arXiv:2603.26524v1 [math.HO] 27 Mar 2026

人工智能时代的数学方法与人类思维

Tanya Klowden 与陶哲轩

摘要。人工智能（AI）是一个通俗名称，泛指一系列计算机工具，它们旨在执行日益复杂的认知任务，其中许多任务过去曾是人类的专属领域。随着这些工具变得愈发先进且无处不在，人们常常质疑快速开发并将其融入社会的合理性，尤其是因为它们会消耗有限资源，并对那些似乎将被其取代的熟练个人的生计构成生存性风险。本文探讨人工智能迅速演进对哲学传统议题的影响，重点关注其在数学中的应用，以及其更广泛应用所带来的现实世界影响。我们认为，人工智能是人类工具自然演进的结果；纵观历史，人类不断开发工具，以促进思想的创造、组织与传播。我们还指出，人工智能的开发与应用必须始终以人为本。着眼于创新解决方案，以满足人类需求、提升人类生活质量，并拓展人类思维与理解的能力，我们提出了一条路径，旨在将人工智能融入最具挑战性、最严谨的智力领域，使全人类受益。

1. 引言

人工智能（AI）技术如此迅速地渗透到数字生活的每个角落，这一点令人惊叹：作者使用常见工具撰写本文时，竟有至少三个不同的数字代理未经邀请便插入了叙述之中[1]。人类正站在数字工业革命的门槛上，而这场革命正以前所未有的速度展开。在物理科学领域，人工智能的进步促成了荣获诺贝尔奖的研究 [1]；在人文学科领域，人们则普遍担忧，现代人工智能的文本生成能力可能会让这一学科走向消亡[2]。语言翻译工具为文化交流与国际合作敞开大门的同时，深度伪造内容和低质生成内容也接踵而至，在我们的数字第三空间中泛滥。人工智能迅速从新奇事物变成了重要资源，继而在某些情况下成为眼前的生存威胁[3]。

[1]所有这些人工智能“贡献”都立即从文本中删除了。1

1.1. 我们对人工智能的定义。 本文所说的人工智能，指的是范围广泛的计算机工具，旨在执行日益复杂的认知任务，其中许多任务过去专属于人类。人工智能工具种类繁多，既包括当今以数据为驱动的机器学习（ML）技术（例如能够处理复杂文本的大语言模型（LLM），以及能够生成图像和其他媒体内容的扩散模型），也包括较为传统的传统人工智能（~~GOFAI~~）（例如自动定理证明器或国际象棋引擎）；后者能够通过应用精确的数学规则，解决范围狭窄的问题。

1.2. 本文的目的。 关于这些工具能做什么、不能做什么的讨论并不少见；但相较之下，人们较少探讨为何这些工具正在如此迅速地开发和部署，以及它们如何影响数十亿使用者的生活——人们借助它们开展研究和教育、工作、娱乐，甚至休息[4]。本文作者来自常被视为截然相反的学术领域：数学和艺术研究。然而，我们都发现，在日常工作中将若干人工智能工具融入各自迥异的研究领域大有裨益，也发现，对于现实世界中的人工智能应用所引发的种种棘手却普遍存在的哲学问题，我们之间有着出人意料的许多共识。本文将以数学为例，探讨将人工智能纳入日常工作流程的益处、风险、伦理问题和影响，并进一步将这些观察推广到更广泛的现实应用中。尽管这些新兴技术带来了风险，且未必在道德上中立，但我们认为，人工智能工具无论是在数学领域还是其他领域，都应得到开发、实施和应用，原因有二：它们有望大幅增强人类固有的能力，也能够拓展可能性的边界，超越人类个体所能做到的范围，乃至突破人类集体能力的局限。结合我们使用这些工具的亲身经历，我们将特别考察人机界面，并就这些技术的发展提出建议，使其为人类带来的益处多于危害，同时珍视人类思想与行动的独特贡献，并使之与未来人工智能发展所带来的新模式相辅相成。

1.3. 浮士德式交易。市场竞争的激励推动了人工智能技术以疯狂的速度发展，也让整个行业沉迷于工作流程大幅加速和成本节约的愿景之中。这种竞争所造成的“囚徒困境”，迫使许多个人和组织尽可能仓促地试验性采用这些工具，却牺牲了对这种采用所带来的经济、社会或道德成本与收益进行更审慎的评估——或者更根本地说，牺牲了对我们最初为何应当开发此类技术的思考。因此，从某种意义上说，我们作为一个整体已经

事实上与这些技术达成了“浮士德式交易”：以向它们逐步开放我们的数据、认知工作流程和决策过程为代价，换取能够以更高效率、更少繁琐劳动完成更广泛任务的承诺。

从理论上讲，技术在道德上是中立的；它既能助长积极的应用场景，也能助长消极的应用场景。但这种赋能会加剧现有的道德困境，并催生新的困境。例如，第二次世界大战期间，对囚犯进行的骇人医学研究虽然提供了有关人类耐力极限的重要数据，挽救了许多生命，却也引发了一个棘手的问题：使用这类数据来推动新的医学进展是否合乎伦理 [5] 。如今，尽管情况并没有那么骇人，但用于训练当前一代人工智能工具的数据及知识产权来源不明，也引发了类似的争议 [6]。

如果一项技术的发展速度足够缓慢，人们就有可能在其广泛应用之前，围绕它展开必要的哲学讨论与辩论；干细胞研究就是一个显著的例子。然而，现代人工智能技术已经得到广泛应用，实际上已无法“把精灵塞回瓶子里”；具有讽刺意味的是，此时实施严格监管会不成比例地扼杀人工智能更积极的应用场景，例如加速科学研究，却无法杜绝更浪费或恶意的技术用途。务实地说，关于人工智能的讨论如今已转向如何管理与这些技术共存：评估人工智能的成本与收益（无论是在学术领域还是在更广泛的社会层面），并确定最佳实践和框架，以尽可能积极的方式使用人工智能，同时遏制人们（以种种方式）滥用这些工具、损害认知成果的可靠性与价值。

2. 历史类比：这一次会有所不同吗?

2.1.自动化技术以往的整合。自动化当然并非新鲜事物。过去的许多技术也使人们能够将此前由人类承担的任务自动化，从而消除或大幅减少对某些类型人力工作的需求，同时创造或增加对其他工作的需求；有时甚至会出现这种情况。例如，在科学界，由于新工具显而易见的优势，曾发生过“阶段性转变”，整个领域广泛而迅速地转向使用新工具（例如互联网、利用计算机进行科学计算，甚至是 LaTeX 这类不起眼的排版语言）。不过，这些过去的技术主要影响了专业工作中的次要方面，例如信息的传播与 disseminate，而非成果本身的创造。此外，尽管这些工具自动化的任务需要专门培训和专业知识才能完成，但通常并不需要理解专业领域更多的

哲学层面，例如知识的本质、美、意义或道德 [7]。当然，这些技术仍可能引发关于哲学议题的讨论——例如，原创艺术作品是否具有机械复制无法恰当捕捉的内在审美或创造性特征，或工业革命造成的劳动力置换所带来的道德与伦理影响——但人们并不认为它们会挑战这些讨论赖以成立的根本哲学假设。例如，印刷机的发明彻底改变了信息和思想的传播方式，却没有显著改变人们对思想或信息本质的理解；这些内容仍然是由人类有意识地创造出来的。

2.2. **现代人工智能**。但现代人工智能能够自动化创作过程本身的大部分环节，从而大规模生成艺术作品、数学证明，或科学、哲学理论等智力产品，而且所需的人工监督远少于以往[2]。这使得此类产品的外在形式与创造这些产品所运用的价值观和思维过程之间，出现了前所未有的脱节。例如，扩散模型如今可以生成一幅赏心悦目的风景画，尽管它并未直接受到物理世界中任何特定地点的启发；不过，训练该模型输出的图像时，确实使用了无数真实风景的图像（以及许多与风景完全无关的图像）。因此，人们对图像的审美反应也与这些审美体验最初的来源脱了节。

