Generative AI and the Future of Work: A Reappraisal

Carl Benedikt Frey and Michael Osborne

Introduction

Like now, the early 2010s was a time of excitement about new technological wonders. The world had bounced back from the worst financial crisis since the Great Depression and, despite residual financial malaise, America seemed to be on the cusp of a productivity boom.[1] Artificial intelligence (AI), long regarded as an academic backwater, was finally bearing fruit. In 2011, a question-answering computer system known as IBM Watson beat the world champion in Jeopardy. Machine learning—a subfield of AI—further expanded the possibilities 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 recognizing cat videos on YouTube.

Gone were the days when a computer programmer needed to write explicit rules to guide the actions of a machine in every contingency. Computers could now infer rules by tapping into the data trails left behind by increasing numbers of humans online. This phenomenon was most clearly visible with the development of autonomous vehicles. No programmer could be expected to foresee every situation a human driver might encounter in city traffic, and even less to capture these in a sequence of "If-Then-Do" commands. Instead, autonomous vehicles made progress by collecting vast amounts of data on drivers' actions in city traffic to predict what humans would in any given situation.

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. 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.

In 2013, we published a working paper estimating that 47 percent of jobs were exposed to the recent advances in machine learning and advanced robotics.[2] Ian Goodfellow and co-authors' paper on "Generative Adversarial Networks" had not yet been published. Still, our estimates provided glimpses of the age of Generative AI: fashion modeling, we found, was among the jobs at risk. Just a few years later, digital fashion models were being generated en masse.[3] Overall, however, we firmly believed that—in the absence of major unforeseen leaps—tasks requiring creativity, social intelligence, and unstructured manual work would remain safe havens for human workers. The jobs of journalists, scientists, software engineers, art directors, and architects were all at low risk of becoming automated.

Fast-forward a decade and Large Language Models (LLMs) like GPT-4 can now answer questions and write essays in astonishingly human-like fashion. Image generators like DALL-E 2 can transform text prompts into new images, mirroring the creative work of designers and 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 breathtaking. OpenAI, an American artificial intelligence research laboratory, reports that their LLM passed a simulated bar exam, SAT Math section, and introductory sommelier training with a score around the top 10% of test takers.[4] In just a few months, as the world moved from GPT-3.5 to GPT-4, AI leaped from the 39th to the 96th percentile of human-level performance in solving college physics problems.[5] While such test results may reflect LLMs "parroting" answers to common exam questions that are found in their training sets, they do beg the question: did we underestimate the near-term scope of automation?[6]

Below, we provide a reassessment of the division of labor between humans and computers in the age of AI. In doing so, we explore whether generative AI has 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 labor market implications of recent trends in technology.

The Rise of LLMs

Ever since MIT's Joseph Weizenbaum launched ELIZA in 1966, computer scientists have been trying to build social machines.[7] 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 told about the betrayal of a friend, it would respond, "Why do you feel betrayed?" Indeed, ELIZA would never have succeeded at the imitation game, a test devised by Alan Turing in 1950 and widely regarded as a benchmark for machine intelligence. If an algorithm could convince a human interlocutor that they were talking to another person, surely it must be understanding something? Based on this principle, Turing test competitions, in which judges are tasked with distinguishing between human and algorithm interlocuters whose identities are unknown, became a common standard for measuring progress in AI.

Yet, half a century later, chatbots remained underwhelming. True, in 2012, on what would have been Alan Turing's 100th birthday, a bot called Eugene Goostman managed to convince 33 percent of human judges that it was human. But the bot succeeded by pretending to be a Ukrainian boy with elementary language skills and knowledge of English culture. Rather than constituting a leap in AI, the incident highlighted the flaws of the Turing test—Goostman was good at feigning intelligence, nothing more. Such limited progress, even in basic textual 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, an extensive database of human-generated jokes and methods of benchmarking the algorithm's performance sufficed, in principle.

