
When Does Better AI Become Economic Progress?

Odysseus Kaloidas
UC Berkeley

Abstract

When does a technical improvement in AI become an economic gain, and who receives that gain? A report may take half as long to draft while checking and correction consume the time saved. The work could still become better, but speed at one stage would not establish higher productivity overall. Adoption, changes in working practices, and benefits to users are separate steps between a better model and a more productive organization. Following those steps explains why technical progress and economic evidence can appear to diverge. The answer depends on what is measured, whose costs are counted, and what would have happened without AI.

1 Adoption Without a Visible Productivity Gain

How can AI be widely used in firms while its economic gains remain difficult to see? In a 2026 four-country survey of nearly 6,000 executives, widespread reported use coexists with most reporting no productivity or employment impact [13]. Productivity concerns output relative to the resources used. Understanding the apparent puzzle requires knowing how firms use AI, which resources they count, and which effects their executives can observe.

A new algorithm, a higher test score, a widely used product, and a more productive firm are different achievements. One can lead to the next, but each transition can fail. Reports about different stages can therefore sound contradictory while both being accurate.

Cho and colleagues argue that organizational conditions can prevent adoption from raising productivity [7]. Lee’s position on predicting labor impact likewise distinguishes outcomes, units, and time horizons [9]. The practical question is where the reported gain was measured and what comparison gives it meaning.

2 From Novelty to Adoption

FlashAttention offers an example at the technical end of this chain. Attention weights relationships among parts of a neural network’s input. FlashAttention computes it more efficiently by reducing data transfers within the chips used for AI computation [8]. The saving can be valuable without immediately changing a firm’s output. Software has to incorporate it, useful work has to benefit, and that benefit has to outweigh the costs of adoption.

HELM, a language-model evaluation suite, assesses models along several dimensions because desirable qualities can conflict [10]. A more accurate model may respond more slowly or require expensive checking. Once integrated into a workflow, the model that won the laboratory comparison can become the less useful choice.

Trial, occasional use, daily reliance, and replacement of a process are different forms of adoption. The 2026 firm survey reports that around 70 percent use AI, but executive use averages roughly 1.5 hours per week [13]. These self-reports describe adoption and perceived outcomes; they do not

identify a causal effect of zero. Modest intensity, difficult measurement, and limited benefits are all compatible with the aggregate description.

Selection can change as a tool improves. METR’s February 2026 developer follow-up encountered participants reluctant to accept assignments without AI, complicating comparisons with its earlier experiment [2]. That reluctance is evidence of perceived value, but it also changes which people and tasks remain in the sample. A behavior that makes adoption more plausible can make its measured effect harder to identify.

3 Complementary Investment and Delayed Returns

Evidence from professional writing and customer support shows why task-level studies matter: they examine work under defined conditions [11; 6]. Extending a finding across an economy requires additional assumptions about task composition, demand, competition, and the use of time saved.

Brynjolfsson and Hitt’s analysis of information technology links business performance to complementary organizational investments and intangible benefits [4]. The productivity J-curve describes an initial period when the cost of building these capabilities can obscure later gains [5]. Training, integration, and process redesign belong to the cost of making a technology useful. A slow aggregate response can therefore coexist with a real component improvement, without guaranteeing that the investment will eventually pay off.

Executives’ expectations of future gains describe their beliefs about this process [13]. They cannot yet tell us whether the investment will repay its cost.

4 Local Savings and the Workflow Boundary

Consider a hypothetical report that previously required four hours of drafting and two hours of review. An assistant halves drafting time but adds one hour of source checking. The six-hour workflow now takes five hours: a one-hour saving, rather than the two hours visible at the drafting stage. If a colleague must spend another hour correcting the report downstream, the total labor saving disappears.

The writer has saved time. The organization may not have. Counting the whole chain changes the answer because review and correction consume the same scarce resource as drafting: people’s working time. Even if total labor falls, we still need to ask whether the report became more useful. Time saved and value created are related, but neither measures the other completely.

5 The Counterfactual Behind an Impact Claim

Suppose firms using AI are more productive than firms that do not use it. Perhaps AI helped. But perhaps the adopters were already better organized or had more suitable work. The missing comparison is what those same adopters would have achieved without the tool. Causal inference calls this a counterfactual: an outcome under an alternative course of events [12]. We observe a firm with or without adoption, not both at once.

