THE GENERATIVE AI REVOLUTION – WHAT WE ALL NEED TO BE THINKING ABOUT

"Generative artificial intelligence presents a philosophical and practical challenge on a scale not experienced since the start of the Enlightenment."

Henry Kissinger
Eric Schmidt
Daniel Huttelocher
ChatGPT Heralds an Intellectual Revolution (WSJ, February 24, 2023)

"No matter how good your utopia you create, if your dystopia is bad enough, it doesn't matter."

Tristan Harris
Center for Humane Technology
The AI Dilemma

Overnight, it seems, while we were focused on countless political and geopolitical developments, ChatGPT burst onto our screens, and ChatGPT as well as a string of terms such as "generative AI," "chatbots" and "large language models," found their way into articles in the non-trade press and now increasingly form parts of daily conversations. Actually, it has been all of a few months.

Some have likened the technology to the most consequential innovation in modern history. It is certainly a game changer. At its core, the technology is machine learning algorithms ("unsupervised" and "semi-supervised") that through the application of "deep learning" use existing online content (text, images and video, for example) to create original (that is, new) content that appears to be human-generated.

The goal of commercial AI labs now is to create artificial general intelligence (AGI), that is, systems that generally are smarter than humans, at any task. While many are extolling the virtues of these developments (and to be clear, the benefits are manifold), others are sounding the alarm.

Today, it appears that even experts do not know where these technologies will take us. And that is the cause for alarm. What has surprised even the AI experts is that the efficacy of these technologies has grown exponentially in a matter of months. GPT-4, for example, was released in November and, according to a UBS research report cited in an article for Reuters, is the fastest-growing consumer application in history. GPT-4 reached 100 million users in its first two months; in contrast, it took TikTok nine months and Instagram 2.5 years to reach 100 million users. According to Similarweb, ChatGPT attracted 1.6 billion users worldwide in March, up from just over 1 billion in February.[1]

ChatGPT is the product of one platform (OpenAI), and there is, unsurprisingly, a race among the platforms to dominate this space. Google's product is the Bard chatbot. Microsoft

[1] This puts ChatGPT on a worldwide basis ahead of Bing and DuckDuckGo, but behind Baidu, Yandex and Google.

announced in February the launch of an AI-powered Bing search engine and Edge browser (based on a next-generation OpenAI large language model).

This briefing note is intended as a primer on the revolution in generative AI that is now upon us. To understand this revolution one needs to understand, at the very least, machine learning, large language models and fine-tuning through human feedback. I also set out below a Lexicon to better navigate the terminology and concepts.

AI Development

Machine learning dates back to the 1930s and 1940s – think back to Alan Turing. Advances in computer science in the 1980s allowed machine learning to flourish outside of laboratories. For some years, machine learning was about predictive models to find and classify content patterns. Then along came generative AI. What has changed is the speed at which machine learning has enabled models to train without human interaction, and to do so on the basis of unimaginable volumes and complexity of data. The more the data, the more the models can do. Most of us have interacted with AI whether we realized it or not. Alexa and Siri are based on AI technology.

According to a 2022 McKinsey Report, efforts to get machines to mimic human intelligence, including machine learning, have more than doubled since 2017, and investment in AI has increased in line with adoption. According to that report, in 2018, in the business sector, manufacturing and risk were the functions in which the largest shares of respondents reported seeing value from AI use. In 2022, the most significant reported revenue effects were found in marketing and sales, product and service development, and strategy and corporate finance, and respondents reported the highest cost benefits from AI in supply chain management.

The latest leapfrog in the technology is best understood in the jump from search engines to generative AI. Over the past 20 years, users would type inquiries into online search engines such as Google or Baidu and browse the multiple entries produced. Users could decide which information to use and generally could identify the source. Now, with the roll-out of ChatGPT, in response to a natural language query (a prompt), generative AI does all the synthesizing of multiple sources, and produces a single response immediately, but the sources are not identified.

