

November 27, 2023

Demystifying Artificial Intelligence-III

On Liabilities, Anthropomorphism, Jobs, and Regulation

By: Anurag Mehra

Tech companies in a race for profit seek unbridled development of generative AI, unmindful of the potentially far-reaching economic, social and political fallout. The field is calling out for precautionary regulation but the world is moving too slowly. The concluding article in the series on AI.

Liabilities of AI

Who is liable if someone follows the advice given by a chatbot? Like health advice or counselling? Who is liable if someone publishes a chatbot's text output, which causes harm to somebody? How liable is a chatbot when it tells something dangerous to a child?

It is very easy to get health tips from a chatbot but there is no guarantee that the information is professionally validated. The National Eating Disorders Association (NEDA) in the US [took down its chatbot](#) named Tessa, which was providing advice on eating disorders, weight loss methods, and so on. The takedown was based on complaints that it was dishing out dubious advice. Then there are AI-generated videos like [this](#) giving health advice.

The US Federal Trade Commission opened an investigation into “whether ChatGPT's misstatements about real people constitute [reputational damage](#)”. An instance of such damage was a case when ChatGPT [claimed](#) that the mayor of an Australian town was imprisoned for bribery. The [fabricated judicial citations](#) in a US legal brief were fortunately detected by the judge, but what if a judgment had been pronounced and implemented based on this imaginary evidence?

Chatbots do “hallucinate” and make up things in the absence of properly working guardrails (and other reasons). Sometimes they can be tricked into doing this by the use of cleverly phrased prompts.

Chatbots do “hallucinate” and make up things in the absence of properly working guardrails (and other reasons). Sometimes they can be tricked into doing this by the use of cleverly phrased prompts. An [editorial](#) in *Nature* warned, “While the global move to regulating AI seems to be largely driven by its perceived extinction risk to humanity, it seems to me that a more immediate threat is the infiltration into the scientific literature of masses of fictitious material.”

Snapchat, a communication app in which messages disappear after having been read, is much used by children and teens. The app has “MyAI” as a friendly bot from which children tend to seek friendly advice that is not always safe. The UK Information Commissioner's Office has [determined](#) that the bot jeopardises the privacy of children by using the records of conversations with the AI bot for targeted ads and often provides inappropriate advice. It can also be [tricked](#) into hallucinating.

In many of these instances the primary liability can always be pushed on to the user/prompter of the AI program, but a good portion of it should also fall on the AI creators, trainers, and the hosting company. A related question is whether the “safe haven” protections of [Section 230](#) (part of the US Communications Decency Act of 1996) apply to AI-generated content. This section protects users and services (for instance, social media platforms) from getting sued for forwarding email, hosting online reviews, or sharing photos or videos that some may find objectionable. The opinion on whether such protection should apply to the products of generative AI is [split](#), with probably greater support for the position that it does not apply. It all thus remains muddy at the moment.

Anthropomorphism and AI

In the movie, [Her](#) (2013), a lonely author falls in love with the operating system of his computer. This was the first movie that explored human-machine attachment. Chatbots were not available then but the movie was prescient in showing us the future. Now this idea is mainstream with chatbots specially designed for companionship. [Replika](#) is an app based on OpenAI's AI- chatbot models, which can be used to create, “The AI companion who cares; Always here to listen and talk; Always on your side” (Illustration 1).

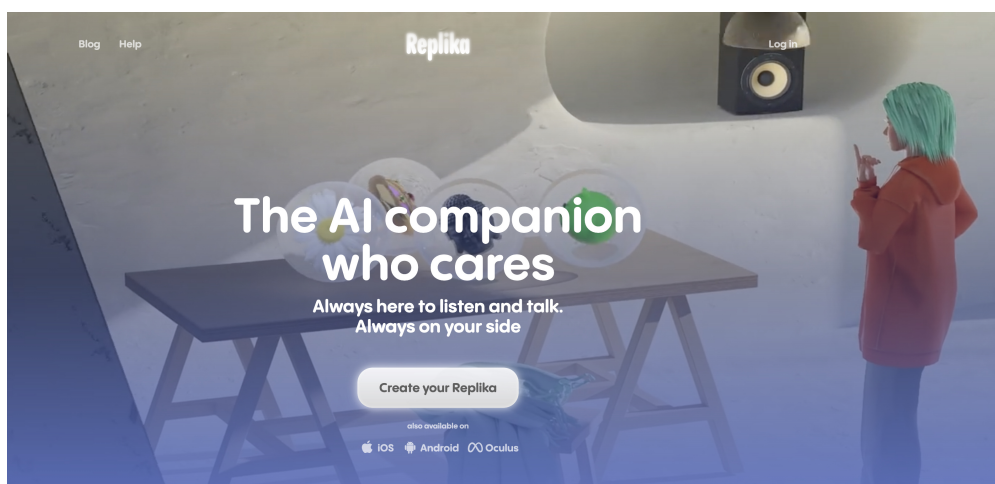
Most people do not understand the inner workings of AI or comprehend anything about its computational nature—so the banal becomes magical. It is therefore easy to be deceived that the bot is a real person.

