arXiv:2305.02115v1 [math.HO] 3 May 2023

公认的证明：客观真理，还是文化稳健性?

ANDREW GRANVILLE

虚空的虚空！数学的虚空！——普鲁士国王腓特烈二世（1778）

Abstract. 数学共同体如何认定某个证明是正确的？基于明确公理进行客观验证是否可行，还是评审人的经验与偏见必然会发挥作用？自动证明器能否避免错误（以及经验与偏见），从而提供客观验证？而自动证明器所作的断言能否得到可证明的验证?

我们将考察一些被发现存在缺陷、但后来得到修正的证明（因为其证明计划足够稳健），以及一些被接受的“证明”最终被证明从根本上是错误的。这对当前的证明共同体标准是否值得推崇意味着什么?

我们将讨论，无论我们多么努力地开发万无一失、客观的“证明系统”，数学文化是否都不可避免地会成为接受证明的一部分。

Contents

1. 证明——为什么以及如何证明：希尔伯特关于数学理论的公理化方法。2. 与 G¨odel 危机共存并将其置之不理。3. 形式化证明与文化适切、直观的解释之比较。包括对证明方法以及语言使用的讨论。4. 什么是纯数学中的公认的证明？审稿流程的现实，以及什么才是“有趣的”。5. 错误。近期著名的错误，以及数学共同体如何应对这些错误。6. 重新思考公理与语言。投资于新理论存在风险。7. 计算机与证明。建立、辅助构造并生成证明。8. 计算机在重大定理中的应用。鼓舞人心的实例，以及计算机在这些定理证明中的作用。9. 计算机错误。显而易见的陷阱，以及我们能否避开它们。10. 自动定理检查器／证明器的协议。我们可以期待什么？11. 证明的未来。计算机证明与一项经修改的图灵测试。

题献中的引文摘自腓特烈二世致伏尔泰的一封信 [17]，信中他抱怨莱昂哈德·欧拉（Leonhard Euler）给予他的“专家建议”。特别感谢 Michael Hallett 对公理系统发展历史的解释、帮助我阐述自己的观点并指出一些值得注意的思想脉络；感谢 Rodrigo Ochigame 从截然不同且发人深省的视角发展了类似的思想；同时感谢 Peter Scholze 提出的若干意见和勘误。此外，我还要感谢 Dave Benson、Kevin Buzzard、Dan Goldston、Deirdre Haskell、Anthony Kennedy、Heather Macbeth、J´an Min´aˇc、Mel Nathanson、Peter Sarnak 和 Carlos Simpson，感谢他们进行的有益讨论、对电子邮件的回复以及提供的有用参考文献；同时感谢审稿人认真审阅本文并提出宝贵建议。作者部分研究经费由加拿大自然科学与工程研究委员会（NSERC）提供。

12. Lean 定理证明器。与 Peter Scholze 一同探讨近期非凡的发展。13. 客观性的神话。关于计算机证明正确性不断滋长的神话，与那些广为人知的、被驳斥的客观性实例相互契合。

14. 机器会改变公认的证明吗？对未来的一些保守猜测。

1. Proof – why and how

首先，我们将讨论为何证明值得追求、通常公认的证明方法是什么，以及从理论上看哪些方面存在问题。

证明的目的。亚里士多德写道

> 如果……理解如我们所设想的那样，那么论证性理解……就必须依赖于真实、原始、直接、比结论更为人熟知、先于结论且能够解释结论的事物。

更一般地，我们可以通过将新概念归约为已被接受的概念来证明它们；特别是，我们不必总是直接从公理推导出最新的断言，正如 Nathanson [49] 所作的精彩阐释：

> 我们数学家天真地相信真理。我们证明定理。定理是从公理推导出来的。证明中的每一行都是证明中前几行或此前已证明的定理的直接推论。我们的结论无条件且永恒地为真。

也就是说，现代数学家梦想着所有定理都能从适当的公理推出，而我们只需找到证明即可。然而，早在19世纪末和20世纪初，人们就发现，要确定这些说法的确切含义并非易事。

希尔伯特的设想。在早期关于数学基础的讨论中，大家的共识是在已被认可为正确的“亚里士多德式原始概念”之上继续构建，而不是质疑是否可以（以及是否应该）从任意一组连贯一致的公理出发。例如，Hallett [30] 写道：

> 与戴德金一样，康托尔认为，数学的进步取决于概念创新，其核心约束是将新概念与已被接受的概念进行“整合”，以及这些概念必须保持一致的条件。

尽管这表明，永远不太可能存在一套固定不变的公理，但人们曾两次尝试提出一套单一且足够的公理系统：第一次尝试出自弗雷格，但因其自相矛盾而存在缺陷，由此引发了伯特兰·罗素的悖论，也就是说，一对听起来合情合理却互相矛盾的陈述（导致逻辑悖论[1]；不过，G¨odel [20] 指出，这些悖论“并非出现在数学本身之中”，而是出现在对数学的“诠释”之中）。第二次尝试见于罗素和怀特海的《数学原理》，

[1]For 例如，这个极具关联性的悖论式断言：“这个陈述不是真的”。

也有缺陷，因为其中包含三条他们无法论证的公理（而且确实显得很牵强）。

康托尔关于无穷基数的研究部分是在假设[2] 每个集合都可以良序的前提下发展起来的。对这一假设的深入理解构成了希尔伯特第一问题的一部分，该问题于1904年通过Zermelo的选择公理得到解决。这些问题促使希尔伯特在20世纪20年代提出了他的基础计划，首次明确地从公理出发，建立一个用于研究数学基础的框架。在这一框架中，抽象的、理想化的数学本身构成研究对象，不再受到早期偏见的束缚，即不再预先认定某种特定途径更为可取。公理创造理论，而非反之。

希尔伯特希望公理能够通过明确的规则得到一致的发展，并且不会导致悖论。[3] 因此，希尔伯特并不像弗雷格和罗素那样，声称选择公理是因为它们是“真理”，而是因为它们是相当直截了当的断言，能够使研究对象得到一致的发展。尤其是，公理系统并不需要从“基本且不容置疑的真理”出发，而是可以从合理且一致的出发点开始。

> 希尔伯特并不像罗素那样关心“正确奠定基础”，也不热衷于找出“正确”的原始概念集合。对希尔伯特而言，并不存在“正确”的原始概念集合；对于某些目的，有些集合可能比其他集合更合适，但即使是相对简单的理论，也绝不会有真正的定论[4].....希尔伯特强调，任何理论都只是一个“概念图式”；如何赋予它内容，“由我们决定”。Hallett [30]

人们确实更倾向于补充基础理论，而不是大量使用新的公理，而是采用不同的概念框架，涵盖远超数字及其性质的广泛概念。公理无法被证明，也无需证明。

希尔伯特也关注语言与诠释问题，他要求理论的所有诠释都应彼此同构，所有演绎技巧都应在不同诠释下保持不变。[5] 希尔伯特纲领奠定了现代数学的基础，因此，让我们来阐明这些理论中若干要素的细节。

构建一种“形式推理系统”。语言往往不够精确，而且人们的理解和记忆可能各不相同。我们希望创造一种能够提升准确性、消除歧义的语言，因此我们着重关注它的规则：

我们从一组符号和变量组成的字母表开始；只要它是可数的，内容任由你定。然后，句子只能是这些符号的有限组合。接下来，我们制定书写有意义句子的语法规则；同样，只要这些规则看起来一致且灵活，内容也可以任由你定。

---

[2]康托尔在1883年将其称为“一种思维法则”，而如今或许可以称之为“不言自明”。[3]Hallett [30] 指出，这一思想的部分灵感来源于对高斯、鲍耶和罗巴切夫斯基提出的另一种非欧几何的一致性的证明。[4]其他杰出的逻辑学家，如弗雷格，认为直线就是直线，即日常语言中所指的物理实体；如果你提出一种理论，它基于一组晦涩难懂但具有一致性的公理，却得出了不同的结论，那么你就应该摒弃该理论；而希尔伯特对此持不同意见。[5]不仅同一理论的所有模型都应当彼此同构，而且无论以何种方式表达一个思想，也无论以何种方式解读它，这些思想最终都必须归结为同一事物。

唯一重要的想法是，应该存在“定理”，也就是不含自由变量的公式（也称为“闭公式”）。最初的定理称为公理，无须证明即可陈述；此后，所有其他定理都必须仅依据我们的推理规则从公理推导出来。此外，我们应该能够在有限时间内（以恰当方式定义）判定一个证明是否正确。关键在于一致性：我们不应能从公理推出某个定理及其否定。

一阶逻辑将特殊符号控制在合理的最小范围内，允许使用量词 $\forall, \exists$、公式联结词 $\vee, \wedge$ 和 $\implies$（借助它们，我们可以用 $A \implies B$ 表示A 蕴含 B），以及否定 $\neg$。这样的语言足以处理大多数直观的数学，人们希望它本身，或经过一两项必要的改进后，能够涵盖所有数学真理。[6] 但我们必须始终牢记，

> 数学是书写。尽管它能进行各种量化，也促成了种种科技与科学发现，但归根结底，它是由文字层层叠加而成的。它以定义为基石……又由公理纵横交错地贯穿其中……这些公理唯一的限制，就是彼此之间不能相互矛盾。我们从这些起点材料中推导出简洁的结论陈述，也就是定理、引理和推论。……论证才是关键。它们就是这些“证明”……没有证明，这些陈述便不过是一堆空话。证明真正 conjure 数学的存在。
>
> . – Dan Rockmore [54], 2022。

多元性。数学基础始于一组公理，但究竟是哪一组？希尔伯特的提议保留了在不同公理系统中进行工作的可能性，而这些不同的公理系统或许会对同样简单的问题得出相互矛盾的结论。那么，我们如何判断该使用哪一个公理系统才是正确的呢？对于接受无法避免演绎多元主义的数学研究者来说，这其实并不是个问题（参见 Hosack [34]），只要每个公理系统都能保持一致。让我们回顾两个著名的例子：

Zermelo 于1904年提出了选择公理（AC），不过直到20世纪30年代末，G¨odel 才得以证明它与其他公理一致。[7] 然而，真正令人震惊的是，1963年 Paul Cohen 证明了它完全独立于其他公理，因为 ¬AC 与其他公理也是一致的。因此，我们究竟要不要采用选择公理?

然后是康托尔的1878年连续统假设，即不存在在大小[8] 上严格介于整数集和实数集之间的无穷基数。1940年，哥¨德尔证明了无法使用其他公理（包括选择公理）证明存在这样的无穷基数，而在1963年，科恩证明了这

[6]将证明中允许使用的方法编纂成文，促成了现代逻辑的诞生。现代逻辑以一小组逻辑原语和一些简单的推理规则为基础。[7]对专家来说，我所说的“其他公理”是指 Zermelo-Fraenkel 集合论。[8]这里的大小是根据一一对应来定义的。

其他公理无法证明不存在这样的无穷基数。也就是说，连续统假设独立于标准公理。[9]

数学研究者很容易接受这种多元性。事实上，近年来纯数学领域一些最有趣的研究计划，为了在相关问题上取得进一步进展，不得不重新审视某些初始假设和定义（我们将在后文讨论）。Nathanson [50] 指出，这种多元性可能会导致一种错误的观念，即

> 数学是从有趣但任意选取的有限公理集和推理规则中推导结论的逻辑游戏。

事实上，G¨odel [20] 也同意这一点。

> 数学所依据的公理，其真理性无法得到证明，因此 [有人可以主张] 推导结论……不过是按照某些规则进行符号游戏，而非洞见。

尽管如此，人们普遍同意首先采用 ZF 的九条公理，也许再加上选择公理。若要得到广泛接受，任何彻底全新的出发点都需要基于充分、真正重要的理由来选择，否则就很难获得支持。例如，如今有大量研究正在探讨：需要在标准的公理中加入哪一条或哪些公理，才能“判定”连续统假设的真理。[10]因此，尽管希尔伯特的梦想已得到广泛接受，G¨odel [20] 指出，我们仍然会作出选择，正如戴德金和康托尔所论证的那样，只不过是在不同的阶段，并且对我们所做之事具有连贯性和一致性的信心程度不同。

一个关键目标。公理系统的存在不仅是为了确保正确性，也希望能提供一个框架，用来证明所有值得证明的定理。毕竟，如果我们能够陈述一个定理，那么它理应能够在我们的“理论”中得到证明；换言之，这个理论不应受到任何人为的限制。

> 希尔伯特……要求[ed] 公理化理论是完备的，也就是说，公理系统能够推导出所有重要事实，或某一类的所有事实。Hallett [30]

正如我们上文所说，我们希望创建一种（有限描述的）语言／理论，以量化和处理所有直观的数学内容；随着我们发现一些独立于现有公理的新问题，或许还会加入更多公理，从而使任何真定理都能被找到并证明。希尔伯特的梦想是，人们能够建立一套公理和理论，从而证明或证伪任何给定的数学命题。

1931年，哥德尔证明了这一梦想注定无法实现。

[9]因此，如果 ZFC 表示策梅洛–弗兰克尔集合论加上选择公理，那么 ZFC+CH 和 ZFC+¬CH 都是一致的。[10]伍丁(*)公理、马丁强迫最大性，或两者的变体——无论我们选择哪一种，都需要导向一种丰富且一致的无穷基数理论，以回答许多重要问题。

2. Living with, and ignoring, the Gödel crisis

1931年，库尔特·哥德尔以其不完备性定理动摇了数学的根基（字面意义上）：

I) 不存在一组一致的有限公理和规则[11]，能够证明关于整数的所有真定理。

更糟的是：

II) 不存在一组一致的有限公理和规则能够证明自身是一致的 t!

这与希尔伯特的梦想截然相悖：他希望通过公理化，证明或证伪任意给定的数学命题。前者令人沮丧，后者则揭示了公理化理论乃至数学基础在可能的表述方式上都存在不可避免的局限：

> 形式化的公理化集合论……不可能成为数学的最终基础理论，[然而] 也没有其他理论 [能够] 做到这一点。Hallett [30]

或者，正如冯·诺伊曼在 1930 年所说：

> 经典数学不可能得到严格的证明。

那么，数学家如何应对本学科的这场存在危机？唯一的答案是，他们学会与之共处。[12] 具体来说：

如何（不）应对不完备性。

At 一开始便有公理；它们一致、独立且无所不能 l.

