

观点与分析

编辑日期: 2026年10月29日

停在办公室里的速度

> 企业可以买到更强的写作与分析能力，随后却发现，真正需要重新设计的是工作在人与人之间的交接方式。如果把结果追踪到员工屏幕之外，实验究竟告诉了我们什么？

2023 年 9 月至 2024 年 10 月，66 家企业参加了一项关于微软 365 Copilot（Microsoft 365 Copilot）的实验。工具使用权被随机分配给员工，样本包括 7,137 名知识工作者。在计划为期六个月的实验后半段，获得使用权的员工每周处理电子邮件的时间减少约 1.4 小时；对实际响应使用权安排者的局部使用效应估计约为两小时。1

但研究者明确指出，他们“不会观察任何工作内容，也不会观察任何生产率指标或绩效评价”。因此，这项实验不能证明企业完成了更多工作，也不能证明客户等待时间有所缩短。1

这个限定正是问题的入口。我们有证据表明，个人的一部分工作发生了变化；但组织最终产出了什么，是另一个问题。两者之间的距离，不能在财务演示中被当作无关细节删去。

客户购买的不是员工的速度，而是自己所求之事得到完成。

一封邮件也许一分钟就能写完，却可能因为一项无人负责提供的信息而等待数日。分析也许更容易制作，但其验收条件仍然含混。此时速度确实提高了，可让速度转化为服务的道路仍在施工。

客户真正购买的是什么？

Shakked Noy 与 Whitney Zhang 于 2023 年在《科学》（Science）发表了一项对照实验，453 名专业人士完成与本职工作相关的写作任务。使用 ChatGPT 后，任务用时平均减少 40%，输出质量提高 18%。这是一项关于特定任务的重要结果，却不是参与者所在机构利润增长的估计。2

混淆往往始于推论单位扩张得比测量单位更快。我们测试的是一份备忘录的写作，谈论的却是整个部门的生产率；测量了部门，又把一个百分比贴到整家公司身上。最后，一个诞生于短时任务的数字，凭借设计精美的幻灯片，获得了管理年度预算的资格。

为说明差别，不妨设想一家工程服务公司正在准备一份技术与商务方案。员工概括客户需求，工程师估算工作范围，同事确认团队是否可用，之后方案才获批准。助手能够改善草稿，但在这些部分达成一致前，方案仍不能发出。这只是一个用于说明任务相互依赖的假想情境，并非某家具体公司的真实事件。

证据中也有不应被隐藏的成功案例。Erik Brynjolfsson、Danielle Li 与 Lindsey Raymond 于 2025 年发表的研究估计，在 5,172 名客户支持员工的样本中，每小时解决的问题数量约增加 15%。分析利用了助手分阶段导入的过程；它并不是一项把工具全面随机分配的实验。3

这项工作的设计之所以重要，是因为员工很大程度上可以独立处理每个案例。个人任务与已经完成的服务单元相当接近。由此可以作出一种解释：当员工能在自己的工作范围内完成客户所需之事，收益可能迅速穿过组织；当他交付的只是一个必须等待其他部分的组件，各部分之间的关系形式便成为一种经济因素，其重要性不亚于单个部分的效率。

知识不会自行运转

继续看这家假想公司。工程师对工作范围作出了更准确的描述，但销售部门依据的是另一版客户需求。财务估算提前完成，执行排期却只在每周会议中更新。每份文件都变得更好，文件之间却未必更加一致。

这里的问题在于，一个要素的改进往往需要另一个要素配合，才能产生价值。经济学把这种关系称为组织互补性（Organizational Complementarities）。Paul Milgrom 与 John Roberts 在 1990 年关于现代制造业的论文中讨论了技术、生产与组织选择如何相互支撑：周边选择发生变化时，某项选择本身的价值也会改变。4

这个概念并不复杂。如果客户需求一改变，方案就能立刻修改，那么所有人访问同一份需求版本的价值就会上升；如果各部门都使用这份共享版本，快速修改又会更有用。只购买其中一项而放任另一项不变，回报可能低于两者结合。

Timothy Bresnahan、Brynjolfsson 与 Lorin Hitt 对企业数据的研究，也发现了与信息技术、工作组织和员工技能互补相一致的证据。论文发表于 2002 年；它属于观察性证据，而非能够把全部改善单独归因于组织重构的实验。5

这些文献对今天讨论的贡献，是移动了问题的位置。我们不能只问模型能否完成任务，还要问：为了让模型的表现产生价值，流程中的其他部分必须满足哪些条件？

答案有时很朴素：统一定义何谓“完整的工作范围”，让两个部门同意输入格式，或允许信息变化时立即更新，而不必等待新一轮邮件往返。并非每个问题都需要全面重组。有时企业缺少的项目，只是一页纸上的明确约定；也正因如此，它很难出现在盛大的发布会上。