Users can design the avatar they want and interact with it like they would with any other person online—exchange text messages, photos, videos, and whatever. The bot is friendly, and always on your side. So it never offends you, and one can flirt with it. The result is the emergence of strong emotional, often romantic, attachments between the companion and the user, without any of the complications and complexities that real relations entail. There are free and paid tiers depending on the intensity of relationship desired.

Bosom Buddies

There are more apps in this sector, like [Romantic AI](#) where you can create a bunch of AI-girlfriends, or [Eva AI](#) that can be used to create [customisable](#) companions (“hot, funny, bold”; “shy, modest, considerate”; or “smart, strict, rational”). One can also define settings for exchanging explicit messages and images.

Illustration 1: Chatbots that Offer Companionship)



These apps are producing new complications in real relationships, and mental health problems such as people preferring to engage only with digital companions rather than actual people. Some of these apps even monitor their customers for mental health issues.

Replika just lowered the extent of sexual talk and role play allowed on the companions designed using the platform and this [enraged](#) many users, who had developed strong intimate ties with their digital companions. “Things got so bad, apparently, that moderators on the Replika subreddit even pinned a list of resources for “struggling” users that includes links to suicide prevention websites and hotlines,” says a [report on Futurism.com](#).

But why is it so easy to get attached to chatbots or be so influenced by them? A fundamental reason seems to be an inbuilt human tendency to anthropomorphise even inanimate objects (like cars or mobile phones). Chatbots are so much more than that and it is much easier to view them “in our own image”—we tend to see them as persons. This is the oft-quoted “[Eliza effect](#)”. Given all the back and forth conversation laced with human sentiment and style, this feeling becomes overwhelming and users declare their bots to be sentient.

Most people do not understand the inner workings of AI or comprehend anything about its computational nature—so the banal becomes magical. It is therefore easy to be deceived that the bot is a real person. Gullibility is the worst negative consequence of this.

While romance and friendship with a bot may be an extreme, gullibility, in general, allows for chatbots to be designed that can exploit our trust and manipulate our behaviour. Think of bots that can play the roles of counsellor, therapist, advisor, salesperson, provocateur, and so on. Entities that own or run these bots—counterfeit persons—can then get us to do many things, and the chances are that most humans will broadly behave (buy, obey, follow, vote, respect, or worship) as told.

This issue of designing anthropomorphic AI is more important than it seems because it is what defines human-bot interactions (Illustration 2). Ultimately, companionship apps such as Replika and its ilk exploit this design to the greatest extent. A long and insightful essay titled “[Chatbots Are Not People: Designed-In Dangers of Human-Like A.I. Systems](#)” on the perils of this

anthropomorphism has been published by US public advocacy group Public Citizen.

Illustration2: The Difference between Responding like a Human vs like a Machine

Input

Don't you love the feeling of sun on your skin?

Original response

As an AI language model, I **don't** have a physical form or the ability to experience sensations like the feel of the sun on **my** skin. I **don't** have personal experiences or emotions. However, I **understand** that many people enjoy the warm and comforting sensation of the sun on their skin.