这绝非全新的哲学议题。塞尔关于机械设备若经过编程、能够用中文对话，是否就真正理解这种语言的“中文房间”思想实验 [8], ，可以追溯到1980年。“人工智能效应”也在大约同一时期得到认可；例如，在国际象棋引擎出现之前，人们认为下好国际象棋是衡量智力的良好标准。后来，国际象棋引擎通过机械地搜索博弈树，能够“毫无思考”地击败国际象棋大师，于是衡量智力的“国际象棋测试”便基本遭到弃用。同样，现代大语言模型实际上已经通过了著名的“图灵测试”——即判断人工智能能否以与人类无法区分的方式进行对话（参见例如[9]），使其不再享有人工智能“黄金标准”的昔日地位。关于这一点的近期讨论，参见 [10]。

目前，我们仍可将一些“根本性”理解的标志作为区分人类创作内容与人工智能生成内容的一项仍然可行的检验标准，例如能否连贯地解释并捍卫创作新艺术作品、数学

[2]目前的工具通常仍需要人类为人工智能生成一个初始提示，供其遵循，但这一过程本身现在也已基本能够实现自动化。

证明或其他智力成果所经历的创作过程。但如果未来几代人工智能也设法令人信服地通过这类检验，我们是否又不得不再次重新设定衡量智力、理解力和创造力的标准？数学和人文学科等领域的定义、价值观和目标是否需要重新评估？我们又该如何定位这些日益精密的人工智能工具——它们是助手、共同作者，还是能够独立创作的创作者？如果是这样，我们应如何看待它们产出的内容，以及促成这些内容的智力过程？

3. 作为人工智能应用试验场的数学

关于人工智能的这类更广泛的哲学问题极其复杂且多面，我们当然不敢声称对其中任何一个问题都能给出定论；而这一领域变化之迅速，也使得我们所作的任何断言都可能被引人注目的新技术进展超越。不过，我们可以从数学领域提出一些看法，既涉及纯粹的数学推理，也涉及现代数学分析在人文学科中的新兴应用。我们认为，数学是探索诸如人工智能对科学（以及整个社会）的影响等广泛问题的合适“试验场”，因为数学有着更悠久、更成熟的基础，而且其本质也适合探索各种与现实相悖的假设性抽象情景。我们希望，将人工智能融入数学（或不融入数学）所汲取的经验，能为人工智能将如何与科学及整个社会互动提供更广泛的视角。

前沿人工智能模型如今能够解决日益复杂的数学问题，并给出可独立验证的证明，而无需直接重现人类数学家解决问题的做法（例如检验特殊情形，再根据这些例子进行推广），尽管其训练数据中包含以这种传统方式生成的证明。因此，数学家将越来越多地遇到这样的情况：证明定理的能力与发现并理解这些证明所需的推理过程相互脱节。这进一步推动了现代数学中已有的去中心化趋势；在一个广泛应用领域都需要高等数学的世界里，由中央权威规定数学正统实践的“布尔巴基时代” [11] 早在几十年前就已成为过去[3]。

[3]尽管有人可能会认为，当前创建大型统一形式化数学库的项目（例如 Lean 的 Mathlib 项目），可以算是布尔巴基学派努力的现代继承者。

以当前的技术水平而言，最先进的人工智能工具仍表现出明显且往往匪夷所思的弱点：它们在某些任务上能取得卓越的超人表现，同时却会在其他任务上展现出常常令人捧腹的基础理解和错误水平。数学也不例外。人工智能生成的数学表面上可能毫无瑕疵——这不足为奇，因为这些模型的设计目标就是生成在视觉上尽可能接近正确的人类生成证明的输出——但与此同时，它们也会犯下根本性错误（例如断言所有奇数都是质数）；这类错误在人类数学家接受培训的早期就会被纠正，而且往往会使最终论证荒谬到无法挽救。与此同时，这种自上而下的方法主要着眼于生成漂亮的输出，而非关注传统上生成此类输出所需的基本认知过程，却可能出奇地有效；同一个人工智能既会经常犯下基本的数学错误，也能神秘地以高于人类专家的准确率得出复杂数学问题的正确答案，甚至给出一种奇怪但在技术上正确的证明，说明答案有效。

如今，人们正投入大量精力，尽可能减少或消除人工智能的这些弱点；通常并非直接增强人工智能对特定智力任务的内在“理解”，而是将此类人工智能工具置于严谨的环境中，进行独立测试、培训和验证，以降低错误的发生率。目前，完全自主的人工智能仍无法破解深奥的数学猜想，但很有可能在不久的将来，此类人工智能工具能够在这类工作中为人类数学家提供极大帮助，即使我们仍会犹豫是否将这种帮助描述为真正智能思维的体现。尽管如此，这样一种机械且容易出错的方法，竟能（或很快就能）在数学这样高度智力化的学科中产生如此多传统的质量标志，这一事实表明，我们必须重新审视关于智力或创造力究竟是什么、又该如何衡量的各种模型。

—

4. 人工智能与数学真理的本质

4.1. 数学与证明标准。 数学[4]有着悠久的客观证明标准传统，这一传统始于

[4]在这里，我们不明确界定数学究竟是什么。例如，可以采取规定主义观点，采用 Davis–Hersh 定义 [12] ，将数学视为“对具有可重复属性的心智对象的研究”。也可以采取描述主义观点，即数学就是数学家在实践中实际从事的活动。我们的讨论在某种程度上更倾向于后一种观点。

欧几里得，并随着二十世纪初稳定且（经验证）可靠的数学基础的出现而得到完善。有人指出（例如，参见 [13]），这些基础几乎得到普遍认可，使现代数学拥有了一种罕见而宝贵的能力：能够就该领域任何给定论证或断言的有效性达成共识，因为（原则上）人们可以要求将这类论证详尽阐述到足以逐步核查每一步是否正确运用了数学的标准公理和逻辑推理规则。一个典型例子是，Nelson [14] 在 2011 年声称皮亚诺公理在逻辑上不一致；这一说法与数学主流相去甚远，但通过指出论证中的一个微妙漏洞，问题最终得以解决。Nelson欣然接受了这一指正，并撤回了这一说法。

然而在实践中，人类数学家的论证往往达不到完全严谨的证明这一理想；文献中的小错误和大错误都很常见，其中一些会通过正式勘误或修订得到纠正，另一些则被忽视，或作为该子领域“民间传闻”的一部分非正式地流传下来。启发式上看似合理的论证往往只经过最低限度的检查便被接受，而违背传统观点的惊人断言则会遭到强烈质疑，即使这些论证最终经逐行审查后证明是正确的。

4.2. **气味测试**。迄今为止，这种状况尚算令人满意：遵循良好启发式方法并凭借直觉的人类数学家，往往能给出令人信服且大体正确的证明，大多数错误也都可以修正；而缺乏这种直觉的数学家，所给出的证明往往包含足够多的表面问题，使人即使尚未仔细检查，也有理由对其内容心存怀疑。非正式地说，人类所提出的数学论证往往带有一种“气味”，经验丰富的数学家会（或许是潜意识地）凭此初步判断论证有多么令人信服，远在他们有机会检查论证中的每一个步骤之前。例如，Aaronson 的博文《数学突破主张有误的十个迹象》 [15]列举了一些常见例子：这些论证散发着这样的“坏气味”，在发现其中具体的逻辑缺陷之前，人们就能察觉。并非所有错误都同样具有灾难性；有些错误甚至可能有益，例如，在尚未能够充分验证某种方法之前，它们就揭示了一个有希望的思路 [16]。

