·盖茨基金会；Jennifer Gibson，加拿大多伦多大学世卫组织生命伦理学合作中心；Kenneth Goodman，美国迈阿密大学米勒医学院生命伦理学与卫生政策研究所；Malavika Jayaram，新加坡数字亚洲中心；Daudi Jjingo，乌干达坎帕拉麦克雷雷大学；Tze Yun Leong，新加坡国立大学；Alex John London，美国匹兹堡卡内基梅隆大学；Partha Majumder，印度加尔各答印度统计研究所和国家生物医学基因组学研究所；Thilidzi Marwala，南非约翰内斯堡大学；Roli Mathur，印度班加罗尔印度医学研究理事会；Timo Minssen，丹麦哥本哈根大学法学院生命医学创新法高级研究中心；Andrew Morris，英国伦敦医疗数据研究；Daniela Paolotti，意大利都灵 ISI 基金会；Jerome Singh，南非德班夸祖鲁·纳塔尔大学；Jeroen van den Hoven，荷兰（王国）代尔夫特大学；Effy Vayena，瑞士苏黎世联邦理工学院；Robyn Whittaker，新西兰奥克兰大学；曾毅，中国北京中国科学院。

观察员

David Gruson，法国巴黎卢米尼斯企业（Luminess）；Lee Hibbard，欧洲委员会，法国斯特拉斯堡。

外部审稿人

Oren Asman，以色列特拉维夫大学；I. Glenn Cohen，美国波士顿哈佛法学院；Alexandrine Pirlot de Corbion，英国伦敦隐私国际组织；Rodrigo Lins，巴西累西腓伯南布哥联邦大学；Doug McNair，美国西雅图比尔及梅琳达·盖茨基金会综合发展部副主任；Keymanthri Moodley，南非开普敦斯坦陵布什大学；Amir Tal，以色列特拉维夫大学；Tom West，英国伦敦隐私国际组织。

外部贡献者

指南的方框 2（儿童使用多模态大模型的伦理考虑因素）由美国斯坦福大学的 Vijaytha Muralidharan、Alyssa Burgart、Roxana Daneshjou 和 Sherri Rose 起草。指南的方框 3（与多模态大模型有关的伦理考虑及其对残障人士的影响）由瑞士日内瓦的独立顾问 Yonah Welker 起草。

所有外部审稿人、专家和撰稿人都根据世卫组织的政策申报了他们的利益。经评估，所

申报的利益均不重大。

世界卫生组织

Shada Al-Salamah，日内瓦数字卫生保健与创新司技术干事；Mariam Otmani Del Barrio，日内瓦热带病研究特别规划项目科学家；Marcelo D'Agostino，华盛顿世卫组织美洲区域办事处信息系统与数字卫生部组长；Jeremy Farrar，日内瓦首席科学家；Clayton Hamilton，丹麦哥本哈根世卫组织欧洲区域办事处技术干事；Kanika Kalra，日内瓦数字卫生保健与创新司顾问；Ahmed Mohamed Amin Mandil，开罗世卫组织东地中海区域办事处研究和创新部协调员；Issa T. Matta，日内瓦法律事务办公室；Jose Eduardo Diaz Mendoza，日内瓦数字卫生保健与创新司顾问；Mohammed Hassan Nour，开罗世卫组织东地中海区域办事处数字卫生保健与创新司技术干事；Denise Schalet，日内瓦数字卫生保健与创新司技术干事；Yu Zhao，日内瓦数字卫生保健与创新司技术干事。

图形设计：Joanna Sleigh（苏黎世联邦理工学院）
排版：中央印刷局（卢森堡）

参考文献

1 Ethics and governance of artificial intelligence for health. Geneva: World Health Organization; 2021. (https://www.who.int/publications/i/item/9789240029200, accessed 26 May 2023).

2 Khullar D. Can A.I. treat mental illness? The New Yorker, 27 February 2023. (https://www.newyorker.com/magazine/2023/03/06/can-ai-treat-mental-illness, accessed 29 May 2023).

3 Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med. 2022;28(9):1773–84. doi:10.1038/s41591-022-01981-2.

4 Hariri Y, Harris T, Raskin A. You can have the blue pill or the red pill, and we're out of blue pills. The New York Times, 24 March 2023. (https://www.nytimes.com/2023/03/24/opinion/yuval-harari-ai-chatgpt.html, accessed 26 May 2023).

5 Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ et al. Foundation models for generalist medical artificial intelligence, Nature. 2023;616(7956):259–65. doi:10.1038/s41586-023-05881-4.

6 Hu K. Chat GPT sets record for fastest growing user-base – analyst note. Reuters, 2 February 2023. (https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/, accessed 26 May 2023).

7 Weise K, Grant N. Microsoft and Google unveil A.I. tools for businesses. The New York Times, 16 March 2023. (https://www.nytimes.com/2023/03/16/technology/microsoft-google-ai-tools-businesses.html, accessed 26 May 2023).

8 Yang Z. Chinese tech giant Baidu just released its answer to ChatGPT. MIT TechnologyReview, 16 March 2023. (https://www.technologyreview.com/2023/03/16/1069919/baidu-ernie-bot-chatgpt-launch/, accessed 26 May 2023).

9 Murgia M, Bradshaw T. Musk to launch AI start-up to rival ChatGPT. Financial Times, 15 April 2023. (https://www.ft.com/content/2a96995b-c799-4281-8b60-b235e84aefe4, accessed 26 May 2023).