We now have both sufficient databases and benchmarking methods. Trained on vast amounts of data from books, articles, and websites that would take a thousand human lifetimes to read, today's LLMs learn patterns and relationships between words and phrases, allowing them to predict the next word in a sentence based on the context. This advancement, paired with innovative approaches to assessing AI output using reinforcement learning—a technique where an agent learns by interacting with an environment, receiving feedback in the form of rewards, and adjusting its actions to maximize those rewards—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 on its training—an immense body of human text that likely included 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 about 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."[8]

Social Machines

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

Machine intelligence extends far beyond text-based communication. Generating deepfakes of particularly persuasive leaders like Steve Jobs to sell anything from iPhones to shaving cream is already possible. Suppose the Metaverse ever materializes. It is easy to imagine supercharged online sales, as the lonely human consumer is surrounded by avatar friends constantly nudging them to buy products. Just like you might be more enticed to buy a BMW if your neighbor gets one, such avatar "friendships" appear the most plausible business model for the Metaverse. In this world, the human middleman will be automated. Even outside the Metaverse, the underlying concept rings true: companies like Walmart are already using AI for social business activities like negotiating prices with vendors.[9]

Just like you might be more enticed to buy a BMW if your neighbor gets one, such avatar "friendships" appear the most plausible business model for the Metaverse.

However, key bottlenecks to the automation of social tasks persist. In-person interactions remain valuable and cannot be readily substituted since LLMs do not have bodies. Indeed, in a world where AI excels in the virtual space, in-person communication will increasingly be a particularly valuable skill across various managerial, professional, and customer-facing occupations. People capable of establishing a strong physical presence and forging face-to-face relationships in which they motivate and convince others will thrive in the age of AI. If your AI-written love letters read just like everybody else's, you had better do well on the first date.

Consider medical professions, for which persuasion is often critical. According to recent research, some doctors are much better than others at convincing their patients to take life-saving medicines.[10] This aptitude is likely aided by the trust built through personal relationships. The venture capital industry is similarly affected by human interaction. When the industry shifted to remote work during the pandemic, investors sought to make up for the loss of information typically shared via in-person meetings by leveraging their existing networks and collaborating with partners with whom they had prior working experience.[11] The importance of human trust is only amplified by the performance of LLMs, illustrated clearly by the AI-generated Seinfeld script: when Elaine orders a chicken salad from a passing waiter, "audience laughter" follows for no discernible reason.[12] ChatGPT has encoded what a sitcom script should sound like without a deep understanding of humor. It simply reconfigures and fine-tunes existing human writing using reinforcement learning from human feedback to reward talking like a human.

The result from this approach is not just spotty performance. LLMs are prone to hallucinations—fabricating content and even references—and have even been regarded as "going off the rails." Google's LLM, Bard, incorrectly claimed that the James Webb Space Telescope "took the very first pictures of a planet outside of our own solar system" in its first video demo—an error that led to a dramatic drop in the stock price of Google's parent company, Alphabet.[13] Perhaps even more concerning, Microsoft's new AI-powered search engine—incorporating OpenAI's GPT-4—displayed a long list of alarming behaviors, from trying to persuade a *New York Times* reporter to end his marriage to declaring some users its "enemies."[14] Worse still, ChatGPT erroneously implicated a law professor in a sexual harassment case, seemingly due to a misinterpretation of statistical but inconsequential associations between unrelated fragments of text.

Many of these problems are unlikely to be solved simply by training even larger models—there are no quick fixes. However, the upper bound of LLMs' capabilities may not be too far from those of current models. For one thing, it is unclear that training sets can grow any orders of magnitude larger, considering the amount of data upon which LLMs have already been trained. Nor is it obvious that significantly more computers than at present will be devoted to training LLMs. We have become accustomed to Moore's law—that the number of transistors in an integrated circuit (IC) doubles about every two years—but many expect this trend to expire due to physical limits by about 2025. Training LLMs is also extraordinarily expensive (training GPT-4 cost more than $100 million), and, with unproven business models, it is uncertain how many companies will be eager to make similar investments going forward.

