Why AI Has Not Made You More Productive, and What to Do About It

Diwas KC, Goizueta Business School, Emory University

Executive Summary

Knowledge workers widely report that AI tools are impressive, yet their days have not gotten shorter, and their output has not grown much, reflecting a pattern that also holds at the level of the broader economy. The usual explanation is that the tools are inadequate or that we are in the early stages of adoption. I argue that the real bottleneck is human. Drawing on fifteen years of studying worker productivity, including studies in emergency departments, telemedicine, and clinical AI documentation, I identify three behavioral traps that absorb the time AI frees up: multitasking, which multiplies the hidden cost of switching between tasks; expansion, in which freed time is reabsorbed into more work rather than fewer hours; and lost focus, as a proliferation of tools erodes the conditions for deep work. The same research points to a practical response. Setting concurrency limits, deciding in advance how freed time will be used, and protecting blocks of focused attention let individuals and managers turn AI's capability into measurable productivity rather than busyness.

Introductory Note

AI tools are widely praised, yet workdays are no shorter and output little changed. Drawing on fifteen years of research on worker productivity, Diwas KC traces this gap to three behavioral traps — multitasking, expansion, and lost focus — and offers practical guardrails for individuals and the managers who lead them.

Ask any knowledge worker about the capabilities of AI productivity tools, and most will tell you these tools are genuinely impressive. They will draft your emails, create your presentation decks, and even produce detailed quantitative analyses. When you then ask if these tools have made them more productive, the enthusiasm, for many, is more muted: they are not much more productive; there is not much more free time, and the workday certainly has not gotten shorter.

These observations reflect findings at the aggregate economy level. Since the release of ChatGPT by OpenAI in 2022, aggregate productivity growth has hovered around 2 to 3 percent

per year, close to baseline levels over the last several decades.[1] These stand in sharp contrast to the forecasts that accompanied the technology's arrival. Goldman Sachs projected that generative AI could raise global GDP by 7 percent, nearly $7 trillion, and lift annual labor productivity growth by roughly 1.5 percentage points over a decade, a boost it compared to those once delivered by the electric motor and the personal computer.[2] McKinsey estimated that generative AI could add the equivalent of $2.6 to $4.4 trillion in value annually.[3]

Careful studies do find real gains at the level of individual tasks. Customer support agents with an AI assistant resolved about 15 percent more issues per hour.[4] Even at the task level the picture is uneven; one randomized study found AI assistance actually slowed experienced developers down, even as they believed they had sped up.[5] Although gains are realized at the level of the task, they dissipate somewhere by the end of the workday.

The most commonly offered explanation for this apparent productivity paradox is that the tools are not good enough yet, or that they are overhyped, or that we are still in the early stages of technology adoption.

I have a different explanation. The bottleneck is not the technology; it is us. The behavioral microfoundations of how we work, coupled with our own cognitive limits, are the true bottleneck. Understanding how we work and incorporating these AI tools more thoughtfully into the design and organization of work is necessary before we can unlock the promised productivity gains. The tools will continue to improve. But the traps described below will not disappear on their own, because they are rooted not in the technology but in how we work.

"The bottleneck is not the technology; it is us."

I have spent the last fifteen years studying worker productivity. Rather than relying on questionnaires and surveys, I have used detailed time-and-motion data that is often generated as a natural byproduct of providing goods and services. I have combined this data with contextual factors including operating conditions, task complexity, and even the effect of peers. This has allowed me to develop a detailed and comprehensive first-hand view of what makes workers improve how they work.

In almost all of my findings that span a wide range of workplaces and industries, from healthcare to retail to manufacturing, worker productivity is largely limited by behavioral and cognitive shortcomings.[6]

AI is no different. Whether AI usage will generate productivity gains or not boils down to the following three traps: multitasking, expansion, and focus.

The multitasking trap: the tendency to take on multiple concurrent tasks, where you repeatedly encounter the setup costs of switching between tasks. The expansion trap: the tendency to fill up any free time with more work. And the focus trap: the inability to muster the time and energy needed to get into a flow state, where you are most productive, and to stay there.

These traps do not require a proliferation of tools. Even a single AI tool can trigger the expansion trap, as the study below shows. But the traps are amplified as the tools multiply: more tools mean more concurrent sessions to juggle and more demands on attention.