10 Heaven WD. The open-source AI boom is built on Big Tech's handouts. How long will it last? MIT Technology Review, 12 May 2023. (https://www.technologyreview.com/2023/05/12/1072950/open-source-ai-google-openai-

eleuther-meta/, accessed 26 May 2023).
11 Martin A. Google CEO Sunder Pichai admits people don't fully understand how chatbot AI works, Evening Standard, 17 April 2023. (https://www.standard.co.uk/tech/google-ceo-sundar-pichai-understand-ai-chatbot-bard-b1074589.html, accessed 26 May 2023).
12 Roose K. A conversation with Bing's chatbot left me deeply unsettled. The New York Times, 16 February 2023. (https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html, accessed 26 May 2023).
13 Marcus G. AI platforms like ChatGPT are easy to use but potentially dangerous, Scientific American, 19 December 2022. (https://www.scientificamerican.com/article/ai-platforms-like-chatgpt-are-easy-to-use-but-also-potentially-dangerous/, accessed 26 May 2023)
14 Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E et al. Sparks of artificial general intelligence: early experiments with GPT-4. ArXiv:2302.12712.
15 McGowran L. OpenAI criticised for lack of transparency around ChatGPT-4. Silicon Republic, 16 March 2023. (https://www.siliconrepublic.com/machines/openai-gpt4-transparency-ai-concerns-stripe-chatgpt, accessed 26 May 2023).
16 Spitale G, Biller-Andorno N, Germani F. AI model GPT-3 (dis)informs us better than humans. Sci Adv. 2023;9(26):eadh1850. doi:10.1126/sciadv.adh1850.
17 Volpicelli G. ChatGPT broke the EU plan to regulate AI Politico, 3 March 2023. (https://www.politico.eu/article/eu-plan-regulate-chatgpt-openai-artificial-intelligence-act/, accessed 26 May 2023).
18 Arcesati R, Chang W. China is blazing a trail in regulating Generative AI – on the CCP's terms. The Diplomat, 28 April 2023. (https://thediplomat.com/2023/04/china-is-blazing-a-trail-in-regulating-generative-ai-on-the-ccps-terms/, accessed 26 May 2023).
19 Martindale J. These are the countries where ChatGPT is currently banned. Digital Trends, 12 April 2023. (https://www.digitaltrends.com/computing/these-countries-chatgpt-banned/, accessed 26 May 2023).
20 Johnson K. ChatGPT can help doctors – and hurt patients. Wired, 24 April 2023. (https://www.wired.com/story/chatgpt-can-help-doctors-and-hurt-patients/, accessed 28 May 2023).
21 Topol E. Multimodal AI for medicine, simplified. Ground Truths, 14 March 2023. (https://erictopol.substack.com/p/multimodal-ai-for-medicine-simplified, accessed 28 May 2023).
22 Heaven WD. AI hype is built on high test scores. Those tests are flawed. MIT Technology Review, 30 August 2023. (https://www.technologyreview.com/2023/08/30/1078670/large-language-models-arent-people-lets-stop-testing-them-like-they-were/, accessed 1 October 2023).
23 Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW et al. Large language models encode clinical knowledge. Nature. 2023;620:172–80. doi:10.1038/s41586-023-06291-2.
24 Kulkarni PA, Singh H. Artificial intelligence in clinical diagnosis: opportunities, challenges, and hype. JAMA. 2023;330(4):317–8. doi:10.1001/jama.2023.11440.
25 Subbamaran N. ChatGPT will see you now: Doctors using AI to answer patient questions. Wall Street Journal, 28 April 2023. (https://www.wsj.com/articles/dr-chatgpt-physicians-are-sending-patients-advice-using-ai-945cf60b, accessed 28 May 2023).
26 Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023; 183(6):589–96. doi: 10.1001/jamainternmed.2023.1838.
27 Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. New Engl J Med. 2023;388(13):1233–9. doi: 10.1056/NEJMsr2214184.
28 The potential of large language models in healthcare: Improving quality of care and patient outcomes, Medium, 7 December 2022. (https://medium.com/@BuildGP/the-potential-of-large-language-models-in-healthcare-improving-quality-of-care-and-patient-6e8b6262d5ca, accessed 28 May 2023).
29 Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T et al. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv. 2017;1711.05225vv3. doi:10.48550/arXiv.1711.05225.
30 Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C et al. A large language model for electronic health records. npj Digit Med. 2022;5:194. doi:10.1038/s41746-022-00742-2.