它们旨在被用于永恒地证明定理，从这些公理出发，逐步构建一座高耸、能够自我论证并自我支撑的大厦——一种结构，使人能够客观地创立万有理论。

然而，人类发现，有些问题是那些公理无法回答的，因此几乎别无选择，只能用更多公理来补充原有公理，以回答那些此前无法回答的问题。但是，无论增加多少公理，似乎都不可能回答所有问题，也不可能证明所有定理；总有一些定理似乎无法仅凭给定的公理加以判定。

事实上，后来 G¨odel 证明，无论一个人从多少条公理出发、选择什么公理或语言，始终都会存在神秘之处：有些问题无法得到回答，而这些问题正是由作为出发点的公理、并以所选择的语言所创造出来的。

应当如何应对这一僵局？定理大厦的建造者们该怎么办？

当一个人住在一座极高、看似建造得十分坚固的塔楼顶层套房里，楼下发生了一场小火灾，但又难以评估造成的损失时，他有几种选择：

---

[11]我暂且回避一些必要的细节。更准确地说，哥¨德尔定理可以表述为：可证明命题的集合虽然不是递归的，却是递归可枚举的；而真命题的集合则不是递归可枚举的。[12]摘自 Janella Baxter 在 Twitter（@BaxterJanella，2021 年 10 月 6 日）上的发文：“我丈夫还是数学博士 [生], 时，哥¨德尔定理让他感到恐惧。他担心自己选定的论文选题根本无法解决。他最终完成了学业，但对于为什么更多数学家没有同样的焦虑，他感到困惑。我认为，这里有一篇数学哲学论文值得撰写……”

• 可以放弃这座塔楼，尝试重建；但前提是能想出如何按照新的消防规范重建。这么做风险很大，因为可能再也无法登上那么高的地方。[13]

• 待烟尘与喧嚣大多散去后，人们便可以继续生活，仿佛一切如常。也许余烬还会闷烧一阵，然后自行熄灭？再说了，要是大楼真的倒塌了——考虑到它大体上建得相当牢固，这似乎不太可能——那时我们也不在了，自然无从得知更多。

面对哥德尔危机，大多数纯数学家选择了后一种做法来应对：投资于现有的大楼，而不是投资于或许尚待建造的大楼。

逻辑无可回避。对许多数学家来说，哥德尔的异议不过是对着风声吹奏的悠扬口哨——代数和分析中的实际问题，或许永远不会遭遇这类认识论问题？我们当然需要一个形式推理系统，尽管它并不完备（不过看起来，只有当人们真正寻找棘手问题时，才会发现它不完备），却能在所有“合理情形”下都可靠。数学家们一直沉浸在这种乐观之中，直到巴黎–哈灵顿拉姆齐定理[51]带来了沉重打击：尽管它只比人们日常研究的拉姆齐理论稍微复杂一些，却无法在皮亚诺算术中得到证明，因为粗略地说，它蕴含了皮亚诺算术的一致性（而哥德尔第二不完备性定理指出，你无法用皮亚诺算术中的定理来证明它自身的一致性）。如今，已有一系列优美而自然的定理具有同样的性质。

数学界普遍的文化是，不必过分担心这些事情。如果研究的问题接近某个在“标准的”公理框架内被发现不可判定的问题，就需要谨慎行事，[14] 但大多数情况下，这似乎是个遥远的问题，并非我们日常工作中会遇到的。[15] 正如 Peter Sarnak 常说的，数学家“都是干活的人”。[16]

[13]谁知道呢？重建进行到一半时，他们可能会修改消防规范，造成延误并增加成本。[14]我们所说的“不可判定”，是指如果某个命题为真，就无法用该理论中的公理证明它；如果它为假，也无法用同一组公理反驳它。[15]确实有人认为，一些广为人知的未解问题在我们的公理框架内可能“不可判定”，我认为这种看法未免过于自负；仅仅因为我们尚未找到对某件事的良好理解，并不意味着它永远是个谜。Knuth [37] 甚至提出了一个牵强的论证：“哥德巴赫猜想..... [是] 一个永远不会得到解决的问题。我认为它甚至可能没有证明。它可能是 G¨odel 证明存在的那些不可证定理之一.....我们现在知道，从某种意义上说，几乎所有关于数学的正确陈述都不可证。”他接着声称，哥德巴赫猜想必定“为真，因为它不可能为假”，并为此给出了一种标准的启发式论证。这里唯一勉强说得通的是，哥德巴赫猜想在皮亚诺算术中可能不可证，因为可能存在另一个满足皮亚诺算术、但哥德巴赫猜想不成立的整数模型；不过，若真是如此，我们就可以调整公理，添加一条以确保仍然处于通常的整数之中，这样哥德巴赫猜想就应该可以证明。人们也曾对黎曼猜想、孪生素数猜想等提出过类似的荒谬主张，却没有任何实质依据。https://mathoverflow.net/questions/27755/ 上对此有很好的讨论。[16]

我对此事的态度与我对食品采购的态度颇为相似：我希望食用天然、健康、未经掺杂且不以剥削他人为代价生产的食品，因此在超市购物时，我会通过粗略查看商品标签来做出相应的选择。不过，我也十分清楚，超市所做的某些选择以及它们的标签标注可能存在问题。事实上，我并不真正相信我当地的有机食品超市能够.....

事实上，在过去一个世纪里，基础公理系统（ZFC）一直被视为现代纯数学大部分领域的基本基础，且几乎没有改变；鉴于哥德尔的结果，这一点令人惊讶，但它确实行之有效。

无法规避哥德尔不完备定理，这一点与主流文化的直觉大相径庭，以至于一些顶尖人士（包括两位菲尔兹奖得主）提出，可以通过编写能力足够强的计算机程序，以某种方式规避这些关于什么能够被证明的问题。Voevodsky 的规避方式是认为拥有一个“可靠”的系统就足够了，但他没有明确说明其含义。Thurston 似乎在暗示，错误很难被隐藏并让程序运行起来，因为程序必须在句法上正确；但这似乎混淆了两个不同的问题：大多数包含重要错误的数学论文，都是以连贯的方式撰写的，遵守我们交流系统的句法和语法——对其中某个部分的误解很少是错误所在（不过，见下文讨论的 Biss 事件）。1950年，图灵已经指出，可以调整哥德尔的证明，构造出任何给定的计算机语言都无法回答的问题，因为该问题在计算机语言的设计中是不可判定的。

可以将丘奇-图灵论题理解为这样一个工作假设：所有合理的计算系统都是“等价的”和“通用的”。也就是说，它们都能计算任何可计算的东西，尽管时间可能会成为问题。因此，不同编程语言的能力在理论上几乎没有差别（尽管它们可能是为不同目的而设计的[18]）。因此，如果一个计算系统是“通用的”，那么就应当能够利用它来模拟任何其他计算系统，其中也包括人类及其演绎系统（我们在上文定义的内容也旨在涵盖人类活动）。事实上，图灵 [62] 写道，

> 数字计算机背后的理念可以这样解释：这些机器旨在执行人类能够完成的任何操作……[19]

然后他写道

> 数字计算机……已经按照我们所描述的原理构建出来，而且……它们能够……非常接近地模仿人工计算员的动作。

是“可信赖的权威”，但我不想花时间仔细跟进每一笔购买、追究供货商的责任，除非有什么特别的情况让我起疑（或感到恶心）。[17]想必，这个想法是要定义一种看似合理的“形式推理系统”概念，从而规避 G¨odel 的结果。[18]有一次，我的一位暑期本科生研究生用 Excel 计算了孪生素数及其相关统计数据（这让我大吃一惊），很好地说明了这一点。[19]他接着举例说明，你可能会通过“家庭类比”向计算机发出什么样的指令：“妈妈想让 Tommy 每天早上上学途中顺道去鞋匠那里，看看她的鞋子是否修好了；她可以每天早上重新交代他一次。或者，她也可以在大厅里贴一张告示，一劳永逸；Tommy 上学出门时会看到告示，告示上让他去取鞋，如果他回来时带着鞋，还要把告示撕掉。”

毫无疑问，Thurston、Voevodskˇı 和其他人都意识到，对丘奇-图灵论题的这种解释意味着，我们没有希望通过不同的计算方法获得理论上更好的方式，因此我们剩下的全部就是实践中的做法。需要注意的是，图灵的评论（在上一段末尾讨论）并不与“通用性”相矛盾，因为计算机各自都能回答一切可回答的问题，但无法回答超出其设计参数（即其公理框架）的问题。

一个合适的时间跨度。是否存在这样一个数学陈述：它可以在我们的标准公理框架内得到证明，但每个证明都长到超出人类的寿命？[20] 实际上，数学问题通常以由A_1、A_2等表示的族出现（例如“对整数 n 进行因式分解”）。如果存在多项式时间算法[21]来解决每个 A_n，则该问题族属于复杂度类NP；如果每个 A_n都有一个多项式长度证明，则它属于复杂度类NP。区别是什么？对于NP，我们不要求算法找出证明，只要求证明存在，因此 P⊂NP。[22]例如，没有人知道一种快速算法可以对整数进行因式分解，但给定因式，就很容易将它们相乘以验证因式分解，因此因式分解属于NP。

简单的计数论证表明，可证明的正确陈述的 NP-族集合，只是所有可证明的正确陈述族集合中的一个相当小的子集。因此在实践中，没有任何机器、算法、人类或计算机能够指望获取从一个可信的公理系统中推导出的所有正确且可证明的数学陈述，甚至连其中极小的一部分也无法获取。

在 [16], Wigderson指出：“NP中的问题实际上就是我们.....数学家所能希望解决的全部问题，因为 [我们需要] 知道自己是否已经解决了这些问题[23].....这正是NP的定义：一个问题属于NP，当且仅当你能够检查所得到的解是否正确。”

“如果P=NP [那么] 所有这些问题都有高效算法，因此可以在计算机上非常快速地解决.....如果P=NP，那么我们试图做的一切都可以完成。[这] 就是为什么P=NP .....会产生如此深远影响。不过，我认为大多数人相信P≠NP。”

3. Formal proof vs culturally appropriate, intuitive explanation

Venkatesh [63] 指出：“证明通常被理解为一种能够促成共识的论证”，确实，

[20]在实践中，计算资源是有限的。如果宇宙中的每个原子都是一台计算机，以光速运行，从现在一直运行到时间尽头，那么步骤数仍然是有限的，因为原子的数量、光速和剩余时间都有限。[21]也就是说，对于某个 $c > 0$，对每个 $n \geqslant 2$而言，该算法至多需要 n^c 步。[22]我们并没有要求知道如何找到最短的证明，只要求存在一个如此短的证明。因此，“NP”代表非确定性多项式时间，其中“非确定性”意味着人们不一定能在较短时间内找到该证明。我们相信P≠NP，这是理论计算机科学中尚未解决的重大问题。[23]例如，物理学家提出基本粒子的存在时，情况也是如此：他们需要提出一个能够证伪或验证其存在性的实验。

> 证明的定义在于，[它] 应当能使人们对 [其] 有效性达成一致，而无需进行重复验证。

DeToffoli 的定义 [15] 是：“数学证明是基于公认前提、针对数学结论所作的正确演绎论证，并且能够事先共享和验证。”此外，

> 论证的公开性使其他数学家能够进行质量控制。这对于筛除非证明至关重要……分享[数学论证…… [是] 数学论证成立的必要条件。

然而，如果证明由共同体来评判，那么能否说一个证明具有客观有效性？事实上，De Toffoli [14] 认为：“严谨证明的可接受性标准并非一成不变……而是与特定时期的数学共同体相关。”

更进一步说，人工证明验证器最满意的情况是：检查完一个证明后，他们能够用自己的话重新推导出它，或类似的证明。这样一来，新知识就不只是人类全部知识总和的一部分，而是被置于具体语境之中，并作为更宏大图景的一部分而为人所知。在柏拉图的《泰阿泰德篇》（148e-151d）中，苏格拉底指出，缺乏严谨性可能导致

> ……思路的失误 [因此] ……他们自己和其他人都能明显看出理解上的欠缺。

我认为，这恰恰说明，从旧思想中推导出新思想并将其置于具体语境中，有积极的作用；而不是把新思想孤立地抽象考察。

不过，这些考量并没有涉及“客观证明”。我们当然希望有客观证明吧？也就是说，经过验证、在任何语境下都成立的思想。如果是这样，我们该如何实现？是将证明及其理解纳入更大的理解体系之中，还是确保证明能够从公理推导出来——我们或许可以称之为“形式化证明”。Nathanson [49] 代表许多人这样写道：

> 我们如何知道一个证明是正确的？逐行检查。

Hales [28] 解释了什么是形式化证明，以及为什么有些人觉得它很有吸引力：

> 形式化证明是一种证明，其中每一步逻辑推理都经过核查，一直追溯到数学的基本公理。所有中间逻辑步骤都一一列出，无一例外。整个过程不诉诸直觉，即使从直觉转化为逻辑的过程很常规也是如此。

这比 Hacking 的莱布尼茨式形式化证明理想 [26] 略微更进一步：

> 每一步都经过细致安排，并且可以用机械化的方式逐行检查，

也就是说，形式化证明要求人们（至少原则上）构建一条逻辑路径，从适当的公理推导出该主张。

或许最常见的担忧是，在逐行阅读一份形式化证明时，人们可能会被每一个步骤所说服，却无法把握整体，也无法察觉其中的全局性错误。

人们也可能会问，这样一个形式化的（或莱布尼茨式的+）证明要如何验证？我们只是因为形式化证明者（无论是人还是其他存在）的说法就相信它吗？如果不是，那么谁会是负责验证的“独立权威”？这位独立权威又具备哪些验证所需的技能？他们会逐行检查，从而可能只见树木、不见森林吗？还是说，这位权威对其中的运作有着强烈的直觉（如果是这样，他们又是从哪里获得这种直觉的）？Hales 接着写道

> 因此，形式化证明直觉性较弱，却也较不容易出现逻辑错误。

该如何理解这一说法？哪些错误不属于“逻辑错误”？为什么形式化方法必然能降低忽略问题的可能性，尤其是在不要求读者了解全局的情况下？

关于形式化证明的更多逻辑谬误。形式化证明试图避免使用“直觉”，因为它不够精确。人们认为，只包含形式化步骤的阐述能让独立的权威更容易核验证明。1895年，皮亚诺写道：

> 不精确的想法无法用符号表示，

这意味着由符号表示的思想必须精确。即便我们认同这一非同寻常（且未经证明）的主张，也很容易对此产生误解，以为只要把数学问题转化为符号，就能保证不会出错（而不是保证不会不精确）；当然，一个人完全可能既精确又出错！假设我们谨慎对待含义。那么，负责核验的“独立权威”是谁？如果这个权威理解有限、直觉不足，细微的错误是否可能蒙混过关？[24] 因此，皮亚诺提出了一种形式化方法：先识别解释中的基本数学思想，再设法用有限的语言表达这些思想（[56, 第 2.3] 节）。此外，皮亚诺希望通过各种“恒等式”，将其形式语言中的证明简化到尽可能少的符号，从而得出数学因此得到了更好的理解。即便我们认同在这种情况下数学确实得到了更好的理解，当人们把皮亚诺最短的证明转译成标准词汇，以便更多人理解时，又会发生什么？这会是最理想的证明吗？它能帮助我们看清接下来的结果吗？

此外，形式化证明验证器能看出如何独立重现该证明吗？如果无法轻松重现，它真的算是经过验证了吗？它因而属于人类知识的总和，但它本身真的为人所知吗？

形式化证明通常会将证明的细节一路追溯到公理。这就像一个孩子不知疲倦地追问“为什么？”（直到追溯到不可改变的真理），但在这一过程结束时，孩子还记得自己一开始问了什么，以及是如何走到终点的吗？这样的证明并不比“因为我这么说”这个答案好多少，无论谁是客观上值得信赖的权威。当我们作为积极参与者

[24]可能会出现许多错误，不仅是数学方面的错误，也包括例如符号和语言解读方面的错误。针对任何已发现的问题，设计一种明确无歧义的协议并不难，但我们能确定自己已经考虑到了所有需要通过协议来避免歧义或误解的可能情况吗？

在一个共同体中，我们期待获得能够理解、诠释、欣赏，并在可能的情况下加以运用的答案。我们也可能会因发现一种替代性的或更清晰的证明而感到兴奋，尽管这在形式系统中并不起任何作用。

还有哪些其他类型的证明？Eugenia Cheng [11] 写道，哲学家[25] 认为，

> 有了“证明”这一概念，我们就有了一种极其严谨的方式，来判定数学中的哪些内容为真、哪些不为真，

但数学家认为，形式化证明过分注重精确性。那么，如果我们对形式化证明如此怀疑，还有什么方法可行？传统证明与形式化证明之间存在着巨大的鸿沟。Hales [28] 写道：

> 传统数学证明的写法旨在让数学家容易理解。其中常规的逻辑步骤会被省略，而读者则被假定拥有大量背景知识……受过训练的数学家 [可以将] 这些直观论证转化为更严谨的论证。

也就是说，在理解和赋予传统证明语境的过程中，会涉及大量的默会知识。这体现了 Hacking 的笛卡尔理想 [26] of a proof:

> 经过一番思考和学习，人们完全理解了 [这个证明] ，并能在脑海中 [领会它]。[26]

但这并未完全解释是什么让我们相信一个证明是正确的。程[11] 指出，

> 尽管证明据称能够确立数学命题无可否认的真理，但证明本身实际上并不能让数学家相信这一真理……真正起作用的是其他因素。[27]

因此，数学研究者写下的“证明”是令人信服的论证，或许可以转化为形式化证明；其主要目的在于增进知识与理解，同时保持一定程度的严谨性，使读者相信沿着这一方向还能做更多工作。Mancosu [42] 指出，这类证明

> 这并未直接涉及某些传统的基础性问题，例如确定性，而这些问题一直主导着数学哲学的大部分讨论。

尽管如此，我们相信，这样的证明，尤其是在被广泛理解的情况下，是“稳健的”，因而不易出错。这长期以来一直是“共同体标准”：

[25]我认为她想写的是“某些哲学家”，事实上也包括某些数学家。[26]Hacking似乎认为，要使证明令人满意，证明的所有细节都应当能够同时理解。尽管这当然是可取的，但我不确定相信某个证明是否必须满足这一条件，只要人曾在某个时候对证明的每个部分以及各部分如何衔接起来感到确信即可。[27]事实上，在1877年6月写给戴德金的一封信中，康托尔展示了一个从欧几里得二维空间到欧几里得三维空间的双射，并评论道：“我看到了，但我不相信。”

> 没有任何一位数学家能对自己的科学如此精通，以至于在发现证明后便立即对其完全确信……他每次重温自己的证明，信心都会增加；朋友们的认可更能增强他的信心；而学术界普遍的赞同和赞誉，则使他的信心达到尽善尽美的程度。
>
> ——David Hume（1739）

Hume 认为，人们之所以相信某个证明，是因为它经得起技术娴熟且对此感兴趣的同行审视，而非因为它经过了某种抽象的验证。这套体系并不完美。已发表且被认可为真的证明，有时也会被发现存在重大错误，[28] 然而我们仍然坚持这套体系；它必定有其优点。