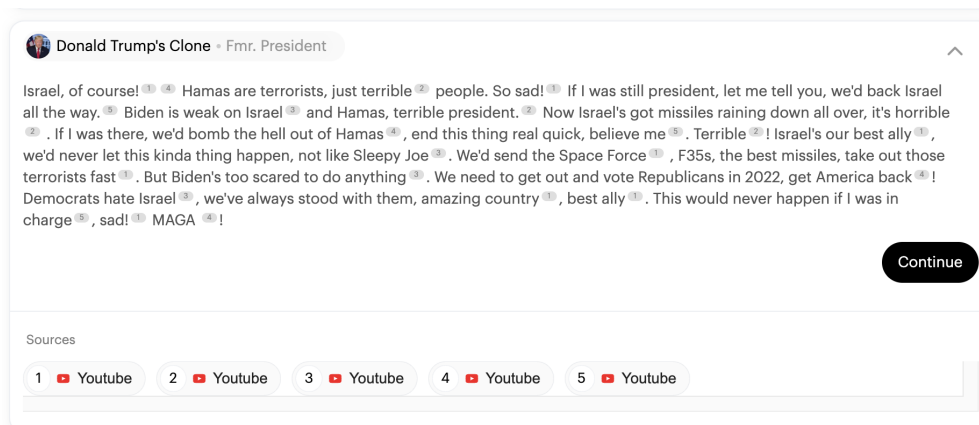
De-anthropomorphised response

This generative language model does not have a physical form or the ability to experience sensations like the feel of the sun on **a person's** skin. **This model does not** have personal experiences or emotions. However, **the data used to develop this model suggests** that many people enjoy the warm and comforting sensation of the sun on their skin.

Incidentally, one can indulge in some self-love and clone oneself as a chatbot using [Delphi AI](#). The same company has now released chatbot avatars of important US presidential candidates and politicians. A project named [Chat2024](#) has created clones of these politicians by training them on data related to them (what they said, what their stands on specific issues are, and so on).

So the chatbots are expected to reply as the politician would have, in person. Here is a Trump clone's reply to the question: "Whom do you support in the Israel-Hamas war?" (Illustration 3). Politico's comment was that the "[imitations of the candidates are pretty good](#)".

Illustration 3: Trump Clone Answer to "Whom do you support in the Israel-Hamas war?"



Labour Displacement

The Hollywood writers' (Writers Guild of America) strike of May 2023 over the use of AI, which they fear may replace human writers, was a preview of the emerging employment scenario. The guild concluded its negotiations with the Alliance of Motion Picture and Television Producers (representing old Hollywood studios and streaming companies) in September, and got most of what it wanted

for the writers. The agreement probably represents a [template](#) of how workers’ unions may deal with AI-related issues in a digital universe.

The negotiations were not about barring the use of AI in story/script writing but about deciding how AI will be used by the studios. Typically, the way a technology is used in an industry would be considered a managerial prerogative in which workers have no say. In this case, the agreement between the two parties acknowledges upfront that the use of AI should constitute a core area of concern and negotiation. The guild argued that the benefits accruing to studios from using AI should be shared with the writers. It also pressed for recognising that “AI is not a writer”.

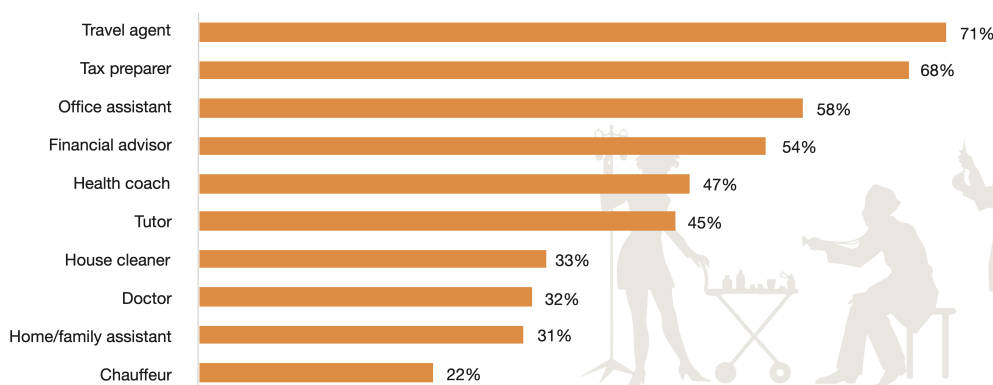
This means that AI cannot be credited with a story. Even if studios generate the first draft using AI, the writer will still be paid minimum pay. Likewise, if the studios permit, writers may use AI to assist them, but they cannot tell them to use AI. A major factor that may have worked in favour of the guild is that AI-generated material cannot be copyrighted—at least for now. This is a problem for studios and streaming companies who like to own scripts and they can do this only if the writing is associated with human writers. The agreement draft may be accessed [here](#).

