Gen AI: too much spend, too little benefit?

The promise of generative AI technology to transform companies, industries, and societies continues to be touted, leading tech giants, other companies, and utilities to spend an estimated ~$1tn on capex in coming years, including significant investments in data centers, chips, other AI infrastructure, and the power grid. But this spending has little to show for it so far beyond reports of efficiency gains among developers. And even the stock of the company reaping the most benefits to date—Nvidia—has sharply corrected. We ask industry and economy specialists whether this large spend will ever pay off in terms of AI benefits and returns, and explore the implications for economies, companies, and markets if it does, or if it doesn't.

We first speak with Daron Acemoglu, Institute Professor at MIT, who's skeptical. He estimates that only a quarter of AI-exposed tasks will be cost-effective to automate within the next 10 years, implying that AI will impact less than 5% of all tasks. And he doesn't take much comfort from history that shows technologies improving and becoming less costly over time, arguing that AI model advances likely won't occur nearly as quickly—or be nearly as impressive—as many believe. He also questions whether AI adoption will create new tasks and products, saying these impacts are "not a law of nature." So, he forecasts AI will increase US productivity by only 0.5% and GDP growth by only 0.9% cumulatively over the next decade.

GS Head of Global Equity Research Jim Covello goes a step further, arguing that to earn an adequate return on the ~$1tn estimated cost of developing and running AI technology, it must be able to solve complex problems, which, he says, it isn't built to do. He points out that truly life-changing inventions like the internet enabled low-cost solutions to disrupt high-cost solutions even in its infancy, unlike costly AI tech today. And he's skeptical that AI's costs will ever decline enough to make automating a large share of tasks affordable given the high starting point as well as the complexity of building critical inputs—like GPU chips—which may prevent competition. He's also doubtful that AI will boost the valuation of companies that use the tech, as any efficiency gains would likely be competed away, and the path to actually boosting revenues is unclear, in his view. And he questions whether models trained on historical data will ever be able to replicate humans' most valuable capabilities.

But GS senior global economist Joseph Briggs is more optimistic. He estimates that gen AI will ultimately automate 25% of all work tasks and raise US productivity by 9% and GDP growth by 6.1% cumulatively over the next decade. While Briggs acknowledges that automating many AI-exposed tasks isn't cost-effective *today*, he argues that the large potential for cost savings and likelihood that costs will decline over the long run—as is often, if not always, the case with new technologies—should eventually lead to more AI automation. And, unlike Acemoglu, Briggs incorporates both the potential for labor reallocation and new task creation into his productivity estimates, consistent with the strong and long historical record of technological innovation driving new opportunities.

GS US software analyst Kash Rangan and internet analyst Eric Sheridan also remain enthusiastic about generative AI's long-term transformative and returns potential even as AI's "killer application" has yet to emerge. Despite big tech's large spending on AI infrastructure, they don't see signs of irrational exuberance. Indeed, Sheridan notes that current capex spend as a share of revenues doesn't look markedly different from prior tech investment cycles (see pg. 15), and that investors are rewarding only those companies that can tie a dollar of AI spending back to revenues. Rangan, for his part, argues that the potential for returns from this capex cycle seems more promising than even previous cycles given that incumbents with low costs of capital and massive distribution networks and customer bases are leading it. So, both Sheridan and Rangan are optimistic that the huge AI spend will eventually pay off.

But even if AI could potentially generate significant benefits for economies and returns for companies, could shortages of key inputs—namely, chips and power—keep the technology from delivering on this promise? GS US semiconductor analysts Toshiya Hari, Anmol Makkar, and David Balaban argue that chips will indeed constrain AI growth over the next few years, with demand for chips outstripping supply owing to shortages in High-Bandwidth Memory technology and Chip-on-Wafer-on-Substrate packaging—two critical chip components.

