The Oxford Martin Working Paper Series on the Future of Work

Generative AI and the Future of Work: A Reappraisal

Carl Benedikt Frey and Michael Osborne

Working Paper No. 2023

Forthcoming in Brown Journal of World Affairs

Generative AI and the Future of Work: A Reappraisal

By Carl Benedikt Frey[1] and Michael Osborne[2]

Forthcoming in Brown Journal of World Affairs

The early 2010s, like now, was a time of excitement about new technological wonders. In the early 2010s, the world had bounced back from the worst financial crisis since the Great Depression. But despite financial malaise, America seemed to be at the cusp of a productivity boom.[3] Artificial intelligence (AI), long regarded as an academic backwater, was finally showing signs of bearing fruit. In 2011, IBM Watson had just beaten the world champion in Jeopardy! And machine learning—a subfield of AI—was expanding the premises of what computers could do. In 2012, a machine learning team from Google trained a deep neural network that, without ever being told what a cat was, proved independently capable of recognising cat videos on YouTube.

Gone were the days when a computer programmer needed to write down explicit rules to guide the actions of a machine in every contingency. Computers could now infer rules themselves, by tapping into the data trails more and more humans were leaving behind online. No programmer could be expected to foresee every situation a human driver might encounter in city traffic; even less, capture these in a sequence of If-Then-Do commands. Instead, progress in autonomous vehicles was being made by collecting vast amounts of data on drivers' actions in city traffic to predict what a human would have done in any given situation.

It was around that time—in 2013, to be precise—that we published a working paper estimating that 47% of jobs are exposed to the recent advances in machine learning and advanced robotics.[4] Ian Goodfellow's and co-authors' paper on "Generative Adversarial Networks" had not yet been published.[5] But there were glimpses of the age of Generative AI in our estimates: fashion models, we found, were among the jobs at risk. And a few years later, digital models were being generated *en masse*. Overall, however, we firmly believed that in the absence of major leaps, tasks requiring creativity and social intelligence, as well as unstructured manual work, would remain safe havens for human workers. The jobs of journalists, scientists, software engineers, art directors, and architects, we documented, were all at low risk of being automated.

Fast-forward a decade and Large Language Models (LLMs), like GPT-4, can answer questions and write essays in astonishingly human-like fashion, just like image-generators, like DALL-E 2, can transform text prompts into new images, possibly replacing designers and

---

[1] Carl Benedikt Frey is the Dieter Schwarz Associate Professor of AI & Work at the Oxford Internet Institute and an Official Fellow of Mansfield College, University of Oxford. He is also Director of the Future of Work Programme and Oxford Martin Citi Fellow at the Oxford Martin School.

[2] Mike Osborne is a Professor of Machine Learning at the University of Oxford, an Official Fellow of Exeter College, Oxford, and a co-founder of Mind Foundry.

[3] This was also true of the Great Depression, see Field, A. J. (2003). The most technologically progressive decade of the century. American Economic Review, 93(4), 1399-1413.

[4] Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation?. Technological forecasting and social change, 114, 254-280. The first working paper version was published in 2013.

[5] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. Communications of the ACM, 63(11), 139-144. This was published as a working paper in 2014.

advertising executives. Not to mention Github's Copilot—an AI-powered "pair programmer", capable of writing code.

The number of domains in which AI has apparently acquired human-level mastery is simply breath-taking. OpenAI report that their LLM passed a simulated bar exam, SAT Math, and Introductory Sommelier training with a score around the top 10% of test takers.[6] And in just a few months, as the world moved from GPT-3.5 to GPT-4, AI made a leap from the 39th to the 96th percentile of human level performance in solving college physics problems.[7] While such test results may largely reflect LLMs 'parroting' answers to common exam questions that are found in its training set,[8] they do beg the question: did we underestimate the near-term scope of automation?

In what follows, we provide a reassessment of the division of labour between humans and computers in the age of AI. Doing so, we explore whether Generative AI has, in fact, changed the rules of the game, threatening to upend the special status of humans in 1) creative, 2) inherently social, and 3) unstructured work. We conclude by discussing the labour market implications of recent trends in technology.

The rise of LLMs

Ever since MIT's Joseph Weizenbaum launched ELIZA in 1966, computer scientists have tried to build social machines.[9] Named after Eliza Doolittle—the protagonist from George Bernard Shaw's 1913 play Pygmalion—ELIZA was the first algorithm able to facilitate some remotely plausible conversation between humans and machines. In the style of Rogerian psychotherapy, ELIZA would take the input it was fed and rephrase it into a question. If you told it about the betrayal of a friend, it would respond, "Why do you feel betrayed?"