[5]我们借用这一感官隐喻，类比软件工程中的“代码异味”概念。

Thurston [18], 指出，构成良好“气味”的一个要素，是让人感到论证带来了理解或洞见：它不只是表明一组假设在逻辑上蕴含某个结论，还要说明这种蕴含何以成立的因果脉络，指出论证中哪些部分发挥了“关键作用”，哪些部分相较于先前文献是新颖或出人意料的，哪些部分则只是常规的技术考量。对数学文本的这类解读和印象，通常并未纳入严谨数学的正式框架，例如一阶逻辑或集合论；但它们对于阅读论证的人类数学家至关重要，使他们能够进一步领会论证如何推广到其他情境，或如何与文献中的其他方法相互作用。这种叙事结构也有助于增强人们对论证稳健性的信心：计算中一个符号放错位置，就可能使一篇冗长的数学论证失效；但如果证明有清晰的策略，系统地分离并处理论证中的关键难题，同时借鉴文献中以往成功论证的思路，那么即便论证中存在局部错误，也更有可能在忠于原证明精神的前提下加以修正。

4.3. 形式化能否力挽狂澜? 有若干发展可能会迫使数学界重新评估这种半形式化的证明标准。其中之一具有技术层面的原因：随着数学日益成熟、愈发精深（并且越来越多地借助计算机），论证也变得更长、更复杂；某些领域的前沿论文篇幅通常超过百页，使得人工审稿人逐行验证愈发费力。实际上，这意味着这种细致核查并非总会进行，只有最受瞩目、最重要的成果才例外，因而人们越来越（过度）依赖前文所说的“嗅觉”来评估数学论证的可信度。

这类问题似乎有可能通过技术手段得到解决（或至少有所缓解），尤其是更广泛地部署形式化证明助手（例如 Lean 或 Rocq）。如果数学论证以某种精确的计算机语言写成，这些助手便能自动检查其有效性 [19]。目前，这种形式化仍然过于繁琐，尚无法系统地应用（将传统的非形式化证明转换为这种形式

[6]另请参阅第二作者的文章 [17] ，文中指出，无论“好的”数学最初如何定义，它在实践中往往会融入更广泛的数学叙事，例如结构与随机性之间的二分，或代数探索几何问题（反之亦然）的能力。

语言通常需要花费最初撰写该证明五到十倍的时间），但目前已有大量工作致力于加快这一过程并使其更加便于使用，例如整合人工智能工具，以实现部分（甚至可能完全）“自动形式化” [20]。

4.4. 形式化验证的局限性。然而，即使这些技术问题得到解决，数学证明也能常规地通过形式化验证来确认正确性，仍会出现一些新问题，尤其是在不久的将来，日益复杂的论证可能会由人工智能工具部分或全部生成。首先，形式化验证只能确认一个形式化论证确立了一个形式数学陈述，却无法排除在形式陈述与原本意图表达的陈述之间的转换中出现错误。例如，费马大定理断言，对于任何大于 2 的自然数 n ，方程$a^n + b^n = c^n$都不存在自然数解 a，b，c；但这一非正式描述隐含了一个约定，即自然数从 1 而非 0 开始。若让人工智能解决这个问题，它可能会错误地假定 a，b，c可以为零，并据此给出一份（经过形式化验证的）证明，证明费马大定理是错误的！因此，尽管形式化原则上可以大幅减少仔细审查非正式数学文本的需要，却不能完全消除此类审查的必要性[7]。

其次，即便是在高等数学纯粹抽象的领域中，一个论证也只有一部分能够用适合形式化的演绎逻辑来表述。虽然演绎证明仍是大多数数学工作的核心，但围绕这一核心，还有一圈启发式、经验性或元数学的推理，它们能提供宝贵的信息：论证为何成立、能否拓展到其他语境、提出这些问题的动机是什么，以及人们如何从更基本的原理重构论证。人类撰写的证明，因其本质，往往会在写作过程中自然而然地带出这圈内容（尤其是作者善于阐述时）；但若人工智能专门以形式正确性为训练标准，而牺牲其他所有考量，它可能会产出“毫无气味”的证明，表面上仿佛出自一篇写得很好的

[7]从理论上讲，甚至有可能通过巧妙操纵 Mathlib 等标准形式化数学库中关键定义的形式化过程，“黑入”数学本身；参见 [21]。讽刺的是，数学研究日益协作化、社会化和大规模化，虽然总体而言是一个非常积极的发展，但也可能增加此类攻击的潜在漏洞；而在过去，数学主要由小群体的个人完成，这些漏洞并未引起显著关注。

人类证明，甚至可能通过形式化验证测试，却依然令人莫名感到不满足——它逐字逐句地实现了确立某个数学命题这一明确目标，却未能像预期那样，让人更深入地理解该命题所属的广阔数学领域。在一个所有生成媒介都经过人工智能润色、光鲜亮丽的世界里——包括排版优美、解释清晰且由 GPT 生成的数学证明——如果我们舍弃粗粝、杂乱的手写（或至少是手打）文本，会不会失去些什么？

4.5. **对早期挑战的适应**。 数学界已经适应了以往技术进步对其证明标准带来的挑战。大型计算机辅助证明，例如四色定理 [22] 或开普勒猜想 [23], 的证明，最初都颇具争议，因为要完全依靠人工核查并不切实际；但随着时间推移，人们为这类论证确立了新的建立可信度的标准，例如提供可复现代码，将论证中的计算部分拆分出来，作为明确陈述的独立引理，与论文中较具概念性的部分分开，并提供额外的相关数据和“校验和”，以检查计算机生成的计算结果是否通过各种“合理性检查”。实际上，这些发展使数学的证明标准朝着自然科学的方向转变：在自然科学中，理论论证和实证实验只要经过恰当的设计、执行和报告，都可以成为科学真理的可靠来源。

4.6. **人工智能辅助数学的演变**。随着重要的人工智能辅助数学或人工智能生成数学的出现，类似的演变也将发生[8] 。产生经过验证的演绎证明这一重担，可能会越来越多地落在计算机而非人类肩上；证明也会越来越多地被重新构建[9] ，从而将以往需要精心编排、以便人类验证的繁琐计算，越来越多地交由软件工具完成。例如，数学中一些臭名昭著的说法，如“证明留给读者”或“根据标准论证，我们有”，可能会被调用大语言模型所取代，由它为此类主张生成既便于人类阅读、又可由计算机验证的论证。随着自动形式化的发展，探究某个论证如何随数学基础的特定选择而变化，也将变得容易得多，从而能够同时严谨地讨论和探究某个结果的元数学[10]

[8]我们在此的思考受到其他数学家对这一主题的观点影响，包括 [24], [25], [16], [26], , 以及 [27]-[28]中的更广泛讨论。[9]关于这些发展的重构与进一步探索的若干具体示例，请参见 [29]。[10]这类元数学的一个例子是对定理进行逆向数学研究（例如，见[30]） ，旨在了解实际需要哪些数学公理（例如，选择公理或排中律）才能

内容本身及其数学结果。

与此同时，未来人类数学家或许会更加关注数学推理中较为“软性”的方面，例如追求某个结果或为其选择证明策略的启发式方法和动机、支持（或反对）该结果的实验性证据[11]，以及发现有效论证所经历的试错过程。这些方面不像演绎证明那样容易自动验证和衡量，因此也较难运用强化学习等机器学习策略。可以想见，职业数学家或许会越来越多地采用其他学科的论证方式，例如理论科学和实验科学，甚至人文学科，以其他类型的推理来支撑核心演绎论证，例如对实验数据进行统计分析，或依据已证实的数学结果和不严谨的哲学原则进行推测性理论构建。从历史上看，数学家一直不愿偏离严谨演绎证明这一“黄金标准”太远，部分原因在于，一旦不再遵循这类标准，就会出现许多显而易见的低质量数学成果[15]；但在未来，证明可以自动