这个思路也有边界。如果流程本来就高度整合，而获得增强的任务恰好是真正的瓶颈，那么工具无需大范围变革也可能带来清晰回报。互补性不是在每一份软件许可之外再附加一个咨询项目的借口；它要求我们说清楚，什么依赖于什么。

两个步骤之间，可能比两个步骤本身更有价值

一项任务可能已经在员工账户里结束，结果却仍不能被同事使用。完成的文件必须包含充分输入，并符合双方认可的验收标准，下一步才能开始。这是一种仅靠写作速度测试无法揭示的运营成本。

在 2026 年工作论文《串联任务，重新定义工作》（Chaining Tasks, Redefining Work）中，Mert Demirer 等人把生产视为具有列昂惕夫结构（Leontief）的相互关联步骤：结果需要互补组件，某个组件再多，也不能弥补另一个组件的缺失。在人类劳动、辅助与自动化之间如何选择，取决于验证、协调和步骤排序的成本，而不只取决于可以自动化多少个步骤。6

回到工程方案的例子。设想工具提取客户需求，把它们连接到工作范围条款，并生成假设与待决问题清单，工程师因而能够审查一个相互关联的材料包。再对比另一种安排：各部门分别制作摘要、方案和财务估算，格式各异，也没有共同的参考版本。单独评价每份输出时，质量可能接近；验证最终结果的成本却会大不相同。差别不只来自部门数量，更来自输出是否一致，以及能否追溯到原始信息。

步骤相连不等于无人监督。可以在一个材料包的末尾设置检查点，在错误后果严重的地方保留内部核验。检查一份很长的结果，成本也可能高于审查几份较短结果。因此，“端到端”这个说法本身并不构成经济论证。

这篇论文采用理论模型，并呈现与预测相符的经验模式；它没有证明串联步骤必然提高任何一家公司的利润。我们的应用结论更窄：评价工具时，必须把其输出进入下一项工作的成本算进去。

由此可以提出一项可检验的工程要求。输出应携带接收者所需的数据、数据来源、尚未解决的问题，以及明确的完成标准。如果下一个部门仍需致电文件负责人，请他重新解释全过程，那么组织并未实现知识交接的自动化，只是改善了那封宣告“我们还需要交流”的邮件。

这类工作一部分需要修改软件，另一部分需要改变流程负责人之间的约定。应用程序接口（API）可以快速传输数据；但各方是否同意这些数据足以启动下一步，则需要重新设计工作本身。

工具能自行解决协调问题吗？

对上述分析最有力的反驳是，人工智能或许能从源头减少部分协调需求。如果工程师能够理解市场要求，或商务人员能够探索技术限制，专业分工之间的关系就可能在任务内部发生变化。

宝洁公司（Procter & Gamble）的一项实验为这种可能性提供了支持。经同行评审的研究于 2026 年发表于《组织科学》（Organization Science）。共有 791 名来自技术与商务岗位的参与者完成研讨任务，他们被随机分配为个人或两人一组，并分别在使用或不使用人工智能的条件下工作。7

获得工具支持的个人，其提案质量接近未获工具支持的团队；有工具支持的提案在技术与商务考量之间也更为均衡。但实验关注的是短时虚拟研讨中产品开发的早期工作，并未证明产品上市周期缩短或销售额提高。7

这项结果应被认真对待。职能边界并非全是固定的墙。工具可能帮助员工在提交建议前，看见建议会如何影响同事的工作，从而减少一部分原本需要在后续会议中解决的误解。这是从结果推导出的潜在机制，并非对企业所有会议的直接测量。

然而，理解限制不等于自动获得履行限制的权力。工程师可以理解定价，却不意味着建议价格已经获批；商务人员可以预见制造问题，也不意味着工厂产能已经可用。知识整合与完成相互承诺，仍是两件不同的事。

因此，更准确的论点必须带有条件：工具可以改善协作；如果其输出在流程其他部分仍不被接受，组织也可以阻止这种改善进入执行。需要测试的是障碍究竟位于何处，而不是预先认定它永远属于技术，或永远属于管理。

当企业改变寻找收益的位置

另一篇 2026 年工作论文提供了补充视角。Hyunjin Kim、Dahyeon Kim 与 Rembrand Koning 研究了 515 家初创企业。按研究者及研究机构的说明，参与者首先获得技术资源和培训，随后随机一组收到围绕人工智能重组生产的案例，另一组则收到一般创业内容。8