But the bigger question seems to be whether power supply can keep up. GS US and European utilities analysts Carly Davenport and Alberto Gandolfi, respectively, expect the proliferation of AI technology, and the data centers necessary to feed it, to drive an increase in power demand the likes of which hasn't been seen in a generation (which GS commodities strategist Hongcen Wei finds early evidence of in Virginia, a hotbed for US data center growth).

Brian Janous, Co-founder of Cloverleaf Infrastructure and former VP of Energy at Microsoft, believes that US utilities—which haven't experienced electricity consumption growth in nearly two decades and are contending with an already aged US power grid—aren't prepared for this coming demand surge. He and Davenport agree that the required substantial investments in power infrastructure won't happen quickly or easily given the highly regulated nature of the utilities industry and supply chain constraints, with Janous warning that a painful power crunch that could constrain AI's growth likely lies ahead.

So, what does this all mean for markets? Although Covello believes AI's fundamental story is unlikely to hold up, he cautions that the AI bubble could take a long time to burst, with the "picks and shovels" AI infrastructure providers continuing to benefit in the meantime. GS senior US equity strategist Ryan Hammond also sees more room for the AI theme to run and expects AI beneficiaries to broaden out beyond just Nvidia, and particularly to what looks set to be the next big winner: Utilities.

That said, looking at the bigger picture, GS senior multi-asset strategist Christian Mueller-Glissmann finds that only the most favorable AI scenario, in which AI significantly boosts trend growth and corporate profitability without raising inflation, would result in above-average long-term S&P 500 returns, making AI's ability to deliver on its oft-touted potential even more crucial.

Allison Nathan, Editor

Email: allison.nathan@gs.com
Tel: 212-357-7504
Goldman Sachs & Co. LLC

Interview with Daron Acemoglu

Daron Acemoglu is Institute Professor at MIT and has written several books, including *Why Nations Fail: The Origins of Power, Prosperity, and Poverty* and his latest, *Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity*. Below, he argues that the upside to US productivity and growth from generative AI technology over the next decade—and perhaps beyond—will likely be more limited than many expect.

The views stated herein are those of the interviewee and do not necessarily reflect those of Goldman Sachs.

Allison Nathan: In a recent paper, you argued that the upside to US productivity and, consequently, GDP growth from generative AI will likely prove much more limited than many forecasters—including Goldman Sachs—expect. Specifically, you forecast a ~0.5% increase in productivity and ~1% increase in GDP in the next 10 years vs. GS economists' estimates of a ~9% increase in productivity and 6.1% increase in GDP. Why are you less optimistic on AI's potential economic impacts?

Daron Acemoglu: The forecast differences seem to revolve more around the timing of AI's economic impacts than the ultimate promise of the technology. Generative AI has the potential to fundamentally change the process of scientific discovery, research and development, innovation, new product and material testing, etc. as well as create new products and platforms. But given the focus and architecture of generative AI technology today, these truly transformative changes won't happen quickly and few—if any—will likely occur within the next 10 years. Over this horizon, AI technology will instead primarily increase the efficiency of existing production processes by automating certain tasks or by making workers who perform these tasks more productive. So, estimating the gains in productivity and growth from AI technology on a shorter horizon depends wholly on the number of production processes that the technology will impact and the degree to which this technology increases productivity or reduces costs over this timeframe.

My prior guess, even before looking at the data, was that the number of tasks that AI will impact in the short run would not be massive. Many tasks that humans currently perform, for example in the areas of transportation, manufacturing, mining, etc., are multifaceted and require real-world interaction, which AI won't be able to materially improve anytime soon. So, the largest impacts of the technology in the coming years will most likely revolve around pure mental tasks, which are non-trivial in number and size but not huge, either.