Emerging Risks

The potential benefits of AI, as noted, are manifold. But, balanced against these benefits are a series of risks that are only beginning to crystalize. Part of the challenge is that, ultimately, the concerns with generative AI, as NYU Professor Gary Marcus points out – are that we simply do not have a handle on the unintended consequences of the technology. I cite some of the more commonly cited risks:

- ***Misleading, inaccurate or wrong outputs***. The models provide coherent answers that mimic human responses, and appear authoritative. But, what if the outputs are factually incorrect, or they miss key items or are out of date because the inputs are stale. (At present ChatGPT has been trained on datasets that only go up to 2021.) Will it become common practice to notify readers whether text is human-generated or AI-generated? What verifications standards or other quality assurance will be applied to outputs? The risks of getting it wrong, even before one factors in the potential for malign interventions such as disinformation or deepfakes, could be significant – be it to reputation or triggering liability.

- ***Biases***. The models are prone to the learn the same biases that beset other algorithm-based technologies, used, for example, to make hiring and credit decisions.
- ***Intellectual property.*** The scrapping of data as inputs may constitute copyright or other intellectual property violations, and potential copyright and other intellectual property infringements by users raise complex legal question of who has the rights to the outputs and who might be liable for what.
- ***Education***. The availability of machine-generated material is likely to have a profound impact on education. Paying another student to write a paper or plagiarising someone else's work is so last century, as now anyone will be able to let the models do the writing, almost instantaneously. That said, there are educators who see positive opportunities to recast learning completely.
- ***Performance evaluation***. The risks associated with education apply equally to enterprises, institutions and government agencies for which communications, and particularly written communications, are critical components of the service offering or the hiring process, and the ability to write is a key performance criterion. While the billable hour may be fast approaching extinction in professional services, the models offer avenues of abuse.
- ***Data protection and privacy.*** The models raise significant challenges for data protection and privacy in business contexts in light of the fact that input data become available (as training data) to answer prompts inputted by others. For example, there is the risk that source code, personal information or sensitive, confidential or proprietary data are used as input data and then become accessible to others. The more an organization's employees make use of the models, the more information the organization is feeding the models.

  An article posted on Dark Reading cites employees cutting and pasting strategy documents into Chat GPT to create a PowerPoint deck, or a professional inputting client or patient names and addresses to draft letters to, or about, them. Cyberhaven estimates in a February report that 11% of data pasted into ChatGPT by employees is confidential (and that at least 4% of employees have used sensitive data as inputs). Cyberhaven also notes that data inputted into ChatGPT (content cut and paste rather than file uploads) typically is not picked up by corporate security products.

  Aside from issues relating to mistakes that the models may produce, lawyers will need to consider the impact on attorney-client privilege and maintenance of client confidences around inputs, as the prompts form part of machine learning. According to OpenAI's FAQs, once data are entered as part of prompts, specific prompts cannot be deleted as part of a user's history – this FAQ then warns, "don't share any sensitive information in your conversations." OpenAI reserves the right to review conversations "to improve our systems," which may have privilege implications, and which incidentally could also arise if a client, rather than the attorney, enters data ostensibly covered by privilege.

While OpenAI states it does not retain information provided in conversations with ChatGPT, ChatGPT does learn from every such conversation and, separately, these conversations are taking place over the internet, which may not be secure.

- ***Environmental.*** Generative AI consumes energy during the training phase as well as during the conversations users have with the model. Training requires computationally intensive tasks, the hardware for which consumes considerable energy and, based on the source, may lead to significant carbon emissions. The climate impact is exacerbated by the multiple iterations of fine-tuning inherent in the training process, and the need for models to be periodically retrained. Once trained, the models are deployed on servers that themselves consume energy and generate their own carbon emissions. Producing and disposing of the hardware used in these processes separately impact the environment. In addition to their carbon footprint, these processes consume massive amounts of water, largely to cool the data processing centers.
- ***Power of big tech***. The models depend on massive amounts of data and the power to process the data. This means that yet greater power will be concentrated among the large tech players.
- ***Labor disruption.*** The models will undoubtedly be disruptive to business models and markets, and while it was expected that AI-driven disruption would have the greatest impact on less-skilled labor sectors, in fact, these models have the greater potential to disrupt professional service sectors, as well as media and advertising and other creative industries. (*See, e.g.*, Goldman Sachs Analyst Report, which estimates that 300 million jobs could be at risk.)
- ***Business model disruption.*** As with other technology advances, there will be winners and losers. I note one, for the time being: news organizations. The models will pull from news sources across the internet, without credit to the source. This also contributes to the amplification of disinformation.