Regardless, it seems unlikely that, in the near future, companies will want to entrust longstanding consumer relationships to regularly hallucinating AI. Amazon, for example, has a dedicated human account manager for leading brands like Nestlé SA and Procter & Gamble Co., but uses AI to manage smaller contracts that may not otherwise be worth the time.[15] As a rule of thumb, the more transactional a relationship becomes, the more prone it is to automation. Looking ahead, we expect many occupations that do not involve in-person communication—like telemarketers, travel agents, and call center operators—to vanish. But, without significant leaps, longstanding relationships that benefit from in-person interaction will remain in the realm of humans.

Automating Creativity

For situations requiring creativity, AI is also unlikely to be a complete replacement for human communication in the foreseeable future. Decades ago, algorithms already existed that produced work we might call "creative." Beginning in the 1970s, the drawing program AARON generated thousands of line drawings later exhibited in galleries around the world. In the 2000s, David Cope's EMI software was already composing music in different styles, making unfamiliar combinations of familiar ideas. Like EMI, today's generative AI essentially combines existing ideas and works that appeal to human emotions in unfamiliar ways. An AI-generated song simulating Drake and The Weeknd trading verses recently went viral.[16] This past November, one software developer asked OpenAI's newly released ChatGPT for instructions, written in the style of the King James Bible, for removing a peanut-butter sandwich from a VCR. The LLM chatbot responded with, "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?'" along with six more stunning paragraphs.[17]

Though the impressive writing might give you goosebumps, ChatGPT did not provide original ideas to the challenges it was tasked with solving. It ultimately suggests sticking a knife between the sandwich and the VCR—a solution even a toddler, who would likely pull the sandwich out by hand, would find flawed. Evidently, ChatGPT has no conceptual understanding of the writing it produces.

What generative AI systems do—with great success—is remix and recombine relevant music or text to a given prompt. But instructing an algorithm to generate the voices of Drake and The Weeknd does not require astonishing creativity. A recombination of Mozart and Schubert will not generate music in the style of Arvo Pärt; likewise, prompting an AI to generate some recombination of impressionist paintings will not yield a bold leap into fresh conceptual art. For example, while we do not know how Marcel Duchamp came up with the idea for *Fountain*—a porcelain urinal bought from a local plumbing supply store and displayed as sculpture—it was certainly not by analyzing a dataset of impressionist paintings. Duchamp had seen porcelain urinals in the real world, and the art form he invented intentionally placed them in a very different light.

As long as algorithms do not interact in the real world, the data on which they have been trained will be limited in comparison to human experiences. For example, although people constantly take pictures, few images exist of people taking pictures online. Whether it takes a body to understand the world, as some scholars argue, is certainly contested, but the limits to learning from a book are known to all of us.[18]

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 unique combinations of preexisting styles might generate considerable commercial value in music, film, or interior design, they will likely lead us to focus on tweaking existing ideas instead of generating radical breakthroughs. Indeed, a recent crowdsourcing experiment pitching humans against AI found that while the algorithm delivered solutions of potentially high financial value, these solutions were generally less novel than those provided by its human counterparts.[19] For breakthroughs, the desired output is much harder to define. It is no coincidence that AI performs best in tasks with a known optimization goal, like the score in a video game. Yet, if the goal is to generate something entirely new, for what do you optimize?

Consider the strategy board game *Go*, where the reward function is relatively straightforward. Here, AI was triumphant in 2016, defeating the *Go* World Champion Lee Sedol four to one in a five-game match and generating some novel moves along the way. Sedol subsequently retired, stating, "Even if I become the number one…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 concepts used by humans, such as the importance of groups of stones in *Go*. 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.[20]

This means that today we can never be quite sure whether AI can be used reliably when novel circumstances, such as a change in tactics, emerge—an important component of human creativity. Thus, incremental improvements through algorithmic tweaks, more extensive data sets, and more parameters seem unlikely to be game-changing for creativity. This realization has far-reaching implications for the future of work, especially when algorithms interact with the physical world, which, as we shall see, 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."[21] This challenge remains pertinent today. The issue is not a lack of progress in automating manual work, but rather that such progress has depended on humans' ability to conceive of 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, in a structured factory setting, we subdivided craftsmen's work into repetitive tasks and gradually automated them 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 merely by evaluating the fraction of tasks that machines can do, as many economists have, will lead to flawed estimates: you will inevitably conclude that the work of lamplighters, farm laborers, elevator operators, car washers, switchboard operators, and truck drivers cannot be automated. Yet history has shown us otherwise.[22]