Consider a hypothetical comparison. Adopting teams complete twelve comparable reports per week, while other teams complete ten. It is tempting to credit AI with the two-report gap. But suppose the adopters would have completed eleven without it because they already had better processes. In that case AI accounts for one additional report; the other reflects a pre-existing difference. The observed comparison alone cannot reveal this division.

Random assignment can address this selection problem by making access to the tool independent of such prior differences, subject to the experiment’s assumptions. Even a good firm-level experiment leaves wider questions. One firm’s savings can change prices or demand for another firm’s work. Effects on competitors, customers, and new businesses may therefore differ from the measured effect on the original adopters.

Acemoglu and Restrepo distinguish work that automation takes over from new tasks that create demand for human labor [1]. Productivity gains, employment, and labor’s income share need not move together. General-purpose technologies can also stimulate complementary innovation [3]. A

benchmark gain consequently cannot determine either the size or the distribution of economic benefit: the tasks created, displaced, and reorganized enter the explanation.

Lee’s call to distinguish outcomes helps explain why apparently conflicting studies can both be informative [9]. A study of the writer’s time, one of the employer’s costs, and one of employment need not give the same answer. They are tracing different consequences of the same change.

Return to the report. A faster draft could yield a shorter working day, a better analysis, more reports, or an extra burden for a colleague. Each is a different economic story. Which one unfolds depends on how people reorganize the work, and what they choose to do with the possibilities a better tool creates.

References

- [1] Daron Acemoglu and Pascual Restrepo. Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, 33(2):3–30, 2019. doi: 10.1257/jep.33.2.3.
- [2] Joel Becker, Nate Rush, Tom Cunningham, David Rein, and Khalid Mahamud. We are changing our developer productivity experiment design. METR, February 24, 2026. URL <https://metr.org/blog/2026-02-24-uplift-update/>.
- [3] Timothy F. Bresnahan and Manuel Trajtenberg. General purpose technologies ‘Engines of growth’? *Journal of Econometrics*, 65(1):83–108, 1995. doi: 10.1016/0304-4076(94)01598-T.
- [4] Erik Brynjolfsson and Lorin M. Hitt. Beyond computation: Information technology, organizational transformation and business performance. *Journal of Economic Perspectives*, 14(4): 23–48, 2000. doi: 10.1257/jep.14.4.23.
- [5] Erik Brynjolfsson, Daniel Rock, and Chad Syverson. The productivity J-curve: How intangibles complement general purpose technologies. *American Economic Journal: Macroeconomics*, 13(1):333–372, 2021. doi: 10.1257/mac.20180386. URL <https://www.aeaweb.org/articles?id=10.1257/mac.20180386>.
- [6] Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. Generative AI at work. *The Quarterly Journal of Economics*, 140(2):889–942, 2025. doi: 10.1093/qje/qjae044. URL <https://academic.oup.com/qje/article/140/2/889/7990658>.
- [7] Won Ik Cho, Seong-hun Kim, and Geunhye Kim. Position: Adopting AI in Practice Does Not Guarantee the Productivity Boost. arXiv:2605.24688, 2026. URL <https://arxiv.org/abs/2605.24688>. ICML 2026 position paper; arXiv version.
- [8] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness, 2022. URL <https://arxiv.org/abs/2205.14135>. arXiv:2205.14135.
- [9] Yong Suk Lee. Position: Predicting AI’s impact on labor is a core machine learning problem, 2026. URL <https://icml.cc/virtual/2026/poster/67177>. ICML 2026 position paper. Official conference abstract consulted.
- [10] Percy Liang, Rishi Bommasani, Tony Lee, et al. Holistic evaluation of language models, 2022. URL <https://arxiv.org/abs/2211.09110>. arXiv:2211.09110.
- [11] Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, 2023. doi: 10.1126/science.adh2586. URL <https://doi.org/10.1126/science.adh2586>.
- [12] Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350.
- [13] Ivan Yotzov, Jose Maria Barrero, Nicholas Bloom, et al. Firm data on AI. Technical Report 2026-3, Federal Reserve Bank of Atlanta, 2026. URL <https://www.atlantafed.org/research-and-data/publications/working-papers/2026/03/24/03-firm-data-on-ai>.