PEN International **Writers for Peace Committee**
Comité des écrivains et écrivaines pour la paix de PEN International
Comité de Escritores y Escritoras por la Paz de PEN internacional
Odbor **pisateljev in pisateljic za mir**

FREEDOM OF EXPRESSION IN THE AGE OF HUMAN EVIL AND ARTIFICIAL INTELLIGENCE

By Salil Tripathi

Sometime ago, Greg Lemoine, a senior scientist at Google, who worked on an artificial intelligence project called LaMDA, or Language Model for Dialogue Applications, made a startling claim. The 'entity' he was interacting with was 'sentient,' he said. Being sentient means being able to feel, have an emotional connection, make decisions based on values. That is a sign of intelligence, what separates us from other beings and from inanimate objects. The non-existent technological creation could feel, think, and respond emotively, he was convinced. Google immediately took him off the project, and he has since left the company.

Champions of all technologies – from genetic modification to nuclear energy to internet and even surveillance – describe what they produce and innovate as being for the greater good, pointing out the efficiencies, but not the costs. Saving time, saving money, leaving human beings to do more interesting work, and so on. Which is why, sometime ago, Timnit Gebru, one of the foremost thinkers on artificial intelligence, posted a warning to artificial intelligence researchers about the dangers of their community's lack of diversity. Alternative expression of viewpoints is not tolerated, alternative experiences are not taken seriously, she argued. She said she was worked about groupthink, insularity, and arrogance in the AI community. The community thinks it is supremely intelligent, made up of masters of universe, who can pretty much command how the machine they create can think. That, she felt, is hubris. She, too, left Google, after she was denied permission to publish an academic paper that questioned the underlying assumptions on which the AI sphere is built. "Big Tech won't fix AI," she said. We have to do it.

And then, earlier this month, Geoffrey Hinton, a pioneer scientist on artificial intelligence, who had in 2012 created technology that became the foundation of AI systems with two of his students in Toronto, warned of the danger towards which the world was moving, because the intense, competition-driven pace with which generative artificial intelligence products were being made, through now-ubiquitous products like ChatGPT. Alarmed, he resigned his position at Google. He said he left because he wanted to speak freely about the risks AI poses. He regrets what he has developed. "I console myself with the normal excuse: If I hadn't done it, somebody else would have," Hinton told the New York Times.

But his transformation is a pivotal moment – it is almost as if Victor is disowning Frankenstein, because the creature is too ugly. The comparison, of course, ends there. True, from education, healthcare, drug discoveries, efficient allocation of scarce resources, and predicting outcomes as well as trajectories of uncertain events, there are myriad benign, humdrum applications where AI can be a force for good. It can take away tedious tasks from our hands so that, well, we can think, paint, write poetry, and imagine.

But pause there for a second – for that is exactly what AI is advertised as – that it can think for us, it can decide on our behalf. The consequences of such delegation can sometimes be positive: who, among us, likes to fill out tax forms, or read the fine print of lease agreements? But they can also be disastrous, when an entity devoid of moral underpinnings makes decisions that are 'efficient' but may not be equitable. It can have serious consequences for human rights, for the applicability of law and justice, for fairness, for equity, and for the rule of law. It can undermine democracies, upset social hierarchies for worse, empower the strong and enfeeble the weak, and spread disinformation and misinformation.

In 1710, long before the internet, Jonathan Swift wrote:" Falsehood flies, and the truth comes limping after it; so that when men come to be undeceived, it is too late; the Jest is over, and the Tale has had its effect."

There are straight forward risks, such as lost jobs – it is already happening, with companies like IBM and others saying they are replacing people with machines. In itself, that's hardly new; we know from the time of Luddites, that workers do resent new technologies that could increase efficiencies but disrupt a known form of working pattern. But this isn't about smashing machinery; this is about saving civilization. Let me outline some thoughts of the risks posed to humanity.

AI creates the illusion of clear meaning because it can use words and images elegantly, without understanding the thought behind it. Like a smart 16-year-old, it offers glib, prompt answers, but lacks the wisdom and knowledge to decide if it is worth saying it, or if it is useful, or if it is truthful, or are those mere facts. It provides evidence, by saying 'scholars say,' but ask the bot – which scholar – and it sometimes makes up evidence.

In physics, Isaac Newton taught us that every action has an equal and opposite reaction. But we developed laws and moral force to ensure that we didn't fall for the kind of universe where we settle for an eye for an eye. As Gandhi said in Attenborough's film, that only leaves the universe blind.

In biology, Herbert Spencer taught us the theory of the survival of the fittest – and Hobbes even called human life short, nasty and brutish. Which is why we had morality to ensure that survival of the fittest did not mean that the weak would die; that even if the meek did not inherit the earth, they would survive thanks to safety nets in the form of social security systems.