ELIZA, it goes without saying, would never have succeeded at the imitation game—a test devised by Alan Turing in 1950, which many came to regard as a reasonable benchmark for machine intelligence. If an algorithm could convince a human interlocutor that she was talking to another person, surely it must be understanding something? And so, Turing test competitions, in which judges are tasked with distinguishing between human and algorithm conversants—whose identities are unknown—became a common standard for measuring progress in AI.

Yet half a century later, chatbots were still underwhelming. True, in 2012, on what would have been Alan Turing's 100th birthday, a bot called Eugene Goostman managed to convince 33% of human judges that it was human. But the bot did so by pretending to be a boy from Odessa with poor language skills and with no understanding of English culture. Rather than constituting a leap in AI, it highlighted the flaws of the Turing test—Goostman was a good bullshit artist, but nothing more.

Such limited progress, even in basic text communication, led us to conclude that jobs requiring human-level social intelligence remained safe from automation—although we noted that for a computer to make a subtle joke, all that was needed, in principle, was a large database of human-generated jokes and methods of benchmarking the algorithm's performance. We now have both. Trained on vast amounts of data—from books, articles, and websites—that would

[6] OpenAI (2023). GPT-4 Technical Report. arXiv:submit/4812508.

[7] West, C.G. (2023). Advances in apparent conceptual physics reasoning in ChatGPT-4. arXiv preprint arXiv:2303.17012.

[8] The data sources LLMs have been trained on remain unknown to the outside world.

[9] Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1), 36-45.

take a thousand human lifetimes to read, today's LLMs learn patterns and relationships between words and phrases, allowing them to make predictions about the next word in a sentence based on the context. This, together with clever approaches to benchmarking AI output—through the use of reinforcement learning to nudge the system in the right direction—has yielded spectacular results.

Consider the following ChatGPT request by one user: "Write the complete script of a Seinfeld scene in which Jerry needs to learn the bubble sort algorithm." To achieve this, the AI drew upon its training, on a vast body of human text that likely includes scripts, to identify the critical "features" of a "Seinfeld script", such that, in its response, the AI assigns greater probabilities to words it finds in sitcom scripts. ChatGPT's eventual response described a scene, set at the Monk's Café, in which Jerry complains how hard it is to learn the bubble-sort algorithm. The AI even came up with a joke. In response to George's remark that "even a monkey" can learn the bubble-sort, Jerry replies: "Well, I'm not a monkey, I'm a comedian."[10]

Social machines

AI may not be the end of comedians or screenwriters. Yet the new generation of chatbots can do many things that previously required human social intelligence. They can analyse the language of negotiators, estate agents, and insurance brokers to identify words and phrases that are likely to be persuasive, leading to higher conversion rates. Meanwhile, face recognition systems are proving able to detect human emotions from facial expressions, just like voice assistants can recognize and respond to human speech patterns and tones. Thus, as we noted in our 2013 paper, the jobs of telemarketers are among those at highest risk of automation.

Machine intelligence, in other words, extends beyond text communication. It is already possible to generate deep fakes of particularly persuasive leaders, like Steven Jobs, to sell anything from iPhones to shaving cream. And if the Metaverse ends up materialising, it is easy to imagine online sales being supercharged, as the lonely consumer might find himself surrounded by avatar friends, nudging their human companion to buy various products in subtle ways. Just like you might be more enticed to buy a BMW if your neighbour gets one, such avatar "friendships" looks like the most plausible business model for the Metaverse. In this world, the human middleman has been automated. But even without the Metaverse, which may or may not bear fruit, this rings true: companies like Walmart are already using AI for social activities, such as negotiating prices with vendors.[11]

It is also true, however, that key bottlenecks to the automation of social tasks still persist. The simple reason is that in-person interactions remain valuable, and such interactions cannot be readily substituted for: LLMs don't have bodies. Indeed, in a world where AI excels in the virtual space, the art of performing in-person will be a particularly valuable skill across a host of managerial, professional and customer-facing occupations. People who can make their presence felt in a room, that have the capacity to forge relationships, to motivate, and to convince, are the people that will thrive in the age of AI. If AI writes your love letters, just like everybody else's, you better do well when you meet on the first date.

