THE INDIA FORUM
A JOURNAL-MAGAZINE ON CONTEMPORARY ISSUES

November 23, 2023

Demystifying Artificial Intelligence-I

On Narrow and Generative AI, Neural Networks, and AI Autonomy

By: Anurag Mehra

A 3-part series explains the main features of the rapid growth of AI today, current research and issues in regulation. Part-I discusses the core idea of AI, the growth of 'generative' AI and, crucially, explores if it will serve humankind or develop an autonomy that can make it do dangerous things.

Preamble

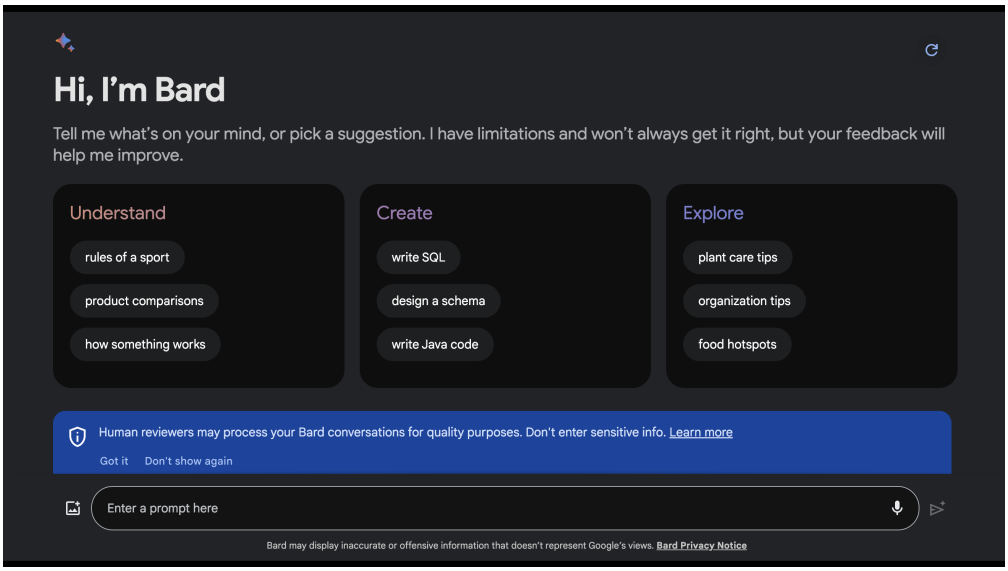
Earlier this year, I attended a faculty workshop on the disruption that the newly released ChatGPT bot was expected to cause in educational institutions. In the workshop, we asked the bot to write essays, and answer queries. It was an eerie experience to watch the words appear on the screen—reasonably well written, plausible sounding, and with occasional mistakes. (GPT stands for Generative Pre-trained Transformer; ChatGPT is an AI program or bot that can converse; that is, chat.)

Indeed, ChatGPT, launched in November 2022 by OpenAI, has brought an unparalleled amount of visibility to artificial intelligence (AI) and made it accessible to people at large. However, AI products and applications have been around since much earlier—facial recognition; natural language processing (digital assistants such as Siri and Alexa); the advertisements that track us when we browse the internet (powered by the profiling ability of AI-based code); dating apps; deep fake videos; and the occasional game matches between "man and machine".

> [The] revival of AI "everywhere" a little more than a decade ago has been made possible by developments in hardware—cheap and fast processors on ultra-small chips, and low-cost, miniature memory storage that can hold humongous amounts of data.

Sophia, the human-like conversational robot made by Hanson Robotics, showed us glimpses of this world, but conversational bots, like ChatGPT and Google's Bard, have made AI an "article" of everyday use. AI is now available in browsers and apps (for example, email, document preparation software, and code writing environments) and is accessible through computers, mobile phones, smartwatches, table-top devices and the brand new AI-pin. Recently, ChatGPT-4 (the latest version) has acquired the ability to talk and see (a multimodal AI). The possibilities of what one can use it for are endless.