The mechanisms behind each trap are not speculation. Each is established through years of detailed measurement of how people actually work, in settings where the stakes are high, and the data are granular. What is new is the argument that these same mechanisms are what stands between AI's capability and our productivity — an explanation for the productivity lapse that, to date, research has yet to examine directly.

The Multitasking Trap

Walk into the emergency department (ED) of a metropolitan hospital and you will find high levels of multitasking, with physicians overseeing the care of multiple patients concurrently.

Multitasking is, at its core, task switching. The brain can only process one complex task at a time, so multitasking is really sequential processing with switching in between. Neuroimaging studies bear this out, identifying a central bottleneck in the brain that forces complex tasks to be processed one at a time.[7] All multitasking activities are sequenced as a series of distinct tasks.

Each switch carries a setup cost: the time and energy needed to re-engage with where you left off. Suppose you are working on task A and then switch to task B. You have to reorient yourself and familiarize yourself with the new task at hand. When you complete task B and return to task A, a similar setup cost is incurred again. Add a third task, and the switching costs continue to

accumulate. The more you switch (A → B → C → A → B → C), the more of your cognitive capacity goes to setup rather than productive work.

But there is another cost as well: the cost of waiting and sitting idle. The optimal level of multitasking trades off these two costs.

To examine these dynamics, I looked at the ED, which has been called a natural laboratory for the study of errors. In the ED, setup costs involve refamiliarizing oneself with a given patient's history and where diagnosis and treatment was last left off; idle cost is incurred when a physician is free while waiting on lab results or consults on current patients. Utilizing this idle time can improve a physician's productivity.

In my study, after looking at over 145,000 emergency department visits, I found that on average the most productive physician maintained a multitasking level of between four and five patients; this level represented the ideal tradeoff between the two costs, resulting in not just productivity but also the highest levels of quality of care. Beyond four to five patients, the setup costs of each patient started to dominate, and below four to five, the idle costs became more prominent. As a result of this dynamic, the net effect of multitasking on performance was an inverted-U-shaped response.[8]

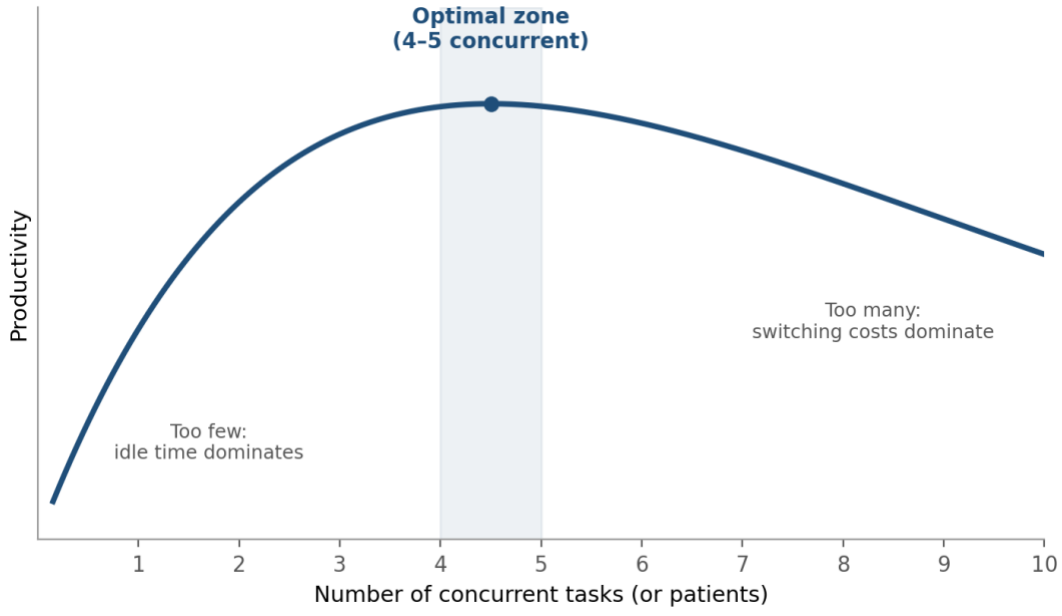


Figure 1. The inverted-U relationship between concurrent tasks and productivity (illustrative). In the emergency department, where productivity means patients treated per shift at a given level of quality, physician productivity peaked at four to five concurrent patients.