To quantify this, I began with Eloundou et al.'s comprehensive study that found that the combination of generative AI, other AI technology, and computer vision could transform slightly over 20% of value-added tasks in the production process. But that's a timeless prediction. So, I then looked at another study by Thompson et al. on a subset of these technologies—computer vision—which estimates that around a quarter of tasks that this technology can perform could be cost-effectively automated within 10 years. If only 23% of exposed tasks are cost effective to automate within the next ten years, this suggests that only 4.6% of all tasks will be impacted by AI. Combining this figure with the 27% average labor cost savings estimates from Noy and Zhang's and Brynjolfsson et al.'s studies implies that total factor productivity effects within the next decade should be no more than 0.66%—and an even lower 0.53% when adjusting for the complexity of hard-to-learn tasks. And that figure roughly translates into a 0.9% GDP impact over the decade.

Allison Nathan: Recent studies estimate cost savings from the use of AI ranging from 10% to 60%, yet you assume only around 30% cost savings. Why is that?

Daron Acemoglu: Of the three detailed studies published on AI-related costs, I chose to exclude the one with the highest cost savings—Peng et al. estimates of 56%—because the task in the study that AI technology so markedly improved was notably simple. It seems unlikely that other, more complex, tasks will be affected as much. Specifically, the study focuses on time savings incurred by utilizing AI technology—in this case, GitHub Copilot—for programmers to write simple subroutines in HTML, a task for which GitHub Copilot had been extensively trained. My sense is that such cost savings won't translate to more complex, open-ended tasks like summarizing texts, where more than one right answer exists. So, I excluded this study from my cost-savings estimate and instead averaged the savings from the other two studies.

Allison Nathan: While AI technology cannot perform many complex tasks well today—let alone in a cost-effective manner—the historical record suggests that as technologies evolve, they both improve and become less costly. Won't AI technology follow a similar pattern?

Daron Acemoglu: Absolutely. But I am less convinced that throwing more data and GPU capacity at AI models will achieve these improvements more quickly. Many people in the industry seem to believe in some sort of scaling law, i.e. that doubling the amount of data and compute capacity will double the capability of AI models. But I would challenge this view in several ways. What does it mean to double AI's capabilities? For open-ended tasks like customer service or understanding and summarizing text, no clear metric exists to demonstrate that the output is twice as good. Similarly, what does a doubling of data really mean, and what can it achieve? Including twice as much data from Reddit into the next version of GPT may improve its ability to predict the next word when engaging in an informal conversation, but it won't necessarily improve a customer service representative's ability to help a customer troubleshoot problems with their video service. The quality of the data also matters, and it's not clear where more high-quality data will come from and whether it will be easily and cheaply available to AI models. Lastly, the current architecture of AI

technology itself may have limitations. Human cognition involves many types of cognitive processes, sensory inputs, and reasoning capabilities. Large language models (LLMs) today have proven more impressive than many people would have predicted, but a big leap of faith is still required to believe that the architecture of predicting the next word in a sentence will achieve capabilities as smart as HAL 9000 in *2001: A Space Odyssey*. It's all but certain that current AI models won't achieve anything close to such a feat within the next ten years.

Allison Nathan: So, are the risks to even your relatively conservative estimates of AI's economic impacts over the next 5-10 years skewed to the downside?

Daron Acemoglu: Both downside and upside risks exist. Technological breakthroughs are always possible, although even such breakthroughs take time to have real impact. But even my more conservative estimates of productivity gains may turn out to be too large if AI models prove less successful in improving upon more complex tasks. And while large organizations such as the tech companies leading the development of AI technology may introduce AI-driven tools quickly, smaller organizations may be slower to adopt them.

Allison Nathan: Over the longer term, what odds do you place on AI technology achieving superintelligence?

Daron Acemoglu: I question whether AI technology can achieve superintelligence over even longer horizons because, as I said, it is very difficult to imagine that an LLM will have the same cognitive capabilities as humans to pose questions, develop solutions, then test those solutions and adopt them to new circumstances. I am entirely open to the possibility that AI tools could revolutionize scientific processes on, say, a 20-30-year horizon, but with humans still in the driver's seat. So, for example, humans may be able to identify a problem that AI could help solve, then humans could test the solutions the AI models provide and make iterative changes as circumstances shift. A truly superintelligent AI model would be able to achieve all of that without human involvement, and I don't find that likely on even a thirty-year horizon, and probably beyond.