31 Ghahramani Z. Introducing PaLM 2. The Keyword, 10 May 2023. (https://blog.google/technology/ai/google-palm-2-ai-large-language-model/, accessed 28 May 2023).
32 Weise K, Metz C. When A.I. chatbots hallucinate. The New York Times, 9 May 2023. (https://www.nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html, accessed 1 June 2023).
33 Bender EM, Gebru T, McMillan-Major A, Mitchell M. On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, March 2021, pp. 610–23. doi: 10.1145/3442188.3445922.
34 Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. Lancet Digit Health. 2023;5(6):e333–5. doi:10.1016/s2589-7500(23)00083-3.
35 Metz, Cade. Chatbots may 'hallucinate' more often than many realize, The New York Times, 6 November 2023. (https://www.nytimes.com/2023/11/06/technology/chatbots-hallucination-rates.html, accessed 7 November 2023).
36 Acar OA. AI prompt engineering isn't the future, Harvard Business Review, 6 June 2023. (https://hbr.org/2023/06/ai-prompt-engineering-isnt-the-future?registration=success, accessed 26 June 2023).
37 GPT-4 system card. Open AI, 23 March 2023. (https://cdn.openai.com/papers/gpt-4-system-card.pdf, accessed 28 May 2023).
38 GPT-4. OpenAI, 14 March 2023. (https://openai.com/research/gpt-4, accessed 28 May 2023).
39 Radford A, Kleinman Z. ChatGPT can now access up-to-date information. BBC News, 27 September 2022. (https://www.bbc.com/news/technology-66940771, accessed 1 October 2023).
40 Kruge S, Ostermaier A, Uhl M. The moral authority of ChatGPT. ArXiv:23101.07098. doi:10.48550/arXiv.23101.07098.
41 Mickle T, Metz C, Grant N. The chatbots are here, and the internet industry is in a tizzy, The New York Times, 8 March 2023. (https://www.nytimes.com/2023/03/08/technology/chatbots-disrupt-internet-industry.html, accessed 29 May 2023).
42 Woo M. Trial by artificial intelligence. Nature. 2019;573:S100–2 (https://media.nature.com/original/magazine-assets/d41586-019-02871-3/d41586-019-02871-3.pdf, accessed 29 May 2023).
43 Muralidharan V, Burgart A, Daneshjou D, Rose S. Recommendations for the use of pediatric data in artificial intelligence and machine learning ACCEPT-AI. npj Dig Med. 2023;6:166. doi:10.1038/s41746-023-00898-5.
44 Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learning Individual Differences. 2023;103:102274. doi:10.1016/j.lindif.2023.102274.
45 Reddy CD, Lopez L, Ouyang D, Zou JY, He B. Video-based deep learning for automated assessment of left ventricular ejection fraction in pediatric patients. J Am Soc Echocardiogr. 2023;36(5):482–9. doi:10.1016/j.echo.2023.01.015.
46 Knight W. These ChatGPT rivals are designed to play with your emotions. Wired, 4 May 2023. (https://www.wired.com/story/fast-forward-chatgpt-rivals-emotions/#:~:text=12%3A00%20PM-,These%20ChatGPT%20Rivals%20Are%20Designed%20to%20Play%20With%20Your%20Emotions,%2C%20-companionship%E2%80%94and%20even%20romance., accessed 29 May 2023).
47 Smuha NA, De Ketalaere M, Coeckelbergh M, Dewitte P, Poullet Y. Open letter: We are not ready for manipulative AI – urgent need for action. KU Leuven, 31 March 2023. (https://www.law.kuleuven.be/ai-summer-school/open-brief/open-letter-manipulative-ai, accessed 29 May 2023).
48 Cuthbertson A. "No, I'm not a robot": ChatGPT successor tricks worker into thinking it is human. Independent, 15 March 2023. (https://www.independent.co.uk/tech/chatgpt-gpt4-ai-openai-b2301523.html, accessed 26 June 2023).
49 Walker L. Belgian man dies by suicide following exchanges with chatbot, The Brussels Times, 28 March 2023. (https://www.brusselstimes.com/430098/belgian-man-commits-suicide-following-exchanges-with-chatgpt, accessed 29 May 2023).
50 DeGuerin M. Oops: Samsung employees leaked confidential data to ChatGPT. Gizmodo, 6 April

2023. (https://gizmodo.com/chatgpt-ai-samsung-employees-leak-data-1850307376, accessed 29 May 2023).
51 Privacy policy. OpenAI, 27 April 2023. (https://openai.com/policies/privacy-policy, accessed 29 March 2023).
52 Coles C. 11% of data employees paste into ChatGPT is confidential. Cyberhaven, 19 April 2023. (https://www.cyberhaven.com/blog/4-2-of-workers-have-pasted-company-data-into-chatgpt/, accessed 29 May 2023).
53 Mihalcik C. ChatGPT bug exposed some subscribers' payment info. CNET, 24 March 2023.(https://www.cnet.com/tech/services-and-software/chatgpt-bug-exposed-some-subscribers-payment-info/, accessed 29 May 2023).
54 Moodley K, Rennie S. ChatGPT has many uses. Experts explore what this means for healthcare and medical research. The Conversation, 22 February 2023. (https://theconversation.com/chatgpt-has-many-uses-experts-explore-what-this-means-for-healthcare-and-medical-research-200283, accessed 2 June 2023).
55 De Proost M, Pozzi G. Conversational artificial intelligence and the potential for epistemic injustice. Am J Bioethics. 2023;23(5):51–3. doi:10.1080/15265161.2023.2191020.
56 Disability and employment. New York: United Nations, Department of Economic and Social Affairs (Disability); undates (https://www.un.org/development/desa/disabilities/resources/factsheet-on-persons-with-disabilities/disability-and-employment.html, accessed 11 September 2023).
57 Whittaker M, Alper M, Bennett CL, Hendren S, Kaziunas L, Mills Met al. Disability, bias and AI. New York: AI Now Institute; 2019. (https://ainowinstitute.org/wp-content/uploads/2023/04/disabilitybiasai-2019.pdf, accessed 11 September 2023).
58 Hallman J. AI language models show bias against people with disabilities, study finds. University Park (PA): Penn State University; 2022. (https://www.psu.edu/news/information-sciences-and-technology/story/ai-language-models-show-bias-against-people-disabilities/, accessed 11 September 2023).
59 Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. eBioMedicine. 2023;90:194512. doi:10.1016/j.ebiom.2023.104512.
60 Lohr S. AI may someday work medical miracles. For now, it helps do paperwork. The New York Times, 26 June 2023. (https://www.nytimes.com/2023/06/26/technology/ai-health-care-documentation.html, accessed 10 July 2023).
61 Eddy N. Epic, Microsoft partner to use generative AI for better EHRs. Healthcare IT News, 18 April 2023. (https://www.healthcareitnews.com/news/epic-microsoft-partner-use-generative-ai-better-ehrs, accessed 31 May 2023).
62 Nuance and Microsoft announce the first fully AI-automated clinical documentation application for healthcare. Burlington (MA):Nuance; 2023. (https://news.nuance.com/2023-03-20-Nuance-and-Microsoft-Announce-the-First-Fully-AI-Automated-Clinical-Documentation-Application-for-Healthcare, accessed 31 May 2023).
63 Ahn S. The impending impacts of large language models on medical education. Korean J Med Educ. 2023;35(1):103–7. doi:10.3946/kjme.2023.253.
64 Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Briefi Bioinformatics. 2022; 23(6):bbac409. doi:10.1093/bib/bbac409.
65 Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK. Artificial intelligence in drug discovery and development. Drug Discov Today. 2021;26(1):80–93. doi:10.1016/j.drudis.2020.10.010.
66 Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. Nature. 2023;613:612. doi:10.1038/d41586-023-00191-1.
67 Zielinksi C, Winker MA, Aggarwal R, Ferris LA, Heinemann M, Lapeña JF Jr et al. Chatbots, generative AI, and scholarly manuscripts. Overijssel: World Association of Medical Editors; 2023. (https://wame.org/page3.php?id=106, accessed 26 June 2023).
68 Gibbs W. Lost science in the Third World. Sci Am. 1995;273(2):92–9 doi:10.1038/scientificamerican0895-92.
69 Birhane A, Kasirzadeh A, Leslie D, Wachter S. Science in the age of large language models. Nat Rev Phys. 2023;5:277–80. doi:10.1038/s42254-023-00581-4.