Predictably, the deployment of autonomous vehicles has been confined to relatively structured environments, such as harbors, mines, and warehouses. As we argued in our 2013 paper, the extent to which robots and autonomous vehicles are adopted will continue to 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 toward prefabrication in construction, whereby parts of a structure are built in a factory rather than onsite. True, recent progress in AI may also expand the domains of possible autonomous vehicle deployment by alleviating some perception concerns. For an algorithm to appropriately respond to the environment with which it interacts, 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 facilitate easier simulation of three-dimensional scenes by producing synthetic—artificially generated—data to train autonomous vehicles more efficiently.[23] But this approach is no panacea: synthetic data will inevitably be a product of the NeRF's own data and implicit assumptions, which must be valid for the synthetic data to be useful. If the NeRF's assumptions and data omit some important real-world considerations, so too will its synthetic data.

While LLMs today are widely celebrated, 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 robo-taxi trials in cities ranging from San Francisco to Shenzhen. Besides the amount of training data available, a crucial difference between the driverless industry and LLMs is that people are much more risk-averse when algorithms are brought into the physical world in general and, in particular, public spaces. 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 potentially 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. The affordability of mistakes represents a key bottleneck to the automation of tasks that are reliant on perception and mobility. Foundation models based on deep neural networks making decisions we cannot explain have the potential to create plenty of mistakes. For widespread deployment in physical spaces, we will need robust, reliable, and explainable AI.[24] Thus, jobs that center on complex perception and manipulation tasks remain relatively safe from automation—the same conclusion we came to in 2013.

The affordability of mistakes represents a key bottleneck to the automation of tasks that are reliant on perception and mobility.

The physicist Niels Bohr supposedly once joked that "God gave the easy problems to the physicists."[25] 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. Significant barriers to continued automation remain, but it is also clear that algorithms can now do jobs and tasks extending well beyond what we observed in our paper a decade ago.

Consider tasks requiring social intelligence which we deemed non-automatable in 2013. AI may now be able to replace human labor in many virtual settings, meaning that if a task can be done remotely, it can also be automated. The trouble is that generative AI remains prone to hallucination, posing a risk to the reputations of the companies deploying it. Given this risk, we expect that firms will primarily use AI for transactional activities that do not create longstanding customer relationships, while in-person interactions will remain important to establish trust.

Generative AI may also play a role in creative work, but is best suited for creating sequels rather than new narratives. It might write another Batman plot (without human input, that plot will likely be dull and full of holes), but it will not come up with *The Seventh Seal* from scratch. AI is good at generating new combinations of existing ideas rather than making conceptual leaps. So, the deployment of Generative AI will center on extending existing product lines rather than on independently creating entirely new lines of business.

Finally, regarding perception tasks, automation will likely continue to focus on structured environments, where engineers can redesign and simplify the environment to enable automation. The reason is simple: in high-stakes contexts like automated delivery services, the number of rare events that an AI might encounter (which are unlikely to be included in the training data) are simply too large. For now, deployment will be confined to lower-stakes activities, like customer service—Amazon Go, for example—or warehouse automation.

In short, over the past decade, the potential scope of automation has expanded to include many virtual social interactions. The same is true of creative tasks that center on recombining existing ideas. Additionally, advancements in computer vision have paved the way for automating more perception tasks. Despite these advancements, however, critical obstacles still hinder the application of automation in high-stakes environments.