Allison Nathan: Your colleague David Autor and coauthors have shown that technological innovations tend to drive the creation of new occupations, with 60% of workers today employed in occupations that didn't exist 80 years ago. So, could the impact of AI technology over the longer term prove more significant than you expect?

Daron Acemoglu: Technological innovation has undoubtedly meaningfully impacted nearly every facet of our lives. But that impact is not a law of nature. It depends on the types of technologies that we invent and how we use them. So, again, my hope is that we use AI technology to create new tasks, products, business occupations, and competencies. In my example about how AI tools may revolutionize scientific discovery, AI models would be trained to help scientists conceive of and test new materials so that humans can then be trained to become more specialized and provide better inputs into the AI models. Such an evolution would ultimately lead to much better possibilities for human discovery. But it is by no means guaranteed.

Allison Nathan: Will some—or maybe even most—of the substantial spending on AI technology today ultimately go to waste?

Daron Acemoglu: That is an interesting question. Basic economic analysis suggests that an investment boom should occur because AI technology today is primarily used for automation, which means that algorithms and capital are substituting for human labor, which should lead to investment. This explains why my estimates for GDP increases are nearly twice as large as my estimates for productivity increases. But then reality supervenes and says that some of the spending will end up wasted because some projects will fail, and some firms will be too optimistic about the extent of the efficiency gains and cost savings they can achieve or their ability to integrate AI into their organizations. On the other hand, some of the spending will plant the seeds for the next, and more promising, phase of the technology. The devil is ultimately in the details. So, I don't have a strong prior as to how much of the current investment boom will be wasted vs. productive. But I expect both will happen.

Allison Nathan: Are other costs of AI technology not receiving enough attention?

Daron Acemoglu: Yes. GDP is not everything. Technology that has the potential to provide good information can also provide bad information and be misused for nefarious purposes. I am not overly concerned about deepfakes at this point, but they are the tip of the iceberg in terms of how bad actors could misuse generative AI. And a trillion dollars of investment in deepfakes would add a trillion dollars to GDP, but I don't think most people would be happy about that or benefit from it.

Allison Nathan: Given everything we've discussed, is the current enthusiasm around AI technology overdone?

Daron Acemoglu: Every human invention should be celebrated, and generative AI is a true human invention. But too much optimism and hype may lead to the premature use of technologies that are not yet ready for prime time. This risk seems particularly high today for using AI to advance automation. Too much automation too soon could create bottlenecks and other problems for firms that no longer have the flexibility and trouble-shooting capabilities that human capital provides.

And, as I mentioned, using technology that is so pervasive and powerful—providing information and visual or written feedback to humans in ways that we don't yet fully understand and don't at all regulate—could prove dangerous. Although I don't believe superintelligence and evil AI pose major threats, I often think about how the current risks might be perceived looking back 50 years from now. The risk that our children or grandchildren in 2074 accuse us of moving too slowly in 2024 at the expense of growth seems far lower than the risk that we end up moving too quickly and destroy institutions, democracy, and beyond in the process. So, the costs of the mistakes that we risk making are much more asymmetric on the downside. That's why it's important to resist the hype and take a somewhat cautious approach, which may include better regulatory tools, as AI technologies continue to evolve.