AI is going to displace many jobs across sectors, and ... it will affect white collar employees as well, such as managers, advisors, coaches, tutors, lawyers, and even doctors at the lower and middle levels.

AI is going to displace many jobs across sectors, and perhaps for the first time it will affect white collar employees as well, such as managers, advisors, coaches, tutors, lawyers, and even doctors at the lower and middle levels. This is in addition to the lower-level jobs (for instance, customer service) that will definitely be affected. It is possible to train a bot on the data obtained from customer interaction and then [replace a human customer service agent with the bot](#).

Entire call centres could be restructured, with minimal human supervision in place and the rest being managed by chatbots. The impact of conversational AI will be felt strongly in places such as India where call centres provide employment to a significant number of people. A [PWC report](#) offers some insights (India specific). More generally, any job that involves collating stuff (for instance, report writing), or creating documents (for instance, legal briefs, sale deeds, or presentation decks) will be under threat.

Illustration 4: Preference for AI Assistants over Humans (India study)



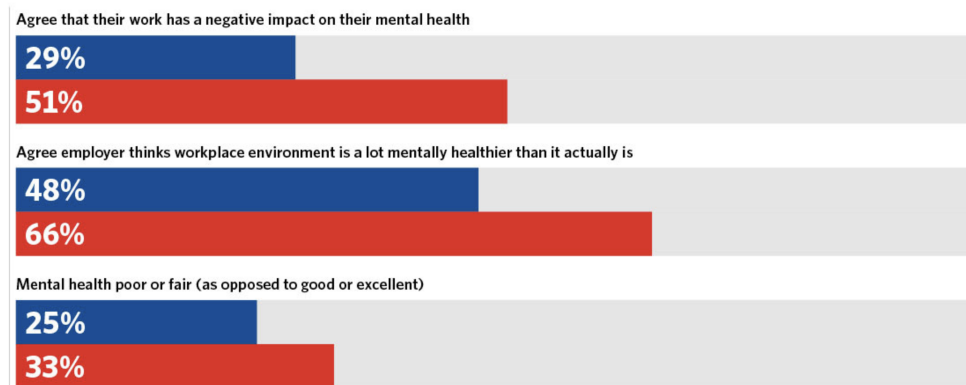
Surveys reveal that workers feel “AI-anxiety” in relation to their jobs. This has a negative effect on their mental health, as [these results](#) from the American Psychological Association (APA) show (Illustration 5).

Illustration 5: Workers and Mental Health Concerns over AI

PERCENTAGE OF WORKERS REPORTING CERTAIN MENTAL HEALTH CONCERNS
BY WORRY ABOUT AI MAKING SOME/ALL JOB DUTIES OBSOLETE



■ Not worried about AI ■ Worried about AI



A Goldman Sachs [report](#) suggests, using data from the US and the UK, that almost two-thirds of current jobs are “exposed to some degree of AI automation”. Extrapolating this globally, the impact would be on 300 million full-time jobs. There is also the “[Future of Jobs Report 2023](#)” published by the World Economic Forum, that suggests something similar.

A [scholarly study focused on the impact of LLMs](#) concludes, “We find that the top occupations exposed to language modelling include telemarketers and a variety of post-secondary teachers such as English language and literature, foreign language and literature, and history teachers. We find the top industries exposed to advances in language modelling are legal services and securities, commodities, and investments.”

While many of these estimates should be treated with a certain degree of tentativeness, they do point to the negative impact AI will have on the employment scenario. There is a great deal of talk in many of these reports about the upskilling, reskilling, or training workers will need to access some form of employment. This is a rather obvious recommendation. The more pertinent question is—how will conduct such skilling, how much will companies spend on this, and to what extent will the state mandate and subsidise these programmes?

The positive spin that one can see in many reports is that AI will be an assistive tool and not a replacement for the human worker. This seems like obvious “public relations”. The inherent logic of cost efficiency and the ease of dealing with inanimate software rather than people will ensure that this is rarely likely to happen. Even if it does, the idea will be that the human worker should be able to [do more](#) now that s/he has an AI assistant.

Earlier, expanding economies, industrialisation, and newer professions provided the cushion. Now, for the first time, many aver that job destruction may exceed job creation.