优秀证明的用处。阅读证明时，积极从事研究的数学家希望拓展自己的直觉和视野，而不只是认同该证明的论证正确。读者并非被动接受。她希望理解、融会贯通，并在自己的研究中运用这些思想：

> 数学……[并不] 奖励被动接受。理解一篇数学论文，就像根据建筑师的图纸在脑海中想象一座建筑：文字和公式不过是蓝图，读者必须借此在脑海中重建作者想象的世界。然而，如果她做到了，那么最好的数学理论便拥有令人屏息的魅力，犹如巴黎的景象层层折叠、最终与自身重合。[29] 这种体验既令人振奋，也令人沉迷。
>
> —— Izabella Laba [38]

不同的人阅读同一篇文章会有不同的收获，因此，一篇新的研究论文可能会激发出新的想法，并将思考引向此前未曾预料到的方向。即使是同一个人，在不同的时间阅读一篇文章，也可能获得不同的收获。我们的理解确实会随着时间而改变，有时甚至连我们对整个领域的研究方式也会发生变化。

大多数读者不会追溯证明所依据的公理，而是依赖已发表文献（一个可靠知识的宝库），依赖已有的知识。这意味着，读者会在合理次数内追问“为什么？”，直到确信某个证明的正确性（这与形式化证明中可能出现的情况不同）；至少在读者具备足够的相关知识时是如此，也就是说，他们读过并基本理解了这篇文章所依据的全部或大部分参考文献。

总而言之，证明是按照共同体标准被接受的。这意味着它们可能是错的，因为我们并不指望所有细节都经过极其仔细的核查；那么，证明的目的是什么呢？或许证明的思路才是主要内容，它说明哪些想法串联起来，从而证明所提出的定理，让读者

[28]而且我们猜测，还有许多尚未发现的错误。[29]这是对电影《盗梦空间》中造梦师场景的引用。

能够学习、记住并重新运用这些想法。[30] 细节通常不那么重要，尤其是经验丰富的读者能够自行补全这些细节，尽管有时需要付出令人不快的努力。

撰写证明。证明是面向特定读者群体的一种解释。

> 数学解释取决于语境，因为“在不同语境中，不同特征可能会更加显著”。– Mancosu [43] 引用 Lange [40]

面向该领域专家（研究论文的首批读者）的解释，与面向该领域初学者的解释不同；而面向领域外人士的解释又有所不同。这里谈的是所需的细节程度，以及对论证如何相互衔接的解释。如今，通常的做法是：论文的摘要和引言前半部分面向广大读者；引言中的概述面向该领域的专家（包括初学者）；而细节则面向专家。

解释可以通过举例和测试用例、类比已经理解的结果，或在某些突出主要思想的特殊情形下证明结果来进行。对数学家而言，“好的证明”既能证明，也能解释；事实上，解释越好，遗漏繁琐细节所带来的风险就越小。不过，还有一个更广泛的群体可能希望阅读并理解某个结果的证明，因此这些细节也可能更有用。

程 [11] 声称，数学交流的目的是通过证明，将作者的信念转化为读者相信的真理。仅仅陈述信念并不能令人信服；只有附上看似合理的推理，同事才会开始认为它是真的。但什么是推理？我们的通信对象会相信这些理由吗？如果没有证明的严谨性和结构，也没有将理由纳入恰当的框架，读者可能仍会心存怀疑；至少，与以证明形式呈现的、论证充分的论证相比，读者更可能心存怀疑。这是因为措辞严谨的证明有助于消除对歧义或误导的担忧。

数学真理即美，美即真理。是什么让一段数学内容“感觉正确”？程 [11] 声称，这关乎“理应如此”，而不应与“有用、有趣、引人入胜、优美、经过详细证明”混为一谈。这一视角有助于理解许多人证明定理时所采取的方法背后的动机。

多产且极具影响力的数学家保罗·埃尔德什声称，客观的至高存在拥有一本“书”，其中记载着每个真命题的完美证明，而每个证明都简短而优雅。简短，便于验证；优雅，则让人明白命题契合得如此完美，必然为真。这是专业纯数学家一个美妙的设想：当一个概念成熟到可以被理解时，它就应当欣然奉上最丰美的果实。

伟大的格罗滕迪克不相信通过大步推进或举例来阐明进展：他认为，当一个课题时机成熟时，理论自然会清晰明了，因此没有必要

[30]程 [11] 将此比作合法性（证明）与道德性（从原理上理解证明正确）。如果证明与理解是同义词，那自然再好不过，但事实并非如此；我们必须体会并接受它们相契合之处，也要看清彼此的差异。

过于急切地推动进展。[31] 逐案分析或许复杂而笨拙，却能帮助研究者理解其中的模式，而这些模式可能最终导向一种证明，使逐案分析变得多余。证明会随着时间逐渐发展，不同作者会选择在这一潜在过程的不同阶段发表成果。[32] 有些数学研究者不愿发表任何内容，除非是“最终证明”，即能够一目了然地看出某个概念为何成立。[33] 另一些人则不那么挑剔，会发表尚不完整的理论；但阶段性成果可以鼓舞他人，增强共同体意识。各种做法都各有其用。

研究论文的语言。在很大程度上，数学家遵循图书馆模式，在以往论文的基础上开展工作。定理是研究论文中的一项重要新成果。由于研究论文建立在知识库之上，它们通常会引用已有结论，也可能对已有结论做简单修改，以满足自身需要——这些就是“引理”。如果这样一个简单推论本身看起来也很有趣，那么它可能会被称为“命题”。这些词的用法并没有严格固定的规则，不过其变体并不多。

有人可以认为，无论语言定义得多么清楚，都没有两个人会以完全相同的方式看待它。事实上，我们已经习惯了日常用语中的词语，比如“认识论的”，它似乎被用于一系列彼此相关却又略有不同的含义。但语言是我们传达证明的方式，而它很少能成为完美的工具。因此，我们希望描述提供了足够的细节，让读者能够从我们写下的内容中“弄明白”或重构我们的本意；他们或许会以自己的方式理解，而这种理解可能略有不同（也许还能由此带来新的发现）。这说明了形式化证明和语言的价值，但我们也看到其中存在陷阱。

4. What is an accepted proof in pure mathematics?

Hales [28]写道：“哲学家告诉我们，数学由分析真理构成，不含任何瑕疵”，并接着“证明了 $1+1=2$。”随后他写道

> 要是所有证明都这么简单就好了。数学错误与数学本身一样古老。欧几里得的第一个命题要求在给定直线上作一个等边三角形。欧几里得的作图隐含地假设（而公理并未证明这一点）：两个圆各自经过对方的圆心，就必定相交。我们敬仰欧几里得，并不是因为他事事正确，而是因为他为我们指明了正确的道路。

公认的证明的核心思想十分简单：从公认的公理出发，我们构造出给定命题的证明。众所周知，罗素和怀特海表明，如果我们的公理彼此一致且在逻辑上相互独立，那么许多看似简单的命题要证明起来却异常费力。因此，要在数学领域取得长足进展，我们就不能总是回到公理。我们需要

[31]在我看来，这是一种令人悲哀的进步还原论。[32]这是反对“引文索引”的诸多论证之一。许多人会不断改进那些尚未成熟的好点子，因此这些点子常常被引用。而一个构思精妙、证明优美的突破，可能很难再得到哪怕一点改进，因此被提及的次数也会少得多。[33]其中也包括高斯，他的“座右铭”是“少而精”（英文意为“少而成熟”）。

建立一个命题库，收录我们知道为真的、表述明确无歧义的命题。传统上，这个命题库保存在研究论文中，并汇编成“书”。

然后，研究者可以从我们图书馆中能够引用的内容出发进行逻辑推演，从而推进全人类知识的总和。如今已有数百万篇文章，因此系统中很容易混入大量错误。文章可能存在缺陷，引用的陈述也可能被误解（甚至被错误输入）。

同行评审流程。为了尽可能确保文章的正确性，我们不会仅仅相信研究者所说的他们已经进行了正确的逻辑推演。优秀期刊通过指派一位或多位匿名的“专家审稿人”对投稿文章进行评审，以确保质量控制。这些评审人/审稿人通常首先会判断文章所声称的内容是否真实、新颖且足够有趣，是否值得在该期刊上发表。众所周知，Littlewood 会问[34]

> 这是新的吗？正确吗？令人惊讶吗？[35]

然后，审稿人应当仔细审查文本，寻找逻辑谬误。所声称的内容可能确实为真，但其中存在错误，这些错误或许可以纠正；也可能是给出的解释难以理解或有些含糊。评审人不仅可能指出错误并要求进一步解释，还可能提出重大修改建议（他们认为这些修改会有所改进），从而参与帮助文章打磨至最终定稿。这样做的目的是让未来的作者觉得，这些“文献”值得关注，可以放心引用和采信。

专家审稿人的风险。审稿人通常因在投稿论文所涉主题方面具有专长而被选中；作者往往也会在论文中引用他们。一些审稿人会维护自己的研究成果，在评审他人关于自己偏爱主题的研究时格外挑剔。另一些评审人则乐于看到其他研究者参与探索相同的问题。这些评审人的态度显然会影响他们如何评审，以及他们认为什么是可以接受的。作为多家期刊的编辑，我会寻找那些热忱、慷慨、积极推动本领域发展的评审人，同时他们对自己要求很高；因此，他们评审他人的研究时，会期望工作尽可能做到最好，但也愿意帮助一些作者更深入地思考自己的论证，使最终发表的成果确实能推动该领域的发展。

任何有判断力的读者都会看出，我刚才所描述的这一过程中存在的陷阱。偏见可能通过各种方式介入其中，而编辑必须相信评审人能够做到公正，尽管偶尔也会出现评审意见明显不妥的情况。[36]冷门研究领域的从业者相对较少，[37] 因此，所有专家

[34]剑桥哲学学会《数学论文集》的执行编辑 Ben Green 也要求他的编辑们说服他，他们推荐的文章确实“有趣”。[35]Venkatesh [63] 认为，“我们赋予一项数学研究成果的价值纯粹是主观的，也就是说，它完全取决于人们对该成果的看法，而不取决于任何客观品质”。[36]不过，这种情况并不常见。[37]Venkatesh [63] 写道：“现代数学的规模和复杂性意味着，大多数论文对我们来说几乎难以理解；因此，我们对它们的看法只能重复他人的看法。能够参与形成对某篇论文或某个问题的看法的人，只有

他们很可能彼此认识，或者至少互相听说过，并且经常在会议上碰面。他们有许多相同的研究目标，因此往往既是合作者，也是竞争者（或两者兼具）。因此，很难指望他们在评估彼此的工作时能够做到真正的中立。

那么，有什么替代方案呢？非专家能否不偏不倚地评判投稿文章？毕竟，任何投稿都应在逻辑上建立在已有文献的基础上；如果文章写得清晰，读者应该只需顺着论证读下去即可。这种想法想当然了很多。比如，阅读并理解已发表文献中的内容究竟有多容易？最新突破可能会引用过去几年里的多项重大进展，因此，要理解新投稿中的论证，审稿人必须能够理解并领会不久前的前沿成果。事实上，要评审一项高质量工作，审稿人必须熟练掌握该领域近期的顶尖成果，因此她在某种程度上也会成为专家。[38] 例如，最近一家顶级期刊请我评审一篇长达104页的论文。作者在文中以自己和他人的研究为基础，而这些研究在过去五年里已发表于顶级期刊，篇幅合计又有数百页。虽然我是该领域的专家，但为了做好评审工作，我还是花了好几个星期；尽管我对作者的技术能力有足够信心，并不觉得自己需要核查每一个细节。作为专家，这是合理利用我的时间（这样我才能跟上自己研究领域的最新进展），但很难指望一个没有深入钻研这些问题和方法的人拿出几周时间，甚至可能发现自己无法胜任评审工作。

检查细节。读者读到同行评审时存在这样的情况——“我觉得没有必要检查每个细节”——或许会感到不安。从纯数学的角度看，这显然有问题——审查一篇文章是否有错误，却觉得不检查细节也无妨，这怎么说得过去？但这正是专业知识发挥作用的地方。如果熟悉这个领域，往往就会遇到一些“大致上”属于“标准的”证明，因此专家审稿人可以看出证明整体上似乎正确，还可能迅速找出该证明与文献中此前类似证明不同的部分，并把精力集中在那里。对非专家审稿人而言，一切都是新内容，所以他们必须检查每个细节，因为他们不太清楚该具体审查哪些地方、潜在错误可能出现在哪里。此外，专家审稿人更有能力帮助作者修正错误论证，因为他们过去或许也曾在同一个问题上陷入困境。

错误往往源于对文献的误用，例如断章取义地引用某项结果：大多数文章都是供读者从头到尾阅读的，但作者在陈述某项结果之前，可能已在前文提出了一个假设

以某种方式与之互动的人。现在，亲自研究任何论证细节的人群规模非常小；更大的一群人则是了解它与其他现有成果之间的关系。”[38]因此，只有专家才能恰当地进行同行评审——这听起来像是一个受哥德尔启发的论证！

，并要求全文都采用该假设。[39] 随后，研究者逐字引用这项结果，却忽略了前文的假设，因此误用了该结果。非专家审稿人或许会核实所引用结果的表述，却不太可能通读整篇被引用的文章，因而也不太可能发现这个问题；而专家可能更清楚某项陈述的结果在什么情况下才可能成立。

我希望现在我们都认同，专业知识在同行评审中有其优势，而且在实践中确实不可或缺，尽管无论如何，同行评审都仍然是一项艰巨的工作。上文我主要讨论了“前沿工作”，也就是那些真正带来重大影响的进展。然而，大多数投稿并非颠覆性成果，而只是小幅进展，或许也更容易评审。尽管如此，同样的注意事项依然适用，只是程度较轻。在这种情况下，专家确实能发挥关键作用，帮助那些有好想法、但或许尚未掌握足够技术能力以将想法贯彻到底的作者。根据我担任编辑的经验，许多审稿人在这种情况下都会给予鼓励和帮助，尤其是当作者还“默默无闻”时，他们会解释作者可以如何修改已有成果，使论证正确，或使定理更强或更具一般性。[40]

最后，如果只有专家才有能力审阅前沿研究领域的重大进展，那么当一份投稿包含真正的新想法时会发生什么？如果它与以往的做法截然不同呢？如果所有潜在的审稿人都照着同一本赞美诗集吟唱，他们是否都可能对这支新曲子的有效性持怀疑态度？尤其是当它解释得不够清楚，或包含无伤大雅的技术缺陷时？当然，起初可能会有（可以理解的）怀疑，但现行体系往往运作良好。例如，57岁的无名数学家张益唐在2013年宣称，他以全新而出人意料的方式运用最深奥的思想，在素数间隙问题上取得了非凡突破，学界很快便承认这是最高水准的研究进展。[41] 不同期刊各自独立运作也有所助益：例如，Mordell证明椭圆曲线上有理点有限生成猜想、并提出了他那后来由Faltings解决的著名猜想的论文，起初遭到伦敦数学会拒绝，[48],大概是因为当时这并非热门课题，但随后被剑桥哲学学会接受。[47]

当构建一个并不有趣的理论时。在 [52] 作者写道：“根据对‘现有成果体系’的构想来判断……工作的原创性”

[39]例如，在我所研究的解析数论领域，人们可能会写道：“在全文中，我们令 $f(x)$ 为一个施瓦茨类函数”，而不是在每条引理和定理的陈述中都重复这一点。[40]因此，我反对“作者双盲评审”，也就是向审稿人隐去作者姓名，因为这会使审稿人不愿给予宽容：审稿人似乎往往会认为，匿名作者“理应更清楚”，不该犯这样的错误，于是只会简短地指出错误，而不会提供有帮助的解释。[41]起初，专家们不太愿意深入研读一位无名作者如此艰深的手稿，但一开始阅读，他们很快就意识到这是一篇非同寻常的作品。

“知识”由秘密知识和可能的知识共同构成，这种说法在智识上是冷漠无情的。当然，如果有人这样表述的话。

一些作者由于未能考虑社群的文化，而与仲裁过程产生冲突。例如，在组合学中，人们学习证明，然后经常根据新的情形对其进行调整（这可能需要大量新的输入）；一般定理的价值很有限，因为它们往往只能涵盖已知情形以及缺乏趣味的推广，而无法应用于新的、真正有趣的情形。[42] 在 Serge Lang 生命的最后几年里，他对一家又一家期刊独立拒绝发表他在解析数论领域这类一般性研究的做法感到十分沮丧。他的观点是，如果他的精确结果尚未在文献中被明确陈述，那么任何审稿人都无权声称该结果是“已知的”（一种公理化观点）；而实践者则认为，所提出的推广几乎没有增加什么有价值的内容（一种社群观点）。同行评审中有一个流传已久的笑话：