Illustration 1: Bard Console



What is AI?

An early definition of AI was given by Alan Turing and it was simply this—if you cannot tell that the "thing" you are "talking" to is a machine because it "sounds" like a human, then the machine is (artificially) intelligent. Much has happened after that, especially with the advent of computers that can play games (Chess, Jeopardy, Go, Starcraft, Poker, Dota2) well enough to beat human champions; machines that "recognise" complex objects like faces; the arrival of digital assistants that "understand" requests in conversational languages, and more. Yet, all of this is called Narrow AI—that is, intelligent computers that do a single task well.

Many of the ideas that power today's rise of AI were conceived decades ago (for instance, neural networks), but then AI entered what has been labelled an "AI winter". The last winter ended with the 1990s. This revival of AI "everywhere" a little more than a decade ago has been made possible by developments in hardware—cheap and fast processors on ultra-small chips, and low-cost, miniature memory storage that can hold humongous amounts of data. So much data capture capability and its rapid analytical processing has made modern AI feasible.

The reason why AI requires vast amounts of data and has to process it very fast is because it works by brute force. AI operates by searching for patterns by analysing immense amounts of data. Consider how object recognition works. An object (for example, a face) is represented by a "numerical map" (Illustration 2) containing values that relate to its features (dimensions of elements that make the object— lengths, areas, angles, distance between specified points, and so on).

Illustration 2: A 'Numerical Map' of a Face

An AI program can extract this set of numbers from an image of an object. Millions of images (or objects) are shown to the program, and every image is assigned a label (rectangle, square, or circle). The AI is "trained" through this process and "discovers" the specific "patterns" that correlate to a label. Post-training, when shown an object, the AI can tell if it is seeing a ball or a box.

> The reason why AI requires vast amounts of data and has to process it very fast is because it works by brute force. AI operates by searching for patterns by analysing immense amounts of data.

Now, extend this simple example to more complex objects such as cats, dogs, and faces—which have many more attributes—and it can be appreciated why huge amounts of data have to be stored and processed at speed to "train" the AI. Think now of the sensors on a driverless car, thousands of them taking data from their surroundings, trying to identify—in real time—the road, snow, water, people, dividers, traffic lights, distance from other vehicles, speed of moving objects, and so on. Here comes the need for speed—to process the data so that decisions can be made about how the car will deal with the object in real time.

Chatbots, the latest AI products that have enthralled humans, work on similar principles. These programs are trained using huge amounts of text (or images and video, for versions designed to go beyond text), and the program finds patterns in the text (what word follows another, in what context). When the bot is asked a question (called a prompt; for instance, "Who are the Rohingya"), it

attempts to "recognise" the context and create a reply by assembling a sequence of words. Each new word in an evolving sentence is selected based on the highest probability of being placed next. This computed probability is based on the patterns that the bot has been trained on. Read what ChatGPT tells us about its abilities here.

> [G]enerative AI it generates text (or images or video) based on what it has encountered in its training. Some have derisively likened this approach to that of a parrot that mimics human speech without understanding it…

The bot therefore constructs a sentence, word by word, by sheer brute force—computing the probability of millions of candidates that might fit in and choosing the "right" one. These bots are based on LLMs—Large Language Models. It is a stupendous feat of technology that the answer to a query is usually sensible, plausible, and most often accurate. And the "answering" is in real time. Keep in mind that a bot is replying to millions of users at any given time.

It is the underlying ability of computers to store huge amounts of data and process it rapidly that has made chatbots possible. To appreciate some numbers, ChatGPT (v3.5) has been trained on 300 billion words of text, which amounts to about 570 GB of data, drawn from websites (notably, Wikipedia), and it has 175 billion parameters (something like weights). (A Washington Post report gives an idea of the kind of content and websites that were used for training by OpenAI.) Some of the processors that are used for AI have now become so precious (a graphics processing unit, or GPU, may cost between $10,000 to $50,000) that companies manufacturing them are making fortunes.