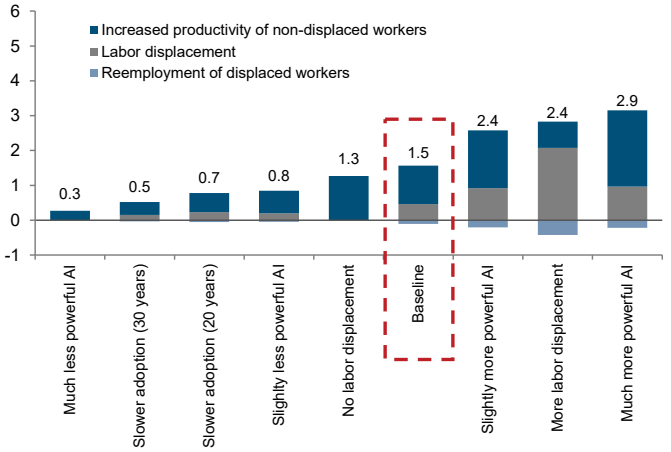
Addressing the AI growth debate

Joseph Briggs addresses the AI productivity and growth debate, arguing that generative AI will likely lead to significant economic upside

We have long argued that generative AI could lead to significant economic upside, primarily owing to its ability to automate a large share of work tasks, with our baseline estimate implying as much as 15% cumulative gross upside to US labor productivity and GDP growth[1] following widespread adoption of the technology.

A significant boost to US labor productivity from generative AI

Effect of AI adoption on annual US labor productivity growth, 10y adoption period, pp



Source: Goldman Sachs GIR.

That said, substantial debate exists around generative AI's potential macro impacts. Studies that assume generative AI will accelerate the development and adoption of robotics or that view recent generative AI advances as foreshadowing the emergence of a "superintelligence", for example, estimate even more upside to productivity and GDP than our baseline forecast. We see such outcomes as possible but premature since they generally assume AI advancements well beyond the frontier of current models.

More notably, MIT economist Daron Acemoglu sees much more limited upside to US productivity and GDP than we expect, with his baseline estimates implying that generative AI will boost US total factor productivity (TFP) by 0.53% and GDP by 0.9% over the next 10 years (see pgs. 4-5). As we take similar approaches to assessing the economic impacts of generative AI, we explore what explains the large differences in our estimates.

Breaking down the differences

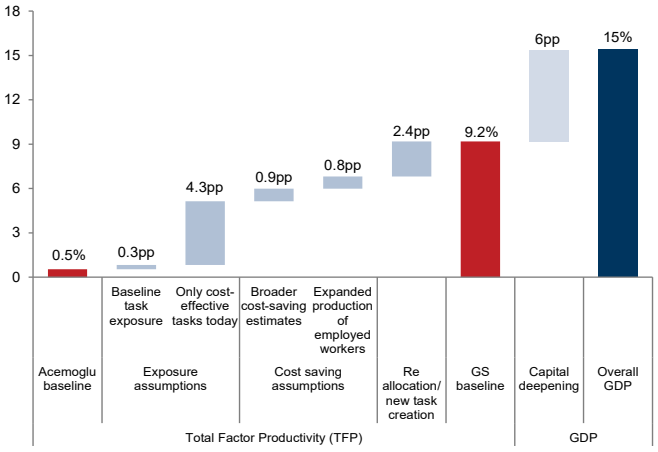
We find two main factors that explain the differences in our estimates versus those of Acemoglu. First, Acemoglu assumes that generative AI will automate only 4.6% of total work tasks, as he estimates that only 19.9% of all tasks are exposed to AI and assumes that only 23% of exposed tasks will be cost effective to automate within the next ten years. In contrast, we assume that generative AI will automate 25% of all work tasks following the technology's full adoption.

Second, Acemoglu's framework assumes that the primary driver of cost savings will be workers completing existing tasks more efficiently and ignores productivity gains from labor reallocation or the creation of new tasks. In contrast, our productivity estimates incorporate both worker reallocation—via displacement and subsequent reemployment in new occupations made possible by AI-related technological advancement—and new task creation that expands non-displaced workers' production potential.