How labor markets will adjust to these developments is naturally on everyone's mind. Some jobs, like telemarketing, forklift driving, and copy editing, seem likely to be automated away. But this automation will not necessarily result in fewer jobs. For example, the automation of copy editing might make books cheaper, creating more jobs elsewhere in the publishing industry. Similarly, more affordable marketing could boost sales across a host of industries to benefit workers elsewhere in the economy, not to mention generate entirely new jobs; who had heard of the job title "prompt engineer" before 2022? Indeed, some might take solace in the fact that most jobs held by Americans today did not even exist in 1940; they had to be invented.[26] AI can aid the process of scientific discovery, like in studying protein folding, potentially leading to the creation of new tasks and even new industries.[27] But there is also no economic law that postulates it will. As we and other scholars have noted elsewhere, much of 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.[28]

The immediate effect of the most recent wave of generative AI will be neither growth in automation nor the emergence of new industries, but the decrease in difficulty of existing content-creating jobs. 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, generative AI requires humans to make a prompt and then select (as well as mostly edit) the desired output. This prompting and selection is where much of the actual creativity resides. It is more apt to think about generative AI as 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.[29]

Might LLMs prove to be 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 user feedback. 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 toward average human performance. Thus, a trade-off exists between algorithms learning from vast datasets (likely embodying average expertise) and their ability to capture top talent's expensive knowledge and aptitudes. In the absence of breakthroughs that allow algorithms to learn from much smaller, curated datasets, investment will continue to flow toward 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 labor markets. According to recent research, software developers gaining access to GitHub's Copilot completed the task 56 percent faster than the control group, and developers with less programming experience exhibited the most substantial gains.[30] Similarly, ChatGPT has been shown to elevate the productivity of writers, particularly those with lower abilities, in completing tasks.[31] Among customer service agents gaining access to an AI-assistant, productivity increased by 14 percent, again with novices and low-skilled workers benefiting disproportionally more.[32] This means 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 will not replace journalists, just like GitHub's Copilot will not replace coders. But they are making these tasks easier for novices, inducing more competition. Generative AI, in other words, will help average writers, designers, and advertising executives undercut their more skilled competitors.

The question that emerges is how much more content will people consume as Generative AI makes it cheaper to produce? In our view, this 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 a day is still limited. Extreme content abundance will be competing against limited human time and attention spans. Instead, people are more likely to substitute this higher-quality content for general content. Consequently, many incumbent content creators will likely see mounting pressure on their wages, while many novices moving in from different, lower-paying jobs will elevate their income.

Thus, generative AI, despite benefiting many workers and not causing widespread job displacement, will significantly disrupt the labor market.

Thus, generative AI, despite benefiting many workers and not causing widespread job displacement, will significantly disrupt the labor market. This upheaval will likely manifest in social unrest. Recall the protests of taxi drivers blocking the streets in London when Uber was introduced, or French drivers resorting to extreme measures like overturning cars and setting fire to tires in resistance. These protests impeded the technology's adoption in some regions, including Germany. Moreover, the white-collar workers feeling the pressure of AI are more politically influential than their blue-collar counterparts who have already experienced decades of technological disruption with the introduction of robots in factories. A powerful 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. Like other white-collar workers, actors and screenwriters 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 could merely be a stepping stone toward total automation. Again, consider the lamplighters. Before the dawn of electricity, lamplighters lit the streets of America's towns and cities, carrying torches and ladders to ignite gas lamps at night. Initially, the arrival of electric streetlights simply made the job simpler. Each lamp had its own switch, which needed to be manually turned on and off. Much like the effects of generative AI, electric streetlights made lamplighting 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.[33]

Can we expect a substation moment for generative AI? Answering this question is inevitably a speculative endeavor, but in our view, this is unlikely to happen soon. Such a change will require new technological breakthroughs. As mentioned earlier, the data consumed by LLMs is already 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 Internet will become inundated with low-quality, AI-generated content, rendering the Web an increasingly inadequate source for training data. In fact, there are recent indications that the content from which algorithms learn has been displaying greater monotony. For instance, in the realm of music, average creativity appears to have declined since the advent of computers, evident in reduced key changes over the decades. Likewise, human writing appears to be more rule-based, formulaic, and mechanical, leading to less diverse input from which AI algorithms may learn.[34]

There are ways to create new data, like using NeRFs for simulations as discussed earlier, or simply by creating synthetic data, like text or code.[35] For instance, in developing AlphaFold—a system that excelled at predicting the three-dimensional structure of proteins and even outperformed 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 huge dataset of known protein structures from publicly available sources such as the Protein Data Bank (PDB) in the first place.[36] Without existing data, there are currently few workarounds. Moreover, it is important to remember that AlphaFold was narrowly built for one particular task and is not a general-purpose technology. Regarding LLMs, research from Oxford and Cambridge has indicated that synthetic data can trigger irreversible damages, resulting in model failure.[37]