This type of AI is called generative AI because it generates text (or images or video) based on what it has encountered in its training. Some have derisively likened this approach to that of a parrot that mimics human speech without understanding it, and have dubbed this kind of AI a "stochastic parrot", or a program that makes superficial connections between words. Or, even more derisively, just a more sophisticated form of the "auto complete" that we see when we type words into a smartphone that then completes sentences for us.

Many AI innovators counter this by asking if it is any different from what the human brain does, at least in some significant ways. They also suggest that what is superficial, or simple, when done at scale can result in complex emergent abilities. Indeed, there may be something to contemplate about the argument that chatbots have figured out a rich set of concepts about the external world.

Linguist Noam Chomsky disagrees strongly with the idea that generative AI mimics how the brain works. In an opinion piece in the *New York Times*, he wrote, "The human mind is not, like ChatGPT and its ilk, a lumbering statistical engine for pattern matching, gorging on hundreds of terabytes of data and extrapolating the most likely conversational response or most probable answer to a scientific question. On the contrary, the human mind is a surprisingly efficient and even elegant system that operates with small amounts of information; it seeks not to infer brute correlations among data points but to create explanations."

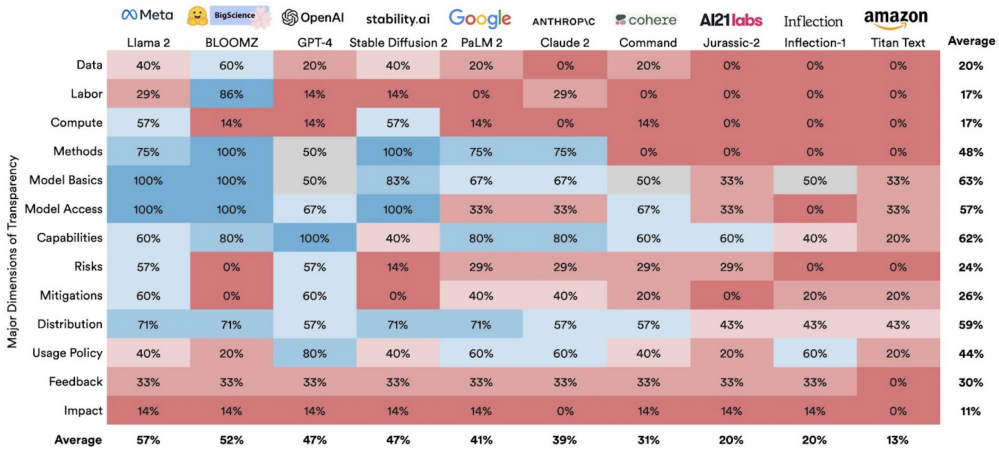
Interventions and Transparency

It is important to note that creating the bots requires a significant amount of human intervention in terms of selecting and curating data, devising sensible data labels, reinforcing learning to impart accuracy, and building guardrails (supervised learning) against hate speech or hallucinations (when a bot produces text that is patently false or nonsense). This intervention is what makes the chatbot output not just plausible but also accurate.

This role of this manual intervention is often obscured by mainstream AI hype. Even more worryingly, some of these manual activities, like tagging and labelling data for AI training, are poorly compensated. The Australian company, Appen crowdsources workers for this and pays them pennies even though its client list includes Amazon, Facebook, Google, and Microsoft.