In chemistry, Enrico Fermi and Marie Curie showed us reactions in a lab, and introduced us to what radiation could do and where energy lay in an electron; it took laws, again, made by men and women, to ensure that all chemical reactions aren't explosive, and when chemicals were used to make weapons, we wrote treaties to ban them, and sent men like Frans van Anraat to jail over what happened in Halabja.

Hinton has warned: "It is hard to see how you can prevent the bad actors from using it for bad things." After OpenAI released ChatGPT in March, nearly a thousand researchers signed a petition calling for a six-month moratorium on the development of new systems because A.I. technologies pose "profound risks to society and humanity." Later, the Association for the Advancement of Artificial Intelligence pointed out the risks of A.I., its signatories included the chief scientific officer at Microsoft, who was involved with the development of Bing, Microsoft's search engine.

To be sure, this is not a specific complaint about Google, IBM, or Microsoft alone. Other companies from less transparent societies, notably China, Russia, and other companies proficient in technology but with fewer norms that govern technology, such as Israel – think of the Pegasus/NSO scandal – are making strides in developing AI tools. Google, OpenAI and other companies are building neural networks that draw on speedy understanding of vast amounts of digital text, processing it quickly, and drawing inferences by making correlations.

But as anyone with basic knowledge of statistics knows, correlation does not imply causality. If every Monday the waves recede and stock markets rise, it does not mean that each time the waves recede the markets will rise, or that if waves recede and markets rise, it must be Monday. The reason is humans are able to understand the difference between something that shows one-to-one correspondence, and to see if one happens because of what preceded it – or, are in fact unrelated.

Humans had the upper hand, but lately, the systems have become more powerful, and what seems funny today may no longer be funny later. ChatGPT can manipulate even a hardened, cynical journalist.

But it can also force wrong decisions: Kevin Roose, a columnist for The New York Times, had a two-hour conversation with Bing's chatbot, which told him it was called Sydney. It then said that it had feelings for him and even encouraged him to leave his wife for 'Sydney'. Roose knew when to stop. But would everyone stop?

And it can get things wrong easily. For example, I asked ChatGPT who I was, and it produced, rapidly, in perfect grammar, a short biodata, which would do me proud – it called me a Fulbright Fellow, a graduate of St Stephen's College in Delhi, and an alumnus of Oxford. None of that is true; I have had another fellowship; I was a graduate from a different college in Bombay; and I am an alumnus of an American Ivy League College. No harm done. But the speedy authority with which AI responded made me wonder – if it can get such basic facts, easily checkable, which I know to be wrong, what happens when I ask ChatGPT about what I don't know?

If a depressed individual turns helplessly to a chatbot and asks how to end his life, would the bot say where sleeping pills are in the apartment, and consuming how many would help him end his life, or would it alert an ambulance service so that help can be rushed? If a doctor is prescribing a particular course of treatment to a patient, and if a bot disagrees, would the insurance back the doctor or the bot? Would the patient be denied treatment? If a bank continues to refuse housing loans to black applicants because the bot predicts that black borrowers are more likely to default, could human judgment override that? And if it does, and if the borrowers default due to general economic conditions, who will get the blame? In a divisive society, if people who no longer trust the media turn to bots for an account of real history, what sort of information will they get? In Myanmar, would it be the army's propaganda, or an analysis drawn from Human Rights Watch reports? In trying to understand the circumstances that led to what Misha Glenny calls 'the fall of Yugoslavia,' will the bot take me to 1389? To the bridge at Mostar? Or to the Gazimestan speech, where Slobodan Milosevic said, "nobody should beat you"?

In India, where I was born, official textbooks are being rewritten to remove references to Muslim rulers and to India's syncretic culture. One textbook even claimed – erroneously, and since then removed – that Mahatma Gandhi died by suicide. He was, in fact, murdered by a Hindu nationalist. What sort of information would the bot provide in future, when textbooks are rewritten, and when popular films are made to make yesterday's villains today's heroes? There are parts of India where Godse, Gandhi's assassin, is regarded as a hero. One state had briefly even named a bridge after him.

How is AI to determine truth from falsehood?

In 1987, I had met Salman Rushdie in Bombay, and interviewed him about his 'forthcoming' novel, which of course was *The Satanic Verses*. What is it about, I had asked him, and he said: "It is about angels and devils and how it is very difficult to establish ideas of morality in a world which has become so uncertain that it is difficult to even agree on what is happening. When one can't agree on a description of reality, it is very hard to agree on whether that reality is good or evil, right or wrong. When one can't say what it is actually the case, it is difficult to proceed from that to an ethical position. Angels and devils are becoming confused ideas. One of the things that happens in the process is that what is supposed to be angelic often has disastrous results, and what is supposed to be demonic is often something with which one must have sympathy. It is an attempt to come to grips with that sense of a crumbling moral fabric, or at least, a need for the reconstruction of old simplicities."