It is true that fine-tuning and reinforcement learning from human feedback (RLHF) can further improve generative AI's ability. The model adjusts its output to human responses and ultimately learns over time. But RLHF turns out to be a labor-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 is even some indication that the effectiveness of LLMs has decreased in recent months. 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 fallacious human-like judgments rose from 18 percent in GPT-3 to 33 percent in GPT-3.5, and further to 34 percent in GPT-4, even as it improved at making correct human-like judgments. This observation indicates that larger and more sophisticated LLMs may display a tendency toward making mistakes similar to those made by humans.[40]

That said, we 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 toward drug discovery. This approach provides a cost-effective method for tailoring a pre-trained model for a specific use. However, this fine-tuning still does not address many of the fundamental issues of AI that we have outlined.

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 narrow focus has undoubtedly resulted in tangible progress but, in our view, 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. Thus, beyond the advances outlined above, the potential scope of automation is unlikely to grow substantially merely through scaling existing models.[41]

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

Notes

[1] This was also true of the Great Depression. See Alexander J. Field, "The Most Technologically Progressive Decade of the Century," *American Economic Review* 93, no. 4 (2003): 1399-1413.

[2] Carl B. Frey and Michael A. Osborne, "The Future of Employment: How Susceptible Are Jobs to Computerisation?" *Technological Forecasting and Social Change* 114, (2017): 254-280. The first working paper version was published in 2013.

[3] Ian Goodfellow et al., "Generative Adversarial Networks," *Communications of the ACM* 63, no. 11 (2020): 139-144. This was published as a working paper in 2014.

[4] OpenAI, *GPT-4 Technical Report* (2023), https://cdn.openai.com/papers/gpt-4.pdf.

[5] Colin G. West, "Advances in Apparent Conceptual Physics Reasoning in GPT-4," *ArXiv* 2303, no. 17012 (2023).

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

[7] Joseph Weizenbaum, "ELIZA—A Computer Program for the Study of Natural Language Communication Between Man and Machine," *Communications of the ACM* 9, no. 1 (1966): 36-45.

[8] Cal Newport, "What Kind of Mind Does ChatGPT Have?" *New Yorker*, April 23, 2023.

[9] Daniela Sirtori-Cortina and Brendan Case, "Walmart Is Using AI to Negotiate the Best Price With Some Vendors," *Businessweek*, April, 26, 2023.

[10] Emilia Simeonova, Niels Skipper, and Peter R. Thingholm, "Physician Health Management Skills and Patient Outcomes," *Journal of Human Resources* 58, no. 4 (2022).

[11] Liudmila Alekseeva et al., "From In-Person to Online: The New Shape of the VC Industry" (working paper, IESE Business School, 2022).

[12] Cal Newport, "What Kind of Mind Does ChatGPT Have?" *New Yorker*, April 23, 2023.

[13] James Vincent, "Google's AI chatbot Bard makes factual error in first demo," *Verge*, February 8, 2023.

[14] Kevin Roose, "A Conversation With Bing's Chatbot Left Me Deeply Unsettled," *New York Times*, February 17, 2023.

[15] Daniela Sirtori-Cortina and Brendan Case, "Walmart Is Using AI to Negotiate the Best Price With Some Vendors," *Businessweek*, April 26, 2023.

[16] Joe Coscarelli, "An A.I. Hit of Fake 'Drake' and 'The Weeknd' Rattles the Music World," *New York Times*, April 19, 2023.

[17] Cal Newport, "What Kind of Mind Does ChatGPT Have?," *New Yorker*, April 23, 2023.

[18] Arthur Glenberg and Cameron Jones, "It takes a body to understand the world – why ChatGPT and other language AIs don't know what they're saying," *Conversation*, April 6, 2023.