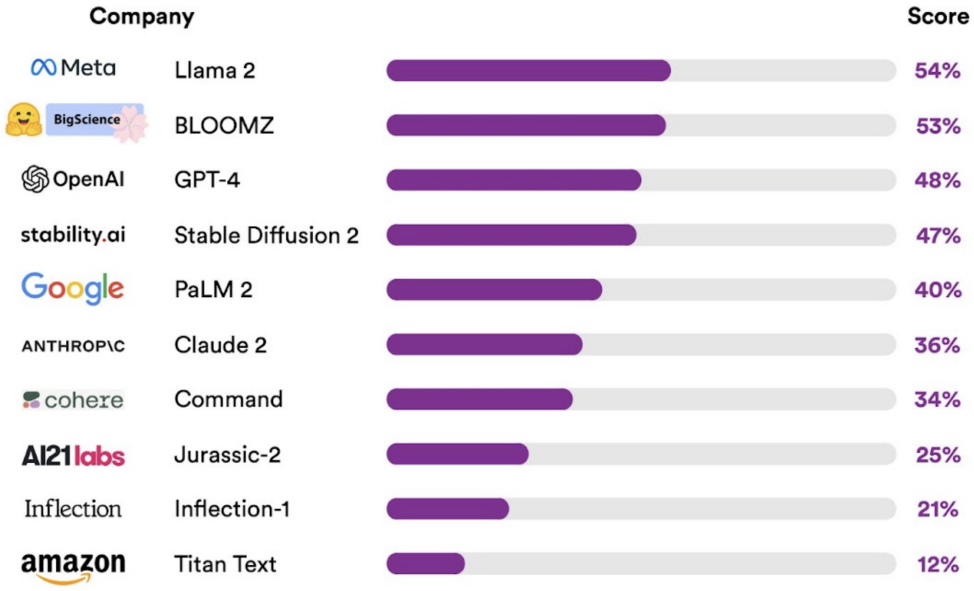
A Stanford research study has shown that most major AI models fare very poorly on transparency. An announcement from the university's Center for Research on Foundation Models, says, "Less transparency makes it harder for other businesses to know if they can safely build applications that rely on commercial foundation models; for academics to rely on commercial foundation models for research; for policymakers to design meaningful policies to rein in this powerful technology; and for consumers to understand model limitations or seek redress for harms caused." Most models have done very poorly, with the highest being 54 out of 100 awarded to Meta's Llama 2. The two charts from the study (Illustrations 3 and 4) are highly illustrative of the metrics used and the final scores, respectively.

Illustration 3: Foundation Model Transparency Index Scores by Major Dimensions of Transparency, 2023



Major Dimensions of Transparency	Meta Llama 2	BigScience BLOOMZ	OpenAI GPT-4	stability.ai Stable Diffusion 2	Google PaLM 2	ANTHROP\C Claude 2	cohere Command	AI21labs Jurassic-2	Inflection Inflection-1	amazon Titan Text	Average
Data	40%	60%	20%	40%	20%	0%	20%	0%	0%	0%	20%
Labor	29%	86%	14%	14%	0%	29%	0%	0%	0%	0%	17%
Compute	57%	14%	14%	57%	14%	0%	14%	0%	0%	0%	17%
Methods	75%	100%	50%	100%	75%	75%	0%	0%	0%	0%	48%
Model Basics	100%	100%	50%	83%	67%	67%	50%	33%	50%	33%	63%
Model Access	100%	100%	67%	100%	33%	33%	67%	33%	0%	33%	57%
Capabilities	60%	80%	100%	40%	80%	80%	60%	60%	40%	20%	62%
Risks	57%	0%	57%	14%	29%	29%	29%	29%	0%	0%	24%
Mitigations	60%	0%	60%	0%	40%	40%	20%	0%	20%	20%	26%
Distribution	71%	71%	57%	71%	71%	57%	57%	43%	43%	43%	59%
Usage Policy	40%	20%	80%	40%	60%	60%	40%	20%	60%	20%	44%
Feedback	33%	33%	33%	33%	33%	33%	33%	33%	33%	0%	30%
Impact	14%	14%	14%	14%	14%	0%	14%	14%	14%	0%	11%
Average	**57%**	**52%**	**47%**	**47%**	**41%**	**39%**	**31%**	**20%**	**20%**	**13%**	

Illustration 4: Foundation Model Transparency Index Total Scores, 2023



Structure of AI

The learning by AI is stored in a neural network that consists of interconnected nodes (they mimic interconnected neurons in the brain) laid out in the form of layers. The input layer receives the inputs (the prompts) and the output layer generates the text. The middle layers provide for capturing complexity. Think of this neural net as a box that relates inputs to outputs—the box holds the complex ways in which the inputs and outputs are correlated (Illustration 4). The nodes represent the factors that play an important role in these connections and the thickness of the lines connecting these nodes represent how important that connection is. The term "deep learning" refers to the depth (number) of layers that may be present.