Consider the medical professions, for which persuasion is critical: according to recent research some doctors are much better than others at getting their patients to take life-saving

[10] For the Seinfeld example, see Newport, C. (2023). What Kind of Mind Does ChatGPT Have? The New Yorker, April 23.
[11] Sirtori-Cortina, D. and Case, B. (2023). Walmart Is Using AI to Negotiate the Best Price With Some Vendors, Businessweek, April, 26.

medicines.[12] And this aptitude is likely to be aided by the trust built through personal relationships. The venture capital industry provides another case in point: when the industry shifted to remote work during the pandemic, investors sought to make up for the loss of soft information by leveraging their existing networks and collaborating with partners with whom they had prior experience.[13] And the importance of human trust is only amplified by the workings of LLMs. A read of the Seinfeld script mentioned above, impressive as it is, illustrates this importance clearly. When Elaine orders a chicken salad from a passing waiter, for no discernible reason, it is met by "audience laughter."[14] ChatGPT has encoded what a sitcom script should sound like without a deep understanding of humour. It simply recombines existing human writing, which is fine-tuned using reinforcement learning from human feedback to reward talking like a human.

The result is not just spotty performance. LLMs, as we all know, are prone to hallucinations—fabricating content and even references—and have been revealed going off the rails. In the first video demo of Google's LLM, Bard, Bard incorrectly claimed that the James Webb Space Telescope "took the very first pictures of a planet outside of our own solar system"—an error that led to a dramatic drop in Google parent Alphabet's stock price. Perhaps even more concerningly, Microsoft's new AI-powered search engine—incorporating OpenAI's GPT-4—displayed a long list of discomforting behaviours, ranging from trying to persuade a *New York Times* reporter to end his marriage to declaring some users its "enemies."[15] Worse still, ChatGPT recently erroneously implicated a law professor in a sexual harassment case, seemingly due to misinterpreting statistical, but inconsequential, associations between fragments of text that were not genuinely related.

Many of these problems are unlikely to be solved simply through training even larger models—there are no quick fixes. Indeed, the upper bound of what current LLMs can do may not be too far from current models. For one thing, it is not clear that training sets can get any orders-of-magnitude larger, considering on how much data LLMs have already been trained. Neither is it obvious that significantly more compute than at present will be devoted to the training of LLMs. We have got used to Moore's law—the empirical law stating that the number of transistors in an integrated circuit (IC) doubles about every two years—but many expect this trend to run out of steam, due to physical limits, by about 2025. Training LLMs is also extraordinarily expensive (the cost of training GPT-4 was more than $100 million), and, with business models still unproven, it is not clear how fast many such investments in future will be made.

Regardless, in the near future, it seems unlikely that companies will want to leave the fate of longstanding consumer relationships in the hands of an AI that regularly hallucinates. Amazon, for example, has a dedicated (human) account manager for leading brands like Nestlé SA and Procter & Gamble Co, but uses AI to squeeze smaller contracts, that may not otherwise be worth the time.[16] As a rule of thumb, the more transactional a relationship becomes, the more prone it is to automation. Going forward, we expect many occupations that don't entail in-person communication—like telemarketers, travel agents, and call centre operators—to

[12] Simeonova, E., Skipper, N., & Thingholm, P. R. (2022). Physician health management skills and patient outcomes. Journal of Human Resources.
[13] Alekseeva, L., Dalla Fontana, S., Genc, C., & Rashidi Ranjbar, H. (2022). From in-person to online: the new shape of the VC industry. Working Paper.
[14] Newport, C. (2023). What Kind of Mind Does ChatGPT Have? The New Yorker, April 23.
[15] Roose, K. (2023). A Conversation With Bing's Chatbot Left Me Deeply Unsettled. New York Times, February 17.
[16] Sirtori-Cortina, D. and Case, B. (2023). Walmart Is Using AI to Negotiate the Best Price With Some Vendors, Businessweek, April, 26.

vanish. But without major leaps, longstanding relationships—benefiting from in-person interaction—will remain in the realm of humans.