This dynamic is at play in the age of AI, where the tendency to start a new task is strong, as we are often offered a barrage of new and easy-to-use AI tools: an agent, a new chat, another tool for analysis. Starting a new task is cheap, or even free. The true setup cost is realized only later, in the form of retrieval: each chat and each agent accumulates context that must be reloaded before work can resume. Because the cost at the start is low, people take on more concurrent tasks. But once the real setup costs are factored in, they are multitasking well beyond their optimal level. The same multitasking dynamic is at play; but because the costs are not visible upfront, people end up multitasking excessively.

The Expansion Trap

In 1955, British historian C.N. Parkinson made an interesting observation: the time taken by public-sector government employees matched exactly the time that they were allotted.[9] The more general form of this finding, that work expands to fill the time available, is formally labeled Parkinson's Law and has been observed across many types of workplaces.

In the age of AI, even if AI tools and automation free up the worker's time, Parkinson's Law advises that one should be vigilant in deploying that free time intentionally, toward either meaningful breaks or more productive uses, lest those savings end up being squandered in simply filling up the time.

In the US healthcare system, in many care settings, providers are required to provide clinical notes for each patient visit. These notes largely serve documentation and financial reporting needs rather than direct patient care. They also impose significant administrative burdens on providers, and are a leading cause of burnout, dissatisfaction, and workforce attrition.

In a recent study, my coauthor and I examined the use of ambient AI scribes, which listen to the patient–provider interaction and craft a fully formatted clinical note that can be entered into the electronic health record (EHR) system. To examine the impact of ambient AI on care delivery, we collected detailed, granular data on the use of provider time across over 2,000 providers, both before and after the adoption of these tools.

We found that the tool generated time savings of about 12 percent in note-taking time, the time the provider would otherwise have spent typing notes into the EHR system. We then looked at total EHR time saved and found that it barely went down at all, by a mere 3 percent. In other words, although physicians saved time in note-taking, they gave up this time to other activities on the EHR system, such as reviewing the notes and orders, and examining the medical history of the patient. The freed-up time was not returned as free time; it remained on the screen.[10]

There is an important caveat here. On the one hand, the freed-up time is sometimes used for genuinely value-added activities: checking the work for errors, conducting additional follow-up analyses, or, in the ambient scribing example, examining the patient's history more thoroughly or verifying and editing the note further. The freed-up time can also be restorative for the provider: other health systems report that these same tools reduce burnout, a benefit that shows up in neither hours nor output.[11] I call this capability-induced expansion: the more powerful your tools, the more your desire to expand the scale and scope of the task. Progress is often the result of expanded capability meeting focused ambition. On the other hand, without deliberate management, expanded capability becomes expanded busyness, which is not the same as productivity.

"Without deliberate management, expanded capability becomes expanded busyness, which is not the same as productivity."

My evidence comes from field-based transactional data. Ranganathan and Ye reach a similar conclusion using a different lens, based on eight months of field-based ethnographic observation at a single U.S. technology company. They find that generative AI intensified work rather than reducing it as employees took on broader scope, worked faster, and extended their hours, often without being asked.[12]

The worker must therefore remain vigilant in how freed-up time is deployed. It should be determined a priori, before the savings have been realized, whether the goal is to generate more leisure time, or to conduct further related work. Absent this, there is the risk that the worker ends up yielding the saved time to non-value-adding, or tangential and undirected work.

The Focus Trap

In his seminal book *Flow*,[13] psychologist Mihaly Csikszentmihalyi defines flow as a state of optimal psychological performance: during this state, individuals experience intense focus and attention, resulting in high productivity. Writers describe the state of flow as where words come naturally to form eloquent prose; musicians experience flow as a heightened state of creative expression; and programmers describe it as allowing them to produce elegant and performant code.

The state of flow does not come naturally. There is effort that needs to be exerted to get into this state. For example, a statistician doing analysis needs to familiarize himself with the details of the data, the manner of data collection, and the empirical specification before he starts to produce effective analysis. Once this activation energy has been expended, it is also necessary to avoid distractions or interruptions to maintain this state of flow.

Allowing workers to achieve flow can yield significant productivity gains. For example, we recently studied the adoption of asynchronous telemedicine practice in the dermatology service line at the Veterans Health Administration. Adoption entailed a workflow change where the arriving patients were sorted into two care paths for referrals: an existing in-person channel and a newly introduced virtual channel. The primary care physicians sorted the patients such that the more complex cases were seen in-person, while the other cases were seen virtually. In comparison to the in-person channel, which had previously involved a mix of different cases, the virtual channel received cases that were simpler and more uniform. Moreover, unlike the in-person channel, there were no interruptions in the virtual channel. The providers worked asynchronously, examining the digital scans of patient conditions and providing their evaluations. The virtual-channel providers were therefore operating in an environment where they were more likely to experience flow.