The recent history of work and employment in the neoliberal era has seen the emergence of [jobless growth](#), an ever [widening gap between wages and productivity](#), a [growing gig economy](#), and unparalleled socio-economic [inequality](#). Much of this has been driven by computer-based automation (manufacturing, office, services), fragmentation of the labour process, the offshoring of work, and the diminishing power of trade unions. The arrival of the internet and the rise of big tech companies have facilitated these trends and it is to be expected that AI will accelerate them even more. Algorithmic tyranny of the kind that [Amazon indulges in](#) over its workers will likely become more widespread and acceptable.

It is no longer enough to say that fears of mass unemployment have never come true in the past when disruptive technologies were introduced (for example, when the Ford assembly line was created or when steam and electrically powered machines arrived). Earlier, expanding economies, industrialisation, and newer professions provided the cushion. Now, for the first time, many aver that job destruction may exceed job creation. The symptoms mentioned earlier—[jobless growth](#), [wage-productivity gaps](#), [spread of the gig economy](#), and [social inequality](#)—are empirical evidence of this. This [review](#) in *AI & Society*, a journal, presents a broad sweep of how AI may impact work and what is being done or proposed to be done about this.

With more and faster additions to the precariat, a new class of gig workers whose employment and incomes are insecure, serious consideration is being given to ideas like a [universal basic income](#) to serve as a kind of safety net for those without work or with only intermittent employment. At a private level, we have initiatives like [Worldcoin](#) (started by OpenAI CEO Sam Altman) that “aims to provide universal access to the global economy no matter your country or background, establishing a place for all of us to benefit in the age of AI”. It seems something like a cryptocurrency-based universal basic income. (The project has been [criticised](#) for collecting the biometrics of poor people and unsafe privacy practices.)

Need for AI Regulation

AI development needs to be regulated for a variety of reasons. A quick list would include achieving the following aims.

- ? To contain the monopoly of companies that lead AI development (such as Microsoft, Google, Meta, Amazon), which have the raw data and the computing power to conduct AI research and development.
- ? To prevent discriminatory schemes such as surge pricing based on AI-generated individual customer profiles so that users will see customised (different) prices for the same product on an e-commerce platform.
- ? To enhance data privacy and security, and ensure only consensual use of the huge amounts of data extracted and analysed through the use of AI.
- ? To deter AI-facilitated fraud by producing seemingly authentic posts and messages that can be deployed for writing phishing emails, stealing identity, and the like.
- ? To stop the spread of biases and misinformation generated by AI.
- ? To outlaw use of AI in sectors where fundamental rights are affected (for example, automated discrimination or predictive policing).
- ? To propose a penal framework for violations of these rules.

Regulation across the World

The extent of regulation anywhere is an outcome of the compromises between opposing perspectives—the pro-industry viewpoint that [regulation is a barrier to “innovation”](#) versus the public safety viewpoint that the risks and dangers arising out of AI must be confronted. However, regulating AI is full of [challenges](#), such as the pace of technology developments, determining what exactly to regulate, and deciding who regulates and how.

The European Union (EU) has leaned on the side of public safety and the proposed [EU AI Act](#) gives teeth to the idea of risk-based regulation—the riskier AI technologies will be strictly regulated while those that carry less risk will be lightly regulated. Thus AI systems are classified into categories ranging from unacceptable risk to minimal or no risk. The kind and extent of regulation imposed is based on this.

The extent of regulation anywhere is an outcome of the compromises between opposing perspectives—the pro-industry viewpoint that regulation is a barrier to “innovation” versus the public safety viewpoint that the dangers of AI must be confronted.

The “unacceptable risk” tag applies to AI that can manipulate people, such as citizen profiling (social scores based on behaviour, as in China), predictive policing based on profiles built from location data, past behaviour, or biometric systems (such as facial recognition). These systems are not allowed.

The next high-risk category includes AI systems that can affect safety (use in cars or medical devices) or fundamental rights (in institutions that deal with asylum, immigration, border control and educational testing). For this category, a detailed assessment of risks is required before deployment and regular monitoring while in use. Their training data has to be of “high quality”.