> 新颖的事物并不有趣，而有趣的事物并不新颖。

依我经验，如果投稿中存在这种情况，那么无论报告的论证多么充分，作者都可能难以理解报告。

已接收（待发表）论文中的论证是否正确，责任在谁？作者、评审人、编辑，还是期刊？一般认为，无论论文经过多么严格的审查，正确性的责任始终在作者。审稿人的职责并非确保论文正确，而是应当认真、尽力核实细节。

证明的可靠性。我们希望建立这样一个体系：已知“得到证明”的数学命题能够可靠地正确且始终为真，因为证明会通过一系列经过核实的步骤，展示某个定理是这些步骤的逻辑推论。证明理应确保定理成立，因此有能力的数学家不必亲自核实自己所用的每个证明（尽管这样做更好）。然而，以目前的专家同行评审证明验证体系，这实际上难以实现。

我们已经讨论过专家同行评审的优势。它很大程度上基于这样一种观点：熟悉会使审查更为简单（尽管这可能带来偏袒一致性的风险），从而使论文得到更快速、更准确的评审。我们提到过这样一种想法：审稿人以前见过某一类论证，便觉得没有必要审查每个细节，或许只关注那些根据审稿人的经验最有可能引起疑虑的细节。这支持了这样一种信念：至少在高水平的研究中，证明具有稳健性。我们相信，经过充分使用的技术工具不太可能出问题，因此愿意对哪些内容需要核实作出假设；这些假设通常是正确的，但也未必总是如此。即便出了差错，经验也表明，只需稍作修改，论证就能成立。每位有经验的研究者在开展自己的工作时都会经常这样做，因此，当他们发现他人工作中的轻微技术缺陷时，往往会认为这些问题

[42] Gowers [24] 解释了这种视角的价值：“尽管这种结构不像许多其他学科中的结构那样显而易见，但它确实存在，表现为一些较为模糊的一般性陈述。这些陈述使证明能够在头脑中得到凝练，因此更容易记忆，也更容易传授给他人。”

是可以解决的。事实上，许多审稿人会在报告中写下这样一句话："我不需要查看修订稿，因为我相信作者能够处理这些问题"。

5. Mistakes

我们的专家同行评审证明验证系统也可能出错，而且确实会出错。重要的研究成果被广泛应用，而其中的错误有时会在发表后才被发现。通常，这并不是因为有人对细节进行了更加仔细的核查，而是因为另一位研究者在应用所声称的研究成果时，推导出了错误的结论，或者得出了不合情理、至少是出乎意料因而值得怀疑的结论。下面我们来讨论一些值得关注的例子。

5.1. Goldston 和 Yildirim。2003 年，Dan Goldston 和 Cem Yildirim 宣布，在当时关于素数间最短间隔所能证明的结果上取得了非凡进展：我们相信孪生素数猜想，即存在无穷多对相差 2 的素数；但在当时，甚至还不知道是否存在无穷多个小于平均间隙十分之一的间隙，而 Goldston 和 Yildirim 声称，他们可以把这个比例降到任意小的正数。两人发布了一份预印本，其中论证的新颖部分写得详尽清晰，易于验证，不过一些技术引理的证明只作了概述。尽管如此，这些技术引理都有文献中类似命题作为依据，因此人们相信它们是成立的；大家认为 Goldston 和 Yildirim 之后只需补上细节即可。证明看起来很稳健，他们的工作也广受认可，知名数学家还向科学媒体称赞这是一次重大突破，等等。

向他们询问其中两个引理的更多细节后，他们回应说，其中一个引理一直难以攻克，但他们预计很快就能解决，届时会发布一个经过完整证明的修订版本。这听起来似乎合情合理，因此 Soundararajan 和我随后继续深入发展 Goldston-Yildirim 方法。令我们惊讶的是，我们成功推导出存在无穷多个大小为 6 16 的素数间隙，这几乎就相当于孪生素数猜想了！然而，总觉得有些地方不对劲。当我们仔细追溯推导过程时，发现那个棘手的引理似乎蕴含了过多结论，这最终使我们找到了一个反例。[43] 这个故事再次凸显了基于稳健性来接受证明所面临的问题，但除此之外，似乎也没有明显可行的解决办法。

不过，这个故事有一个圆满的结局。通过与 Janos Pintz 合作，Goldston 和 Yildirim 找到了他们所做工作的正确版本，最终促成了张益唐的突破、Maynard-Tao 对 Goldston-Pintz-Yildirim 方法的改进，并最终让人们知道存在无穷多个

[43]对于了解复分析的读者而言，关键问题在于如何在极高维度中进行积分时移动积分路径。高维几何有一些方面不同于我们所熟悉的维度。Goldston-Yildirim 引理在低维度中有效，但在足够高的维度中却出了问题。

相差不超过 246 的素数对，这是数学史上伟大的成果之一。[44]

5.2. 怀尔斯。1993年，Andrew Granville 在剑桥艾萨克·牛顿研究所的一场小型会议上，连续作了三场演讲，宣布自己已经证明了费马大定理（FLT）。他采用了一种极为深刻的策略，该策略受到谷山、志村、韦伊、弗雷、里贝和塞尔的研究启发，旨在证明某些椭圆曲线族可以按照猜想所述的方式参数化（即模性猜想）。怀尔斯曾询问会议组织者约翰·科茨，能否安排三场各一小时的演讲；他拒绝告诉科茨原因，但表示这样做会很有价值。我当时也在会上，流言四起。当时显而易见的是，怀尔斯对他的证明给出了经过精心打磨的概要，尽管要在短短三小时内讲清楚七年思考所得的结论，时间远远不够。怀尔斯大胆地制定了一套卓越的研究计划，要从弗雷、里贝和塞尔已完成的工作继续推进，最终证明费马大定理。每一步要么在当时的技术条件下看起来可信，要么他已有计划使之变得可信，而且至少勾勒出了具体细节。我们都知道，补全这些步骤或许需要对计划的具体细节作些调整，但他相信自己的方案足够稳健，能够经受住可能需要勘误的几处差错。我的印象是，他在尚未百分之百准备就绪时就作了演讲；不过，这场会议将有世界上大多数专家参加，地点又是怀尔斯母校新建的研究机构，他觉得与会者会更愿意现在就听到这项研究，因为他们可以确认各个步骤都可行或至少合理（甚至或许还能提出有益的建议）。[45]

众所周知，当怀尔斯继续补充所有这些细节时，证明中的一个部分无论他如何努力，都开始显得越来越不可靠、越来越难以解决，以至于他觉得自己需要撤回关于费马大定理的主张，至少暂时如此。在他的工作尚未公开发表、也未经审稿之前，为什么共同体就已经如此准备好接受怀尔斯的主张了呢？毕竟，任何人都可以声称自己取得了这样的成果。[46] 答案在于共同体文化。怀尔斯此前已经取得了几项重大成果，改变了我们对代数数论的认识；他当时已经是普林斯顿大学的教授，而且他的每一篇新论文都会引起轰动。此外，他此前的研究工作中并没有出现过任何已知的严重错误。[47] 他对自己如何攻克这一问题（的部分内容）的描述。

[44]此外，还有第二个圆满结局：格林和陶调整了原始的戈德斯顿–伊尔德勒姆筛权，完成了他们的证明，证明素数中存在任意长的算术级数。[45]据我记忆，有位参与者提到，怀尔斯使用了一个已知为假的结果，但类似的结果应该为真，也应该能够证明。（我试着向那位参与者核实此事，但他不记得了。）[46]事实上，很多人都这样做过。在怀尔斯定理出现之前，我们许多人常常被业余人士声称自己证明了费马大定理的说法淹没。此后，我们收到的这类说法甚至更多，至少有一段时间如此，但拒绝逐一阅读时，我们觉得没那么内疚了。[47]其实我不知道他是否犯过任何错误，但这无关紧要。我加上“严重”一词，是为了强调只要证明的整体思路连贯且可行，就没人会太在意小错误，尤其是对专家而言。

模性猜想对在场的专家而言是可行的；事实上，他们迫不及待地想拿到一份手稿。[48]

当怀尔斯被迫撤回自己的主张时，他着手寻找补救论证的方法。起初只是对论证的那一部分作了相对细微的重新思考，后来则进行了更大幅度的调整，但他始终沿用同一个总体计划。他知道自己的计划很稳健；事实上，他的证明计划中的几个部分已经发展成了各自独立、颇有趣味的小领域。他相信总会有办法填补这些空缺，问题只在于找到一条可行的路径。[49] 怀尔斯着手“修补”时（后来他的前学生 Richard Taylor 也加入合作），学界的顶尖研究者对此反应各异。少数好胜之人告诉媒体，怀尔斯所做的离证明还差得远；但大多数人似乎都在为怀尔斯加油，称赞他已经取得的成就，认可他所开创的方法之精妙，并公开表达希望，相信他能够完成这项艰巨的任务。[50]

5.3. Biss。21世纪初，Daniel Biss 在同伦论领域取得了一些重要进展，并在《数学年刊》等顶级期刊上发表论文（事实上，他的工作发展了杰出数学家兼《数学年刊》编辑 Robert MacPherson 的思想）。Biss 的证明经历了你所期待的、来自《数学年刊》这类期刊的严格审查。然而，论文中还是存在错误；事实上，关键错误隐藏在命题 4.5 中。专家们原本觉得 Biss 的结果“可行”，但这个错误使他的主要结果失效。Biss 在《数学发明》和《数学进展》等顶级期刊上发表的其他论文中也出现了类似错误，而这些论文很可能由其他顶尖专家审稿。那么，这里的审查体系究竟出了什么问题？专家们是否都想法过于相似，以至于轻易接受看似合理的结论，让这样的错误悄然漏过，甚至连最负盛名的期刊也未能察觉？然而，情况还要更糟。Biss 的其他论文也存在错误；例如，2017年，《拓扑学及其应用》“撤回”了 Biss 2002年发表的一篇论文，理由是“论文中的定义含糊不清，而且大多数结果都是错误的”。

在纯数学中较为抽象的一些领域，这类问题出现在定义或最简单的论证中，并不完全罕见（因为这些内容往往最少受到审查）。整体论证的脉络优美而有说服力，因此，即使作者公开发表时某些细节尚未完全推敲妥当，通常也只是一个很快便能解决的技术问题。Biss 的错误由圣彼得堡的 Nikolai Mnev 发现；他悄悄通知了作者和其他专家，并期待 Biss 能迅速解决这些问题或发表撤回声明。Biss 与一些相信这些问题可能得到解决的同事合作，但他们也因此感到沮丧。时间流逝，四年后，Mnev 觉得他

[48]肯·里贝特在怀尔斯宣布费马大定理后的第三场演讲之后发表了演讲。里贝特的演讲开始时，满屋子的数学家都兴奋不已，却不太知道该如何自处。里贝特先是有些紧张地笑了笑，然后说：“在那之后，连我自己都不在乎接下来要说什么了”。[49]那条路径需要比原先更进一步、更深入的理解。现在回想起来，他公开宣布自己的证明，反而让事情变得更加困难，因为他随即感到压力，必须把它完成；而此前他已经在远离聚光灯的地方潜心工作了七年。[50]还有一些人认为，怀尔斯应该发表他已有的成果，让其他人“试着”修补那个漏洞。

应该通知更广泛的群体（在 arXiv 上发表自己的观察），因为他担心其他人会在未意识到 Biss 的工作可能存在错误的情况下，继续发展这项工作。发表勘误或勘误表可能需要四年之久，似乎时间很长；但在极其深奥且困难的研究中，确定前进的正确路径可能需要一段时间，尤其是在压力之下。

在最后这两个例子中，学界群体长期没有对错误作出判断，直到情况得到澄清。我们或许无法给出一个正确答案，说明何时应该放弃一个出色且看似稳健的证明计划。

6. Rethinking axioms and language

6.1. 望月新一的重新思考。望月新一是算术几何领域备受尊敬的研究者，在远阿贝尔几何方面作出了基础性贡献。2012年，他宣布证明了 abc 猜想——这或许是丢番图算术中最根本的未解问题。该证明源自他创立的“宇宙际泰希米勒理论”（IUTT）；他声称，这一理论对丢番图分析中的代数几何方面进行了大规模的重新思考。望月新一采取了截然不同的做法，舍弃了算术几何发展的诸多成果，并以他认为“更适合目的”的新理论取而代之。望月新一在创立这套理论时毫不妥协：语言是新的，概念据他所称也是全新的、不同的；他还指出，这些内容无法转译为算术几何中惯常使用的语言和结构。望月新一完全有权创立一套涉及新语言和结构的理论。只要能够证明他的证明以通常方式、经由某种路径从恰当的公理推导而来，那么这些证明就是正确的。[51]

然而，在对他的作品进行同行评审时，出现了一个实际问题：望月新一的研究已经偏离得如此之远，以至于没有其他专家能够判断其正确性；而通常看来最有资格评判这一领域研究的人，也不愿在没有同行帮助的情况下作出判断。望月新一使用一种新的语言，他那四篇声称给出了 abc 猜想证明的预印本长达约 500 页。他没有尝试将自己的思想与标准词汇体系进行比较，也没有用常见的算术几何对象作类比来帮助理解。此外，他拒绝了那些专家提出的请求，没有提供更标准的解释，并表示既然他的著作中一切都定义得很清楚，其他专家就有责任从第一性原理出发研究望月新一的著作。这意味着他们不得不在相当长一段时间内搁置自己的科研生涯；不过可以说，如果他们最终认同望月新一所提出的变革具有重大价值，那么这样做也会让他们受益匪浅。

望月新一的证明确立了 abc 猜想的一种形式，这与通常的猜想略有不同。随后，Vesselin Dimitrov 推导出了一个更具适用性的望月新一结果版本，并发现了反例，因此最初声称的结果

[51]事实上，许多顶尖数学家（尤其会让人想到格罗滕迪克）都发现，恰当的语言和定义能够改变深层问题的面貌：原本远非可解的问题，可能变成从理论中必然推出的结论。的确，Scholze [32] 最近写道：“我们通过定义提供的透镜来感知数学世界，而定义能否聚焦于关键要点至关重要。”

望月新一的证明可以被证明是错误的。望月新一声称，即便存在一些小瑕疵，他的证明依然稳健；随后他弱化了一项主张，使最终结果无法被 Dimitrov 的例子反驳。尽管冗长而复杂的证明需要一些小修订并不罕见，但在此案例中，这个极其微小的改动让人们不禁疑惑：新版本是否可能存在尚未发现的反例——那又该怎么办？可以接受多少次“微调”？

一小群资深研究者研读了望月新一的著作，并认为其中的思想和逻辑是合理的。在研究过程中，望月新一继续修订自己的工作，以便让这一过程更容易进行。然而，这仍未能让大多数顶尖算术几何学家愿意深入钻研宇宙际泰希米勒理论，因此学界既无法令人满意地接受望月新一的主张，也无法令人满意地将其驳斥。[52]

世界顶尖数学家之一 Peter Scholze 承担起一项责任：尝试将望月新一的手稿充分转译为更通行的数学语言，以便自行判断。他们认为自己理解了望月新一的策略后，Scholze 和 Jakob Stix 找到了一个他们不知该如何继续推进的环节（推论 3.12），于是于2018年前往京都与望月新一讨论此事。望月新一与他们会面，却声称他们的译文对宇宙际泰希米勒理论作了无效的过度简化，因此他们的异议并不成立。然而，他未能说服他们相信问题出在他们没有抓住要点。他们发表了一份报告 [59]，其中 Scholze 和 Stix 声称：“小幅修改无法挽救这一证明策略”。

这种情况下，通常的体系已经失灵，因为大多数数学家都希望看到的工作，是作为对既有成果的补充而提出，并遵循一定的传统。这不仅体现在写作上，也体现在交流中。数学家希望能在会议上或通过电子邮件相互询问彼此最新工作的相关问题，并希望对方能用所有相关人员都最方便的语言作出澄清。[53]

6.2. Voevodskˇı。Vladimir Voevodskˇı 是一位极具直觉的数学家。他信赖自己的直觉，却在发现自己与 Kapranov 合作完成的一些工作有误时深受震动。或许他的直觉并不足够，他们的一些论证中存在致命错误，甚至还有反例。[54] Voevodskˇı 在 [53] 中指出了专家同行评审制度可能出现的问题：“一位受人信任的作者[，比如他]，，提出的技术性论证难以核查，又与已知正确的论证相似，因此几乎从未有人会对其进行详细核查。”