Differences in these assumptions explain over 80% of the discrepancy between our 9.2%[2] and Acemoglu's 0.53% estimates of increases in TFP over the next decade[3]. The remaining 20% of the gap reflects differences in cost savings and marginal productivity assumptions. For instance, Acemoglu assumes 27% cost savings based on two studies that he considers the most representative of AI's real-world impact, but cost savings would rise to 36% if the full set of studies were considered. We are also more optimistic that AI will raise non-displaced workers' output, largely because we expect AI automation to create new tasks and products.

Differences in macro estimates mostly reflect differences in assumptions around tasks that can be profitably automated and the reallocation of labor to new tasks

Reconciling estimates of AI impact on GDP: Acemoglu (2024) vs. GS (2023), %



Source: Goldman Sachs GIR.

More widespread AI automation ahead

So, whose estimates regarding the share of automated tasks and new task creation—will more likely prove correct?

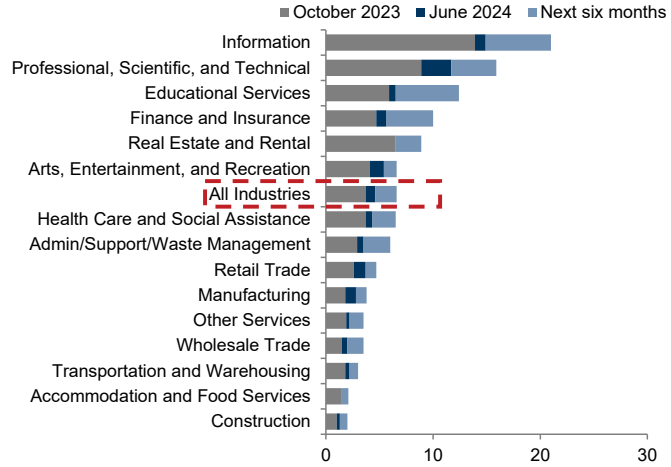
We are very sympathetic to Acemoglu's argument that automation of many AI-exposed tasks is not cost effective today, and may not become so even within the next ten years. AI adoption remains very modest outside of the few

[1] Our GDP estimate assumes that the capital stock evolves to match increased labor potential, which seems broadly validated by the sizable investment response aimed at facilitating the AI transition.
[2] This figured is calculated by multiplying the labor share of output, 62%, by our 15% estimate of the AI upside to labor productivity and growth.
[3] The quantitative contribution of different channels to the discrepancy between Acemoglu's and our estimates depends on the order that they are considered in, with differences in exposure assumptions explaining more of the gap if differences in cost savings assumptions are considered first and vice versa. To reduce this sensitivity, we consider both orderings and present the average contributions.

industries—including computing and data infrastructure, information services, and motion picture and sound production—that we estimate will benefit the most, and adoption rates are likely to remain below levels necessary to achieve large aggregate productivity gains for the next few years. This explains why we only raised our US GDP forecast by 0.4pp by the end of our forecast horizon in 2034 (with smaller increases in other countries) when we incorporated an AI boost into our global potential growth forecasts last fall. When stripping out offsetting growth impacts from the partial redirection of capex from other technologies to AI and slower productivity growth in a non-AI counterfactual, this 0.4pp annual figure translates into a 6.1% GDP uplift from AI by 2034 vs. Acemoglu's 0.9% estimate.

AI adoption remains modest on average across industries

Share of US firms using AI by sector, %



Source: Census Bureau, Goldman Sachs GIR.

That said, the full automation of AI exposed tasks that are likely to occur over a longer horizon could generate significant cost savings to the tune of several thousands of dollars per worker per year. The cost of new technologies also tends to fall rapidly over time. Given that cost-saving applications of generative AI will likely follow a similar pattern, and that the marginal cost of deployment will likely be very small once applications are developed, we expect AI adoption and automation rates to ultimately far exceed Acemoglu's 4.6% estimate.