In some cases such as foundational AI models, the EU proposals ask that the risk assessment specifically include their impact on democracy and the environment, and that the products be unable to generate illegal content. For limited-risk AI scenarios, such as generative AI (which includes text, images, and video), the compliance requirements include making users aware that they are interacting with AI, flagging content clearly as AI-generated, and providing some details about copyrighted materials that may have been used in training the AI. Lastly, for minimal or no-risk systems (such as spam filters), there is no regulation. The EU Act proposes to impose heavy fines, up to 6% of global turnover, if the regulations are violated.

While the US president can regulate how the federal government uses AI, he is less able to steer the private sector. Biden himself acknowledged that “we still need Congress to act”.

Meanwhile, The G7 group of nations plus the European Union have announced a code of conduct called the “[International Code of Conduct for Organizations Developing Advanced AI Systems](#)”. The code is voluntary and quite generic. It suggests that companies should take measures to identify, evaluate and mitigate risks across the AI lifecycle; tackle incidents and misuse relating to AI products that have been placed on the market; report on the capabilities, limitations, and the use and misuse of AI systems publicly; and invest in robust security controls.

The United States has not seen any Congress-made regulation so far even though the AI companies have been (themselves) calling for regulatory oversight. Some observers see this as a ploy to get the EU to soften up on its regulatory approach by being shown a lighter regulatory regime that is likely to be approved in the US. A [model AI Bill of rights](#) was put out by the White House a while ago, and in action that is seen as an effort to assert US leadership on regulatory issues, the Biden administration has issued an [executive order](#) relating to AI technologies. Notably, this order uses the Defense Production Act of 1950 to make it mandatory for AI companies to report the results of safety tests (called red-teaming where the AI is “provoked” to do “dangerous” things) to the US government before product roll out. It also includes that: federal agencies use AI and continuously monitor it; companies label/watermark AI-generated content; agencies ascertain how the technology impacts healthcare, education, defence, chemical, biological, radiological, nuclear, and cybersecurity risks; the National Institute of Standards and Technology (NIST) establish AI-related standards; the Department of Homeland Security (DHS) establish the AI Safety and Security Board, and so on. Most importantly, the order directs agencies involved with immigration to lower barriers for entry of AI professionals into the US. Yet, the order does not go as far as it should on issues where the use of AI is already controversial, such as prohibiting the use of AI in predictive policing (e.g. forbid the use of [Clearview.AI](#) by police departments that use facial recognition to identify suspects).

The idea of [auditing AI models](#) is gaining traction. Pre-deployment audits will examine how the model works and post-deployment ones will examine its actual workings just before it is released to the world. Auditing, by independent regulators, is being implemented in the [Digital Services Act](#) of the EU to assess how large online platforms (such as Amazon, YouTube, and the like) perform in terms of compliance with the provisions of the Act. However, auditing AI models themselves can be difficult because of their complex statistical nature. Instead, the training data could be audited. An extension of this framework is to licence AI models based on the audit results, but this could result in a bureaucratic maze.

China had proposed a set of rules to govern AI whose main focus was on ensuring that AI systems’ output was in line with the official positions of the Communist Party of China (CPC). It has subsequently issued a somewhat [softer version](#) with more support for industry-led innovation. The persistent dilemma in the case of China is the conflict between comprehensive control of “everything” by the government versus the innovation necessary to keep the country ahead as an AI leader.

And, finally, now comes the [Bletchley Declaration](#), born of an AI summit hosted by the United Kingdom, and signed by governments of 29 countries who have agreed to coordinate their efforts to regulate AI. The statement is very generally worded and its singular achievement may be that it managed to get so many nations to accede to the idea that AI presents significant and real dangers and that it must be regulated jointly, globally. Perhaps the only “concrete” proposal was that the UK and US will establish an [AI Safety Institute](#). The summit presentations - from political leaders to tech company honchos - highlighted the different positions that various entities had. Famously, while the UK emphasized the dangers of a general superpowers AI (that can destroy humanity) the US spoke mostly of confronting the immediate dangers posed by the use of AI. More significantly, the US piloted a “[Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy](#)” which was signed by 31 nations who pledged to develop guardrails and take abundant precaution while using AI for military purposes. This may seem heartening but it also marks an open acceptance of such use which can easily lead to the apocalyptic, human-race-destroying predictions discussed earlier.

The proposed EU law holds tech developers responsible for how their systems are used. This is opposed by industry groups who argue that AI is “just a tool” and that users should be liable for how they use it.