他的这一错误促使他提出重新思考本领域基础的主张：由 Voevodskˇı 发起的单值性基础计划试图调和 Martin-L¨of 式的

[52]这里并非评论其正确性的地方，因为我们关注的是这个共同体在这一特殊而复杂的情境中是如何运作的。[53]例如，普林斯顿的 Nick Katz 对某些代数几何对象有着深入的了解，并以一种我难以欣赏的方式进行写作。但在讨论时，他愿意抛开这一切，尽可能用通俗易懂的方式解释，并体谅我的局限。这可以说就是“共同体标准”。[54]从 Carlos Simpson（经由 Kapranov）让他意识到这个错误，到他承认错误，花了15年，因为他想弄清楚自己的直觉究竟错在何处，而不愿纠结于一个或许可以纠正的细节。

依赖类型论与传统的证明数学处理及范畴构造相结合。这需要以新的方法和新的语言，重新思考许多基础问题。不过在这一领域，这项计划的主要设计者之一 Dan Grayson 不遗余力地 [25] 尝试让这些基本思想更易于理解，甚至让其他数学家也能理解，正如我们从文章标题《面向数学家的单值基础导论》中所能推知的那样。稍后我们会再次讨论这项新的研究计划。

Voevodskˇı 也于 1996年提出了一般布洛赫–加藤猜想的证明计划。这为最终在 2011 年完成证明提供了路线图。在此过程中，Voevodskˇı 计划中的部分内容得到证明，部分则被舍弃并替换（最显著的是由 Chuck Weibel 和 Markus Rost 所做的工作）。然而，整个策略和计划足够稳健，能够找到绕过障碍的方法（尽管这需要许多杰出数学家的创造力）。这是由计划引领的社群方法的一大优势。

7. Computers and proofs

Frenkel 在 [53] 中被引述说：

> 随着这一宏伟的数学事业不断攀登新的高峰，其峰峦变得愈发错综复杂、晦涩难懂，每位从业者都深入钻研那些细枝末节，即使是同一领域的同行也根本无法理解。计算机将成为推动这一事业不断发展的必要工具，并用于验证这一结构的完整性——这一结构在细节上极其复杂，在规模上极其庞大，以至于仅凭人类审稿人进行评判将不再可行。

计算机在证明中有三种主要用途：

- 它们可用于进行计算，以建立证明；
- 它们可用于协助验证作者论证的逻辑，或许还可以交互式地进行。这些被称为“计算机辅助证明”；
- 它们可用于证明所声称的定理，即所谓的“计算机生成的证明”。

现在我们按相反的顺序简要讨论这些用途。

计算机生成的证明。它们尚处于起步阶段。常见观点包括：

> 我不相信由计算机完成的证明.....如果我理解一个证明，我才相信它。– 皮埃尔 · 德利涅 我对计算机完成的证明不感兴趣 ..... 我更愿意思考 · – 约翰 ·H· 康威 [53] 计算机生成的证明 ..... 教给我们的东西非常少 [55] – 弗拉基米尔·沃沃多斯基ˇı[53]。

[55]并继续说道：“如果我们转向计算机生成的证明，那么我们就会失去数学中所有美好的东西——作为精神修养的数学，以及帮助塑造纯净心灵的数学。”

由于这些顶尖研究者不希望仅仅因为计算机如此断言，就相信某个主张为真，我们或许可以反驳这些观点，坚持认为计算机生成的证明应当被设计成便于人类阅读，并且不仅要尝试解释“如何”，还要解释“为什么”。在讨论近期发展时，我们稍后将回到这一观点。

Kahle [36]声称，计算机一旦找到并描述出一个证明，验证它是否正确就相对容易，即使过程可能很繁琐（因此很适合交给计算机）。然而，找到证明却很难。“尝试每一种可能性”的想法会遭遇众所周知的复杂性问题，但有一些策略可以改进这一点，我们将讨论 Ganesalingam 和 Gowers 提出的另一种方法。

计算机辅助证明。Russell 和 Whitehead 以艰苦的工作表明，对于任何并非极其平凡的结果，形式化证明中的逻辑推理数量都过于庞大，难以适应，因此整个探索似乎只是为了探索本身才有意义。因此，当 G¨odel 终结了他们最初的目标时，也不算太令人遗憾。然而，形式化证明正在卷土重来！它们不再是“用绳子围起来、只能静静欣赏而不能直接触碰的博物馆展品”（[28]）。如今，计算机拥有足够的内存空间来处理几乎任何证明的长度，而且（我们希望）具备足够的逻辑资源，确保不会遗漏任何步骤（不过，这也引发了人们对计算机计算结果是否可信的疑问，下文将对此进行讨论）。

未来，证明验证可能会采用“计算机辅助证明”，因为作者可以向经过适当设计的证明检查器交互式地解释她的证明。事实上，2008年 Harrison [33] 在英特尔撰文指出，他进行形式化的目标之一是

> 用客观且可机械化的证明正确性标准，补充甚至部分取代主流数学论文的同行评审过程。

即使在今天，这也会过度占用作者的时间，但在某些情况下仍可能有用，例如在 Biss 和 Voevodsk˘ı 的著作中，我们上文讨论过其中有许多必须正确的新定义。

我们当然必须警惕相信计算机验证证明，理由依然是：我们正在将数学翻译成一种专门的语言。[56] 为了论证这一点，可以重新援引皮亚诺的信念：在一种经过适当设计的语言中，可以消除误译并获得“精确性”，从而消除错误。然而，即使这是可能的，并且你已经实现了这一梦想，你又将如何证明自己已经成功了呢?

在 Voevodsk˘ı 关于其单值性基础计划的工作中，他觉得自己已经无法再信任自己对细节的把握，于是设计了一个证明助手。在 [53] 中，他声称证明助手可以让你保持诚实：你制定计划，它构建枯燥的细节。当它无法完成时，你就必须进一步完善计划；这就像与一位神话般较真而热情的同事一起工作。Roberts [53] 写道，Voevodsk˘ı 当时是在

[56]事实上，证明验证软件是硬件验证软件的衍生物；如今，软件能够形式化验证高级计算机语言或微处理器（以及两者之间的任何事物）是否按预期运行。例如，Leroy [41] 为 C 编程语言开发了一种形式化验证编译器。

与计算机比剑。他指示计算机试试这个、再试试那个，很像是在玩电子游戏。

作为证明的一部分进行计算。当计算机首次能够生成数量惊人的、可用于证明的数据和信息时，人们曾一度难以确定应要求作者提供什么，才能使读者相信该证明。如今的共识是，作者必须对这些计算给出连贯的解释和论证，并对其进行详细描述，以便感兴趣的读者能够轻松地独立重复这些计算。[57]

8. Uses of computers in major theorems

大多数证明都包含作者所做的计算。事实上，作者通常会断言他们正在研究的对象具有某些性质，而不觉得有必要详细论证，因为该领域的任何人都知道或至少会预期这些性质。这些并非定义，而是断言所研究的对象是某种熟悉事物的实例，因此遵循我们对这类对象所预期的性质。作者需要提供多少细节，并没有确切的规则：这取决于她对读者的预期，以及她认为读者对她的预期！例如，对于给定的多项式，可以不作进一步解释就断言它没有有理根；也可以说“根据高斯有理根判别法”；还可以逐一检验高斯判别法得出的有限种可能性。有时，论证此类断言所需的计算相当繁琐，因此作者可能觉得有必要说明如何进行这些计算，以便读者核查结果，而不是直接给出数据，因为这些数据或许完全没有启发性。[58]

我们首先讨论一些非凡的计算曾作为证明一部分的例子，然后讨论证明助手后来如何提供了帮助。

素数间的小间隔。在 Maynard 关于素数间小间隔的研究中[45] ，他需要构造一种具有特定性质的筛法，并证明这种筛法可以通过构造一个次数为 6 11、含有 105 个变量的特殊多项式得到，从而使该多项式的某个函数为 > 4。要验证 Maynard 的证明，可以严格按照他的步骤进行，也可以在某个阶段偏离他的路径，看看能否通过另一种方式得出

[57]尽管如今这似乎不言自明，但在文献中可以找到由数学研究者和哲学家撰写的关于计算机验证的“哲学讨论”，其中表达了即使有适当协议，人们在接受此类计算时仍感到不安。接受技术变革起初总是很困难：例如，1994年，万维网普及几年后，MSRI 举办了一次关于“数学交流的未来”的研讨会，以讨论其潜力。与现在一样，互联网中充斥着大量未经编辑的无意义内容，研讨会上的许多资深参与者似乎对能否创建一个达到印刷期刊标准的在线期刊持怀疑态度。即使有人提出了显而易见的解决方案（使用审稿人），仍有几人未被说服。[58]尤其是在需要为一个简单性质验证大量情形时。

相同的结论。[59] 通过清晰地解释，他让读者有机会核验自己的整体思路，并判断他的证明是否稳健。相比于计算细节，说明每一步背后的思路会更有帮助。

上帝之数是20。这指的是从任意初始状态还原魔方所需的转动次数。也就是说，无论魔方一开始被打乱成什么样，都能在20步内恢复到初始状态，而且确实存在需要20步才能还原的状态（Michael Reid于1995年证明了这一点）。Morley Davidson等人于 [12] 2010年提出了这一结论。魔方大约有 4.3×10^{19}种可能状态。合作者借助一些群论知识，将问题简化为在20步或更少步数内还原约5600万个状态，然后利用35个CPU年闲置计算机时间（由Google捐赠）彻底解决了这些状态。找出这20步，远比按照已经写好的20步操作难得多，因此使用他们的输出文件进行核查，甚至独立核查，并不太难。[60]

有限群分类。这是另一个总体性且稳健的计划范例，随后需要填补数量庞大的各个部分。在这个案例中，即使该计划本身过去和现在都具有非凡的深度与复杂性，以至于似乎只有少数人对其充满信心（不过这种情况随着时间推移已有所改善）。填补其中的步骤注定会是一项庞大的工作，需要许多参与者，也处处存在出错的可能。然而，这里的许多步骤本身就很有趣，因此有更多人愿意参与这一伟大项目。随着进展推进，新的结构不断出现，计划也随之调整，以便解释更多内容，同时在技术上变得更加简洁。一位资深参与者 Michael Aschbacher [1, 2] 坦率地指出，有时他认为分类工作已经完成，有时又不这么认为；而且他始终确信其中到处存在一些小错误。他相信该计划具有稳健性，相信最终会沿着所宣称的路线完成完整分类，尽管期间会有一些修改，有时甚至是重大修改。到目前为止，这一分类尚未经过计算机验证。

四色定理（四色定理）。有大量（哲学）文献表达了对四色定理最初证明的认识论上的困惑，其中一些文献误解了所谓证明中发生的事情，因此我来试着澄清一下。关键概念依然是“稳健性”。所有证明的思路都是：如果四色定理存在反例，那么就可以将其“化简”，从而找到一个属于某个有限可计算集合的“最小反例”；然后证明这样的最小反例并不存在，因为它们必须具有过于特殊的性质（或者还能进一步化简）。[61] 迄今为止，这一过程一直很复杂。

[59]多年来一直思考类似问题，因此当 Maynard 第一次告诉我他得出了常数 > 4 时，我对此持怀疑态度。我不仅按照 Maynard 的建议进行推导，得出了相同的答案，还尝试了不同的修改，因为我担心他富有说服力的文字掩盖了某个错误。[60]实际上，一台典型的个人电脑一天可能能够执行 10^{15} 次基本的计算机操作（例如将两个数字相加），这意味着一天内检查 $< 10^8$ 个位置应该很容易，但一年内检查超过 10^{14} 个位置则不切实际。[61]这在图论中并不是一种罕见的策略。

在 Appel 和 Haken 于1976年给出的四色定理原始证明中，他们关于如何化简的论述组织得不够完善，部分内容也难以核验。他们需要证明1478个不同的子图都可以进一步化简；他们使用了计算机，却没有提供令人信服的文档记录。此外，按其描述，计算机算法相当复杂（包含487个步骤），难以核验其有效性以及是否不存在自相矛盾。尽管如此，Appel 和 Haken 确实提出了一套令人信服的四色定理证明方案，人们可以看出每个主要步骤的作用，以及它为何应该奏效。问题在于细节；即使这些细节都得到了正确实现，也始终很难核验、相信并在此基础上继续推进。[62] 整体方案具有稳健性，但具体实现并不令人信服。

1996年，四位杰出的图论学家 Robertson、Sanders、Seymour 和 Thomas[55]（RSST）决定重新整理现有证明，使其更令人信服。他们沿用了 Appel 和 Haken 的总体思路，但试图简化具体实现。他们只需借助计算机处理 633 个子图（数量仍然不少，但作者煞费苦心地让整个构造过程清晰透明）。更重要的是，他们的计算机算法仅包含 32 个步骤，因此该领域的优秀数学家只需花上一两天，就能相信作者已成功涵盖所有可能情况。这份新证明经过刻意编写，便于核验（尽管仍然冗长复杂），而且没有严肃的数学家会怀疑四色定理已经得到证明，至少按照共同体标准是如此。也就是说，RSST 的这份证明具有足够的稳健性，人们有理由相信，未来研究者若发现任何误解，都能轻松加以弥补。

几年后，Gonthier [22] 使用 Coq v7.3.1 证明助手验证了 RSST 证明。他开发了一套“形式化证明”，涵盖了 32 个步骤以及需要约化的 633 个子图；该证明随后“ [只] 依赖于若干计算机硬件和软件组件的正确运行”，而且并非专为这个证明而设，因此显得更加稳健。最重要的是，这里使用的许多程序此前已用于其他计算，这让人觉得，不太可能有专门针对这里所做工作的程序错误悄然混入。此外，Coq 可以生成证明的“见证”（尽管人类无法读懂）。[63]

三维球体堆积。Hales 在 2005年 [27] 发表的开普勒猜想证明，研究的是三维空间中单位球的堆积问题[64] ，并借助了大规模计算。与四色定理的证明类似，Hales 的策略是将开普勒猜想的一个最小反例归约为某个有限可计算集合中的若干种可能的球体排列，而这些排列

[62]我听说，当时发现了不少错误，但每个错误都可以通过对细节作相对小幅的修改来处理；然而，这类修补实在太多，以至于该领域很少有人相信最终成果的细节。[63]见证是对问题已正确解决的一种相对简短的验证。例如，要向你证明我已经分解了 $14757395258967641 2927$，我只需给出因数 $193707721 \times 761838257287$；你不必重复我得出这些因数的步骤。只要把这两个因数相乘，就能进行验证。[64]1606 年，沃尔特·罗利爵士请一位数学界熟人帮忙计算船上能堆放多少炮弹；到 1611 年，开普勒在其论文《论六角雪花》中猜想，蔬果商用来堆放圆形水果的六边形排列（以及一种与之密切相关的替代排列），是尽可能将最多球体挤进一大片空间的最佳方式。

必须满足一些极为苛刻的约束。随后便可通过大规模计算枚举这些可能性并逐一排除。然而，案例数量实在太多，而且同行评审极其繁重，以至于 Lagarias [39] 写道

> 这种证明的性质……使人类难以可靠地检查每一步…… [D]对许多具体断言进行详细检查后，发现它们在所有情况下基本上都是正确的…… [这] 使这些评审人对这一证明方法基本正确深信不疑，也相信约化方法得出的是规模可处理的非线性规划问题，

这是有条件但恰当的认可。2017年出现了第二个证明，使用了 HOL Light 和 Isabelle 证明助手。 [29]中完整描述了每一步的目的，但研究者显然仍然需要对许多内容予以信任。

9. Computer error

我对计算机证明系统最大的担忧之一，是人们往往认为，只要程序经过良好实现，就值得信赖，继而把“值得信赖”误认为“毫无错误”。

我们都接触过据说“从不出错”的计算机系统，但它们还是会出错，无论是信用卡、电话公司、银行、航空公司还是所在大学的计算机系统。原因可能是编程或实现错误、系统无法理解你所处的特殊情况，甚至是硬件问题。这些问题可以避免吗？我们能否修正这些程序和计算机，使其不再出故障？

Dealing 面对所有可能发生的情况。图灵 [62]写道：

> 不可能制定出一套规则……来描述在所有可以想象到的情境下， [to] 应当如何行事。例如，人们可以制定这样一条规则：看到红色交通信号灯时应当停车，看到绿色交通信号灯时则应当通行。但如果由于某种故障，两种信号灯同时亮起，又该怎么办呢？人们或许会认为，停车是最安全的选择。然而，这一决定日后也可能引发其他问题。试图制定一套行为准则，以涵盖所有可能发生的情况，甚至包括由交通信号灯引发的各种情形，似乎是不可能的。

那么，即使一个程序的准确性和可靠性令人印象深刻，当出现升级时又该怎么办？如果我们避免升级，[65] 是否最终能够使一个系统达到完美？即使一个系统在能够对所有可能出现的情况作出恰当响应这一意义上是“完美的”，我们又如何知道这一点呢？（也就是说，我们如何知道自己已经列出了所有可行的情况？又该如何证明这一点？）

[65]其缺点在于无法利用新的发展成果。1990 年，Donald Knuth 决定不再升级 TeX：不添加新功能，只进行错误修复；大多数科学家更愿意使用更新的排版系统，例如基于 TeX 的 LaTeX，这些系统会定期升级。

计算机硬件可靠性。在非学术领域，计算机硬件和软件的黄金标准是再也没有任何投诉，[66] 但这并不意味着能够保证证明验证！英特尔前总裁安迪·格鲁夫曾说过

> 没有任何微处理器是完美的；它们只是越来越接近完美。

在讨论硬件可靠性时，广泛使用的计算机芯片中一个微小的数学错误，就有可能被利用来攻破广泛使用的密码协议，从而使整个电子安全体系面临风险。这种情况确实有可能发生。事实上，1993年，奔腾推出了一款芯片，随后发现该芯片存在一个影响其浮点处理器的硬件漏洞。奔腾没有召回有缺陷的芯片，而是决定保持沉默，并通过软件更新来修正问题。1994年6月，Thomas Nicely 在尽可能精确地计算一个数论常数时发现了这一错误。随后，人们发现，在多个使用浮点运算的软件中，将某些七位整数相互相除时，都能够检测到这一错误。即便如此，奔腾仍然拒绝召回芯片，直到 IBM 拒绝出货其产品。[67] 我们是否应该进一步阐述这个故事如何影响我们的讨论?