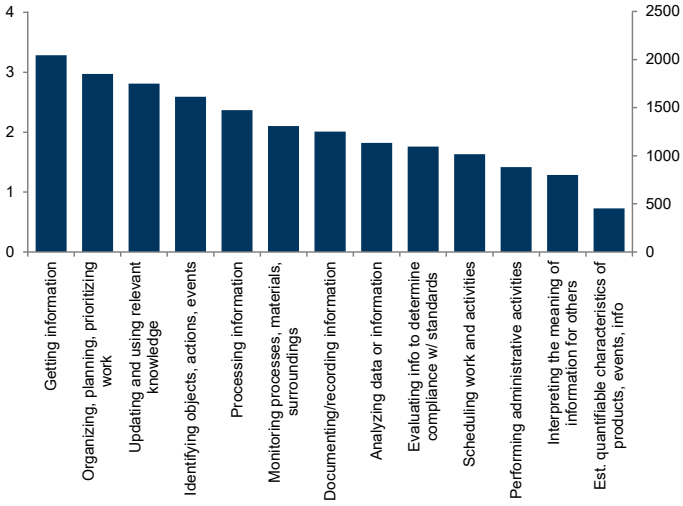
Labor reallocation and new task creation on the horizon

We also disagree with Acemoglu's decision not to incorporate productivity improvements from new tasks and products into his estimates, partly given his questioning of whether AI adoption will lead to labor reallocation and the creation of new tasks. The historical record provides strong evidence that economic growth stems mainly from technology-driven reallocation of resources and expansion of the production frontier, and we anticipate that AI will raise output both by raising demand in areas where labor has a comparative advantage and by creating new opportunities that were previously technologically or economically infeasible.

This dynamic clearly played out following the emergence of information technology—which created new occupations like webpage designers, software developers, and digital marketing professionals and indirectly drove demand for service sector workers in industries like healthcare, education, and food services—and is visible over a much longer horizon in recent work by MIT economist David Autor and coauthors. Using Census data, they find that 60% of workers today are employed in occupations that did not exist in 1940, with their estimates implying that the technology-driven creation of new occupations accounts for more than 85% of employment growth over the last 80 years.

Automation of work tasks should generate significant economic value, particularly as costs decline

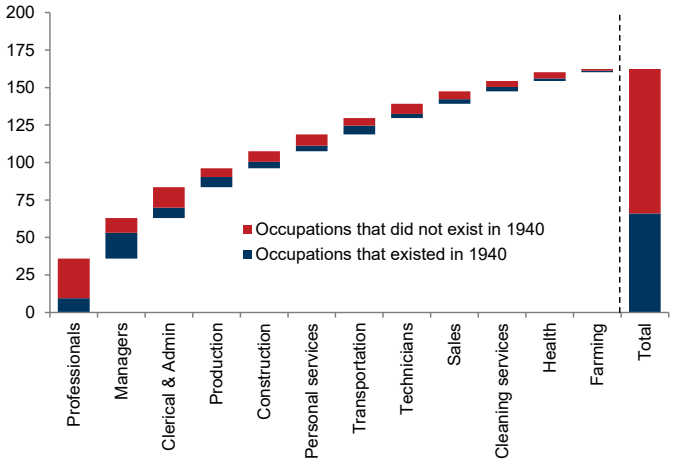
Value of automating work task categories per worker, % of time (lhs), $ (rhs)



Source: Goldman Sachs GIR.

Technological creation of new opportunities is a main driver of employment and economic growth

Employment by new and pre-existing occupations, millions



Source: Autor et al. (2022), Goldman Sachs GIR.

Accordingly, while we believe that Acemoglu's relatively pessimistic assessment of generative AI's economic potential highlights valid concerns that the macroeconomic impacts could be more backloaded than is commonly appreciated, we maintain that generative AI's large potential to drive automation, cost savings, and efficiency gains should eventually lead to significant uplifts of productivity and GDP.

Joseph Briggs, Senior Global Economist

Email: joseph.briggs@gs.com *Goldman Sachs & Co. LLC*
Tel: 212-902-2163