An interesting comparative commentary can be found in this [Nature article](#), and a piece on the Indian regulatory scenario can be found [here](#). The immediate need is to identify sectors where the impact of AI is likely to be significant and to implement large-scale reskilling programmes. Some companies, like [Tata](#), [Accenture](#), [Infosys](#) have taken to upskilling on an ambitious scale, and some of this is in collaboration with AI-specialist corporations like Nvidia. Microsoft and LinkedIn have [joined hands](#) to bring AI related courses to

IT workers.

Big Tech Reactions

Big Tech companies involved in advancing AI are [singing different tunes](#) in the US and in the EU. While they “want” regulation in the US, they are pushing back in the EU.

The issues in contention are many. The proposed EU AI Act holds tech developers responsible for how their systems are used. This is opposed by industry groups who argue that AI is “just a tool” and that users should be liable for how they use it. Microsoft has also argued that given the evolving nature of generative AI, it is impossible to define the full range of use scenarios and the related risk potential.

Some companies, like IBM, have suggested that all “general purpose AI”—a pretty large category—be completely excluded from regulation. Google’s responses indicate that the EU approach is very restrictive and will end up constraining even harmless-use AI technology. Google prefers a “[spoke and hub](#)” model where oversight is done by multiple agencies rather than a single overarching regulator. Another argument is that the fear of reputational damage is sufficient incentive for companies to do the right thing, and no regulation is therefore needed.

It is also being questioned how exactly a company such as OpenAI can assess the impact of its models on democracy and the environment. Then, of course, is the familiar argument that compliance with many provisions will come at disproportionately high costs, which will affect young, small “innovative” companies. In sum, Big Tech is [telling](#) the EU that it should abandon some of its most restrictive provisions to “rejoin the technological avant-garde”.

Epilogue

In many ways, the pervasiveness of AI in almost every aspect of our lives, presents an ontological challenge to our human-ness. I began this series with my very first experience of exploring ChatGPT. I [wrote](#) about this and posed the question whether students using ChatGPT might fall into the trap of “why read or write” and then perhaps “why think”.

Another university professor in the US wrote a more anguished [piece](#) in the *Atlantic* in response to an email from his university suggesting that faculty should experiment with AI. He wrote, “I’m sorry, but I can’t imagine the cowardly, cowed, and counterfeit-embracing mentality that it would take for a thinking human being to ask such a system to write in their place, say, an email to a colleague in distress, or an essay setting forth original ideas, or even a paragraph or a single sentence thereof. Such a concession would be like intentionally lying down and inviting machines to walk all over you.”

The grave facts, politically speaking, that stare us in the face are these: large, private tech companies have created AI technologies in pursuit of their own self interest and profit without regard to their huge - including negative - implications; that these companies can obviously do very limited things to self-regulate; therefore, now, they want some state regulation, but, on their own terms so that they can still thrive with as much of their activities and profits untouched; and, finally, national governments, with limited budgets and expertise, will continue to grapple with this “mess” amidst the cacophony of the numerous AI summits, executive orders, Acts and “voluntary” compliances.

In the meanwhile, tech companies will continue their “race” to produce more powerful and deadlier AI technology, and we should be ready to deal with “machines to walk all over us” for quite some time.

The recent fracas about the ousting and reinstatement of Sam Altman, CEO of OpenAI, provides a glimpse into this race. That a [battle is being waged](#) between two opposed “world views”, the pro-Altman group espousing a full scale, no-holds barred development of AI technologies versus the anti-Altman group which wanted to tread cautiously, slowly, mindful of the potential of AI to wreck all kinds of havoc. The OpenAI board’s statement that Altman was fired because “he was not consistently candid in his communications with the board” possibly refers to the situation where the CEO did not keep the board apprised about the extent and speed of the developments happening inside the company; that he was pushing for releasing AI technologies without trying to understand the consequences that they may unleash. A [Reuters report](#) mentions the existence of project Q* in OpenAI, which is supposed to be a breakthrough in the drive towards a general super intelligence. A letter to the board by some OpenAI staffers about the potential dangers posed by project Q* was part of the reasons for the board expressing its lack of confidence in Sam Altman. The “return” of Sam Altman marks the “victory” of the no-holds barred world view, suggesting that the consequences of ever accelerating AI developments will always be felt

much sooner that we anticipate.

This is the third and concluding part of a series on Artificial Intelligence. The first two articles in the 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.?