Automating creativity

That AI is unlikely to be a complete human replacement in the foreseeable future is also true of creativity. Decades ago, there were already algorithms doing work we might call "creative." Beginning in the 1970s, AARON, a drawing program, was generating thousands of line-drawings, being exhibited in galleries around the world. And David Cope's EMI software was composing music in different styles already in the 2000s, making unfamiliar combinations of familiar ideas. Like EMI, today's Generative AI is essentially making unfamiliar combinations of existing ideas and works that appeal to human emotions. An AI-generated Drake and The Weeknd song, simulating both artists trading verses, recently went viral.[17] And this past November, just after OpenAI's release of ChatGPT, one software developer asked it for instructions for how to remove peanut-butter sandwich from a VCR in the style of the King James Bible. "And he cried out to the Lord, saying, 'Oh Lord, how can I remove this sandwich from my VCR, for it is stuck fast and will not budge?'", it responded, along with six other stunning paragraphs.[18]

But while reading the conversation might give you goosebumps, ChatGPT did not provide a very original idea to the challenge it was tasked with solving. Though the writing was impressive, it ultimately suggests sticking a knife between the sandwich and the VCR—something even a toddler would figure is not going to work very well. Indeed, unlike a child, which would rather pull the sandwich out, ChatGPT has no conceptual understanding of what it is talking about. What Generative AI systems do—with great success—is remixing and recombining music or text that is relevant to a given prompt. But instructing an algorithm to generate the voices of Drake and The Weeknd, for example, does not require astonishing creativity on the algorithm's part. And, just like a recombination of Mozart and Schubert won't generate music in the style of Arvo Pärt, prompting an AI to generate some recombination of impressionist paintings won't yield a bold leap into fresh conceptual art. Marcel Duchamp's Fountain—a porcelain urinal bought from a local plumbing supply store—provides a case in point. While we do not know how Duchamp came up with the idea, it certainly wasn't by analysing a dataset of impressionist paintings. Duchamp had seen porcelain urinals in the real world, and the new artform he invented placed them in a very different light.

A remaining bottleneck for creative AI is that, as long as algorithms do not interact in the real world, the data on which they have been trained will be limited, and so will their "experiences." There are few pictures of people taking pictures online, yet somebody must have taken them. Whether it takes a body to understand the world, as some scholars argue, is certainly contested, but the limits to book learning are known to all of us.[19]

More fundamentally, even if algorithms could experience the real world the way humans do, what sort of prompt would Duchamp have given to generate his Fountain? While combinations of existing styles might generate considerable commercial value, in, let's say, music, film, or interior design, it is also likely to lead us down the road of focusing on tweaking existing ideas rather than generating radical breakthroughs. Indeed, a recent crowdsourcing experiment, pitching humans against AI, found that the algorithm delivered solutions of potentially high

[17] Coscarelli, J. (2023). An A.I. Hit of Fake 'Drake' and 'The Weeknd' Rattles the Music World. New York Times, April 19.

[18] Newport, C. (2023). What Kind of Mind Does ChatGPT Have? The New Yorker, April 23.

[19] Glenberg, A. and Jones, C. (2023). It takes a body to understand the world – why ChatGPT and other language AIs don't know what they're saying. The Conversation, April 6.

financial value, but these were generally less novel than those provided by its human counterparts.[20] For breakthroughs, the desired output is simply much harder to define. It is no coincidence that AI does best in tasks where we know what we want to optimise for, like for the score in a video game. Yet if the goal is to generate something entirely new, for what do you optimise?

Consider the game of Go, where the reward function is relatively straightforward. Here AI was triumphant in 2016, defeating the World Champion Lee Sedol in a 5-game match four to one, generating some novel moves along the way. Sedol subsequently retired, stating that: "Even if I become the number one," he said, "there is an entity that cannot be defeated." But this year, humans made an astounding and unexpected comeback. As it turns out, deep-learning-driven AI does not understand all of the important concepts used by humans, such as the importance of groups of stones. And, by exploiting new tactics, to which the AI had not previously been exposed in training, a human amateur beat the AI convincingly, albeit with the help of a computer.[21]

What this means is that, today, we can never be quite sure if AI can be used reliably when novel circumstances emerge, such as for a change in tactics—an important component of human creativity. And so, incremental improvements through algorithmic tweaks, bigger data, and more parameters, seem unlikely to be game-changing for creativity. This, as we shall see, has far-reaching implications for the future of work, especially when algorithms interact with the physical world, which has stymied the driverless car industry.