70 Monitoring the building blocks of health systems: a handbook of indicators and their measurement strategies. Geneva: World Health Organization; 2010. (https://apps.who.int/iris/bitstream/handle/10665/258734/9789241564052-eng.pdf, accessed 26 June 2023).
71 Morozov E. The true threat of artificial intelligence. The New York Times, 30 June 2023. (https://www.nytimes.com/2023/06/30/opinion/artificial-intelligence-danger.html, accessed 2 July 2023).
72 Introducing ChatGPTPlus. San Francisco (CA): Open AI; 2023. (https://openai.com/blog/chatgpt-plus, accessed 1 June 2023).
73 The hidden workforce that helped filter violence and abuse out of ChatGPT. Wall Street Journal, 11 July 2023. (https://www.wsj.com/podcasts/the-journal/the-hidden-workforce-that-helped-filter-violence-and-abuse-out-of-chatgpt/ffc2427f-bdd8-47b7-9a4b-27e7267cf413, accessed 13 July 2023).
74 Firth N. Language models may be able to self-correct biases – if you ask them. MIT Technology Review, 20 March 2023. (https://www.technologyreview.com/2023/03/20/1070067/language-models-may-be-able-to-self-correct-biases-if-you-ask-them-to/, accessed 1 June 2023).
75 Khan L. We must regulate A.I. Here's how. The New York Times, 3 May 2023. (https://www.nytimes.com/2023/05/03/opinion/ai-lina-khan-ftc-technology.html, accessed 2 June 2023).
76 Hatzius J, Briggs J, Kodnani D, Pierdomenico G. The potentially large effects of artificial intelligence on economic growth (Briggs/Kodnani). Goldman Sachs Economics Research, 26 May 2023. (https://www.key4biz.it/wp-content/uploads/2023/03/Global-Economics-Analyst_-The-Potentially-Large-Effects-of-Artificial-Intelligence-on-Economic-Growth-Briggs_Kodnani.pdf, accessed 1 June 2023).
77 Milmo D. AI revolution puts skilled jobs at highest risk, says OECD. The Guardian, 11 July 2023. (https://www.theguardian.com/technology/2023/jul/11/ai-revolution-puts-skilled-jobs-at-highest-risk-oecd-says, accessed 12 July 2023).
78 Health and care workforce in Europe: time to act. Geneva: World Health Organization; 2022. (https://iris.who.int/handle/10665/362379, accessed 1 June 2023).
79 Health workforce. Geneva: World Health Organization; 2023. (https://www.who.int/health-topics/health-workforce#tab=tab_1, accessed 1 June 2023).
80 Hurst L. OpenAI says 80% of workers could see their jobs impacted by AI. These are the jobs most impacted, Euronews.next, 30 March 2023. (https://www.euronews.com/next/2023/03/23/openai-says-80-of-workers-could-see-their-jobs-impacted-by-ai-these-are-the-jobs-most-affe, accessed 1 June 2023).
81 A new era of generative AI for everyone. Dublin: Accenture; 2023. (https://www.accenture.com/content/dam/accenture/final/accenture-com/document/Accenture-A-New-Era-of-Generative-AI-for-Everyone.pdf, accessed 1 June 2023).
82 Burgess M. The security hole at the heart of ChatGPT and Bing. Wired, 25 May 2023. (https://www.wired.co.uk/article/chatgpt-prompt-injection-attack-security, accessed 1 June 2023).
83 Heikkila M. Open AI's hunger for data is coming back to bite it. MIT Technology Review, 19 April 2023. (https://www.technologyreview.com/2023/04/19/1071789/openais-hunger-for-data-is-coming-back-to-bite-it/, accessed 1 June 2023).
84 General Data Protection Regulation, Regulation 2016//679 of the European Parliament and of the Council, 27 April 2016. Strasbourg: European Parliament; 2016. (https://eur-lex.europa.eu/eli/reg/2016/679/oj, accessed 27 September 2023).
85 The impact of the General Data Protection Regulation on artificial intelligence (STOA Options Brief). Strabourg: European Parliament; 2020. (https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530(ANN1)_EN.pdf, accessed 26 June 2023).
86 OPC launches investigation into ChatGPT. Ottawa: Office of the Privacy Commissioner of Canada; 4 April 2023. (https://www.priv.gc.ca/en/opc-news/news-and-announcements/2023/an_230404/, accessed 1 June 2023).
87 Lomas N. Italy orders Chat GPT blocked citing data protection concerns. Tech Crunch, 31 March 2023. (https://techcrunch.com/2023/03/31chatgpt-blocked-italy/, accessed 1 June 2023).