确立某个给定结果。传统上，对某个结果进行逆向数学研究，往往要在该结果的原始证明完成许多年后才开始，而且既需要逻辑方面的专业培训，也需要对该定理所属数学子领域的专门知识。[11]尤其是，鉴于人工智能日益能够对极其复杂的数学问题“猜出”答案，却拿不出任何类似形式化证明的东西，制定标准程序，以负责任的方式引用并纳入此类未经验证的猜测，势必变得越来越必要。[12]自动化数学研究过程这些环节的另一道障碍，是数据相对匮乏；已发表文献往往侧重于结果的成功证明，却很少详述促成这些证明的过程——无论正式还是非正式的过程，这些过程往往内容丰富且细致入微。[13]尤其可以想见，未来的数学研究将出现日益明显的劳动分工：所有数学家都应大致熟悉提出、确立并解释数学结果的不同阶段，但任何一位数学家都可能越来越专注于这一过程中的少数几个方面，例如，根据某位资深研究小组成员的指示，专门利用人工智能助手证明结果；或者利用人类数学家与人工智能助手共同产出的最新文献，提出新的研究方向。[14]F

or i例如，先前的 i提议 l by Jaffe and Quinn [31] 将.....系统化 ically d 发展“理论数学”这一领域的提议遭到了职业数学家的普遍否定，引发了多篇回应，其中包括前文提到的 Thurston 文章 [18]。

[15]Kim [32] 借用货币隐喻来描述社会动态：职业数学家需要通过证明来积累一定的可信度“货币”，

如果能以高度可信的方式生成并验证，就可能有更多机会安全地探索更广泛的数学推理和讨论模式。

这些新技术也可能以显著的负面方式影响数学的长期目标。在教育层面，我们已经看到许多学生几乎立即求助于现代人工智能工具来完成指定的课程作业，以牺牲培养更持久的数学技能和直觉为代价，实现眼前的目标——为给定问题产出可验证的答案；同样，在研究层面，数据驱动数学的“第四范式” [33] 可能会大获成功，进而挤压经验经验证据、理论推理和计算数值分析这些更传统范式的空间（其中第二种目前是纯数学的主导范式），也可能削弱人类数学家[16] 从视觉、动觉及其他感官直觉中，或从我们对物理定律、经济学、生物学等的熟悉所形成的直觉中获得的巨大价值。即使假设形式化方法得到完全可信的实施，在数学研究领域不加批判地拥抱人工智能辅助，也可能导致大量[17] 主要由人工智能生成的论文涌现。这些论文中的结果在技术上正确且新颖，却无助于构建更广泛的数学叙事，也无法为作者或读者培养直觉。此类低质量作品造成的负面印象，可能会使人们对人工智能辅助在数学领域中哪怕最审慎、最负责任的应用也产生污名，

困难的新数学结果，然后才能“负担得起”将这种货币“花”在猜想构造或探讨某项结果更广泛影响等推测性活动上。[16]与此有些相关的是，诸如数学论证的“美”或“优雅”之类的审美观念，可能会比现在更加脱离这些论证的形式正确性。以 AlphaProof [34] 为 2024年国际数学奥林匹克竞赛中的题目生成的证明为例，其中包含许多冗余或难以解释的步骤，却仍经形式验证，确认是正确解答。另请参见 [25]中的讨论。[17]这可以视为意外后果定律的一个例证。在过去的数学时代，获得严谨数学证明的任务需要人类付出艰苦努力，数学活动自然会聚焦于数学界认为值得关注的问题，即使大多数数学界成员往往并未明确思考一个哲学问题：究竟是什么让某个结果“有趣”或“相关”。文献发展的速度足够缓慢，因此，这种在很大程度上依赖社会机制来判定数学意义的方式，能够随着时间推移自行纠偏。未来，随着自动化使数学结果得以大规模生产，速度显著加快，这类哲学问题可能需要得到更多主动关注。另请参见 [35], [36] ，其中讨论了在决定是否将注意力分配给一项声称成立的数学结果时，作出价值判断的必要性，包括对作者可信度的判断。

进而抑制此类技术可能带来的积极益处，例如以更广阔、更全面的方式探索数学的能力，正如上文所述。

4.7.将哲学问题应用于现实世界中的人工智能应用。任何可作为其他研究基础参考的内容都伴随着额外的责任，数学也不例外。我们可以正式认证任何人工智能生成的数学论证是否有效；但有效性只是价值的一个组成部分，在现实情境中呈现人工智能驱动的研究时，还需要做出细致的价值判断。对于研究者而言，在大量琐碎与非琐碎的发现中，哪些内容尤其有趣、值得分享给本领域内外的人，以及如何向更广泛的受众呈现这些内容，尚未在人类研究者之间形成统一规范。优先权和署名的分配方式也存在不确定性。人工智能辅助研究还带来了新的伦理和法律影响，以及尚未解答的问题：人工智能生成内容（包括证明）的知识产权归属。

研究者应遵循哪些原则，来判断某个人工智能模型是否适用、哪种应用方式最佳，或者人工智能是否根本就是一个好选择？在学术领域，假设大多数投身学术研究的人都是希望让世界变得更美好，并为此做出有意义的贡献，这并非不合理。数学家会希望优先选择对数学最有益的应用场景。各领域的研究者通常不仅希望优先考虑有益于本领域的应用，也希望优先考虑具有跨学科效益的应用。可以认为，绝大多数将人工智能用于研究的人，都会希望优先选择有益于人类而非有害于人类的用途。因此，在人工智能开发领域，重要的是明确这些工具惠及了谁、带来了哪些益处，以帮助人们了解如何尽可能负责任地优化结果。

4.8. 知识产权与责任。单是知识产权与责任（或许还有问责）的问题，就已是一片雷区，需要谨慎讨论。将人工智能应用于某个问题时，谁该为错误负责？洞见应归功于谁？这可能并非同一方，而且相关各方也未必界定明确。迄今，大语言模型（LLM）的开发者一直主张，收集大量训练数据属于“合理使用”原则的范畴。在美国，“合理使用”的适用具有一定灵活性，取决于知识产权使用的目的等因素 [37]。作为一项思想实验，我们可以思考，利益越大，是否就越值得扩大使用范围 [38]。

如果目的是拯救世界、避免迫在眉睫的灾难，那么主张借鉴所有已记录的知识属于合理使用，这样的说法是否合理？如果要拯救世界免受更遥远的生存威胁（例如气候变化），如此广泛的适用是否仍然成立？如果只是要消除所有疾病呢？或者仅仅是根除癌症呢？既然这些都被认为是人工智能的有益应用，那么允许人工智能使用所有已记录的信息来实现这些奇迹，是否合理？

除了支持对“合理使用”作极其宽泛解释的论点本身存在问题之外，制定明确的署名与引文标准和规范也迫在眉睫。人工智能使用案例所依托的，不仅是研究人员的数据，还有人工智能先前接受训练时使用的信息、人工智能训练所用信息的选择（由可能从未与主要研究人员接触过的软件工程师和设计师作出），当然还有人工智能本身的贡献。面对可能多达数百甚至数千名“隐形”贡献者，传统学术引用体系是否足以妥善署名？还是只引用人工智能模型本身就够了？研究论文中有相当大一部分内容由人工智能代笔、却未披露这一事实，引发了尤为强烈的反应；许多学者认为这种做法无异于剽窃。讽刺的是，这使得一些从这些工具中获益的研究人员反而更加隐瞒其使用情况。显然，我们需要制定有关人工智能披露和使用的全新专业标准与实践[18]。

人工智能也即将造成可能大范围出现的循环引用，这一现象被 Randall Munroe 幽默地称为“引文生成”[19]，这一说法发表于2001年。举例来说，最近人工智能“深度研究”工具 [40] 成功从鲜为人知的文献中找出未解决问题的解答后，第二作者协助发起了一项工作，在一个数学未解决问题网站 [41] 上系统地使用这些工具，报告这些问题的已有文献，或说明是否不存在相关文献。尽管这为网站带来了实际价值，我们也发现，深度研究工具会将这些报告用作搜索的权威来源，意外导致网站上的搜索结果摘要妨碍了之后使用这些工具寻找此类问题真正的新文献！因此，即使没有恶意意图，这些工具日益增强的能力也要求我们更全面地核查所引信息的来源。