Moravec's paradox

In 1988, Hans Moravec noted that "it's comparatively easy to make computers exhibit adult-level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception or mobility."[22] This challenge, in our view, remains pertinent today. The point is not that there hasn't been any progress in automating manual work, but that such progress has depended on the ability of humans to come up with clever ways of restructuring work to enable its automation. For example, we did not automate the jobs of medieval craftsmen by inventing robots capable of replicating their exact manual procedures. Ultimately, we subdivided craftsmen's work into repetitive tasks, in a more structured factory setting, and gradually automated those tasks, one at a time. Nor did we automate away the jobs of lamplighters by building robots capable of carrying ladders and climbing lampposts. Hence, attempts to assess whether a job is automatable by merely looking at the fraction of tasks that can be done by machines, as many economists have, will lead to flawed estimates: you will inevitably conclude that the work of lamplighters, farm labourers, elevator operators, car washers, switchboard operators, and truck drivers cannot be automated. Yet history has shown us that such occupations have indeed been automated.[23]

Predictably, so far, the deployment of autonomous vehicles has been confined to relatively structured environments, like harbours, mines, and warehouses. And as we argued in our 2013 paper, to what extent robots and autonomous vehicles will be adopted will continue to

[20] Boussioux, L., N Lane, J., Zhang, M., Jacimovic, V., & Lakhani, K. R. (2023). The Crowdless Future? How Generative AI Is Shaping the Future of Human Crowdsourcing. Working Paper.
[21] Waters, R. (2023). Man beats machine at Go in human victory over AI. Financial Times, February 17.
[22] Moravec, H. (1988). Mind children: The future of robot and human intelligence. Harvard University Press.
[23] Frey, C. B. (2019). The technology trap: Capital, labor, and power in the age of automation. Princeton University Press.

depend on the ingenuity of engineers in reconfiguring the environment in which the technology operates. Amazon Robotics' use of stickers to guide robots around warehouses provides one such example, as does the push towards prefabrication in construction.

True, recent progress in AI may also expand the domains of possible deployment by alleviating some concerns of perception. Indeed, for an algorithm to respond to the environment in which it is interacting, it needs some "understanding" of the objects it might encounter. How would a driverless car, for example, respond to a snowman standing in the middle of the road? Improvements in computer vision may be important in this regard—for instance, advances in Neural Radiance Fields (NeRF) might make it easier to simulate 3D scenes, producing synthetic data, so as to train autonomous vehicles more efficiently, in simulation.[24] But at a same time, this approach is no panacea: synthetic data is unavoidably going to be a product of the NeRF's own data and implicit assumptions—only if those data and assumptions are valid can the synthetic data be useful. If the NeRF's assumptions and data omit some important real-world consideration, so too will its synthetic data.

LLMs today are widely celebrated, while the driverless industry is often ridiculed for failing to live up to its early promises. Yet autonomous vehicles have also seen considerable recent progress, as evidenced by numerous robotaxi trials, from San Francisco to Shenzhen. Besides the amount of training data available, a crucial difference, relative to LLMs, is that we are much more risk averse when algorithms are brought into the physical world in general, and into public spaces in particular. As noted, LLMs are prone to hallucination. However, the consequences of ChatGPT making up the references for an essay seem minor when pitched against the potential devastating consequences of a driverless car hallucinating in traffic. While fake text and images can likely be edited or deleted, fatal traffic accidents cannot be reversed.

This highlights a broader point: AI—in its current form—is less likely to be deployed in higher-stakes contexts, like driving, than in lower-stakes activities, like customer service or warehouses. A key bottleneck to the automation of perception and mobility tasks, in other words, is that we cannot accept mistakes. And foundation models based on deep neural networks, whose decisions we cannot explain, have the potential to creating plenty of mistakes. For widespread deployment in physical spaces, we will need robust, reliable, and explainable AI.[25] For now, jobs that centre on complex perception and manipulation tasks remain relatively safe from automation, as we deemed that they were in 2013.

The future of work

The physicist Niels Bohr supposedly once joked that "God gave the easy problems to the physicists."[26] While the laws of physics are time invariant, and apply across time and space, boundary conditions in social sciences are not timeless. The same is true of engineering, which has steadily expanded our means of automating work into previously inconceivable domains, with new and unpredictable implications for workers and society more broadly. As noted, significant bottlenecks to automation remain, but it is also clear that there are jobs and tasks that algorithms can do now that go well beyond what we observed in our paper a decade ago.

[24] Barron, J. T., Mildenhall, B., Verbin, D., Srinivasan, P. P., & Hedman, P. (2023). Zip-NeRF: Anti-aliased grid-based neural radiance fields. arXiv preprint arXiv:2304.06706.
[25] Marcus, G. (2020). The next decade in AI: four steps towards robust artificial intelligence. arXiv preprint arXiv:2002.06177.
[26] Frey, C. B. (2019). The technology trap: Capital, labor, and power in the age of automation. Princeton University Press.