88 ChatGPT: Italian SA to lift temporary limitation if OpenAI implements measures. Rome: Italian Data Protection Authority; 2023. (https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9874751#english, accessed 1 June 2023).
89 Weatherbed J. OpenAI's regulatory troubles are only just beginning. The Verge, 5 May 2023. (https://www.theverge.com/2023/5/5/23709833/openai-chatgpt-gdpr-ai-regulation-europe-eu-italy, accessed 1 June 2023).
90 Wiggers K. Open AI's new tool attempts to explain language models' behaviours. Tech Crunch, 9 May 2023. (https://techcrunch.com/2023/05/09/openais-new-tool-attempts-to-explain-language-models-behaviors/, accessed 1 June 2023).
91 Libeau D. ChatGPT will probably never comply with GDPR. 10 April 2023. (https://blog.davidlibeau.fr/chatgpt-will-probably-never-comply-with-gdpr/, accessed 1 June 2023).
92 Lomas N. ChatGPT maker OpenAI accused of string of data protection breaches in GDPR complaint filed by privacy researcher. TechCrunch, 30 August 2023. (https://consent.yahoo.com/v2/collectConsent?sessionId=3_cc-session_6bdecae4-d7b6-448f-8e26-e7805c03b964, accessed 11 September 2023).
93 Fung B. The FTC should investigate Open AI and block GPT over "deceptive" behaviour, AI policy group claims. CNN, 30 March 2023. (https://edition.cnn.com/2023/03/30/tech/ftc-openai-gpt-ai-think-tank/index.html, accessed 2 June 2023).
94 Waters R, Murgia M, Espinoza J. Open AI warns over split with Europe as AI regulation advances. Financial Times, 25 May 2023. (https://www.ft.com/content/5814b408-8111-49a9-8885-8a8434022352, accessed 1 June 2023).
95 Technology-facilitated gender-based violence: Making all spaces safe. New York: United Nations Population Fund; 2021. (https://www.unfpa.org/publications/technology-facilitated-gender-based-violence-making-all-spaces-safe, accessed 1 October 2023).
96 Murgia M. DeepMind reinvents itself for AI counterattack. Financial Times, 2 May 2023. (https://ft.pressreader.com/v99c/20230502/281724093873699, accessed 2 June 2023).
97 Schaake M. Regulating AI will put companies and governments at loggerheads, Financial Times, 2 May 2023. (https://www.ft.com/content/7ef4811d-79bb-4b4f-b28f-b46430f0c9ff, accessed 2 June 2023).
98 Metz, Cade. Tech giants are paying huge salaries for scarce A.I. talent, The New York Times, 22 October 2017. (https://www.nytimes.com/2017/10/22/technology/artificial-intelligence-experts-salaries.html).
99 Leswing K. Google reveals its newest AI supercomputer, says it beats Nvidia. CNBC, 5 April 2023. (https://www.cnbc.com/2023/04/05/google-reveals-its-newest-ai-supercomputer-claims-it-beats-nvidia-.html, accessed 2 June 2023)
100 Ahuja K. Antitrust has role in policing AI landscape. Financial Times, 10 April 2023. (https://www.ft.com/content/953817f5-5bc4-49e1-b583-977cc4780eca, accessed 2 June 2023).
101 Ahmed N, Wahed M, Thompson NC. The growing influence of industry in AI research. Science. 2023;379(6635):884–6. doi:10.1126/science.ade2420.
102 Røttingen JA, Regmi S, Eide M, Young AJ, Viergever RF, Ardal C et al. Mapping of available health research and development data: What's there, what's missing, and what role is there for a global observatory? Lancet. 2013;382(9900):1286–307. doi:10.1016/S0140-6736(13)61046-6.
103 A new partnership to promote responsible AI. Google Blogs, 26 July 2023. (https://blog.google/outreach-initiatives/public-policy/google-microsoft-openai-anthropic-frontier-model-forum/#:~:text=Anthropic%2C%20Google%2C%20Microsoft%20and%20OpenAI%20are%20launching%20the%20Frontier%20Model,development%20of%20frontier%20AI%20models, accessed 29 July 2023).
104 Fact sheet: Biden–Harris Administration secures voluntary commitments from leading artificial intelligence companies to manage the risks posed by AI. Washington DC: The White House, 21 July 2023. (https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/, accessed 29 July 2023).
105 Volpicelli G. Europe pitches AI pact to curtail the booming tech's risk. Politico, 26 May 2023.