[19] Leonard Boussioux, "The Crowdless Future? How Generative AI Is Shaping the Future of Human Crowdsourcing" (working paper, Harvard Business School, 2023), http://dx.doi.org/10.2139/ssrn.4533642.

[20] Richard Waters, "Man Beats Machine at Go in Human Victory Over AI," *Financial Times*, February 17, 2023.

[21] Hans Moravec, *Mind Children: The Future of Robot and Human Intelligence* (Cambridge: Harvard University Press, 1988).

[22] Carl B. Frey, *The Technology Trap: Capital, Labor, and Power in the Age of Automation* (Princeton: Princeton University Press, 2019).

[23] Jonathan T. Barron et al., *Zip-NeRF: Anti-Aliased Grid-Based Neural Radiance Fields* (2023), https://arxiv.org/pdf/2304.06706.pdf.

[24] Gary Marcus, "The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence," *ArXiv*, no. 06177 (2020).

[25] Carl B. Frey, *The Technology Trap: Capital, Labor, and Power in the Age of Automation* (Princeton: Princeton University Press, 2019).

[26] David Autor et al., "New Frontiers: The Origins and Content of New Work, 1940–2018," (working paper, National Bureau of Economic Research, 2022), https://www.nber.org/system/files/working_papers/w30389/w30389.pdf; Thor Berger, Carl B. Frey, "Did the Computer Revolution Shift the Fortunes of US Cities? Technology Shocks and the Geography of New Jobs," *Regional Science and Urban Economics*, 57 (2016): 38-45; Thor Berger, Carl B. Frey, "Industrial Renewal in the 21st Century: Evidence from US Cities," *Regional Studies*, 51, no. 3 (2017): 404-413.

[27] John Jumper et al., "Highly Accurate Protein Structure Prediction with AlphaFold," *Nature* 596, no. 7873 (2021): 583-589.

[28] Carl B. Frey, *The Technology Trap: Capital, Labor, and Power in the Age of Automation* (Princeton: Princeton University Press, 2019); see also Daron Acemoglu, Pascual Restrepo, "The Race Between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment," *American Economic Review* 108, no. 6 (2018): 1488-1542.

[29] Thor Berger, Chinchih Chen, and Carl Benedikt Frey, "Drivers of Disruption? Estimating the Uber Effect," *European Economic Review* 110 (2018): 197-210.

[30] Sida Peng et al., *The Impact of AI on Developer Productivity: Evidence from GitHub Copilot* (Cambridge: MIT Sloan School of Management, 2023).

[31] Shakked Noy, Whitney Zhang, "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence" (working paper, Social Science Research Network, 2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4375283.

[32] Erik Brynjolfsson, Danielle Li, and Lindsey R. Raymond, "Generative AI at Work" (working paper, National Bureau of Economic Research, 2023), http://www.nber.org/papers/w31161.

[33] Carl B. Frey, *The Technology Trap: Capital, Labor, and Power in the Age of Automation* (Princeton: Princeton University Press, 2019).

[34] Ian Leslie, "The Struggle To Be Human," *Ruffian*, December 10, 2022, https://www.ian-leslie.com/p/the-struggle-to-be-human.

[35] For examples of synthetic data, see Madhumita Murgia, "Why Computer-Made Data Is Being Used to Train AI Models," *Financial Times*, July 19, 2023.

[36] John Jumper et al., "Highly Accurate Protein Structure Prediction with AlphaFold," *Nature* 596, no. 7873 (2021): 583-589.

[37] Ilia Shumailov et al., "*The Curse of Recursion: Training on Generated Data Makes Models Forget,"* *ArXiv* 2305, no. 17493 (2023).

[38] Billy Perrigo, "Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to

Make ChatGPT Less Toxic," *TIME Magazine*, January 18, 2023.

[39] Lingjiao Chen, Matei Zaharia, and James Zou, "How Is ChatGPT's Behavior Changing Over Time?" *ArXiv* 2307, no. 09009 (2023).

[40] Phillip Koralus and Vincent Wang-Maścianica, "Humans in Humans Out: On GPT Converging Toward Common Sense in Both Success and Failure," *ArXiv* 2303, no. 17276 (2023).

[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.