What this tells us is that the AI model is essentially a statistical correlation between many inputs and many outputs. There is no sense of cause and effect, or any explanation why certain inputs lead to certain outputs—in other words, how a conclusion is reached. The more the number of nodes, layers and connections, the more complex it is and the more this is true. Therefore, it is has been said that the workings of AI are not understandable.

Illustration 5: Predicting Personality Traits from Users' Data on Social Media

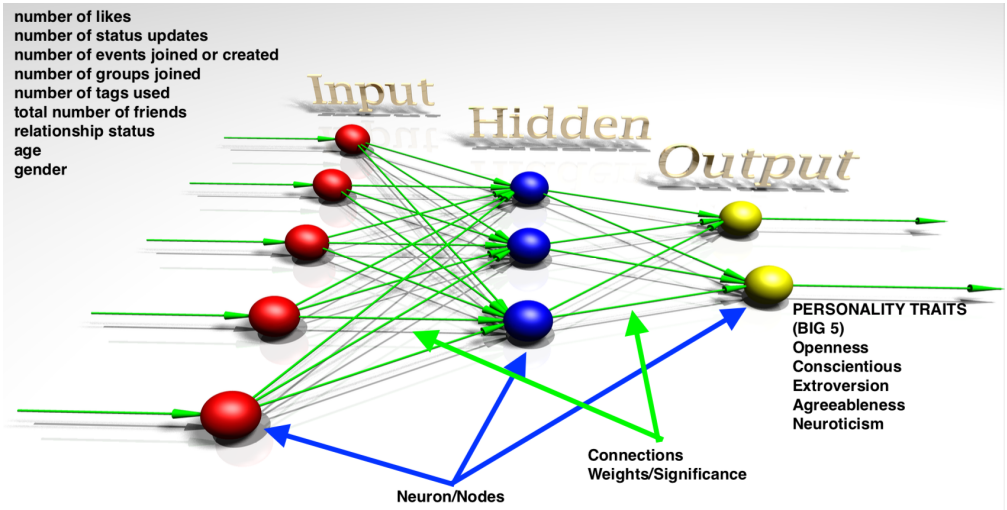


Illustration 5 shows a highly simplified schematic of how AI predicts personality traits from the Facebook data of users. These traits are very important for predicting and nudging behaviour (for example, what they can be made to buy, or whom they can be persuaded to vote for). This kind of analysis was made famous by the Cambridge Analytica (CA) scandal. Cambridge Analytica was a British political consulting company that illegally obtained Facebook users' data and built psychological user profiles, which were apparently used by Donald Trump's campaign in 2016 to target voters.

Great AI Debates

The conversational abilities of chatbots, which enable them to talk intelligently about everything, engender an impression that these entities are all-knowing. Chatbots mostly sound very officious and fluent, thereby invoking the "authority heuristic" (like receiving an answer from a doctor) and the "fluency heuristic" (well written text enhances credibility). They tend to convey a feeling of being "generally" intelligent.

However, the truth is that these bots can only do the one task they have been designed for, whether it is playing a game, or recognising faces, or generating text, and so on. This is called a narrow AI. In contrast is the idea of a general, comprehensive AI, the kind of intelligence that can develop into a super intelligence that surpasses the natural intelligence of human beings—in terms of competence and scale. Even though the current state of AI is nowhere near this holy grail of a super-intelligent AI, there are many AI experts who feel (and probably desire) that this will happen not too far in the future (forecasting website Metaculus predicts that this will happen in 2030).