Consider, first, tasks requiring social intelligence, which we deemed non-automatable in 2013. As a general rule, it now looks like AI may be able to replace human labour in many virtual settings, meaning that if a task can be done remotely, it can also be potentially automated. The trouble is that Generative AI remains prone to hallucination, posing a risk to the reputation of the companies deploying it. Given this risk, we expect that firms will primarily use AI for transactional activities, which do not build on creating long standing customer relationships, and for which in-person interactions remain important to establish trust.

Second, Generative AI has a role to play in creative work. But it is best suited for creating sequels rather than new narratives. It might write another Batman plot (though, without human input, that plot will likely be dull and full of holes), but not The Seventh Seal. AI is good at generating new combinations of existing ideas, rather than making conceptual leaps. So the deployment of Generative AI, in our view, will centre on extending existing product lines rather than independently creating entirely new lines of business.

Finally, when it comes to perception tasks, over the coming years, automation is likely to continue to focus on structured environments. The reason is simple: in high-stakes contexts, like automated delivery services, the number of rare events an AI might encounter, which are unlikely to be in the training data, are simply too large. For now, deployment will be confined to lower-stakes activities, like customer service—e.g. AmazonGo—or warehouse automation, where engineers can redesign and simplify the environment to enable automation.

Over the past decade, in short, the potential scope of automation has expanded in that many virtual social interactions can now be automated. The same is true of creative tasks that centre on recombining existing ideas. Additionally, advancements in computer vision have paved the way for automating more perception tasks. However, despite these advancements, critical obstacles still hinder the application of automation in high-stakes environments.

How labour markets will adjust to these developments is naturally on everybody's minds. Some jobs, like those of telemarketers, forklift drivers, and copy editors, seem likely to be automated away. But this does not necessarily mean fewer jobs. The automation of copy editing, for example, might make books cheaper, creating more jobs elsewhere in the publishing industry. Similarly, cheaper marketing could boost sales across a host of industries to the benefit of workers elsewhere in the economy. Not to mention the entirely new jobs that might emerge in response. Who had heard of the job title “prompt engineer” before 2022? Indeed, some might take solace in the fact that most jobs done by Americans today did not even exist in 1940; they had to be invented.[27]

AI can aid the process of scientific discovery, like in protein folding, potentially leading to the creation of new tasks and even new industries.[28] But there is also no economic law that postulates that it will. In fact, as we have noted elsewhere, along with other scholars, much of

[27] Autor, D., Chin, C., Salomons, A. M., & Seegmiller, B. (2022). New Frontiers: The Origins and Content of New Work, 1940–2018 (No. w30389). National Bureau of Economic Research; Berger, T., & Frey, C. B. (2016). Did the Computer Revolution shift the fortunes of US cities? Technology shocks and the geography of new jobs. Regional Science and Urban Economics, 57, 38-45; Berger, T., & Frey, C. B. (2017). Industrial renewal in the 21st century: evidence from US cities. Regional Studies, 51(3), 404-413.

[28] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589.

the history of technology and work can be seen as a race between technologies that create new types of work and automation technologies that replace existing ones.[29]

The immediate effect of the most recent wave of Generative AI, though, will not be either automation or new industries but the transformation of existing content-creating jobs, making them easier to perform. As we argued in 2013, AI has considerably expanded the potential scope of automation. But thinking of the most recent wave of Generative AI—which, it must be noted, is merely a subfield of AI—as an *automation* technology is, in its current form, a mistake. For one thing, it requires humans to make a prompt and then select (as well as mostly edit) the desired output, and this prompting and selection is where much of the actual creativity resides.