In our analysis, we found that the adoption of telemedicine in fact led to a significant improvement in productivity for the VA, largely due to the sorting process, allowing the telemedicine-channel physicians the opportunity to deal with similar cases in a less distracting environment.[14]

AI tools can inhibit the realization of flow states in two ways. First, having multiple AI tools means that cognitive capacity is spread thin across multiple tasks. As a result, it is difficult to devote the time and attention necessary to achieve activation and get into the state of flow.

Second, AI tools, by design, prompt more frequent monitoring, alerts, and user inputs. All of this need for intermittent attention detracts from achieving and maintaining flow states.

Each of these traps by itself can diminish productivity. However, with AI productivity tools, these traps often tend to occur together, further compounding the productivity declines.

For example, if the use of an AI tool has saved you time, Parkinson's Law might lead you to take on additional AI-related tasks to fill up the time. Because these tools are easy to use and involve minimal setup costs, there is the tendency to multitask with several tools. Each task incurs the setup costs, which dissipate valuable time and attention. Consequently, you may find yourself spread too thin to generate the activation energy needed for flow state. Even if you manage to get into it, the additional frequent demands for time and attention will kick you out of any flow state. As a result, any time that was initially freed up is quickly dissipated by an increasing number of unfocused, and maybe even tangential, tasks.

What Should You Do About It?

These traps can be managed through the following practices, summarized in Table 1.

Trap	What happens	The fix
Multitasking	Starting AI tasks is cheap, so you take on too many; the setup cost is paid later, when each session's context must be reloaded	Set concurrency limits — for yourself, and for your team
Expansion	Freed time is quietly reabsorbed into more work rather than returned as output or shorter days	Decide in advance where freed time will go
Focus	A proliferation of tools and alerts prevents reaching, and keeping, a flow state	Protect focused time structurally: batch engagements, silence alerts

Table 1. The three traps and their fixes.

Set Concurrency Limits

Setting a concurrency limit beforehand provides the necessary guardrails. In the ED study, the optimal multitasking level varied quite a bit across physicians; although four to five was the average, some performed best at three, and others at eight. In other words, each physician had a different threshold, because for them the switching costs were different.

The same is true for knowledge workers managing AI tools more broadly. A practical signal that you have crossed your ideal threshold is that you are spending more time re-locating context, rereading a chat or reconstructing where an analysis was left off, prior to resuming productive work. At that point, the overhead of adding a new tool has exceeded the incremental benefit.

For managers, the same principle applies at the level of the team. Concurrency limits need not be left to each individual: managers can cap the number of concurrent projects, tools, and AI initiatives assigned to their teams. The limit will differ across individuals; the point is that there is one, and that someone is responsible for enforcing it.

Escape the Expansion Trap

If using an AI tool ends up saving you time, it would be prudent to decide how to spend that saved time beforehand. You may choose to either take a break, shorten the workday, pick up a new task, or enhance the output on the shortened task. This should be planned before the time savings have been realized, otherwise you may end up yielding the time to non-productive and non-value-adding tasks. Again, because the new AI tasks are so easy to begin, there is a tendency to fill up the freed-up time with tangential and often unproductive tasks. You then have neither the free time, more output, nor the enhanced output.

We saw this play out with the ambient scribing tool, whose initial promise was to free up screen time for more face-to-face time with the patient. But while note-taking time fell, total EHR time barely moved. The freed-up minutes were absorbed by other tasks rather than returned as shorter days.

The broader lesson is to claim the time deliberately, at the moment it is created. When an AI tool hands you back twenty minutes, name where those minutes go before you start the next thing, whether to the next patient, to addressing patient concerns, or out the door at the end of the day. We repeatedly find that freed time that isn’t assigned a purpose in advance gets quietly reabsorbed into the work. A simple test: would you have chosen this activity for the freed time in advance? Reviewing a patient's history more thoroughly passes the test.

This decision, too, need not be left to the individual. When a manager introduces an AI tool, the deployment plan should state where the expected time savings will go: to more throughput, to higher quality, or to shorter days. The telemedicine gains at the VHA did not come from

physicians individually resolving to focus; they came from managers redesigning the work. The same deliberateness should govern how freed time is spent.