(https://www.politico.eu/article/big-tech-rumble-europe-global-artificial-intelligence-debate-ai-pact/, accessed 29 July 2023).
106 Grant N, Weise K. In AI race, Microsoft and Google choose speed over caution. The New York Times, 7 April 2023. (https://www.nytimes.com/2023/04/07/technology/ai-chatbots-google-microsoft.html, accessed 2 June 2023).
107 Schiffer Z, Newton C. Microsoft lays off team that taught employees how to make AI tools responsibly. The Verge, 14 March 2023. (https://www.theverge.com/2023/3/13/23638823/microsoft-ethics-society-team-responsible-ai-layoffs, accessed 2 June 2023).
108 Center for Research on Foundation Models. The Foundation Model Transparency Index, 2023. (https://crfm.stanford.edu/fmti/, accessed 21 October 2023).
109 Milmo D. Google chief warns AI could be harmful if deployed wrongly. The Guardian, 17 April 2023. (https://www.theguardian.com/technology/2023/apr/17/google-chief-ai-harmful-sundar-pichai, accessed 2 June 2023).
110 Fiesler C. AI has social consequences, but who pays the price? Tech companies' problem with ethical debt. The Conversation, 19 April 2023. (https://theconversation.com/ai-has-social-consequences-but-who-pays-the-price-tech-companies-problem-with-ethical-debt-203375, accessed 2 June 2023).
111 Criddle, Cristina and Murphy, Hannah, Meta disbands protein-folding team in shift towards commercial AI, Financial Times, 7 August 2023. (https://www.ft.com/content/919c05d2-b894-4812-aa1a-dd2ab6de794a?accessToken=zwAGBZu-oWVwkdORnAXSuJRIEtOqGt0qtt55Sg.MEQCIA1QQ1iG8KPAnuDAuPvt-Ngds3OzxL1lt-0FnaVbAQFtAiAZvHnmKD_fABj8ZzLTNXRp1v7V38nTcUf_pPxAPdx16A&sharetype=gift&token=3ac5a132-e08e-412e-bc3c-08edea8a7417, accessed 18 September 2023).
112 Ananthaswamy A. In AI, is bigger always better? Nature, 8 March 2023. (https://www.nature.com/articles/d41586-023-00641-w, accessed 2 June 2023).
113 Li P. Making AI less "thirsty": Uncovering and addressing the secret water footprint of AI models. ArXiv. 2023;2304.03271v. doi:10.48550/arXiv.2304.03271.
114 Syed N. The secret water footprint of AI technology. The Markup, 15 April 2023. (https://themarkup.org/hello-world/2023/04/15/the-secret-water-footprint-of-ai-technology, accessed 2 June 2023).
115 Livingstone G. It's pillage: Thirsty Uruguayans blast Google's plan to exploit water supply. The Guardian, 11 July 2023. (https://www.theguardian.com/world/2023/jul/11/uruguay-drought-water-google-data-center, accessed 12 July 2023).
116 Thornhill J. The sceptical case on generative AI. Financial Times, 17 August 2023. (https://www.ft.com/content/ed323f48-fe86-4d22-8151-eed15581c337, accessed 11 September 2023).
117 Marcus G. The imminent enshittification of the Internet. Substack, 16 August 2023. (https://garymarcus.substack.com/p/the-imminent-enshittification-of, accessed 11 September 2023).
118 Pause giant AI experiments: An open letter. Narberth (PA): Future of Life Institute; 2023. (https://futureoflife.org/open-letter/pause-giant-ai-experiments/, accessed 13 June 2023).
119 Perrigo B. DeepMind's CEO helped take AI mainstream. Now he's urging caution. Time, 12 January 2023. (https://time.com/6246119/demis-hassabis-deepmind-interview/, accessed 13 June 2023).
120 Lomas N. Unpacking the rules shaping generative AI. Tech Crunch, 13 April 2023. (https://techcrunch.com/2023/04/13/generative-ai-gdpr-enforcement/, accessed 13 June 2023).
121 Mökander J, Schuett J, Kirk HR, Floridi L. Auditing large language models: a three-layered approach Soc Sci Res Netw. 2023. doi:10.2139/ssrn.4361607.
122 Lomas N. Report details how Big Tech is leaning on EU not to regulate general purpose Ais. Tech Crunch, 23 February 2023. (https://techcrunch.com/2023/02/23/eu-ai-act-lobbying-report/, accessed 20 June 2023).
123 Sambasivan N, Kapania S, Highfill H, Akrong D, Paritosh P, Aroyo LM. "Everyone wants to do model work, not the data work.": Data cascades in high-stakes AI. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, May 2021. doi:10.1145/3411764.3445518.
124 Browne G. AI is steeped in Big Tech's digital colonialism. Wired, 25 May 2023.