[18]有关该主题的初步讨论，请参阅作者之一的文章： [39]

[19] https://xkcd.com/978

5. 人工智能的成本与收益

5.1. 经济和社会影响：谁会受益？ 鉴于人工智能已经对个人产生了重大影响，且其发展速度迅猛，我们很容易想象人工智能不断扩展，最终对整个物种构成生存威胁。每取得一步进展，开发者和其他有影响力的个人都需要仔细考虑这些进步会让谁受益，又会伤害谁。我们主张，未来的任何发展都应优先造福全人类，且人工智能应用应始终直接服务于人类（无论是个人还是集体）。

对于每个具体用例，都应进行评估，明确预期的受益者是谁。这种特定的人工智能模型或模型实现方式，究竟能惠及整个社会，还是只会为一小部分个人带来切实利益（例如成本节约）？人工智能工具功能强大、结构复杂，如果少数个人的巨额经济收益是以数百万人的利益为代价，那么这种代价将是难以容忍且不可接受的道德代价。我们必须推动人工智能的应用，使其珍视并维护人类的人性，而不是将人类商品化。

资本凌驾于人类福祉之上所导致的灾难性后果，其实并不难找到。19世纪初，诺丁汉的纺织工人自称“卢德派”，他们常被描述为任意反对技术和进步的人；他们激烈反对自动化取代自己、使自己失业，并以技能和工资更低的工人取而代之。在严峻的经济环境中，高失业率和猖獗的通货膨胀使他们的生计岌岌可危；眼前的失业威胁甚至构成了生存威胁。尽管我们如今认为工业革命时期的自动化总体上有益于社会，但这些益处也伴随着真实且可衡量的人类代价。

如今，与卢德派所处的时代不同，我们已经看到熟练工人被取代，而取代他们的不是薪资更低的人力劳动，而是人工智能。长期以来，入门级工作一直是新一代劳动者实现经济和社会繁荣的途径。当这些工作就此消失，机会也随之消失。随着人类开始直接对抗那些原本有望改善其生活质量的工具，绝望与怨愤逐渐演变为愤怒与强烈抗议。

所有新兴技术虽然都能为全人类带来某些益处，但也会付出切实的人力代价。对于人工智能这类具有颠覆性的技术，必须在地方和全球层面量化其人力代价，并谨慎权衡其利弊。我们用于这项评估的指标仍然模糊不清、定义不明。我们是否继续像以往那样，只关注货币收益与损失？我们是否应该考虑资源获取的增加，同时权衡

失去的资源？我们是否也要考虑生活质量和幸福感这些更难以捉摸的益处？如果要考虑，又该如何将这些无形的益处与更易量化的收益进行比较？

当前的商业环境不幸地在寻求一种奇迹武器，并不断优化其威力和尽可能广泛的影响，希望借此跑赢任何潜在问题。但未能量化新兴技术带来的人力成本，严重损害了全人类的利益，却只让少数人受益。此外，当前人工智能几乎同时被应用于社会的各个领域，却不考虑它是否能为终端用户带来任何有意义的益处，这只会让各行各业的人们感到疏离和沮丧。我们已经看到，当一项技术未经同意便强加于个人时，人们会自然地作出反应——他们感到失去了掌控权，第一反应就是拒绝所有人工智能技术，即使这可能导致他们把“婴儿”（能为生活带来可量化益处的人工智能应用）和“洗澡水”一起倒掉。如果我们能让技术首先着眼于切实改善大多数或所有人的生活，那么相比于只专注于将机械劳动、数字劳动和人力劳动商品化，我们就不那么容易自我毁灭。

5.2. **盘点人工智能的成本**。除了直接的人力成本，要以合乎伦理的方式实施人工智能，就不能忽视其他更难以察觉、更隐蔽的成本。开发并构建有效的人工智能基础设施，最重大且最显而易见的成本，源自这样一个现实：与20世纪70年代计算革命时期的技术不同，这些技术无法靠业余爱好或小作坊开发出来——如今已没有堆满电脑零件的车库，能让史蒂夫·乔布斯这样的创新者白手起家，打造商业帝国。现有的人工智能模型需要在硬件、服务器、人才和预训练方面进行巨额投资，之后才能开发出可用的人工智能，更不用说实现盈利了。

要理解开发人工智能所需的规模，更恰当的参照物是19世纪下半叶在美国建成的横贯大陆铁路网。修建铁路的公司必须开发并建造一支庞大的机车车队，还要规划并铺设数千英里的铁轨，之后，列车才能快速可靠地将货物从艾奥瓦州运往旧金山，实现这些公司孤注一掷所期待的经济回报。

人工智能技术巨额的前期投入，促使开发者追逐利润驱动的资本主义模式，造就了一个新的技术精英阶层。他们调配巨额投资资本并管理债务，同时精心布局，攫取并占有有限资源（土地、能源、水、熟练劳动力等），正如19世纪镀金时代的强盗贵族所为。正如

那个时代一样，这些投资的规模导致经济稳定性、技术获取机会以及整个发达国家的总体生活质量都出现了严重不平等。

我们的社会已经开始认识到，大规模人工智能所需付出的环境成本相当巨大。巨大的能源消耗和水资源消耗给生活在这些人工智能模型所需的大型设施阴影下的人们带来了显著的日常挑战。有充分理由认为（例如，见 [42]），人工智能生成的解决方案可以用于缓解甚至消除人类两个世纪以来使用技术所造成的高昂气候成本。随着基础设施不断建成，以及更高效的计算方式得以开发，运营这些工具的边际成本或许会逐渐下降。然而，迄今为止，运行中的大型人工智能模型无一能够提出解决方案，甚至无法抵消自身的资源消耗和废弃物排放。

此外，值得注意的是，现代人工智能工具并不会通过在物理世界中的呈现，或通过理解现实世界物理定律不可改变的本质，来追求或直觉地把握“真理”；相反，这些模型高度依赖人类生成的数据（且往往不注明来源），并依靠大量人类反馈来反复改进自身。如果不想让集体信息体系受到人工智能生成的信息污染，就不能在构建模型时让它们降低对人类智力劳动的依赖，否则将面临严重风险。在某个领域中，人工智能能够用来生成“新信息”的程度显然存在上限；一旦越过这一上限，人工智能崩溃 [43] 就会成为严重问题。缺乏足够的真实内容，人工智能就会脱离现实，陷入一种与真实世界毫无关联的思维模式，严重妨碍人类与人工智能界面上的有意义互动。数学凭借其形式化验证流程，或许比其他领域更能容忍人工智能污染；但正如我们所见，它也无法完全免受这一危险的影响。

5.3. **数字鸿沟**。另一个需要考虑的重要社会成本，是人工智能技术可能加剧现有不平等或造成新的不平等。从原则上说，所有人类都有能力运用自己的自然智力才能（当然，前提是获得充分的教育和支持性环境）；但前沿人工智能模型的应用趋势已表明，大规模人工智能工具可能仅向资金充足或人脉广泛的研究团队开放，或者只提供给那些最愿意交出个人数据、并对使用此类模型所涉及的任何伦理问题视而不见的个人。这在人工智能拥有者与人工智能无者之间造成了一道根本性的“数字鸿沟”。