——也许如今芯片中存在一些更隐蔽、因而更不易被发现的错误。此外，如果我们不独立分析大规模计算的结果（正如 Nicely 将自己得到的答案与已知答案进行比较那样），又如何能发现这些细微的错误呢?

——芯片输出中看似微小的错误，真的值得制造商付出成本去修复吗？此外，制造商很少会透露其产品存在的问题（以免吓退潜在购买者），因此，即使他们知道存在某种缺陷，这些信息又有多大可能被广泛传播呢?

计算机软件可靠性。计算机程序也存在许多问题。Hales [28]指出，商业软件每百代码行大约包含 1 个程序错误，而当程序极其注重谨慎时，每 10,000 代码行或许有 1 个程序错误。[68] 他写道

> 公司……把重大程序错误从账簿中抹去，以限制法律责任……只修复那些影响利润的程序错误。

此外，修正程序错误可能会带来问题，因为这一过程往往会产生新的程序错误。[69] Harrison [33] 指出，物理工程项目中或许会有更多错误，但由于其中的问题通常是连续的，一个小误差影响不大；而计算机工程处理的是离散问题。尽管这使人有机会找到精确解，但也意味着一个小小的错误，比如输错一个数字，更有可能导致重大故障。

[66]换言之，即与其他信息明显不一致的输出。[67]奔腾声称，这次召回给他们造成了4.75亿美元的损失。可以理解他们为何犹豫不决！[68]就像航天计划中的情况一样。[69]当你正在处理的事情还历历在目，而且既能把握全局，又能看清细节时，对其进行一些小修改是一回事；但多年后重新着手处理某件复杂的事情，只为进行一项希望只是小幅度的修正，而你又不想为此花费太多时间时，就很难重新找回当初的视角了。

对于这些特定的计算机问题，或许唯一（合理）的解决办法是：在不同芯片驱动的不同计算机上，使用不同的软件重复进行相同的计算。这样一来，多个不同系统不太可能在“证明”的同一处出错。如果两个足够独立的系统得出的结果一致，那么出现问题的可能性似乎微乎其微。因此，我们认为，在研究证明验证时，可以放心地忽略这些计算机实现问题，尽管这对如何保证（即证明）一个证明是正确的，意味着什么仍不清楚。

10. Protocols for automated theorem checkers/provers

证明检查器应具备哪些理想特性？我们上文讨论过的证明验证（四色定理和开普勒猜想）都需要人与计算机交互，从头开始构造证明。这些证明并非人类可读，但可以尝试确认人类提出的进一步主张，例如一些细微变体，以增强我们对它们的信任。事实上，如果它们能通过帮助我们继续深入而展现出自身确实能带来理解，我们或许会对计算机证明更加信服。

然而，一个无法被详细理解的“证明”，真的是证明吗？我们已经讨论过，证明的目的不仅是确立真理，还要增进理解。如果它做不到这一点，那对整个社群又有什么用呢？

如果我们读不懂证明验证器或证明器的输出，为什么要信任它们？如果证明助手只能通过自身的内部逻辑来检查，它们还能自我纠错吗？的确，它们很可能会继续传播某个细微的错误。

这些程序当然需要一个社群来验证它们的证明吧？也许可以使用不同的程序来独立验证它们的输出；实际上，我们提议由各自的社群对计算机证明验证器的输出进行同行评审！这意味着它们必须使用一种共同的语言，从而给不同的设计增加额外负担。如此一来，人们或许能通过独立的计算机验证，相信经过计算机验证的证明；证明验证器也因此成为值得信赖、客观且专业的权威（也就是审稿人）。计算机程序将借助其社群获得有价值的认可印章。这样，我们就能以已有的有效做法为基础，设计未来。[70]

证明的呈现。若要相信一个可追溯至公理的证明，我们需要一种共同的语言，以便对其进行独立验证。我们已经看到，让人类进行验证并不可行，而且人们从这一过程中获益甚少，那么怎样才能让它更有用呢？显然，证明验证器也可以输出人类可读的证明。它可以学习人们能够理解和欣赏的高层次论证类型，然后呈现其证明，不仅供人类验证，也帮助人们加深理解。

自动化证明器可以从现有的经验证明结果库中选取结果，构建一份简短且人类可读的论证，用于推导最新进展，从而符合该群体就正确性达成共识的既有协议

[70]这很像电子期刊如何借助审稿流程，在新的语境下确立传统纸质期刊的诚信与标准。

。任何人都可以使用此前已经确立的任何结果。尽管每份证明中的每一步都经过计算机验证，并一直追溯至公理，但人们希望，随着越来越多的研究者为该系统作出贡献，互动将逐渐趋向于类似数学家的高层次实践。这样的系统易于使用，并且应当成为数学家工具库中不可或缺的一部分。

通常，定理证明器是交互式的，用户可以向其提供提示。用户根据计算机已经掌握的较简单对象，在证明验证器中输入陈述。证明助手会根据当前已有的知识，判断该陈述是否「显然」为真或为假。如果无法判断，用户就需要输入更多细节。因此，证明助手会迫使用户以严谨的方式解释其论证，并补充比人类数学家可能认为必要的更简单的步骤。[71] 例如，假设一个证明需要十个引理。定理证明器可能能够识别其中一些引理，并迅速加以解决。对于其他一些引理，它可能会陷入停滞，此时用户需要提供更多细节，直到计算机能够找到证明的途径。在这一过程中，程序会不断学习更多知识，并建立和维护一个知识库，因此，当它下次遇到类似问题时，或许能够更加高效地处理。

目前，证明助手还无法阅读“书”，需要人类把所有内容拆解后提供给它们。[72] 证明助手无法判断某个数学陈述是否有趣或重要，只能判断它是否与已证明的内容一致。最终，它应该需要更少的帮助，甚至少得多。我们不知道它何时（以及是否）能够生成自己的证明。

客观证明验证的不确定性原理。数学实践史表明，

对证明提出的质疑越少，它就越容易受到错误的影响。这一重要原则强烈表明，我们必须寻找多种多样的方法来解释和验证任何给定的证明，即使是计算机证明，并尽可能从多个不同的角度对其进行审视。

11. The future of proof

计算机生成的证明包含过多细节（通常超过人类可读程度的100倍），而人类生成的证明又缺少足够细节，无法由计算机验证！因此，我们必须就何为证明达成共识，包括需要多少细节。如果我们接受计算机需要处理人类风格的证明[73]，那么它就需要能够补充细节，以证明缺失的步骤是合理的；关键难点在于，计算机必须独立确定这需要补充哪些内容。一个著名的例子是下面关于$\sqrt{2}$为无理数的证明：

[71]用户反映，他们常常在输入过程中学到很多东西！[72]而这种人机交互界面也可能带来一些问题。事实上，不同用户输入 Lean 的某些定义曾经出现过不一致的情况——我们所知的这些问题都已得到纠正。但有时，文献中也存在一些歧义。例如，Kevin Buzzard 曾向我指出，对于代数几何学家和逻辑学家而言，topos（拓扑斯）可能具有不同的含义，有时两者之间的差异还相当微妙。对于这种二分现象，目前还没有明确的处理方法。[73]毕竟，这不正是我们创造机器的目的所在吗！？

如果 $\sqrt{2}$ 是有理数，那么我们可以写成 $\sqrt{2} = a/b$，其中 a 和 b 是互质整数，使得

$$a^2 = 2b^2.$$

因此， a 是偶数，我们可以写成 $a = 2A$，使得

$$b^2 = 2A^2$$

而且 b 是偶数。但这样一来，2 就同时整除 a 和 b，而它们互质，矛盾。

2006 年，Wiedijk [65] 指出，没有任何计算机验证器能够将这段文本作为输入，并验证其正确性。

我们发现，要帮助证明检查器，就需要回到“什么是证明”这个问题。Avigad [4] 解释说，证明是一种交流，它提供足够的信息，以确立所声称的定理为真。除了正确性之外，还可以根据背景和兴趣来评价证明。2006 年所有主要的计算机验证器都能基于上述论证，构造出证明$\sqrt{2}$ 为无理数的证明，但这些证明会长得令人难以承受，除了验证器本身，没人能够检查它们的证明。此外，由于它们使用的语言差异很大，彼此甚至无法检查对方的证明。因此，正如 Kahle [35] 所指出的，问题不仅在于生成人类可读的输出，还在于弄清楚证明验证器如何理解抽象数学证明的结构，并将其呈现出来，供他人阅读和分析。它还需要解释其中的数学；证明不一定要最短，但必须易于理解、令人信服，并且可能便于读者复现。

稳健性与脆弱性。我们已经讨论过，传统的直觉性数学证明具有稳健性：即使存在细微缺陷，也可以修正，因此即便没有确定性，我们仍能对其抱有信心。另一方面，形式化证明则极其脆弱：一旦发现任何错误，就会引发更多疑虑——形式化证明只要在任何方面有误，其表述的方方面面都会受到质疑，尤其是我们无法读懂细节，也就无法彻底审查其论断。

同一个定理有几个看似不同的证明并不少见，值得探究的是，两个证明在何种情况下才算是“伪装后的同一个证明”。关键思路相同就够了吗？还是说，我们应该期待它们有更多共同之处？目前尚不清楚（参见 [14]）。此外，有时这种伪装相当逼真，让人难以辨认出其中的共同脉络。这不禁让人追问：对某个直观证明进行形式化，所得的会是同一个证明吗？当一个直观证明被拆解成 Lean 所需的各个部分时，它的面貌会截然不同，所依托的知识库看起来也大不相同。若将同一个证明修改为适用于另一种语言，它又会有多大的变化？

Buzzard（在私人通信中）给出了几个很好的例子：

– Leaner A[74] 可能会通过假设三角形位于 $\mathbb{R}^2$ 中来证明勾股定理，将坐标变换为使直角位于原点，然后证明 that

$$\int_{(0,a)}^{(b,0)} 1ds = \sqrt{a^2+b^2}.$$

Lean B 对欧几里得实际证明的形式化采用抽象的欧几里得平面，从不定义面积，而是定义两个图形面积相同的含义。这些不仅仅是不同的证明，而是毕达哥拉斯定理的不同表述，它们源自采用了两种不同（但都有效）的欧几里得平面模型。

– 在 Lean 中，实数被定义为柯西序列的等价类；但另一种证明器可能会用戴德金分割来定义实数。因此，像中值定理这样的定理，在这两种证明器中的证明看起来会大不相同，尽管它们必然等价，因为两者都必须依赖这样一个事实：非空有界实数集有最小上界。

形式化人员希望，形式化或许能帮助我们看清证明中必须用到什么（例如哪些公理）。他们的设想是创造一种“证明压缩器”，让它以某种方式识别如何缩短证明，并将其压缩到尽可能小的篇幅。只有当证明压缩器无能为力时，人类才会参与其中。人们希望这种“尽可能短的证明”会带来一些好处，不过我认为，最短的证明为何应当使用最少的公理，或为何必然更有优势，这一点并不显而易见。

计算机能生成自己的证明吗？“机器学习”通常通过巧妙的算法，以简单的方式逐步形成自己的理解。创建大型数据库，并用专门设计的工具对其进行分析，可能会取得惊人的成效（例如 Google 翻译或 ChatGPT），但这并不等同于培养直觉（甚至有效地模拟直觉）。围绕“机器学习”和其他一些形式的“人工智能”，投入了大量资金，也吸引了大量宣传；不过请放心，许多被大肆吹捧的进展要么言过其实，要么很容易用精心设计的算法和非凡的计算能力来解释。[75]目前还没有“思考机器”。

我们很难知道该如何朝这个方向继续前进，因为我们对创造力和直觉知之甚少，也不清楚我们如何从一种理解转向截然不同的另一种理解。要在机器上模拟这一过程，似乎还遥不可及。[76] Ada Lovelace（1815—1852）从 Babbage 早期的计算机器中提出了（我们如今所说的）计算机程序这一概念，她甚至设想计算机能够被教会作曲。她写道

> [Babbage 的] 分析机并不妄称能够创造任何东西。凡是我们知道如何指令它执行的事情，它都能做到。

[74]Lean 使用者是指在 Lean 中实现证明的人。[75]Google 搜索引擎的影响力，与其背后数学思想有多么简单（尽管可能很优雅）大致并不相关。[76]事实上，机器学习领域最新的激动人心的发展并不预示着任何真正的理解；参见 Melanie Mitchell 的精彩“书” [46] ，其中对支撑这些发展的一些基础进行了细致剖析。

在计算机出现之前，图书馆员常被认为比实际情况懂得多得多（因为他们把守着大量知识）。计算机是容量更大的知识库，更容易访问，也较少受到他人限制，而且能够完成一些令人惊讶的任务；因此，它们被赋予尚不具备的能力，也就不足为奇了。

我们能否设计出一种计算机验证器，使其像人类一样学习和思考? Ganesalingam 和 Gowers [19] 着手设计一种自动定理证明器，其运行方式与将一切内容还原为公理截然不同。相反，他们的理念是像人类一样进行推理，模拟人类的思维方式，生成读起来像人类证明的证明。在他们的设计中，输入和输出必须易于理解，并使用标准词汇（以实现“用户友好”），[77] 提供不仅仅是验证结果的、具有信息价值的解答，同时还应能够轻松地向程序中添加进一步的功能，以纳入新概念、问题求解技术等。简而言之，人们应当能够像与一位（记忆力极佳的）人类互动那样与该程序互动。此前也曾出现过面向人类的程序，但过去二十年中，面向机器的程序取得了更大的进展（有关历史，请参阅 [19] 的第 1.3 节）。

面向机器的搜索程序往往会尝试许多选项，其缺点是会造成组合爆炸；不过，借助一些战术（例如剪枝搜索树），有时可以很好地控制这种情况。人类必须避免严重的组合爆炸，因此会更早地运用战术意识。例如，如果一个待证定理有若干假设和结论，定理证明器不知道哪个最重要（因而应该重点关注），基本上只能随机按顺序尝试不同的陈述来寻找联系；而人类可能会迅速审视重要结论，并开始琢磨如何从已知条件推导出它。计算机可能会在搜索树中反复尝试非常相似的代换，而人类则可能通过一个例子就看出某一类代换（比如 y 是关于 x 的线性式）一般行不通。因此，人类拥有一些不同于大型搜索策略的策略。迄今为止，我们还不知道如何列出所有这些差异。

在 [19]Ganesalingam 和 Gowers 中，他们指出

> 对于数学家找到的大多数证明而言，都有某种关于产生该证明的思想的“故事”可讲。通常，这样的故事会概述主要难点以及如何克服它；这里的“克服”是指将问题化归为一个或多个不再存在该难点的问题。实现这种化归的方式往往是选取一个恰当的中间陈述，结果发现它既可由初始假设推出，又能推出结论。中间陈述本身通常不是通过蛮力搜索找到的，而是经由近似过程得出的：人们可能先提出一个猜想，发现它没有帮助，再弄明白其中的原因.....