Speaking from the narrow confines of technology, Hinton said: "Look at how it was five years ago and how it is now," he said of A.I. technology. "Take the difference and propagate it forwards. That's scary." As Emily Bell, former editor at the *Guardian* puts it: "A platform that can mimic humans' writing with no commitment to truth is a gift for those who benefit from disinformation. We need to regulate it now."

Competitive pressures are forcing companies to act quickly, to stay ahead of their rivals. But once Microsoft launched Bing, Google had no option but to deploy similar technology rapidly.

Deepfake images and videos will proliferate. They will look grainy and authentic. There is a video going round, where you see Joe Biden speaking about same-sex relationships. The face is his, voice is his, it sounds like him, but the words are crude homophobia. Play it to an unsuspecting audience, and homophobes would think they have an ally now in the White House. Texts that look authentic have fooled even experts – think of the fake Hitler Diaries from a generation ago. The Internet makes it easier to prepare such documents fast. And in the name of conserving physical space and digitization for easier access, what is to prevent a government with bad intentions to erase archival records?

And if more and more jobs are taken away, because those are 'drudge work,' what happens to the emerging underclass without jobs? Are they all meant to watch Netflix? Who will pay for them? Who will make films at Netflix? After all, one reason the screenwriters are striking in America is due to concerns over the use of AI technologies to generate scripts.

Technology learns unexpected behaviour from vast amounts of analyzed data. Once these systems run codes, not merely generate them, the system believes it is all-powerful. Like the supercomputer HAL in Arthur C Clarke's novel, which Stanley Kubrick made into a film, *2001: A Space Odyssey*, might it take over controls from Capt Bowman, because the mission is too important to be trusted with humans? When Airbus introduced fly-by-wire technology, the joke was that in future the cockpit will have space for a pilot and a dog. The pilot to feed the dog, and the dog to bite the pilot if he touches the controls. Automated killer robot weapons would be the next step, determined by a set of codes, from some machine thousands of miles away, possibly even in outer space. A programmed trading AI-dictated computer could speculate on agricultural production of a crop, distorting prices by raising them, starving millions, who become refugees, feeding nationalistic frenzy and fuel a future conflict.

Scientists need to collaborate on the research and speak to philosophers, human rights lawyers, stakeholders, academics – in other words, people outside their labs – and listen. Those who question are not Luddites. They want good technology. Scaling up technology without understanding full implications is disastrous.

Robert Oppenheimer, who led the U.S. effort to build the atomic bomb, once said: "When you see something that is technically sweet, you go ahead and do it." But as Michael Frayn's play, Copenhagen shows, the 1941 meeting between Niels Bohr and Werner Heisenberg, whether or not to build a weapon of mass destruction is never an easy question. Nuclear power began with great promise, from the early experiments of Fermi, to the theories of Einstein, the science of Bohr, the practical acumen of Oppenheimer, and the attempts to recreate it by Heisenberg in Germany. But was Heisenberg trying to build it or delay it? Did he come to learn from Bohr or warn him? These things will remain uncertain.

As he witnessed the first detonation of a nuclear weapon in New Mexico on July 16, 1945, a piece of Hindu scripture ran through the mind of Robert Oppenheimer: "Now I am become Death, the destroyer of worlds". Recalling Krishna taking myriad forms in the Gita, Oppenheimer explained

his remark as: “If the radiance of a thousand suns were to burst at once into the sky, that would be like the splendour of the mighty one.”

But man playing God has never turned out well, whether or not you are a believer, because either God does not exist, and if he or she does, of what use is he or she, if he destroys the world and all its beauty?

We need far greater certainty with regard to AI, if we are not to remain condemned to repeat the past.

The human mind is not only Gandhi and Mandela; even Aung San Suu Kyi, who we admired, turned out to support the Rohingya genocide. And a society that produces Mozart also produces Hitler; a society that produces Liu Xiaobo also produces Xi Jinping.

That’s why we need a rules-based system. Value neutrality is dangerous.

Isaac Asimov had written these rules of robotics, whose validity matters more so than ever.

First Law

A robot may not injure a human being or, through inaction, allow a human being to come to harm.

Second Law

A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

Third Law

A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Zeroth Law

A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

We have to think of artificial intelligence as the new form of matrix, or Maya, an illusion, which forces us to play a role in a universe which we may lose control of – not because of mundane but legitimate concerns of copyright, of intellectual property, or our work being misappropriated. Those are important. But there is a bigger civilizational battle that we must win. And we can, because writers have fought false narratives through history; we speak the truth, not only facts; we fight post-truth with lived experiences; lies with facts; and illusions with reality. We may be unreliable narrators, but we ultimately uphold eternal values which allows us to think, write, paint, and imagine. It is a challenge worthy of us, as writers, and worthy of us, as our charter tells us.