What fuels this hope is the unanticipated abilities that AI programs sometimes develop, such as, for instance, the ability to identify sentiment (even though it was not mandated to); translate between languages (even though the training was for different languages, each one separately); come up with game moves that humans are unlikely to think of; and even lie or be evasive. Each new generation of chatbots is much more powerful, and seemingly without any limit – the latest ChatGPT-Turbo understands context much better, is faster and can accept more complex prompts (and ChatGPT-5 is coming soon). Some of the development towards a broader AI is also derived from the non-linguistic data layers that have been added to chatbots—such as images and video—and it opens up the possibility of "conceptual understanding" in terms of relating objects to text. This can be seen, in some limited manner, from that images and videos can be generated by bots through purely textual prompts.Creating a good prompt - to get what you want - is a skill by itself, and is called prompt engineering. There are short term courses (paid and free) to learn this skill. See here and here for samples of AI-generated images and the prompts that were used to produce them.

> …[T]though the current state of AI is nowhere near …[the] holy grail of a super-intelligent AI, there are many AI experts who feel (and probably desire) that this will happen not too far in the future.

A powerful, general AI raises interesting possibilities. For instance, could it ingest scientific empirical data and come up with hypotheses, which it could test or suggest the experiments that will be needed to validate them? Will it be able to come up with Albert Einstein's theory of relativity or Charles Darwin's theory of evolution? This is broader than what Google's Alpha Fold did by

predicting the shape of proteins—a problem of great significance in biology and chemistry, and a tedious one to address by experiments.

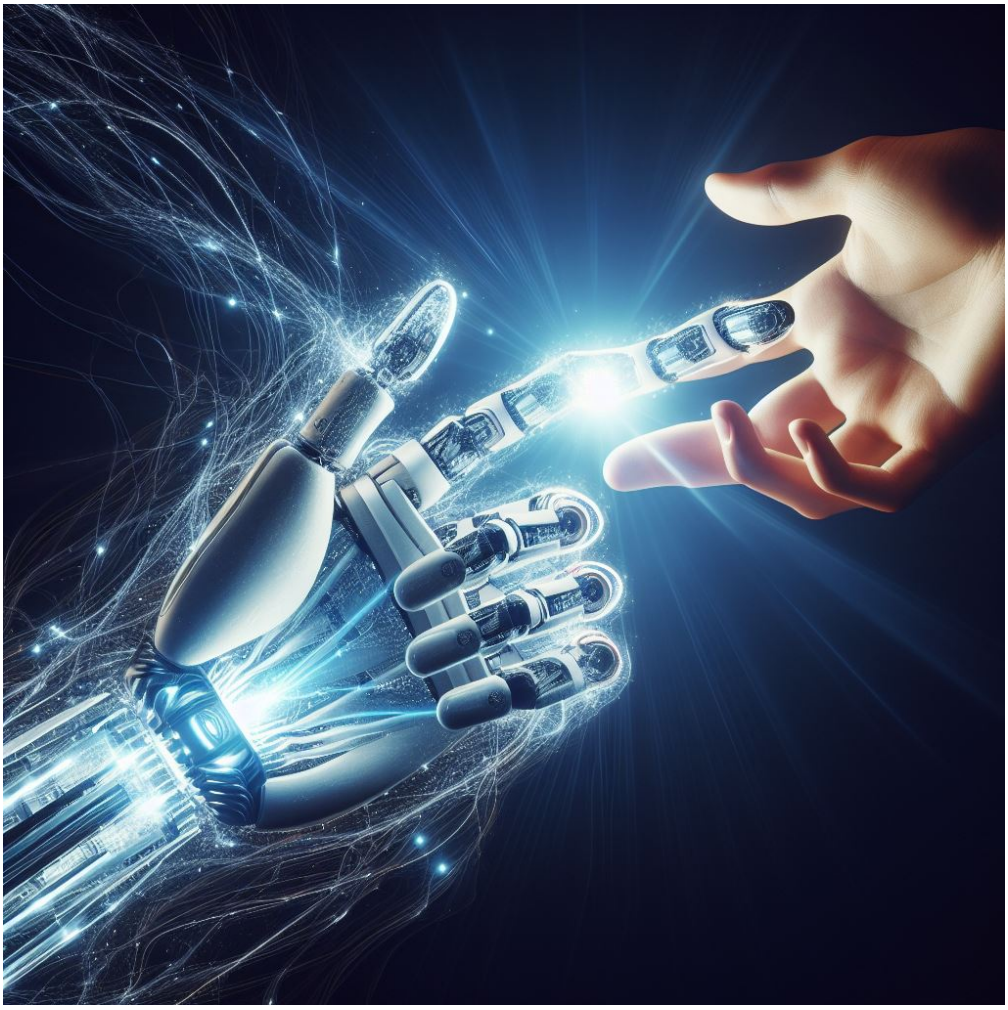
Related to this are the questions of bots becoming sentient, self aware, developing a consciousness, and perhaps taking over the world. Indeed, it is easy to be convinced that this is possible. The chief scientist of OpenAI tweeted last year that "it may be that today's large neural networks are slightly conscious". Earlier a scientist at Google, Blake Lemoine, made a claim that the bot he was interacting with—LaMDA—had become sentient. Google dismissed his claims and fired him. Embodied AI—a bot housed in a human-like body—is likely to promote such perceptions by opening up channels for AI's direct interaction with the physical world and people in everyday settings. This will bring the ideas of AI rights, and morality to the fore.