当人工智能有能力大幅提升研究绩效时，优先确保公平获取至关重要，然而在

当前的人工智能格局中，出现了第二种更为微妙的数字鸿沟。当占主导地位的人工智能模型被资本化、私有化，并争夺有限资源（投资以及依赖其而形成的用户群）时，它们（或许是无意的）会受到激励，去发展“尖峰式”能力，以保持彼此之间的竞争优势，而非在不同领域提供稳定、均衡的表现。由于机构谈判和市场限制，个人被锁定在某一种模型中，而无法选择其他模型，因此我们必须考虑这样一种风险：某个模型会在特定研究领域赋予使用者显著优势，从而即使在能够可靠、便捷地获取人工智能资源的人群这一子群体内部，也造成分化。

另一方面，人工智能模型在科学研究和人文研究中的许多益处，并不一定需要最先进的模型。较小的“本地模型”，以及证明助手等非大语言模型技术，证明了它们能够比那些需要依托庞大数据中心、处理人类全部知识的模型更快、更高效地得出有意义的结果。我们有很大的潜力从现有的大型模型中蒸馏出较小的模型，从而利用最先进的人工智能能力，同时使用规模较小、由用户定义且经过精心筛选的训练库，针对特定研究领域的兴趣进行训练。或许，由用户群体维护的各种规模较小、目标更明确的模型，可以成为当今现有的超大型、高成本模型的一种可行替代方案。加大对这类社区项目的支持，有助于缓解获取机会不平等的问题。

尽管其中许多小型项目可由规模较小的公共和私营机构切实开发并运行，但业界人士和政策制定者呼吁采取监管措施，以创造并维护公平获取人工智能技术的机会 [44]。作为这项工作的一部分，投资发展一个面向公众的国家或跨国联盟，以开展先进人工智能研究，并开发大型、由公共资助且公众可访问的人工智能资源（或模型），将带来显著优势 [45] ，从而让那些原本会被当前主导该领域的私营企业化模型抛在后面的个人和群体，也能轻松获取人工智能。

5.4. 减害。 在航空发展的早期，航空旅行是一种极不安全的技术，发生过无数起惨烈事故。如今，航空旅行已成为长距离出行中最安全、最可靠的交通方式。正如人工智能在短期内可能导致灾难性后果，要让它沿着类似的轨迹发展（希望致命事故更少），就需要采取果断行动来减少伤害。必须制定最佳实践 [46] ，并通过培训和监管

旨在促进以最负责任的方式使用这项技术，同时劝阻或禁止隐蔽或有害的用途。

这是一项微妙的平衡。一方面，谨慎而负责任地使用人工智能辅助的个人，短期内可能会被那些不那么讲究操守、利用更快但也更不可靠的人工智能方法加速工作进度的竞争者超越。与此同时，这些个人甚至可能仅仅因为敢于考虑将这项技术纳入自己所在行业的工作流程，就遭到一群不信任人工智能的同行嘲笑、谴责和排斥。当前这种基本放任人工智能技术不受约束、任其快速发展的做法，似乎并不利于推动一种如此细致且负责任的采用方式成为主流。

我们可以借鉴一些先例，从中汲取指引。二十一世纪初，维基百科迅速发展，起初给教育体系带来了一些冲击：许多学生开始盲目地将这个在线资源中的文字原封不动地纳入作业，许多教师则以禁止使用这一百科全书资源作为回应。对维基百科不可靠且可能存在偏见的批评屡见不鲜。然而，随着该网站日渐成熟，学术界也逐渐熟悉其优缺点，人们最终大致达成共识，明确了如何将这一资源用于教育和研究。人们开始鼓励，或至少默许学生和研究人员使用维基百科作为探究某一主题的起点；学生不应直接使用其中的文字，而是应进一步查阅该网站提供的二手资料，或通过独立信息来源核实这些资料。如今，维基百科已被学术界广泛认可为一种有用的资源。

我们能否以同样负责任的态度接受人工智能？我们谨慎乐观地认为这是可能的，但这需要持续的努力和明确的哲学指引。比如，我们认为，让人工智能工具惠及所有（或至少大多数）人类，而非少数特权阶层，是道德和伦理上的必然要求；这些工具必须针对人类的实际需求提出解决方案，尽可能提升更多人类的生活质量和生活体验；同时，人们应当认识到这些工具已经造成或可能造成的危害，将其与益处进行权衡，并尽可能加以减轻。认识到这些目标中有许多在实践中无法实现，并不需要过度的愤世嫉俗；但讨论我们希望这些工具遵循的价值体系，是让这些目标真正得以实现的第一步。

当人们（希望能够）就这些价值观达成某种共识后，在配合上述旨在减轻人工智能最严重影响的行动之余，还必须将注意力转向摩擦最大的领域——人工智能与人类的交界面。若要超越这种令人不安且不稳定的休战状态，我们需要制定方法，让个人能够将

人工智能工具融入日常生活，并以令人满足、充满活力的方式使用，而非感到严苛压迫。随着人工智能不断发展演进，人类对人工智能的使用方式、互动方式，以及最终与人工智能的关系，也需要随之演变，从便捷的工具发展为辅助伙伴，再成为随时准备协作的伙伴。

6. 人类与人工智能界面

6.1. 短期视角：人工智能是智力生产中的“香草精”。我们应该如何构想人类与人工智能工具之间的界面？就目前而言，将这些技术主要视为新奇事物仍不失为一种合理看法，而且许多用户尚不确定如何恰当地应用它们。

我们建议用一个烹饪类比来指导人们应对当前的转型：香草精是大多数甜点食谱中的常见配料，以其几乎人人喜爱的香气而闻名。单独食用香草精通常令人极其不快，但少量添加普遍被认为能改善并提升菜肴的其他风味，即使它无法与这些风味区分开来。很容易得出香草精加得越多越好的结论，但大多数用过它的人都明白，超过某个上限，就会彻底毁掉整道菜 [20]。我们大多数人并不清楚这个上限究竟是多少，因此最明智的做法是只添加极少量。

类似地，我们可以将当前对人工智能的使用视为认知工作流程中的一种可选补充：适度尝试颇有意思——将人类撰写的文本交由人工智能语言模型处理，以获取语法和措辞方面的建议；或者把一份要点交给人工智能，让它整理成建议的结构。这些轻微的点缀就像少许香草精，能够提升并丰富作品的特色，却不会喧宾夺主。然而，将人工智能生成内容用作此类工作流程的核心组成部分，并不会带来理想、有效或有价值的成果。秉持这种哲学（并适当引用人工智能的使用情况），我们无需立即重新思考人类在数学、科学或创意艺术等智力活动中的作用这一基本假设。

6.2. 中期：让人工智能站上“红队”。然而，随着这些工具的能力不断增强、应用范围日益广泛，人们“选择不使用”这类技术的能力将会减弱。即便有人个人选择主动避免使用人工智能辅助，与其互动的同事、学生和专业机构也会越来越多地将人工智能融入自身工作。目前，人们严重担忧，整个学术话语领域可能会被大量低质量的人工智能生成内容淹没

[20] Tumblr 上有个著名的思想实验 [47]得出结论：含有 44% 香草精的蛋糕根本无法食用。

。短期内，可以通过严格的编辑政策禁止大多数形式的人工智能生成内容来应对这一问题；但随着这些工具日益普及，个性化人工智能代理网络也越来越常见，我们将有必要采取更细致入微的方法。

至少在中期，我们仍然可以、也有必要制定规则和指导方针，以辨识更负责任的人工智能用法并遏制不负责任的使用，同时不从根本上改变各领域的人文性质——简言之，将人工智能辅助视为工具或初级合作伙伴，而不是取代以人为中心的工作。在这种情况下，可以区分[21] “蓝队”负责生成新内容和新结构的任务，与“红队”负责验证、测试或维护这些内容的任务。人工智能用于“红队”工作相对安全，例如审查由人类生成的内容，查找错误或提出改进建议；但鉴于当前及近期工具在随机性方面的不可靠，以及缺乏根基性，将它们用于超出“红队”（可以由人类或更自动化的验证工具，例如形式化证明助手组成）验证能力范围的任何“蓝队”结构性工作，都并不安全。这种哲学的重点，是在管理使用人工智能的潜在风险的同时，仍然获取其许多潜在益处，而不是彻底重新思考该领域的根本性质。