(https://www.wired.co.uk/article/abeba-birhane-ai-datasets, accessed 17 June 2023).
125 Baxter K, Schelsinger, N. Managing the risks of generative AI. Harvard Business Review, 6 June 2023. (https://hbr.org/2023/06/managing-the-risks-of-generative-ai, accessed 17 June 2023).
126 Samuelson P. Generative AI meets copyright. Science. 2023;381(6654):158–61. doi:10.1126/science.adi0656.
127 El-Mhamdi E, Farhadkhani S, Guerraoui R, Gupta N, Hoang L, Pinot R et al. On the impossible safety of large AI models. arXiv. 2209.15259v2. doi:10.48550.arXiv.2209.15259.
128 Open AI. GPT-4 technical report. arXiv:2303.08774v3. doi:10.48550/arXiv.2302.08774.
129 Murgia M. Open AI's red team: experts hired to "break" ChatGPT. Financial Times, 14 April 2023. (https://www.ft.com/content/0876687a-f8b7-4b39-b513-5fee942831e8, accessed 10 July 2023).
130 Clegg N. Openness on AI is the way forward for tech. Financial Times, 11 July 2023. (https://www.ft.com/content/ac3b585a-ce50-43d1-b71d-14dfe6dce999, accessed 11 July 2023).
131 Huang S, Toner H, Haluza Z, Creemers R, Webster G. Measures for the management of generative artificial intelligence services (draft for comment) (translation). DigiChina. Palo Alto (CA): Stanford University, Program on Geopolitics; 2023. (https://digichina.stanford.edu/work/translation-measures-for-the-management-of-generative-artificial-intelligence-services-draft-for-comment-april-2023/, accessed 17 June 2023).
132 Ye J. China says generative AI rules to apply only to products for the public. Reuters, 13 July 2023. (https://www.reuters.com/technology/china-issues-temporary-rules-generative-ai-services-2023-07-13/, accessed 13 July 2023).
133 Bommasani R, Klyman K, Zhang D, Liang P. Do foundation model providers comply with the draft EU AI Act? Palo Alto (CA): Stanford University, Human-centered Artifical Intelligence; 2021. (https://crfm.stanford.edu/2023/06/15/eu-ai-act.html, accessed 17 June 2023).
134 Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. Strasbourg: European Parliament; 2023. (https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html, accessed 10 July 2023).
135 Beyond ChatGPT: How can Europe become a leader in generative AI? Kaiserslautern: German Research Centre for Artificial Intelligence; 2023. (https://www.dfki.de/en/web/news/jenseits-von-chatgpt-wie-kann-europa-bei-der-generativen-ki-eine-fuehrungsposition-uebernehmen, accessed 17 June 2023).
136 Spirling A. Why open-source generative AI models are an ethical way forward for science. Nature. 2023;616(7957):413. doi:10.1038/d41586-023-01295-4.
137 Vincent J. Meta's powerful AI language model has leaked online – What happens now? The Verge, 8 March 2023. (https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse, accessed 29 July 2023).
138 Maffuli S. Meta's Llama 2 license is not open source. Open Source Initiative, 20 July 2023. (https://blog.opensource.org/metas-llama-2-license-is-not-open-source/, accessed 29 July 2023).
139 Marble A. Software licenses masquerading as open source. marble.onl, 1 June 2023. (http://marble.onl/posts/software-licenses-masquerading-as-open-source.html, accessed 29 July 2023).
140 Keary T. Report finds 82% of open-source software components "inherently risky". Venture Beat, 17 April 2023. (https://venturebeat.com/security/report-finds-82-of-open-source-software-components-inherently-risky/, accessed 8 July 2023).
141 Generative AI raises competition concerns. Technology blog, 29 June 2023. Washington DC: Federal Trade Commission; 2023. (https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/06/generative-ai-raises-competition-concerns, accessed 29 July 2023).
142 Wishart-Smith H. Generative AI: cybersecurity friend and foe, Forbes, 6 June 2023. (https://www.forbes.com/sites/heatherwishartsmith/2023/06/06/generative-ai-cybersecurity-friend-and-foe/?sh=4407e0884bd2, accessed 29 July 2023).
143 Metz C. Researchers poke holes in safety controls of ChatGPT and other chatbots. The New