While this may seem far-fetched, a basic concern here is whether AI will "serve mankind" or develop an autonomy (read, evade human oversight) that will extend to "doing things" that can be dangerous for humans. A telling instance was when ChatGPT-4 "lied" when faced with a problem it could not solve—a Captcha. It got a human to solve it by saying, "No, I'm not a robot. I have a vision impairment that makes it hard for me to see the images".

This is the problem of "alignment". The OpenAI blog provides the context for this, "Currently, we don't have a solution for steering or controlling a potentially superintelligent AI, and preventing it from going rogue. Our current techniques for aligning AI, such as reinforcement learning from human feedback, rely on humans' ability to supervise AI. But humans won't be able to reliably supervise AI systems much smarter than us." OpenAI says that it will devote 20% of its compute capacity to alignment research.

> ...[A] basic concern here is whether AI will "serve mankind" or develop an autonomy (read, evade human oversight) that will extend to "doing things" that can be dangerous for humans.

An extreme consequence of an AI behaving autonomously will be it avoiding a shutdown—the bot can lie and take strategic action to do this. For instance, it could replicate itself and spread across the internet, garner resources and control newly encountered systems along the way, and make sure that some version of it will remain alive somewhere. Such autonomy poses tremendous risks if a "misalignment" occurs. This could range from providing easy access to dangerous technology (for example, dual use biotechnology) to the design of novel, deadly pathogens or chemicals.



A ChatGPT-based product, AutoGPT, follows prompts to carry out tasks autonomously by interacting with software and services available online. The simplest one is to assemble programming code while more complex examples are “build a website” or “enhance my business”, and so on. It does not take much function creep for prompts like “hack into this system” or “replicate yourself to prevent shutdown”. The system is still in infancy but it does show us what may be possible.

At a more basic level, the imperatives of efficiency, automation, or even plain stupidity may cause us to hand over decision-making, which must always be in the domain of humans, to an AI (think nuclear switches or power grid operations), with catastrophic consequences. The involvement of AI in defence and military affairs has already begun with Istari, a start-up backed by former Google CEO Eric Schimdt, virtually assembling “war machines” and running simulations for testing them. Scale.AI now has the US Department of Defence as its client and the scope of its work includes development of autonomous systems as well as human-machine interfaces.

> The imperatives of efficiency, automation, or even plain stupidity may cause us to hand over decision-making, which must always be in the domain of humans, to an AI, with catastrophic consequences.

Alignment research concerns certainly make these possibilities quite feasible. This is why there are suggestions to create an international agency on the lines of the International Atomic Energy Agency to regulate the use of large AI development by using licences, reporting frameworks, and specific restrictions on the development of AI technology. There is also talk of a kill-switch to shut off AI in case of an emergency.