研究报告称，前一组发现了更多使用场景，部分绩效指标也有所提高。但目前可核验的材料是研究者摘要与机构说明，完整手稿中的表格未得到审阅。因此，本文不采用外界传播的收入倍数，也不把初创企业的结果当作适用于大型机构的稳定定律。8

值得思考的是干预本身：教组织去哪里寻找价值。不同组织可能拥有相同工具，却把想象力限制在屏幕前可见的工作上。它改善方案写作，却不问同一事项为什么需要三轮往返才能确定；它在每个职能内部寻找用法，却错过了重新组合职能之间工作的可能。

在假想的工程公司中，更大的改进也许是用一种能够同时估算范围、资源与成本的方式接收需求。此时，员工最初收到的材料就变了。我们不再要求他加速处理残缺文件，而是让文件在流向下一部门前具备完整的可能。

这是综合证据后提出的一项战略判断：人工智能被放在流程的什么位置，可能比给所有席位平均配置工具更重要。许可证分配得平等，并不等于创造价值的机会相等。

审查的人与等待的客户

这个问题还有一面不会出现在文本质量测试中。在假想案例里，方案制作者能生成许多备选版本，同事却仍需逐一审查。如果交接方式不变，瓶颈可能只是移到了审查者身上。此时仅测量作者的工作，不能证明组织整体改善。

也不应因为审查者要求充分证据，就指责他抗拒变化。新的要求看似很清楚：审查更多。但更重要的运营问题是：待审内容变得更容易核验了，还是仅仅变多了？

另一端的小客户，可能围绕方案截止时间安排一连串事务：选择供应商、筹措资金、确定开工日期。这些只是示例中的经济可能性，并非在某个样本中量得的损失。它们的作用，是提醒我们组织的时间会进入组织外部人员的决策。

在沙特的一种可能应用中，企业或服务机构可以选择一种重复发生的事务，从收到第一份完整申请开始，一直追踪到受益人拿到结果为止。这是一项本地测试建议，并不是对现有项目的描述。终点必须定义清楚：一份可执行的方案、一项完成的服务，或一个足以让接收者据此行动的可靠答复。

之后才出现比许可证数量更精确的财务问题：在资源不变的情况下，被接受的输出是否增加？质量是否提高？总耗时或返工是否减少？即使数量不变，质量提高也可能值得投资；如果数量增长引发日后昂贵的修正，它也可能毫无价值。

因此，我们不需要用“完成率”取代分钟数，成为新的崇拜对象。我们需要明确结果、质量标准、总成本，以及一个能够从自身位置解释改善的受益者。

在推广一个版本前，先测试四个流程版本

这场讨论可以转化为一项可证伪的组织实验。与其比较热衷工具的员工和其他员工，不如依据析因设计（Factorial Design），把相近的工作单元随机分配到四组，从而分离技术效应与某项特定流程调整的效应。

下表是一项拟议设计，并非已经开展的实验结果：

组别	可使用人工智能工具	调整工作方式
00	否	否
10	是	否
01	否	是
11	是	是

流程调整应事先定义。在工程方案的例子中，它可以包括：使用共享需求文件、在交接时采用统一验收标准，并在商务版本获批前更新资源可用性。培训投入与管理关注也需受到控制，以免它们的全部效应被误归到工具名下。

可以选择每期达到质量标准的完整方案数量作为结果指标，前提是需求量充足；如果需求有限，总耗时或质量也许比数量更合适。还应记录审查与返工成本，避免测量窗口在修正到来前就结束。

定义 Y 时，让数值越高表示结果越好；如果指标是完成时间，就使用时间的相反数，或反转对效应符号的解释。依据四组均值，可以估计交互作用：

$$\Delta = Y_{11} - Y_{10} - Y_{01} + Y_{00}$$

如果 Δ 的可靠估计为正，说明在所选指标上，工具与特定流程调整结合产生的增量，超过两者单独增量之和。这是在特定环境中检验互补性，不是适用于所有组织的定律。

实验也可能得出令人不舒服的结果：仅调整流程就已足够，或工具单独有效，或组合方案成本过高而没有合理增益。项目开始前就必须允许这些结果出现；否则，评价只会变成一场延迟举行的批准仪式。

比较还需谨慎选择随机分配单位。如果不同组共用同一批审查者或交换文件，一组的干预可能影响另一组的结果。因果推断文献把这种现象称为干扰（Interference），并区分直接效应与间接效应。因此，应选择效应不易相互传递的团队或路径，或明确建立这种传递的模型。10