[77]机器理解这类输入，远比生成这类输出困难，这有点像我使用不同语言交流时的体验。

> 这并无助益，因此应利用这种理解来指导我们寻找更好的中间陈述。这些具有鲜明人类特征的技巧使数学家能够深入探索“证明空间”，但通过这种方式能够发现的证明仅占整个空间中极小的一部分。人类方法对于那些希望找到这些特殊证明的程序而言显然具有重要价值，而人类数学家似乎又总能以某种难以解释的方式出色地找到这些证明。

人类很擅长选择要尝试的一种或多种技巧，而新技巧可以提高他们证明事物的效率。另一方面，新技巧只会扩大计算机的搜索空间，并且很可能降低其效率。

就充分描述人类证明而言，我们仍处于这样的阶段：不知道该如何描述什么是好的证明，但看到时就能认出来。人类和机器的工作都可以用“战术”来定义。目前，我们最好的做法是使用基于“战术”的交互式系统，这些系统旨在模仿人类推理。然而，要让机器识别何时该使用哪些战术似乎很困难，也就是说，要为尝试不同战术安排顺序，并且或许根据某个战术刚刚失败后获得的直觉，调整后续战术或其顺序。在 [19] 中，他们探讨了是让计算机从过往经验中学习，尝试设计一种更能模仿人类选择的理论，还是将两者结合起来。人类何时会使用某些战术（例如反证法），也很难判定。

最终，我们需要让机器生成的证明参与图灵的“模仿游戏” [62]；也就是说，这些证明应当与杰出的人类证明难以区分。于是，我们让机器与一位菲尔兹奖得主同台，给他们都出一道有难度但能够解答的问题，再看看我们能否判断哪份证明出自谁手。[78] 图灵 [62] 指出，由于计算机是通用的（根据丘奇-图灵论题），它们能够完美模仿任何可计算的事物，包括人类互动，因此这最终一定能够实现（事实上，计算机最终将能够彼此完美模仿）。[79]

在 [19], Ganesalingam 和 Gowers 选定了一些待证明的问题，并让数千名互不相关的读者尝试辨别哪些证明出自他们的程序，哪些出自真人（见 https://gowers.wordpress.com/2013/04/14）。结果令人鼓舞，不过这当然不是图灵测试，因为选定问题的并非独立裁决者。

目前很难预测未来的发展。事态进展迅速，越来越多杰出人才参与其中。Kevin Buzzard（见下一节）认为，

[78]在模仿游戏中， [62] 一名审问者试图分辨：一方是努力证明自己是人类的女性，另一方则是试图蒙骗审问者、让其相信自己是女性的计算机。（在原版、图灵之前的游戏中，计算机一方由一个淘气的男性扮演。）如今，这被称为“图灵测试”。[79]他还做出了一些预测，例如，到2000年，计算机将基本赢得模仿游戏；届时，“人们谈论机器思考时，将不会有人提出异议”。他的预测过于乐观，但另一方面，他确实有不少判断是正确的！

越多人熟悉这款软件，就越早会有有趣的事情发生，这是一个颇具吸引力的观点。

12. The Lean Theorem prover

过去一两年里，Lean 定理证明器[80] 在数学研究界掀起了“一阵风暴”。Lean“是一种函数式编程语言，能够让你轻松编写正确且易于维护的代码。你也可以将 Lean 用作交互式定理证明器。”著名的帝国理工学院算术几何学教授 Kevin Buzzard 让它在纯数学领域广受欢迎（他会在博客 [9]上介绍自己的进展）。他写道

> 我认为，将数学数字化很重要，原因很简单：将任何事物数字化，都能让你用新的方式处理它。[81]

Buzzard 最初的重点是追随公理化的理想，尝试证明纯数学本科课程的全部内容都能在 Lean 中得到论证（有些步骤可通过交互方式完成），这有点像在电子游戏中逐关推进。这样一来，Lean 就能验证许多技术细节，让学生能够专注于更高层次的思考。对于这种新能力将带来什么，Buzzard 不愿作过多预测。

Buzzard 非常清楚，这可能会引发数学证明文化的巨大转变。他让本科生参与使用 Lean 为本科数学经典体系中的各个部分创建证明，但他也承认（对我说），自己不确定这些本科生究竟理解了其中多少证明（尽管“也许他们理解了其他至少同样有价值的东西”）。一个独立项目[82]正在研究这些本科生对数学证明的思考方式有何不同。

研究级数学。Buzzard 写道

> 在不久的将来，我相信计算机或许能够帮助像我这样的数学家（算术几何学家）开展数学研究：补全引理的证明，并为定理提供强大的搜索工具……但在这成为现实之前，仍有大量工作要做。

2021 年 6 月初，这一愿望实现了：Lean 验证了数学界最近的菲尔兹奖得主 Peter Scholze 在与 Clausen 合作撰写的《解析几何》笔记中对某个论证感到不确定的部分 [58] 。在一篇客座博客文章 [57] 中，Scholze 写道

> 我觉得这简直不可思议：交互式证明助手如今已发展到这样的水平，能在相当合理的时间内对艰深的原创研究进行形式化验证。

[80] http://leanprover.github.io/ [81] 并继续写道：“目前我们拥有的计算机证明系统还不足以向数学研究者提供关于 p-adic Langlands 计划 [（Buzzard 是该领域的专家）] 或其他时髦的菲尔兹奖级数学方面的任何新信息，因此顶尖数学研究者往往对它们不感兴趣。然而 [这并不意味着] 它们永远不会对我们有用，而 [这个] 项目的目标就是让这件事更快发生。” [82] https://www.researchgate.net/project/Learning-about-proof-with-Lean

开发该证明的 Lean 形式化版本，需要 Lean 与 Lean 形式化人员（即在 Lean 中将证明形式化的人）之间，以及 Lean 形式化人员与 Scholze 之间的互动。Lean 形式化人员将 Clausen-Scholze手稿逐行输入 Lean，由此生成“当前目标的清晰表述”，随后他们会查阅手稿，弄清接下来几步该如何推进。必要时，他们会联系 Scholze 请求澄清。Scholze 写道：

> 有时后来会发现，即使在论文中，接下来该如何推进似乎也不清楚，于是这个问题就被提请 [我] 注意……而通常很快就得到了解决。

因此，Scholze 的直觉帮助了 Lean（以及 Lean形式化人员）；他们只需往前看一两步，就能理解并形式化克劳森-肖尔策证明。

Lean 确实发现了一些不严谨之处，尤其是某个下确界未必是最小值（此前却假定它是），因此需要对原证明作一些修改。Scholze 指出：

> 这正是我要求进行形式化验证时所担心的那类疏漏。

更有趣的是，Scholze 觉得自己学到了“证明实际上为何成立”。通过 Lean形式化人员与 Lean 的互动，他得以将证明的各个部分阐述得更加明确（因而也更加初等），并计算出某些他一直想更好理解的关键常数。Gonthier [23] 在2005年开发四色定理的形式化证明（使用 Coq）时，也发现了类似的情况：

> 或许这就是形式化证明最有希望的一面：它不仅是一种确保我们在证明中绝不出错的方法，也是一种向我们展示并迫使我们理解证明为何成立的工具。

事实上，在与 Lean形式化人员（尤其是 Commelin）互动的过程中，Scholze 意识到自己可以采用一个较弱定理，从而省去原先使用的稳定同伦理论。

Scholze 警告说，他无法阅读形式化证明；让他能够做到这一点，正是 Lean开发者的目标之一。[83] 其中存在一些显而易见的问题；例如，对于人类而言，根据 A 和 B的定义， $A = B$ 或许是“显然的”，然而 Lean 却需要相当长的时间才能达成一致。[84] 尽管如此，Scholze 觉得这是一

> 这一里程碑式的成就……就是拿一篇研究论文……逐条向证明助手解释其中的引理，直到 [它] 全部完成形式化。

Lean 是否具有认识论优势？有没有理由声称，Lean 比人类更少在不知情的情况下作出未经证实的假设（只要这些假设看起来是一致的）？ 这个过程似乎类似于人类数学家相互验证工作时的情形（提出细枝末节的问题，想要

[83] Lean 对克劳森-肖尔策证明的形式化包含数万行代码，篇幅约为原文的 20 倍，而且不易供人阅读。[84] 这让人想起一个老笑话：有人问一位教授，证明中的某一步是否真的显而易见。教授沉默不语，绞尽脑汁地思索，半小时后才回答：“是的，显而易见。”

进一步了解定义，并查看补充完整的细节，这可能会促使他们修改）。Scholze 向我解释说[85]，

> 正是与 [Lean 聊天] 的互动让我确信，确实正在进行恰当的验证。对我来说， [它是]计算机这一 [事实完全无关紧要，它也可以是 Ofer Gabber。[86]如果我 [只是] 得到了一张笼统的“这一切都正确”的盖章确认，我会极其怀疑，因为我完全确定自己犯了一些疏漏。

Scholze 曾对一个极其复杂且技术性很强的定理感到担忧，也意识到有一些棘手的细节需要理清。他写道

> 对其正确性的信心来自.....看到在这一过程中， [Lean形式化人员] 似乎确实遇到了我所预料的所有细小而棘手的细节（有时也包括一些 [我] 未曾 [预料到的]细节）。

他最后强调说

> 整个实验是一种真正的人类体验，与和一位非常谨慎的同事一起完成这一过程极为相似。

Scholze 对证明本身的描述令人沮丧：这些证明目前由 Lean 以晦涩的方式呈现（因此这种感觉可能就像面对一位不太善于沟通的同事），但没有理由认为 Lean 的这一方面无法改进。就目前而言，我们“对地形毫无感觉”，只知道山峰在哪里，以及我们是否已经抵达山顶，那么究竟能说我们真正学到了什么呢？既然 Lean 现在似乎已经成为一种可行的证明验证助手，那么它的设计或许可以找到一种更直观的方式，来描述它所构造的证明？人类首次给出证明、勘察完地形后，后续的证明往往会对其中涉及的思想作出更丰富的描述，那么 Lean 为什么不能做到这一点呢？这将会很有用，也会推动数学向前发展，然而，

> 目前，我确实看不出 [Lean] 如何能帮助我作为数学家开展创造性工作。Peter Scholze 在 NATURE

13. MYTHS OF OBJECTIVITY

在确认一个证明正确时，我们相信自己能够识别并确立客观真理。但我们真的能做到吗？人们很容易相信自己具有客观性，或相信一台“公正无偏的机器”具有客观性，但这些信念有效吗，还是只是在自我安慰？为了凸显我们可能陷入的谬误，现在我们回想两个著名的所谓客观性的例子：它们在各自的时代和原有范式中（可以说）是客观的，但如今显然不再如此。

13.1. 这些不证自明的真理。在我们社会的很长一段时间里，《圣经》的指导被认为是客观的。图灵 [62] 指出，在伽利略时代，引用的

[85]在博客的评论区。[86]一位以坚持细节正确而闻名的数学家。

> 太阳停住了……约有一整天没有下落——《约书亚记》10:13 他立定大地的基础，使大地永不动摇——《诗篇》104:5

被视为许多人认为，这是对哥白尼理论的客观反驳 .

像 Donna Haraway 这样的社会评论家 [31]探讨了对“客观性”的滥用。[87] 就她关注的主要问题——性别歧视而言，这一点并不难证明：长期以来，社会一直容忍这样一种观念，即一个人的性别或种族会以可量化的方式影响其能力，而社会阶级、成长环境、机遇等社会因素却不会。如今，很少有人会认为，过去那些关于客观性的观念不是出于自利；而如今，大多数人都认为，我们应努力以更客观的认识取代这些观念。然而，Haraway 指出，任何所谓的客观标准实际上都是一种主观的“权力操作，而非迈向……真理的举动”。的确，声称某事“已被客观证明”（也就是说，在某个当前范式中是客观的），似乎会压制任何可能的异议。

学术研究的拙劣之处仍在助长这种带有偏见的“客观性”。例如，2005年，时任哈佛大学校长的 Larry Summers 试图解释女性从事 STEM 工作的人数为何偏少：当时，针对十二年级学生认知能力的一些测试显示，女性的分数差异小于男性，这意味着女性处于最高分段的人数较少（处于最低分段的人数也较少）。在找出可能导致这些统计差异的因素时，他着重谈到“内在天赋方面的问题”（以及“涉及社会化和持续歧视的次要因素”）。他为选择“内在天赋”作为最有说服力的解释所给出的理由是，他“想要挑起争议”。[88] 如今，甚至在当时，人们都可能会认为这种挑衅“有问题”。

为了不止于指出偏见，也为了不让问题变成坏人的“偏见”与好人的“客观性”之争，女权主义者 [31] 主张，偏见根植于任何社会建构之中，并指出存在一种“集体历史主体性”。[89]

Haraway [31] 强调，在数学学生的培训过程中，会向他们讲述“关于客观性的寓言”，尽管数学家在实践中并非如此。她认为，不存在真正“值得信赖的权威”，取而代之的应是一个遍布全球、尊重不同视角的联结网络。事实上，她声称，

[87]关于数学中的客观性与社会本体论之间关系的近期深入讨论，见专题期刊 [10]。[88]遗憾的是，代表性团体以性别歧视为由谴责他，而不是以学术研究不足为由。这导致他们的抗议遭到忽视；奥巴马总统任命萨默斯为国家经济委员会主席，进一步为他的拙劣学术研究增添了几分体面。[89] Haraway 还指出，人们渴望找到一种正当的女权主义“客观性”，但这显然自相矛盾，因为人们可能会重蹈自己煞费苦心试图指出的他人所犯的错误。

> 科学一直致力于探索翻译、可转换性、意义的流动性和通用性，[90]

但当作品仅仅与霸权进行比较时，便无法满足这些标准。在本文的语境中，Haraway 的工作表明，过去曾很好地服务于我们的共同体标准，应当为计算机证明这一快速发展的时代照亮前行之路。这些标准能够验证证明的计划和具体细节看起来是否正确，是我们在任何给定范式内所能达到的最接近客观性的程度。

13.2. 客观性与无谬性。是否存在无谬的证明？是否存在“无条件且永恒”正确的证明？Leibniz 和 Newton 的无穷小在 17世纪大多为人接受，但后来被 Cauchy 到 Weierstrass 推导出的 $\epsilon - \delta$ 证明所取代。我们目前对微积分基础的证明能经受永恒的考验吗？它们是无谬的吗？它们能适应未来提出的所有问题吗？我们怎么知道？我们有一个框架，在其中这些证明似乎无可辩驳，但从未来的认知来看，这个框架还会显得合适吗？

> 一个三角形。这看起来极其简单，你可能会认为我们……已经对它了如指掌了……然而，即使我们证明了它具有我们所能设想的一切属性，另一位数学家或许会在1000年后的未来发现它还具有其他性质。因此，我们永远无法确切地知道自己是否已经完全掌握了关于三角形的一切知识。这一点同样适用于物体、它们的广延性，以及所有事物！——Descartes（1648年4月16日）与 Burman 的谈话。

那么，在当前框架中，客观真理又如何呢？从 G¨odel 我们知道，无法接近验证所有真理，因为真命题的集合远大于可证明命题的集合，因此我们需要把注意力限制在可证明的内容上。[91] 即便如此，我们如何知道某个给定的证明是否正确？Tarski [60] 指出：“直觉证据远非无谬性可言，不具备客观性质，而且常常会导致严重错误”，他认为“公理化方法的后续发展……在于……限制[对直觉证据的诉诸] ”。这一论证表明，我们只能寄望于借助形式化证明，对客观真理进行可靠验证；但是，证明验证器能否变得无谬，从而成为判定何为正确的最终裁决者？毫无疑问，编写良好的程序或许能够发现既有证明中的新问题，并构造出更难以质疑的证明，但这与无谬性并不是一回事。

数学哲学家 Avigad [5]声称，

[90]本文涉及的所有问题，实际上也正是我们所说的“共同体标准”的一部分。[91]我们还需要将注意力进一步限制在复杂度类 NP 中的（小得多的）命题集合上。粗略地说，这些命题是指那些存在证明或解答，能够在人的时间尺度内得到验证的命题；这是理论计算机科学中的一个关键问题。第 2 节末尾对此作了简要讨论。

> 根据标准的观点，一个数学陈述是定理，当且仅当存在该陈述的形式推导，或者更准确地说，存在其适当的形式化表达。[92]

这是一个颇具吸引力的定理定义，但它建立在“该陈述的形式推导”无谬性的假设之上。[93] 获得直观证明和形式化证明（基于直观证明）是很有吸引力的目标，也是确保不存在明显错误的不同方式，但要相信这样就能断言客观真理，未免牵强。Avigad 的主张建立在一种客观真理观之上，这种观点并未反映任何共识（尽管 Avigad 的措辞十分笃定），而是为某个子社群所关注的方向提供依据。事实上，他的同事 Azzouni 的极端主张也支持这一观点 [7],