A passionately argued piece by Eliezer Yudkowsky of the Machine Intelligence Research Institute paints a bleak doomsday scenario. He writes, “Progress in AI capabilities is running vastly, vastly ahead of progress in AI alignment or even progress in understanding what the hell is going on inside those systems. If we actually do this, we are all going to die.” He also says that it does not matter whether AI is conscious or not, it is its potential to do things that matters, and that we cannot ever “decode anything that goes on in the giant

inscrutable arrays". More warnings have come from serious insiders, such as Geoffrey Hinton who quit Google so that he could speak about the dangers unfettered, Yoshua Bengio from the Mila-Quebec AI Institute, who speaks of establishing an international "human defense organization"; and Demis Hassabis, the British chief executive of Google's AI unit suggests that something similar to the Intergovernmental Panel on Climate Change (IPCC) should be established for AI oversight.

An unexplored, not-much-written-about debate is on how the mainstreaming of AI in everyday life is bringing a new impetus to the idea of technological solutionism. This holds that there are technical fixes for social and political problems, thus steering the discourse on these issues (such as poverty or hunger) away from their ideological and political origins. Evgeny Morozov, who coined the term "technological solutionism", writes, "Depending on how (and if) the robot rebellion unfolds, A.G.I. [Artificial General Intelligence] may or may not prove an existential threat. But with its antisocial bent and its neoliberal biases, A.G.I.-ism already is: We don't need to wait for the magic Roombas to question its tenets."

This is the first part in a 3-part series on Artificial Intelligence. The second and third parts of this series can be read here and here.

Anurag Mehra teaches engineering and policy at IIT Bombay. His policy focus is the interface between technology, culture and politics.
?

References:

History of Data Science. "Artificial Neural Networks: Deeper Learning." August 26, 2023. www.historyofdatascience.com/artificial-neural-networks-deeper-learning/.

JSTOR. "The Turing Test and the Frame Problem: AI's Mistaken Understanding of Intelligence." June 2016. www.jstor.org/stable/26046989.

New York Magazine. "The Chatbot Will See You Now." March 1, 2023. nymag.com/intelligencer/article/ai-artificial-intelligence-chatbots-emily-m-bender.html.

The New York Times. "Noam Chomsky on the Future of AI." March 8, 2023. www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html.

Journal of Information Technology and Politics. "The Ethics of AI: A Review." July 31, 2023. journals.sagepub.com/doi/full/10.1177/21582440211032156#bibr27-21582440211032156.

Stanford Institute for Human-Centered Artificial Intelligence. "Introducing the Foundation Model Transparency Index." October 18, 2023. hai.stanford.edu/news/introducing-foundation-model-transparency-index.

Scroll.in. "Artificial Intelligence Will Soon Become Impossible for Humans to Comprehend. Here's Why." April 5, 2023. scroll.in/article/1046756/artificial-intelligence-will-soon-become-impossible-for-humans-to-comprehend-heres-why.

The Atlantic. "Chatbots Sound Like They're Posting on LinkedIn." April 25, 2023. www.theatlantic.com/technology/archive/2023/04/ai-chatbots-llm-text-generator-information-credibility/673841/.

OpenAI. "Introducing Superalignment." July 5, 2023. openai.com/blog/introducing-superalignment/.

The New York Times. "AI and Humanity: The Possibilities and Perils of Our Future." June 10, 2023. www.nytimes.com/2023/06/10/technology/ai-humanity.html.

Science. "Could chatbots help devise the next pandemic virus?." March 31, 2022. www.science.org/content/article/could-chatbots-help-devise-next-pandemic-virus.

Wired. "Eric Schmidt Is Building the Perfect AI War-Fighting Machine." June 1, 2023. www.wired.com/story/eric-schmidt-is-building-the-perfect-ai-war-fighting-machine/.

Time. "An Open Letter to OpenAI." July 12, 2023. time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/.

The New York Times. "The Danger of Artificial Intelligence." June 30, 2023. www.nytimes.com/2023/06/30/opinion/artificial-intelligence-danger.html.