A more apt way of thinking about Generative AI is analogous to Uber and its impact on taxi services. With the advent of GPS technology, knowing the name of each street in New York City was no longer a valuable skill. So when Uber rolled out across the United States, average drivers with little knowledge of the cities in which they operated took full advantage. The result was not fewer jobs, but more intense competition, which reduced the incomes of incumbent drivers. In joint work with Lund University's Thor Berger, our lab found that drivers hourly earnings fell by 10% when Uber entered a given city.[30]

Might LLMs prove a GPS for language? It is worth reiterating that LLMs consider the probability that a human would have used a word, reassessing this probability through feedback from users. As noted, today's LLMs are incredibly data-hungry, and given that they need to be trained on large parts of the internet, rather than on relatively scarce material written by experts, LLMs tend to converge towards average human performance. Thus, there is a trade-off between algorithms learning from vast datasets (embodying likely average expertise), and the ability to capture the expensive knowledge and aptitudes of top talent. In the absence of breakthroughs that allow algorithms to learn from much smaller curated data, however, investment will continue to flow towards algorithms built for average human creativity. Like peer-review, these AIs aim for consensus rather than for novelty by design. Put differently, LLMs compete with average human talent rather than with superstars.

AI's average aptitudes, in turn, have implications for labour markets. According to recent research, software developers gaining access to Githubs Copilot completed the task 56% faster than the control group and developers with less programming experience exhibited the most substantial gains.[31] Similarly, ChatGPT has been shown to elevate the productivity of writers, particularly those with lower abilities, in composing tasks.[32] And among customer service agents gaining access to an AI-assistant, productivity increased by 14%, again with novices and low-skilled workers benefiting disproportionally.[33] What this means is that many more people can do the job adequately. Just like Uber reduced barriers to entry in taxi services, many more people will engage in creative work. ChatGPT won't replace journalists, just like Github's Copilot won't replace coders. But they are making these tasks easier for

[29] Frey, C. B. (2019). The technology trap: Capital, labor, and power in the age of automation. Princeton University Press. See also Acemoglu, D., & Restrepo, P. (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. American economic review, 108(6), 1488-1542.

[30] Berger, T., Chen, C., & Frey, C. B. (2018). Drivers of disruption? Estimating the Uber effect. European Economic Review, 1(10), 197-210.

[31] Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). The impact of ai on developer productivity: Evidence from github copilot. arXiv preprint arXiv:2302.06590.

[32] Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. Available at SSRN 4375283.

[33] Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). Generative AI at work (No. w31161). National Bureau of Economic Research.

novices, inducing more competition. Generative AI, in other words, will help average writers, designers, and advertising execs undercut their more skilled competitors.

The question that emerges, of course, is how much more content will people consume as Generative AI makes it cheaper to produce? This, in our view, is somewhat akin to asking: how much more time would you spend on Netflix if it was cheaper and the content was better? The answer is probably not that much—the length of your day is still limited. Extreme content abundance will be competing from limited human time and attention spans. You are probably more likely to substitute away for some content towards better content. Consequently, many incumbent content creators are likely to see mounting pressure on their wages, while many novices moving in from other less well-paid jobs will add to their income.

Thus, Generative AI, despite not causing widespread job displacement, and even benefiting many workers, will bring significant disruptions to labour markets and possibly even lead to social upheaval. We can recall the protests of taxi drivers blocking the streets in London when Uber was introduced, or the French drivers resorting to extreme measures like overturning cars and setting fire to tires in resistance. These protests managed to impede the technology's adoption in some regions, including Germany.

Moreover, the white-collar workers currently feeling the pressure are more politically influential than their blue-collar counterparts who have already experienced decades of technological disruption with the entry of robots into factories. A striking example of this is the joint strike by Hollywood screenwriters and actors against the use of Generative AI, which resulted in the shutdown of TV and film production—the industry's first collective strike in over sixty years. Similar to other white-collar workers, they are better positioned to resist technologies that threaten their livelihoods, setting the stage for potential conflicts that may slow down the widespread adoption of Generative AI.

The future of AI

Task simplification, of course, could merely be a stepping stone towards total automation. Consider the case of the lamplighters. Before the dawn of electricity, lamplighters lit the streets of America's towns and cities, carrying torches and ladders to ignite the gas lamps at night. When electric street lights first arrived, it simply made the job simpler. Each lamp had its own switch, which needed to be turned on and off manually. Much like what we have seen with Generative AI, this made the job so easy that lamplighters soon faced more competition. Even children could easily switch the lights on and off during their daily commutes to school. But it was not long until streetlights were controlled from substations and the demand for lamplighters dramatically dwindled.[34]

Could we expect a substation moment for Generative AI in the near future? Answering this question is inevitably a speculative endeavour, but in our view, this is not likely to happen in the near future, as it will need new technological breakthroughs. As mentioned earlier, the data consumed by LLMs has already been substantial, and it is not feasible to dramatically increase training sets by many orders of magnitude. Additionally, there are valid reasons to anticipate that the Web will become inundated with low-quality AI-generated content, rendering the Web a progressively poorer source for training data. In fact, there are indications that, in recent times, the content from which algorithms learn has been displaying greater monotony. For instance, in the realm of music, since the advent of computers, average creativity appears to have declined, evident in reduced key changes over the decades.