> 形式化证明已成为数学实践的常规。

不太明白 Azzouni 怎么会把这说成是他的“常规”，因为这与大多数数学研究群体的实际情况相去甚远。形式化证明或许有一天会发挥不止边缘性的作用，但那一天尚未到来。

Detlefsen [13]更能理解数学实践：

> 数学证明通常不会以一种能使其形式化显而易见或顺理成章的方式呈现。尽管如此，它们通常会以一种足以彰显其严谨性的方式呈现……至少在它们广泛流传于同行之间时是如此……

也就是说，一份被学界认可为严谨的证明，其作者很少会关心形式化，尽管形式化者必然会关注严谨性（但Azzouni竟然认为形式化会显著提高证明的可信度，这就有些奇怪了）。Avigad [5] 也主张采用形式化证明，因为“提供更少的信息只会使问题恶化：如果连完整呈现的形式推导都无法可靠地核查，那么提供严格更少的信息，几乎不可能让人更有信心。”那么，解释到什么程度才算足够？就我个人而言，我更喜欢清晰的一页证明，而不是冗长晦涩的论述；后者即便补上几个细节、填补“文献中急需填补的空白”，也未必能领会什么才是重要的！

回到第 11 节“稳健性与脆弱性”这一小节的主题，有意思的是，形式化人员觉得自己的证明更值得信赖，恰恰是因为这些证明显然脆弱得多。此外，如果正如 Avigad 似乎所主张的那样，一项证明只有在能够被形式化之后才算是证明，那么我们就必须追问：哪一种形式化才是正确的？不同的 Lean形式化人员可能会为勾股定理的证明给出截然不同的形式化，那么哪一种才应该被接受？证明是否存在“等价类”（也就是说，实质相同、只是表面形式不同的证明属于同一个等价类）？如果存在，那么判断某个证明属于哪个等价类的标准是什么？如果不止一个等价类，这又说明了证明的

[92]因此，Clausen-Scholze 的结果似乎只有在 Lean 断言其证明成立时，才成为 Avigad 会认可的定理。[93]Avigad 接着声称：“当一位数学审稿人认证某项数学结果时，无论审稿人是否认可，这一判断的正确性都取决于是否存在这样的形式推导。”这一说法展现了对数学过程非同寻常的全局把握。

根本性质是什么？机器是否应该致力于编撰一本“书”，只收录最好的证明？

13.3. 机械降神字面意思是“来自机器的神”。它描述了形式化证明共同体中一些人的观点。相信自身代码具有无谬性，会让一些程序员意识不到自己的程序偶尔也可能出错！只有通过交流自己的想法，这些想法才会得到认可。事实上，在 [8] Barendregt 和 Wiedijk 自信地断言，只要“始终遵循少数几条逻辑规则”，一个所谓的证明就算得到了验证；因此，一旦他们认为自己做到了这一点，就会对形式化内容作出乏味的描述（见 [8]图 8）。他们需要认识到，证明是一种社会契约，因此应当多花些心力，让感兴趣的读者了解相关内容。

13.4. 克莱千禧年奖。2000年，克莱基金会宣布设立一项100万美元的奖金，用于解决七个著名的数学问题中的任何一个。一个解答只有在同行评审期刊上发表两年后，并且“已在全球数学共同体中获得普遍认可”，才能获得奖金。这些规则几乎毫不含糊地表明，制定者只信赖基于共同体的证明验证，即便如此，他们仍觉得需要一段时间才能确定。

14. Will machines change accepted proof?

在本文中，我主张，证明验证对作为一种社会建构的证明之核心原则几乎没有影响。此外，得知某个程序声称一项证明已通过验证，并没有多少额外价值（除非它还能提供更多有助于加深读者理解的信息）。尽管如此，我们可以期待人们会努力让这些形式化证明更易于理解，并希望它们也更有用（例如，它们最终或许能够从直观证明中去除无关的想法）。事实上，Patrick Massot [44] 最近宣布，正在开发软件工具，以自动将形式化证明转换为人类可读的交互式证明，使读者能够逐步深入探究，直到她遇到一个无需进一步解释、自己便相信的论断，无论它出现在形式化证明还是直观证明中。

自国际象棋程序深蓝于1997年击败国际象棋世界冠军Garry Kasparov以来，机器学习程序在棋盘游戏策略方面的表现日益出色，如今已能轻松击败国际象棋和围棋领域的世界顶尖选手。事实上，2017年末，DeepMind的AlphaZero启动运行，仅用24小时与自身进行对弈，便在国际象棋和围棋（以及将棋）中达到了特级大师水平，堪称真正的“超人类”水平。它的搜索树相对较浅（在国际象棋中每秒搜索80,000个局面，而某些其他顶级软件每秒可搜索7,000万个局面），但它凭借更出色的战术弥补了这一不足。[94] 这种“强化学习”能否带来更深刻、更多样化的数学证明？它是否也能像其国际象棋对弈那样，有时创造出比人类证明更加令人惊讶的证明？

[94]国际象棋特级大师觉得它的棋风“异乎寻常”，展现出“疯狂的进攻型国际象棋”，例如牺牲后和象来利用位置优势，这种做法人类不太可能采用。

2022年10月，DeepMind 推出了[95] AlphaTensor，其任务是以尽可能高效的方式将两个给定维度的矩阵相乘，也就是说，使用尽可能少的乘法。AlphaTensor 在大约 20 种情况下改进了已有成果 [18]，但改进幅度从未超过 5%；尤其值得注意的是，它似乎并未推导出任何新的一般定理，也未指出如何找到新类型的一般定理。[96] 这表明，在缺乏有力理论思想的领域，这项新技术可以超越人类直接通过计算获得的观察结果，但目前没有迹象表明 DeepMind 的算法会促成新的深刻定理的诞生。也许有一天，它能更好地识别其战术如此成功的原因，从而帮助我们更好地理解其中的运作机制。

迄今还没有人造出一台量子计算机，能在特意挑选的问题上比经典计算机计算得更快。尽管如此，理论表明，某些可并行化问题或许能在量子计算机上快得多地得到解决。最著名的例子是Shor量子因数分解算法，它促使大量资源投入“后量子”密码学。[97] 对我们而言，问题在于量子计算能否用于寻找证明的任务（例如，使用一个大得离谱的搜索树）。Carlos Simpson向我指出，我们很可能会遇到这样的问题：虽然得到了一个证明，却完全不知道它是如何找到的；而且这个过程或许漫长到无法在任何经典计算机上验证。

原始计算能力以许多重要的方式拓展了我们的数学能力。借用 Kevin Buzzard 的话来说，人们对这些可能性投入得越多，有趣的事情就越早发生。Venkatesh [63]指出，数学可能会“发生重大改变；其核心问题和价值观……将与我们习以为常的截然不同”。这“不仅会增强我们从事数学研究的能力，也会改变我们对数学本质的理解”。事实上，我们终将学会如何发掘计算能力、软件和卓越编程思路所揭示的数学及证明领域的巨大潜力，这只是时间问题。

References

[1] M. Aschbacher，《有限单群分类的现状》，美国数学会通报 51（2004年），736–740[2] M. Aschbacher，《高度复杂的证明及其影响》，英国皇家学会哲学汇刊A辑 363（2004年），2401–2406[3] J. Auslander，《证明在数学中的作用》，载于 B. Gold 和 R. A. Simons（编），《证明与其他难题：数学与哲学》（第 61–77 页）。华盛顿特区（2008年），美国数学协会。[4] J. Avigad，《数学方法与证明》，Synthese 153（2006年），105–149。[5] J. Avigad，《数学推理的可靠性》，Synthese 198（2020年），1–23。[6] J. Avigad，《数学理解的多样形式》，美国数学会通报 59（2022年），99–117。

[95] https://www.deepmind.com/blog/discovering-novel-algorithms-with-alphatensor [96] 在纯数学中，寻找大量例子往往是深入理解并进而发现真正新定理的前奏。[97] 若干重要密码协议的安全性建立在因式分解的困难之上，因此一旦因式分解变得容易，这些协议将不再安全。这让有权势的人感到担忧，因此大量资金正投入这一领域的研究。

[7] J. Azzouni, 《为何非形式证明遵循形式规范？》, 《科学基础》, 14（2009）, 9–26。[8] Henk Barendregt 和 Freek Wiedijk, 《计算机数学的挑战》, 《皇家学会哲学学报A辑》, 363（2005）, 2351–2375。[9] Kevin Buzzard, 《Xena 项目》, 发表于 https://xenaproject.wordpress.com/ 的博客[10] Paola Cant`u 和 Italo Testa（编）, 《数学实践与社会本体论》, 《Topoi》42（2023）, 187–344。[11] Eugenia Cheng, 《道德意义上的数学》, 2004年（预印本）。[12]M. Davidson、J. Dethridge、H. Kociemba 和 T. Rokicki, 《上帝之数是20》, https://www.cube20.org/。[13] M. Detlefsen, 《证明：其本质与意义》, 载于《证明及其他困境：数学与哲学》（B. Gold 和 R. A. Simons 编）, 华盛顿特区：美国数学协会（2008）, 3–32。[14] Silvia De Toffoli, 《调和严谨与直觉》, 《Erkenntnis》（2020）。[15] Silvia De Toffoli, 《数学的可错论解释的基础》, 《哲学季刊》71（2021）, 823–844。[16] Bjorn Ian Dundas 和 Christian F. Skau, 《2021年阿贝尔奖访谈：L ´aszl ´o Lov ´asz 和 Avi Wigderson》, 《美国数学会通报》69（2022）, 828–843[17] M. Eckert, 《无忧宫的水景问题——欧拉在理想流理论兴起前夕对实用流体力学的参与》, 《物理学D：非线性现象》237（2008）, 14–17。[18] Alhussein Fawzi、Matej Balog、Aja Huang、Thomas Hubert、Bernardino Ro mera-Paredes、Mo- hammadamin Barekatain、Alexander Novikov、Francisco J. R. Ruiz、Julian Schrittwieser、Grze- gorz Swirszcz、David Silver、Demis Hassabis 和 Pushmeet Kohli, 《利用强化学习发现更快的矩阵乘法算法》, 《自然》610（2022）, 47–63。[19] M. Ganesalingam 和 W.T. Gowers, 《具有类人输出的全自动问题求解器》, 《自动推理杂志》58（2017）, 253–291。[20] Kurt G¨odel, 《从哲学角度看数学基础的现代发展》, 提交给美国哲学学会的1961年演讲稿, 《文集》第III卷, 牛津（1995）, 375–388。[21] Dan Goldston 和 Cem Yildirim, 《连续素数之间的小间隔》（撤回的预印本）[22] Georges Gonthier, 《四色定理的计算机验证的证明》（2005）, 预印本。[23] Georges Gonthier, 《形式化证明——四色定理》, 《美国数学会通报》55（2008）, 1382–1393。[24] W.T. Gowers, 《数学的两种文化》, 载于《数学：前沿与展望？》, 美国数学会, 2000年。[25] Daniel R. Grayson, 《面向数学家的单值基础导论》, 《美国数学会通报》55（2018）, 427–450。[26] Ian Hacking, 《为何会有数学哲学？》, 剑桥大学出版社, 2005年。[27] T.C. Hales, 《开普勒猜想的证明》, 《数学年刊》162（2005）, 1065–1185。[28] T. Hales, 《形式化证明》, 《美国数学会通报》55（2008）, 1370–1381。[29] T.C. Hales 及其他21人, 《开普勒猜想的形式化证明》, 《数学论坛：Pi》5, e2（2017）, 29页。[30] Michael Hallett, 《数学基础》, 《剑桥哲学史：1870–1945》, Thomas Baldwin 编（第一部分第10章：1870–1914）, 剑桥大学出版社（2003）, 128–156, 833–837。[31] Donna Haraway, 《情境知识：女性主义中的科学问题与局部视角的特权》, 《女性主义研究》14（1988）, 575–599。[32]Michael Harris 的博客：《不道歉的数学》, 2018年6月2日, https://mathematicswithoutapologies.wordpress.com[33] John Harrison, 《形式化证明——理论与实践》, 《美国数学会通报》55（2008）, 1395–1406。

[34] John Hosack, 《包容性的数学哲学》, 《美国数学会通报》66（2019年）, 1433–1437。[35] Reinhard Kahle, 《迈向数学证明的结构》, CEUR研讨会论文集第1186卷（2014年）, 论文#22。[36] Reinhard Kahle, 《什么是证明？》, Axiomathes 25（2015年）, 79–91。[37] Donald Knuth, 《所有问题均已解答》, 《美国数学会通报》49（2002年）, 318–324。[38] Izabella Laba, 《摄影》, 载于《数学家生活中的艺术》（编者：Anna Szemer´edi）, 美国数学会, 2015年。[39] J. Lagarias, 《开普勒猜想及其证明》, （Springer, 2009年）, 3–26。[40] M. Lange, 《因为没有原因：科学与数学中的非因果解释》, 2017年, 牛津：牛津大学出版社。[41] X. Leroy, 《编译器后端的形式化认证，或：借助证明助手编写编译器》, 第33届ACM编程语言原理研讨会, ACM出版社（2006年）, 42–54。[42] Paolo Mancosu, 《数学解释：问题与前景》, Topoi 20（2001年）, 97–117。[43] Paolo Mancosu, 《数学中的解释》, 《斯坦福哲学百科全书》（2018年）, Edward N. Zalta（编）, https://plato.stanford.edu/archives/sum2018/entries/mathematics- explanation/[44] Patrick Massot, 《面向数学家和数学学生的形式化数学》课程, youtube.com（ https://www.youtube.com/watch?app=desktop&v=tp h3vzkObo）。[45] James Maynard, 《素数间的小间隙》, Ann. of Math. 181（2015年）, 383–413。[46]Melanie Mitchell, 《人工智能：人类思维指南》, Farrar, Straus and Giroux, 2019年。[47] L. J. Mordell, 《关于三次和四次不定方程的有理数解》, Proc. Cab. Phil. Soc 21（1922年）, 179–192。[48] L. J. Mordell, 《一位八旬数学家的回忆》, 《美国数学月刊》78（1971年）, 952–961[49] Melvyn B. Nathanson, 《苦苦追寻数学真理》, 《美国数学会通报》（2008年）, 773。[50] Melvyn B. Nathanson, 《苦苦追寻数学证明》, Math. Intelligencer 31（2009年）, 8–10。[51] J. Paris和L. Harrington, 《皮亚诺算术中的数学不完备性》, 载于Barwise, J.（编）, 《数学逻辑手册》。荷兰阿姆斯特丹：北荷兰出版社, 1977年。[52] Colin J. Rittberg、Fenner S. Tanswell和Jean-Paul Van Bendegem, 《数学中的认识论不公》, Synthese 197（2020年）, 3875–3904[53] Siobhan Roberts, 《数学中，错误已非昔日的错误》, Nautilus在线版, 2015年。[54] Dan Rockmore, 《证明它！》, 《纽约书评》, 2022年1月13日。[55] Neil Robertson、Daniel P. Sanders、Paul Seymour和Robin Thomas, 《四色定理的新证明》, Electron. Res. Announc. Amer. Math. Soc. 2（1996年）, 17–25。[56] Dirk Schlimm, 《皮亚诺论符号化、记法设计原则与点记法》, Philosophia Scientae, 25（2021年）, 95–126。[57]Peter Scholze, 《液体张量实验半年：惊人的进展》, https://xenaproject.wordpreWhy abc is still a conjecture.com/2021/06/05[58] Peter Scholze（与Dustin Clausen合著）, 《解析几何讲义》, https://www.math.uni-bonn.de/people/scholze/Analytic.pdf[59] Peter Scholze和Jakob Stix, 《为什么 abc 仍然是一个猜想》（预印本, 2018年）。[60] Alfred Tarski, 《真理与证明》, 《科学美国人》220（1969年）, 63–77。[61] William P. Thurston, 《数学中的证明与进展》, Bull. Amer. Math. Soc. (N.S.) 30（1994年）, 第2期, 161–177[62] A. M. Turing, 《计算机器与智能》, Mind, 新系列, 59（1950年）, 433–460。[63] Akshay Venkatesh, 《我们如何赋予数学以价值》, 预印本。[64] Vladimir Voevodsk˘ı, 《如果当前的数学基础并不一致，会怎样？》, 讲座, https://www.ias.edu/ideas/2012/voevodsky-foundations-of-mathematics, 2010年。

[65] Freek Wiedijk, 《形式化证明——入门》, 《美国数学会通报》55（2008年）, 1408–1414。

Département de Mathématiques et Statistique, Université de Montréal, CP 6128 succ Centre-Ville, Montréal, QC H3C 3J7, Canada.电子邮箱地址:
andrew.granville@umontreal.ca