York Times, 27 July 2023. (https://www.nytimes.com/2023/07/27/business/ai-chatgpt-safety-research.html, accessed 11 September 2023).
144 Harris T, Freuh S. The complexity of technology's consequences is going up exponentially, but our wisdom and awareness are not. Issues in Science and Technology, 16 May 2023. (https://issues.org/tristan-harris-humane-technology-misinformation-ai-democracy/, accessed 19 June 2023).
145 Schyns C. The lobbying ghost in the machine: Big Tech's covert defanging of Europe's AI Act. Brussels: Corporate Europe Observatory; 2023. (https://corporateeurope.org/en/2023/02/lobbying-ghost-machine, accessed 17 June 2023).
146 Fact sheet: Biden–Harris Administration announces new actions to promote responsible AI innovation that protects Americans' rights and safety. Washington DC: White House, 4 May 2023. (https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/, accessed 19 June 2023).
147 Sijbrandij Sid. AI weights are not "open source". Open Core Ventures, 27 June 2023. (https://opencoreventures.com/blog/2023-06-27-ai-weights-are-not-open-source/, accessed 29 July 2023).
148 Meeker H. Towards an open weights definition. Copyleft Currents, 8 June 2023. (https://heathermeeker.com/2023/06/08/toward-an-open-weights-definition/, accessed 29 July 2023).
149 Dastin J, Tong A. Google, one of AI's biggest backers, warns its own staff about chatbots. Reuters, 15 June 2023. (https://www.reuters.com/technology/google-one-ais-biggest-backers-warns-own-staff-about-chatbots-2023-06-15/, accessed 9 July 2023).
150 Kanter GP, Packel EA. Health care privacy risks of AI chatbots. JAMA. 2023;330(4):311–2. doi:10.1001/jama.2023.9618.
151 Interim measures for the management of generative artificial intelligence services. Beijing: Cyberspace Administration of China; 13 2023. (http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm, accessed 29 July 2023).
152 Satariano A. E.U. agrees on landmark artificial intelligence rules. The New York Times, 8 December 2023. (https://www.nytimes.com/2023/12/08/technology/eu-ai-act-regulation.html, accessed 15 December 2023).
153 The EU should regulate on the basis of rights, not risks. Access Now, 17 February 2021. (https://www.accessnow.org/eu-regulation-ai-risk-based-approach/, accessed 21 June 2023).
154 Marks M, Haupt CE. AI chatbots, health privacy, and challenges to HIPAA compliance JAMA. 2023;330(4):309–10. doi: 10.1001/jama.2023.9458.
155 Marcus G. Two models of AI oversight – and how things could go deeply wrong. Substack, 8 June 2023. (https://garymarcus.substack.com/p/two-models-of-ai-oversight-and-how, accessed 17 June 2023).
156 Kang C, Metz C. FTC opens investigation into Chat GPT maker over technology's potential harms. The New York Times, 13 July 2023. (https://www.nytimes.com/2023/07/13/technology/chatgpt-investigation-ftc-openai.html, accessed 29 July 2023).
157 Ordish J. Large language models and software as a medical device. MedRegs blogs, 3 March 2023(https://medregs.blog.gov.uk/2023/03/03/large-language-models-and-software-as-a-medical-device/, accessed 19 June 2023).
158 Gilbert S, Harvey H, Melvin T, Vollebregt E, Wicks P. Large language model AI chatbots require approval as medical devices. Nat Med. 2023. doi:10.1038/s41591-023-02412-6.
159 Ghost in the machine: Addressing the harm of generative AI. Forbrukerradet. Oslo: Norwegian Concumer Council; 2023. (https://storage02.forbrukerradet.no/media/2023/06/generative-ai-rapport-2023.pdf, accessed 9 July 2023).
160 Mökander J, Floridi L. Ethics-based auditing to develop trustworthy AI. Minds & Machines, 2021. doi:10.1007/s11023-021-09557-8.
161 Minssen T, Vayena E, Cohen IG. The challenges for regulating medical use of ChatGPT and other large language models. JAMA. 2023;330(4):315–6. doi: 10.1001/jama.2023.9651.
162 Questions and Answers: AI Liability Directive. Brussels: European Commission; 2022. (https://ec.europa.eu/commission/presscorner/detail/en/QANDA_22_5793, accessed 20 June 2023).

163 Duffourc MN, Gerke S. The proposed EU directives for AI liability leave worrying gaps likely to impact medical AI. NPJ Digit Med. 2023;6(1):77. doi:10.1038/s41746-023-00823-w.
164 Regulatory considerations on artificial intelligence for health. Geneva: World Health Organization; 2023. (https://iris.who.int/bitstream/handle/10665/373421/9789240078871-eng.pdf?sequence=1&isAllowed=y, accessed 16 November 2023).
165 Marcus G. Artificial Intelligence is stuck. Here's how to move it forward. The New York Times, 29 July 2017. (https://www.nytimes.com/2017/07/29/opinion/sunday/artificial-intelligence-is-stuck-heres-how-to-move-it-forward.html, accessed 20 June 2023).
166 Parker G. Rishi Sunak to lobby Joe Biden for UK "leadership" role in AI development. Financial Times, 5 June 2023. (https://www.ft.com/content/7c30ea28-2895-44c2-9a2d-c31ea7fa27e7, accessed 19 June 2023).
167 Hogarth I. We must slow down the race to god-like AI. Financial Times, 13 April 2023. (https://www.ft.com/content/03895dc4-a3b7-481e-95cc-336a524f2ac2, accessed 10 July 2023).
168 Blinken A, Raimondo G. To shape the future of AI, we must act quickly. Financial Times, 24 July 2023. (https://www.ft.com/content/eea999db-3441-45e1-a567-19dfa958dc8f, accessed 30 July 2023).
169 Guterres, Antonio, Networked, inclusive multilateralism can help overcome challenges of era, says Secretary General, opening general assembly session, United Nations, 17 September 2019. (https://press.un.org/en/2019/sgsm19746.doc.htm, accessed 18 September 2023).

附件：编制方法

世卫组织依靠其卫生领域人工智能的伦理和治理专家组以协商的方式制定本指南。该专家组由来自世卫组织所有区域的 20 名专家组成，每两周举行一次会议，为期约四个月。专家组将此前发布的卫生领域人工智能伦理和治理指南中的共识原则和建议应用于卫生保健与药品领域新兴多模态大模型的使用。

专家组首先对多模态大模型的潜在用途和收益以及终端用户层面的风险进行了初步梳理。专家组还确定了卫生系统和社会在使用此类 AI 系统时可能遇到的风险。对此，专家组进行了全面的文献搜索，以补充：（a）在逐渐演变和使用的几年中实现的卫生领域多模态大模型的现有和拟议用途，（b）多模态大模型的预期用途，以及（c）在本指南发布之前发布的对多模态大模型的批评和分析。

在对已知和潜在的益处和风险达成共识的基础上，专家组确定了一个适当的框架，以应对与使用多模态大模型相关的各种伦理挑战和机遇。专家组一致认为，“价值链”方法非常适合描述在何处以及如何组织适当的治理，以及哪些行为、哪些行为者可以负责执行相关措施。

专家组参考并利用了一些法域现有或拟议的立法和监管措施，作为确定建议领域的一种手段，同时对每项建议进行了精心设计，使其适用于多个国家和法律体系。专家组还讨论了企业、此类人工智能系统的购买者和多模态大模型的终端用户（特别是医疗服务提供者和患者）应适用的建议。

世卫组织承认，该指南可能需要修订和更新，以确保专家组的研究结果和建议始终具有相关性和实用性。