Europol released its Tech Flash Report entitled “The Impact of Large Language Models on Law Enforcement,” which identified the potential exploitation of ChatGPT and other LLMs by criminal elements in the following areas:

- ***Online fraud and social engineering***, and in particular the ability of LLMs to re-produce language patterns to impersonate the style of speech of specific individuals or groups, which can be used at scale to mislead. LLMs enable phishing and online fraud, without the telltale grammatical, spelling or syntax mistakes that traditionally are associated with phishing emails, to be created faster and at low cost.
- ***Disinformation***, as LLMs can produce, with minimal effort, convincing, authentic sounding text at scale and speed (after being fed prompts loaded with disinformation); and
- ***Cybercrime***, as ChatGPT can produce code across a number of programming languages.

The report notes that there are workarounds to bypass safety features (moderation tools) intended to prevent malicious use. The workarounds include “prompt engineering,” which involves the refinement of the precise way a question is asked to influence the output. While this tool creates versatility, it can also be abused, and as one loophole is closed, others are found. One particular vulnerability is the potential to manipulate ChatGPT to generate malicious (hacking) codes.

Writing in The Atlantic in September 2020, Renée DiResta wrote that as a result of AI, the supply of disinformation soon will be infinite, triggering an urgent need for verification of purported authors of content and validation of trustworthiness. As Emily Bell, director of the Tow Center for Journalism at Columbia University, writing last month in The Guardian noted, generative AI models have no built-in commitment to the truth. Moreover, they produce responses with no attribution and with no explanation of, or citation to, sources. Bell also notes that while deepfake audio and videos can be easily debunked, there is the potential to create confusion and exhaustion at scale by "flooding the zone" with disinformation.

Industry Warnings

The competition to create "ever-larger unpredictable black-box models with emergent capabilities," in turn led tech leaders, in an open letter published March 22, to call for a six-month pause in the development of AI *beyond the latest version of ChatGPT* released last month, GPT-4. The open letter calls on governments to intervene if the pause is not adhered to.

These AI experts recognize that AI systems with "human-competitive intelligence can pose profound risks to society and humanity," necessitating guardrails. They admit they may be wrong in their and the wider community's view that, as there is neither a legal nor technological barrier to inhibit the development of AGI, AGI will be reality and it will be such sooner than expected. At the same time, they concede that if they are not wrong, the "implications for our society are profound."

The necessary guardrails have not materialized even though "AI labs have been locked in an out-of-control race to develop and deploy ever more powerful digital minds that no one – not even their creators – can understand, predict, or reliably control." In short, AI systems should be developed "only once we are confident that their effects will be positive and their risks will be manageable." Even OpenAI recognizes the risks of "misuse, drastic accidents an societal disruption" inherent in their mission to create AGI.

The pause should be used to draw up a set of safety protocols. The open letter also acknowledges that developers must work with policymakers to "dramatically accelerate" the development of AI governance systems, and calls for, among other things, the establishment of AI-oversight authorities, watermarking systems to distinguish human-generated from synthetic content, auditing and certification, liability for AI-generated harm and well-resourced institutions "for coping with the dramatic economic and political disruptions (especially to democracy) that AI will cause."

Last month, the Center for AI and Digital Policy (CAIDP) filed a complaint with the Federal Trade Commission (FTC) calling for an investigation of OpenAI and ChatGPT, on the ground that GPT-4 is not "transparent, explainable, fair or empirically sound, or accountable" (the FTC position) and calling on the FTC to establish a moratorium along the lines called for in the open letter. Among other things, the complaint makes reference to the 2019 OECD Principles on Artificial Intelligence as well as the 2018 Universal Guidelines for Artificial

Intelligence adopted by the International Conference on Data Protection and Privacy Commissioners.

The European Consumer Organization (BEUC) has called on EU and member state authorities to launch an investigation similar to that called for in the CAIDP complaint. The Deputy Director of the BEUC stated that waiting for the EU's proposed Artificial Intelligence Act would be harmful given the massive take-up of ChatGPT in only a few months and the attendant concerns around deception and manipulation. Among other concerns flagged by the BEUC are harmful financial advice, biased scoring for consumer credit and insurance, misleading recommendations to consumers.