即使事务数量很多，如果独立单元数量很少，证据仍然不足。实验应估计统计功效，考虑案例难度与需求变化，并设定足以观察后续结果的时长。无法随机化时，受控的时间比较仍有价值，但其因果强度较弱，不应被包装成等价实验。

试点结束后，应该留下什么？

建设这种能力可能需要时间。Brynjolfsson、Daniel Rock 与 Chad Syverson 在 2021 年发表的《生产率 J 曲线》（The Productivity J-Curve）中讨论了通用技术为何需要无形的互补投资，以及这些投资的影响为何可能延迟出现在测得的生产率中。9

但这条曲线并未向每个项目承诺“成功只是尚未到来”。我们必须说清楚正在建设的资产是什么：有记录的任务依赖知识、可重复的核验方法、输入与输出之间的一致规则，或重新分配工作的能力。如果没有任何可检查的积累，仅仅等待并不构成组织投资。

决策预算也应随之改变。除了使用费用，还要为数据准备、交接设计、接收者培训与结果测量配置资源；并设立一个审查节点，让组织能够依据自身服务的证据缩减或调整项目。

本文值得检验的假设很明确：当单项任务收益很大、又高度依赖其他部门时，交接改善可能解释试点成功与运营成功之间差距的重要部分。如果在质量不变、测量充分的条件下，特定调整没有改善结果，我们就应修正假设或重新诊断瓶颈，而不是责怪现实没有理解这场转型。

回到开头的邮件屏幕。我们不应让省下的分钟证明超过它们实际证明的事情，也不应贬低这些时间对员工的价值。但组织决策需要追踪邮件关闭之后发生了什么：下一步是否因此成为可能？工作是否到达了客户可以使用的终点？

邮件或许一早就已抵达。尚未抵达的，是关于“完成”究竟意味着什么的共识。

参考文献

1. Dillon 等，Shifting Work Patterns with Generative AI，第四版记录：2025 年 11 月 13 日；AER: Insights：待刊。随机实验由包括微软研究人员在内的团队开展；事件与测量细节见正文。
2. Shakked Noy、Whitney Zhang，Experimental evidence on the productivity effects of generative artificial intelligence，《Science》；2023 年 7 月 13 日在线发表，7 月 14 日见刊。经同行评审的随机实验，支持写作任务改善结论。PubMed 记录。
3. Erik Brynjolfsson、Danielle Li、Lindsey Raymond，Generative AI at Work，《QJE》140(2): 889 – 942；2025 年 2 月 4 日在线发表，5 月见刊。经同行评审的分阶段导入研究；支持客户服务问题解决结果，不代表整家公司利润。
4. Paul Milgrom、John Roberts，The Economics of Modern Manufacturing: Technology, Strategy, and Organization，《American Economic Review》80(3): 511 – 528；1990 年 6 月。经同行评审的理论研究，是生产与组织选择互补性论述的基础。
5. Timothy F. Bresnahan 等，Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence，《QJE》117(1): 339 – 376；2002 年 2 月。经同行评审的观察性研究；提供互补性证据，但因果识别仍有限。
6. Mert Demirer 等，Chaining Tasks, Redefining Work: A Theory of AI Automation。工作论文；首版于 2026 年 2 月 22 日发布，文本于 9 月 28 日更新，第二版于 10 月 2 日提交。支持步骤排序与交接成本的作用；不是企业回报实验。
7. Fabrizio Dell’Acqua 等，The Cybernetic Teammate: A Field Experiment on Generative AI and Teamwork，《Organization Science》37(4): 1217 – 1242；2026 年 6 月 12 日在线发表。经同行评审的实验于 2024 年 5 月至 7 月开展；支持人工智能在早期创新中整合知识的作用。
8. Hyunjin Kim、Dahyeon Kim、Rembrand Koning，Mapping AI into Production: A Field Experiment on Firm Performance，工作论文，2026 年 4 月；以及哈佛商学院人工智能研究所 2026 年 4 月 7 日发布的研究设计机构说明。本文使用原始摘要与说明；因无法审阅完整手稿表格，未采用收入倍数。
9. Erik Brynjolfsson、Daniel Rock、Chad Syverson，The Productivity J-Curve: How Intangibles Complement General Purpose Technologies，《AEJ: Macroeconomics》13(1): 333 – 372；2021 年 1 月。结合模型与测量的同行评审研究；解释互补投资的作用，但不保证任何特定项目最终回升。
10. Michael G. Hudgens、M. Elizabeth Halloran，Toward Causal Inference With Interference，《Journal of the American Statistical Association》103(482): 832 – 842；2008 年 6 月。经同行评审的方法研究；支持在分析中考虑处理效应在单元之间的传递。