Protect Focused Time

Protecting focus is less about discipline in the moment and more about structuring your work so that interruptions can't reach you in the first place. Flow is fragile and takes time to enter; a single alert can reset the clock. The goal is to arrange your tools and your time so that deep work is more likely to occur.

Recall the telemedicine physicians at the VHA. Their productivity improved because the work itself was reorganized around them, with similar cases batched and interruptions removed, so the conditions for focused work were built into the structure rather than left to individual willpower.

The same logic applies to AI tools, which are designed to ping, prompt, and await your reply. Left at their defaults, a handful of active tools will fragment your day. The fix is structural: turn off the alerts, batch the engagements, and set aside protected blocks where only one or two things have your attention.

Resolving the Paradox

In sum, the productivity promised by AI tools has not been achieved for many. The reason is not that the tools are too weak, or that we are early in the adoption curve. The reason is that the bottleneck was never the tools. It has always been us: our cognitive limits, and the behavioral tendencies that govern how we work.

This is why the gains will not arrive on their own. Counterintuitively, it is by placing more guardrails on how we use AI — limiting the number of concurrent tasks, deciding in advance where freed time goes, protecting blocks of focused attention — that we stand to gain the most from it. The constraint is what unlocks the capability.

"The constraint is what unlocks the capability."

Time and again, I find that managing ourselves is often harder than managing the technology. When it comes to using AI to improve productivity, the bottleneck isn't the artificial neural network. It's the human one.

Author Bio

Diwas KC is the John M. and Lucy Cook Chair in Information Strategy and Professor of Information Systems and Operations Management at the Goizueta Business School, Emory University, with a secondary appointment at the Rollins School of Public Health. His research examines worker productivity, technology adoption, and healthcare delivery, and has appeared in leading journals in management and operations. diwas.kc@emory.edu

Endnotes

1. Bureau of Labor Statistics, U.S. Department of Labor, "Productivity and Costs, Fourth Quarter and Annual Averages 2025, Revised," March 24, 2026, https://www.bls.gov/news.release/prod2.nr0.htm

2. Goldman Sachs Research, "Generative AI Could Raise Global GDP by 7%," March 2023.

3. McKinsey Global Institute, "The Economic Potential of Generative AI: The Next Productivity Frontier," McKinsey & Company, June 2023.

4. E. Brynjolfsson, D. Li, and L.R. Raymond, "Generative AI at Work," *The Quarterly Journal of Economics* 140, no. 2 (2025): 889–942.

5. J. Becker, N. Rush, E. Barnes, and D. Rein, "Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity," METR Report, July 2025.

6. D.S. KC, *Worker Productivity in Operations Management*, Foundations and Trends in Technology, Information, and Operations Management 13, no. 3 (2020): 151–249.

7. P.E. Dux, J. Ivanoff, C.L. Asplund, and R. Marois, "Isolation of a Central Bottleneck of Information Processing with Time-Resolved fMRI," *Neuron* 52, no. 6 (2006): 1109–1120.

8. D.S. KC, "Does Multi-Tasking Improve Performance? Evidence from the Emergency Department," *Manufacturing & Service Operations Management* 16, no. 2 (2014): 168–183.

9. C.N. Parkinson, "Parkinson's Law," *The Economist*, November 19, 1955.

10. R. Gurlek and D.S. KC, "Generative AI and Physician Productivity," working paper, Emory University, 2026, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6755362

11. K.D. Olson, D. Meeker, M. Troup, T.D. Barker, V.H. Nguyen, J.B. Manders, C.D. Stults, V.G. Jones, S.D. Shah, T. Shah, and L.H. Schwamm, "Use of Ambient AI Scribes to Reduce Administrative Burden and Professional Burnout," *JAMA Network Open* 8, no. 10 (2025): e2534976.

12. Aruna Ranganathan and Xingqi Maggie Ye, "AI Doesn't Reduce Work—It Intensifies It," *Harvard Business Review*, February 9, 2026.

13. M. Csikszentmihalyi, *Flow: The Psychology of Optimal Experience* (New York: Harper & Row, 1990).

14. S. Tushe, D.S. KC, H. Ding, and H. Yeung, "Multi-channel Healthcare: Impact of Asynchronous Telemedicine Adoption on Patient Flow," *Manufacturing & Service Operations Management* 27, no. 1 (2025): 59–74.