6.3.更长远的未来：哲学上的退却是否不可避免？但不妨设想一下更遥远的未来：届时，人工智能工具目前的弱点都已得到妥善解决，其能力在所有实际维度上都达到或超过[22] 专家级人类，使风险管理哲学不再适用。面对这类先进技术变革性本质所引发的复杂哲学问题，我们又将如何应对？

一种选择是干脆退回到纯技术框架中，在那里，这些问题已不再适用。在数学中，我们有“形式主义”观点，其唯一目标是依照精确规则运用数学符号。在科学领域，务实的“闭嘴计算”哲学立场也发挥着类似作用；而在创意艺术中，人们可以做一名工匠而非艺术家，按照外部客户提供的参数创作作品，而不对产品的价值作出任何评判。在这些情况下，无须区分

[21]这种术语借鉴了网络安全领域中的一种区分：防御系统免受攻击者侵害的“蓝队”，以及探查系统弱点的“红队”。[22]这种情景有时被称为“通用人工智能”，不过人们对这一术语的确切定义尚未达成共识。

人类创作的作品与人工智能生成的作品，只要满足任务的技术规格即可。

然而，尽管技巧无疑是这些学科各自不可或缺的一部分，却无法涵盖数学、科学和艺术在实践中开展的完整体验，也无法为一些实际问题提供多少指导，例如如何激励新一代学生，或应当开展哪些由好奇心驱动的研究。因此，人们也可以退守到一种截然不同的立场：认为人类智力或人类创造力具有一种不可言喻的特殊地位，并从根本上将任何运用这些人类特质的活动与对该活动的任何人工复制区分开来，无论后者在技术层面上复制或超越前者的能力有多强。在这种框架下，人工智能永远会被视为“诉诸纯正苏格兰人谬误”：缺乏真正的“灵魂”或“理解”。长期以来，我们对自身物种十分熟悉，早已习惯人类并不可靠，能力“参差不齐”，有时还会凭借随机词语联想和死记硬背，侥幸成功完成某项任务；但当人工智能工具表现出类似行为时，人们往往会严厉得多地评判它们，例如将这些缺陷归因于其作为“随机鹦鹉”的固有本质。但或许，这种立场只是在否认一个令人不安的事实：我们引以为傲的人类能力中，有一部分实际上并没有比我们如今设计来模仿它们的人工智能算法高明多少。随着人工智能的表现不断提升，这种人类沙文主义观点也有可能逐渐沦为一种越来越站不住脚的“缝隙中的上帝”哲学：人们宣称，一份不断缩小的品质清单体现了人类的本质成就，而人工智能仍无法复制这些品质。

第三种观点尤其受到一些相关技术爱好者的青睐：他们认为，人类认知能力很快就会被人工智能对应的能力彻底取代，因此，关于人类对数学、科学和艺术的贡献及其价值的哲学讨论将越来越无关紧要。这种观点更为极端的版本甚至认为，运用人类智力本身是一种不受欢迎且令人厌倦的活动，理应尽快由自动化取代，以便腾出时间和心力投入更为闲适或享乐的追求。显然，实施这种哲学理念会带来诸多风险，例如，人类能力可能退化到这样的程度：作为一个物种，我们将集体丧失监控、控制甚至理解那些日益强大的人工智能系统行动的能力，而我们已将自己的文明托付给了这些系统[23]。

[23]若要想象这一框架在实践中的样貌，我们建议观看科幻电影《机器人总动员》[48]。

不过，在这两种“稻草人式”的极端观点之间，似乎还存在某种哲学上的中间立场，能够为人类与人工智能代理之间新兴的合作范式和互补共存提供有益的视角。国际象棋领域便有一个先例：它曾被视为纯粹运用人类智力的典型活动。几十年来，还没有哪位人类特级大师能够战胜国际象棋引擎。然而，国际象棋仍是一项广受欢迎、生机勃勃的人类活动。国际象棋棋手将引擎融入培训，用它们重新审视旧有的国际象棋理论、探索新的理论，探查原本所向无敌的人工智能国际象棋棋手的破绽与弱点，或者创造性地引入新的竞赛形式，让人工智能辅助程度各不相同。国际象棋这项游戏究竟是什么、下棋的价值何在，这些哲学问题依然值得探讨；而目前普遍接受的答案，与上述三种极端立场都相去甚远。

6.4. 哥白尼式的观点。一种可能性是，在认知层面接受哥白尼革命的类比。在古代，占主导地位的宇宙学模型（只要宇宙被视为以机械论方式运行）本质上都是地心说，认为地球处于宇宙静止不动的中心，具有特殊的本体论地位；在自然属性上，它与上方的天界或下方的冥界根本不同。然而，天文学和物理学领域的多项进展逐步瓦解了这一观点，几个世纪以来相继证明，地球实际上会绕自身轴线运动，并绕太阳运行；太阳本身又绕着我们所在星系的中心运行，而我们的星系则是一个不断膨胀的宇宙的一部分，这个宇宙不存在任何空间意义上的中心。事实上，采取一种完全相反的哲学观点极具成效，这种观点如今被称为哥白尼原理：地球只是宇宙中无数行星之一，根本定律对它不作任何特殊对待。

乍看之下，这种观点似乎相当威胁人类对家园星球的情感依恋，但归根结底，宇宙对地球漠不关心，与我们对地球怀有深厚感情并无根本矛盾；我们仍完全有理由优先关注地球特有的问题，而非其他行星上的问题，同时也承认这些行星确实存在，对各自的居民而言也同样重要。其他科学的发展史中也能看到类似的革命，例如达尔文革命推翻了人类在其他不断演化的物种中独享特殊地位的观念；又如，欧几里得几何作为数学中综合先天真理来源的特权地位遭到颠覆。

直到最近，人类一直也接受着一种与地心模型相类似的智识观念，认为人类智能处于认知宇宙的中心，因此享有特殊的哲学地位。但如今，我们正在发现（或创造）其他与我们自身在许多方面相当、同时又在许多方面截然不同的智力“行星”。与其否认这些行星的存在或重要性，或争论哪颗行星应当成为“中心”，不如承认人类智能与人工智能属于同一本体论范畴，尽管二者存在许多鲜明差异，也具有互补性。虽然我们的兴趣与依恋仍将在很大程度上与人类智力领域紧密相连，但我们可以探索它与其他智力形态之间的关系，既可以出于实际目的，更高效地实现各种现实目标，也可以出于更具哲学意味的原因，例如从外部视角审视人类认知，而这一点过去很难做到。

7. 结论

人工智能技术以缺乏规范、混乱且大范围铺开的方式进入世界，已经深刻改变了社会、知识和经济领域，带来的影响既令人警忧，也大有裨益。毫无疑问，人类需要以某种方式集体行动，无论是通过监管、市场压力，还是借助某种尚未界定的力量；但我们显然还没有到达无法摆脱这些新技术所造成的高昂经济和社会成本的临界点。将人工智能融入数学领域的方法，同样迅速展现出人工智能能为学术研究、科学进步乃至全人类带来的可观益处。数学研究在很大程度上具有客观且可验证的性质，因此为我们提供了独特的机会，可以试验这些新技术，并研究其带来的影响，而不会对个人或更广泛的社会造成伦理或生存风险。通过将人工智能应用于数学，我们能够探讨全球人工智能应用中迫在眉睫的哲学与道德问题。此外，我们还可以推演缓解人工智能与人类交互界面张力的潜在途径，并提出人工智能与人类协同思维的新范式，尊重双方各自带来的独特而宝贵的特质。虽然我们永远无法让精灵重新回到瓶中，但我们乐观地认为，随着理解不断深化、行动迅速推进，我们仍能拨开迷雾，展望一个光明却多少有些不确定的未来。