Regulatory Responses

Generative AI, thanks to the roll-out of ChatGPT, has caught the attention of regulators and policymakers.

- In the European Union, the European Commission had proposed a Regulation in 2021 to be known as the Artificial Intelligence Act, but as POLITICO reported last month, ChatGPT has upended the EU effort and is forcing policymakers back to the drawing board. Ironically, as POLITICO reports, German lawmakers used ChatGPT to draft the speeches they gave in support of the need to rein in the technology.
- The Chinese Cyberspace Administration has proposed for public comment "Administrative Measures for Generative Artificial Intelligence Services.
- The US National Telecommunications and Information Administration (part of the Commerce Department) has requested public comment on "self-regulatory, regulatory, and other measures and policies that are designed to provide reliable evidence to external stakeholders—that is, to provide assurance—that AI systems are legal, effective, ethical, safe, and otherwise trustworthy." The National Institute of Standards and Technology (also part of the Commerce Department) issued its Artificial Intelligence Risk Management Framework in January and the White House Office of Science and Technology issued a Blueprint for an AI Bill of Rights in October.
- Both the Chair of the Federal Trade Commission and the Assistant Attorney General for the Antitrust Division of the Department of Justice have spoken publicly about the urgent focus on competition in the AI sector.
- Last month, the Italian data protection regulator, the Garante, ordered OpenAI to temporarily stop processing data of ChatGPT users due to alleged GDPR violations.

Axios reported last week that Senate Majority Leader Chuck Schumer is leading a congressional effort to draw up legislation to regulate AI, focused on transparency.

Concluding Thoughts

We are very much in uncharted territory, and early days of that voyage. There are myriad benefits to generative AI, but there are also myriad risks, and we, as a society, will be far better off if those risks can be addressed before it is too late. The good news is that there are sufficient voices in the industry that are calling for governance and guardrails, and we have the benefit of seeing the consequences of technology driven by a lack of transparency and accountability. (Imagine how much better we all would have been if we had better understood, for example, the power of social media algorithms to amplify hate or the adverse impact on young people of access to social media on ubiquitous smartphones at too early an age.)

In the meantime, it is hard to imagine any enterprise, institution or government agency that does not need to be thinking about what generative AI means for them, both from a defensive standpoint as well as the opportunities that if missed, in the case of businesses, for example, may well benefit competitors. To gain the benefits though, many of us will need to balance those benefits against the attendant risks cited above, and the ease with which those risks can be addressed by society at large will likely depend upon a combination of regulation and self-regulation.

Lexicon

Artificial Intelligence, or AI: Machines that mimic understanding natural language, recognizing images, problem solving and decision-making of the human mind, and can be applied to predict and automate tasks historically done by humans. AI encompasses a number of subfields, including machine learning.

Machine learning, or ML: A subset of artificial intelligence that focuses on the use of data and algorithms to imitate the way humans learn. In short, algorithms derive knowledge from data to predict outcomes. In other words, these computer programs recognize patterns in data and make predictions. There are three main categories of machine learning, that is three ways of training and producing algorithms:

- **supervised machine learning**, which uses data that is tagged with a label (labelled data) to train algorithms to predict outcomes (in effect the desired output is already known). This is "supervised" in the sense that humans are teaching the model what to do;
- **unsupervised machine learning**, which analyses unlabelled data sets, identifying hidden patterns of similarities and differences without human intervention, and then creates its own data clusters. This model in effect makes its own predictions based on massive amounts of data available to it. Here the desired output is not known; and
- **semi-supervised machine learning**, which uses a mix of labelled and unlabelled data to train algorithms.

The "fine-tuning" of the algorithms is undertaken leveraging the foregoing – this is known as "transfer learning." In effect, the process applies knowledge gained while solving a particular task to a related task.

Deep learning: A subset of machine learning that attempts to simulate the behavior of the human brain to recognize patterns and make decisions. It does so by training neural networks and accessing massive amounts of data. Machine learning is more dependent on human intervention to learn, while deep learning automates elements of the feature extraction process, eliminating the need for some human intervention.

Chatbot: A computer program designed to stimulate conversation with human users, allowing such users to interact with digital devices as if they were interacting with other humans. These programs can answer questions, write essays or code, or otherwise generate text that they are asked to do after a user enters a prompt. Chatbots are driven by artificial intelligence (AI) and natural language processing (NLP). Chatbot technology, for example, is used in smart speakers.