[34] Frey, C. B. (2019). The technology trap: Capital, labor, and power in the age of automation. Princeton University Press.

Likewise, human writing appears to be more rule-based, formulaic, and mechanistic, leading to less diverse input for AI algorithms to learn from.[35]

It is true that there are ways to create new data, using NeRFs for simulations, as discussed, or simply by creating synthetic data, like text or code.[36] For instance, in developing AlphaFold—a system that excelled at predicting the 3D structure of proteins, even outperforming human researchers in competitions like CASP (Critical Assessment of Structure Prediction)—DeepMind incorporated some of the model's own forecasts into the training data, scaling up the dataset. But ultimately this depended on having a very a large dataset of known protein structures from publicly available sources in the first place, such as the Protein Data Bank (PDB). Without such data, there are currently few workarounds. And it is important to remember that AlphaFold was purpose-built for one particular task—it is not a general purpose technology. With regard to LLMs, research from Oxford and Cambridge has indicated that synthetic data can trigger irreversible damages, resulting in model failure.[37]

It is also true that fine-tuning and reinforcement learning from human feedback (RLHF) can further improve the Generative AI's ability, as the model adjusts its output to human responses and so learns over time. But RLHF turns out to be a labour-intensive task. A recent investigation by TIME revealed that OpenAI delegated some of this work to Kenyan workers earning less than $2 an hour.[38] There's even some indication that the effectiveness of LLMs has decreased in recent months. And one interpretation is that interactions with users have made these systems worse, implying that RLHF, in its current form, has hit a wall.[39] Meanwhile, other studies show that the rate of human-like fallacious judgments rose from 18% in GPT-3 to 33% in GPT-3.5, and further to 34% in GPT-4, even as it got better at correct human-like judgements. This observation indicates that larger and more sophisticated LLMs may display a tendency towards making mistakes similar to those made by humans.[40]

That said, we do expect near-term improvements from fine-tuning, as businesses start to leverage foundational models like GPT-4, utilizing more specialized datasets for specific tasks. For example, companies training a customer service bot will have data from genuine customer inquiries, offering examples of effective responses, just as pharmaceutical companies will have data to enable fine-tuning towards drug discovery. This approach offers a cost-effective method for tailoring a pretrained model for a specific use. However, fine-tuning doesn't address many of the fundamental issues of AI that we have underscored.

When we published our paper in 2013, the AI field was relatively diverse, featuring a wide array of methods. However, since then, the focus has shifted to methods that demand extensive computational power and data, such as deep learning. This has undoubtedly resulted in tangible progress, but, in our view, this narrow focus is likely to encounter diminishing returns. For one thing, Generative AI still tends to generate erroneous or fantastical outputs, and, without further innovation, the problem of hallucinations will persist.

---

[35] Leslie, I. (2022). The Struggle To Be Human. https://www.ian-leslie.com/p/the-struggle-to-be-human

[36] For examples of synthetic data, see Murgia, M. (2023). Why computer-made data is being used to train AI models. Financial Times, July 19.

[37] Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2023). The Curse of Recursion: Training on Generated Data Makes Models Forget. arXiv preprint arxiv:2305.17493.

[38] Perrigo, B. (2023). Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. TIME Magazine, January 18.

[39] Chen, L., Zaharia, M., and Zou, J. (2023). How Is ChatGPT's Behavior Changing over Time? arXiv:2307.09009v1.

[40] Koralus, P., & Wang-Maścianica, V. (2023). Humans in humans out: On GPT converging toward common sense in both success and failure. arXiv preprint arXiv:2303.17276.

Thus, beyond the advances outlined above, the potential scope of automation is unlikely to substantially grow merely through scaling existing models.[41]

In conclusion, while we expect AI to continue to surprise us, and for many jobs to be automated away, in the absence of major breakthroughs, we also expect the bottlenecks we outlined in our 2013 paper to continue to constrain our automation possibilities for the foreseeable future.

[41] Hallucinations are potentially fixable when there are clear benchmarks of truth. For example, does an LLM-generated reference actually exist? The algorithm can just search the web for it. But in most instances, such straight-forward benchmarks do not exist.