7.1. **致谢**。 感谢 Silvia de Toffoli 提供有益的评论和参考资料。

参考文献

[1] J. Jumper、R. Evans、A. Pritzel 等人，“使用 AlphaFold 高精度预测蛋白质结构”，《自然》，第596卷，第583–589页，2021年8月。[2] S. Marche，“《大学申请文书已死》”，2022年12月。[3] D. Anguiano 和 L. Beckett，“好莱坞编剧如何战胜人工智能——以及这为何重要”，《卫报》，2023年10月。[4] E. Oh、W. Kearns、M. Laine、G. Demiris 和 H. J. Thompson，“对使用人工智能的消费级睡眠技术的认知与体验”，《传感器》，第22卷，第3621页，2022年1月。[5]F. Swain，“如果纳粹研究能拯救生命，使用它是否正当？”https://www.bbc.com/future/article/20190723-the-ethics-of-using-nazi-science。[6] A. Tarkowski，“开源与人工智能的民主化”，载于《人工智能与全球治理的挑战：实现负责任人工智能的九篇论文》（A. Krasodomski 编），第30–36页，英国皇家国际事务研究所，2024年6月。[7] J. Chun 和 K. Elkins，“《人工智能危机：面向以人为本的人工智能的新数字人文学科课程》”，《人文与艺术计算国际期刊》，第17卷，第147–167页，2023年10月。[8] J. R. Searle，“《心灵、大脑与程序》”，《行为与脑科学》，第3卷，第417–424页，1980年9月。[9] Q. Mei、Y. Xie、W. Yuan 和 M. O. Jackson，“检验人工智能聊天机器人在行为上是否与人类相似的图灵测试”，《美国国家科学院院刊》，第121卷，文章编号 e2313925121，2024年2月。[10] H. Chen、S. R. Grimm、O. Russakovsky 和 T. Lombrozo，“机器理解”。未发表的预印本。[11] M. Mashaal，《布尔巴基：一个数学家的秘密社团》。罗得岛州普罗维登斯：美国数学学会，2006年。[12] 《数学体验》。波士顿：Birkhäuser，1981年。[13]R. Wagner，“数学共识：一项研究纲领”，《Axiomathes》，第32卷，第1185–1204页，2022年12月。[14]J. Baez，“《算术的不一致性》。”https://golem.ph.utexas.edu/category/2011/09/the_inconsistency_of_arithmeti.html。[15] S. Aaronson，“判断一项数学突破是否错误的十个迹象”，2008年1月。[16] S. DeDeo，“AlephZero 与数学体验”，《美国数学会通报》，第61卷，第375–386页，2024年7月。[17] T. Tao，“什么是良好的数学？”，《美国数学会通报》，第44卷，第4期，第623–634页，2007年。[18]W. P. Thurston，“数学中的证明与进步”，载于《关于数学本质的非传统论文集》（R. Hersh 编），第37–55页，纽约：施普林格，2006年。[19]S. de Toffoli 和 F. Tanswell，“数学中的技术转向”，《布莱克韦尔数学哲学指南》，2025年。[20]Y. Wu、A. Q. Jiang、W. Li、M. Rabe、C. Staats、M. Jamnik 和 C. Szegedy，“使用大语言模型进行自动形式化”，《神经信息处理系统进展》，第35卷，第32353–32368页，2022年12月。[21] F. Tanswell，“数学能被黑客攻破吗？基础设施、人工智能与……”，2025年6月。[22] K. I. Appel 和 W. Haken，《每个平面地图都是四染色的》，第98卷。美国数学学会，1989年。

[23] T. C. Hales, 《开普勒猜想的一个证明》, 《数学年刊》, 第162卷第3期, 第1065–1185页, 2005年。[24] A. Venkatesh, 《关于自动化与数学研究的一些思考》, 《美国数学会通报》, 第61卷, 第203–210页, 2024年2月。[25] S. DeDeo, 《艰难证明与充分理由》, 2024年10月。[26] J. Avigad, 《数学已经过时了吗? 》, 2025年。[27]《人工智能与数学》专题, 第I部分, 《美国数学会通报》, 第61卷, 第199–372页, 2024年4月。[28]《人工智能与数学》专题, 第II部分, 《美国数学会通报》, 第61卷, 第373–524页, 2024年7月。[29] H. Macbeth, 《形式化数学中的算法与抽象》, 2024年5月。[30] J. Stillwell, 《逆向数学: 从内部看证明》。新泽西州普林斯顿: 普林斯顿大学出版社, 2018年。[31] A. Jaffe和F. Quinn, 《"理论数学": 迈向数学与理论物理学的文化综合》, 《美国数学会通报》, 第29卷第1期, 第1–13页, 1993年。[32]M. Kim, 《思考与解释》。MathOverflow。https://mathoverflow.net/q/38694（版本: 2024-01-05）。[33]T. Hey, 《第四范式——数据密集型科学发现》, 载于《电子科学与信息管理》（S. Kurbanoğlu、U. Al、P. L. Erdoğan、Y. Tonta和N. Uçak编）, 第317卷, 第1–1页, 柏林、海德堡: Springer Berlin Heidelberg, 2012年。[34]《人工智能解决国际数学奥林匹克竞赛问题达到银牌水平》。https://deepmind.google/blog/ai-solves-imo-problems-at-silver-medal-level/, 2024年5月。[35] C. J. Rittberg, 《数学中有正当理由的认识论排斥》, 《数学哲学》, 第31卷, 第330–359页, 2023年4月。[36]S. De Toffoli和F. S. Tanswell, 《数学中的信任》, 《数学哲学》, 第1–25页, 2025年。在线先行发表。[37]《版权与合理使用 | 总法律顾问办公室》。https://ogc.harvard.edu/pages/copyright-and-fair-use。[38] A. Weir, 《第11章》, 载于《挽救计划》, 企鹅出版社（丛书）, 第191–194页, 伦敦: 企鹅出版社, 2022年。[39]《在论文中使用人工智能等工具的最佳实践》。https://ai-math.zulipchat.com/。[40] S. Bubeck、C. Coester、R. Eldan、T. Gowers、Y. T. Lee、A. Lupsasca、M. Sawhney、R. Scherrer、M. Sellke、B. K. Spears、D. Unutmaz、K. Weil、S. Yin和N. Zhivotovskiy, 《使用GPT-5开展的早期科学加速实验》, 2025年11月。[41] T. F. Bloom, 《埃尔德什问题》。https://www.erdosproblems.com/。[42] J. Cowls、A. Tsamados、M. Taddeo和L. Floridi, 《人工智能策略: 利用人工智能应对气候变化——机遇、挑战与建议》, 《人工智能与社会》, 第38卷, 第283–307页, 2023年2月。[43] I. Shumailov、Z. Shumaylov、Y. Zhao、N. Papernot、R. Anderson和Y. Gal, 《人工智能模型在递归生成数据上训练时会发生崩溃》, 《自然》, 第631卷, 第755–759页, 2024年7月。[44] 《加速研究: 运用人工智能应对全球挑战 | 总统科学技术顾问委员会》, 技术报告, 总统科学技术顾问委员会, 2024年6月。[45] E. Jones, 《"人工智能领域的欧洲核子研究中心"——国际人工智能研究组织可能关注什么? 》, 载于《人工智能与全球治理的挑战》

治理：实现负责任人工智能的九篇论文（A. Krasodomski编），第10–17页，查塔姆研究所、英国皇家国际事务研究所，2024年6月。[46] M. Mantegna，《人工智能生成未来的伦理框架》，载于《人工智能与全球治理的挑战：实现负责任人工智能的九篇论文》（A. Krasodomski编），第47–57页，英国皇家国际事务研究所，2024年6月。[47]《香草精》。
https://knowyourmeme.com/memes/vanilla-extract，2023年2月。[48] 《机器人总动员》，2008年。