According to IBM, historically chatbots were text-based and programmed to reply to a limited set of questions with answers that had been pre-programed by developers. Thereafter, chatbots integrated more rules and NLP that enables users to experience a conversation. In the latest iteration, AI chatbots use natural language understanding (NLU) to discern user needs and then apply elements of AI (machine learning and deep learning) to discern what the user is trying to accomplish. Chatbots have no emotions and they cannot discern whether or not they are making sense.

Generative AI: Technology that creates content, including text, images, audio, video, simulations and computer code (the outputs), by identifying patterns in large quantities of training data (the inputs), and then creating original material (*e.g.,* completely new data, new insights and new content) that has similar characteristics, from the existing data. This is based on the deep learning model. Generative AI need not be limited to creating new text; it can create designs, imagines, videos and much more. Examples include ChatGPT for text and DALL-E (a combination of artist Salvador Dalí and the PIXAR robot WALL-E) for images.

ChatGPT: A text-generating chatbot developed by OpenAI, trained using RLHF. GPT stands for generative pre-trained transformer, a family of large language models. The training in the acronym refers to the process of teaching a digital system to recognize patterns and make decisions based on available input data. Pre-trained means that the models have already been trained. The chat means just that – one interfaces with the chatbot through natural language prompts.

OpenAI was founded as a start-up in 2015, and the original paper on generative pre-training transformer models was posted on the OpenAI website in June 2018. The inflection model, GPT-3, was announced in 2020 as a language model trained on massive online datasets. According to estimates, GPT-3 was trained on approximately 45 terabytes of text – that is, according to a McKinsey featured insights report, about one million feet of bookshelf space or a quarter of the Library of Congress. In November, the next iteration AI chatbot based on GPT-3.5 was released – this was ChatGPT.

GPT-4 was released March 14. GPT-4 is a "multimodal" model, meaning it can answer questions about pictures as well as text, because it can accept both text and images as inputs. There are both free and premium versions of these models. It can answer in less than a second.

Hallucination: The core outputs of large language models are responses to sentences and questions that mimic human-generated responses. The models understand structure and patterns of the vast amount of data on which they have been trained, but they do not understand the meaning behind human language. As the models are designed to produces answers, if the models do not know the exact response, they can make up answers, that on their face appear coherent. Inventing facts is called a "hallucination."

Large language model, or LLM: A type of neural network that learns skills, including generating written text, conducting conversations and writing computer code. LLMs use deep learning techniques such as transformer models to understand, interpret and generate text based on vast amounts of input data available online. In effect, an LLM predicts the next course of action by analyzing the data inputs that came before it, generating proper syntax, context and semantics. According to Alex Hughes, writing for the BBC's Science Focus, ChatGPT-3 was trained using 570GB of online data, representing 300 billion words. It has 175 billion parameters.

Natural language processing, or NLP: Techniques used by LLMs to understand and generate human language. These methods often use a combination of machine learning algorithms, statistical models and linguistic rules. The goal of NLP is to generate text that mimics human writing.

Neural network: A mathematical system, inspired by and modelled on the human brain, that learns skills by finding statistical patterns in data. It in effect mimics the human brain through a set of algorithms and consists of layers of interconnected noted or neurons: The first layer receives the input data, and the last layer outputs the results. A neural network with three or more layers is considered to be a deep learning algorithm. This is what allows the models to process natural language.

Reinforcement Learning from Human Feedback, or RLHF: Trains AI models to find the best outcome by trial and error, using a combination of automated rewards and human feedback that creates reward signals. The reward signals improve the performance of the model. Initially, the AI model is trained using supervised machine learning to predict correct actions or outputs based in the inputs given. Human trainers then provide feedback on performance, creating reward signals. Algorithms then incorporate the reward signals, optimizing performance. This process is repeated iteratively.

Transformer model: A neural network architecture that does not have to analyze words one at a time but can look at an entire sentence at once, thus enabling models to capture context and long-term dependencies in language.

* * *

Mark S. Bergman
7Pillars Global Insights, LLC
Washington, D.C.
April 23, 2023

No portion of this briefing note was prepared using ChatGPT.