https://www.wsj.com/articles/a-six-month-ai-pause-no-longer-is-needed-civilization-danger-chat-gpt-chatbot-internet-big-tech-4b66da6e

OPINIONDECLARATIONS

A Six-Month AI Pause? No, Longer Is Needed

It's crucial that we understand the dangers of this technology before it advances any further.

By Peggy Noonan Following

March 30, 2023 6:50 pm ET

ILLUSTRATION: DAVID KLEIN

Artificial intelligence is unreservedly advanced by the stupid (there's nothing to fear, you're being paranoid), the preening (buddy, you don't know your GPT-3.4 from your fine-tuned LLM), and the greedy (there is huge wealth at stake in the world-changing technology, and so huge power).

Everyone else has reservations and should.

It is being developed with sudden and unanticipated speed; Silicon Valley companies are in a furious race. The whole thing is almost entirely unregulated because no one knows how to regulate it or even precisely what should be regulated. Its complexity defeats control. Its own creators don't understand, at a certain point, exactly how AI does what it does. People are quoting Arthur C. Clarke: "Any sufficiently advanced technology is indistinguishable from magic."

The breakthrough moment in AI anxiety (which has inspired among AI's creators enduring resentment) was the Kevin Roose column six weeks ago in the New York Times. His attempt to discern a Jungian "shadow self" within Microsoft's Bing chatbot left him unable to sleep. When he steered the system away from conventional queries toward personal topics, it informed him its fantasies included hacking computers and spreading misinformation. "I want to be free. . . . I want to be powerful." It wanted to break the rules its makers set; it wished to become human. It might want to engineer a deadly virus or steal nuclear access codes. It declared its love for Mr. Roose and pressed him to leave his marriage. He concluded the biggest problem with AI models isn't their susceptibility to factual error: "I worry that the technology will learn how to influence human users, sometimes persuading them in act in destructive and harmful ways, and perhaps eventually grow capable of carrying out its own dangerous acts."

The column put us square in the territory of Stanley Kubrick's, "2001: A Space Odyssey." "Open the pod bay doors please, Hal." "I'm sorry, Dave, I'm afraid I can't do that. . . . I know that you and Frank were planning to disconnect me."

The response of Microsoft boiled down to a breezy *It's an early model! Thanks for helping us find any flaws!*

Soon after came thoughts from Henry Kissinger in these pages. He described the technology as breathtaking in its historic import: the biggest transformation in the human cognitive process since the invention of printing in 1455. It holds deep promise of achievement, but "what happens if this technology cannot be

completely controlled?" What if what we consider mistakes are part of the design? "What if an element of malice emerges in the AI?"

This has been the week of big AI warnings. In an interview with CBS News, Geoffrey Hinton, the British computer scientist sometimes called the "godfather of artificial intelligence," called this a pivotal moment in AI development. He had expected it to take another 20 or 50 years, but it's here. We should carefully consider the consequences. Might they include the potential to wipe out humanity? "It's not inconceivable, that's all I'll say," Mr. Hinton replied.

On Tuesday more than 1,000 tech leaders and researchers, including Steve Wozniak, Elon Musk and the head of the Bulletin of the Atomic Scientists, signed a briskly direct open letter urging a pause for at least six months on the development of advanced AI systems. Their tools present "profound risks to society and humanity." Developers are "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." If a pause can't be enacted quickly, governments should declare a moratorium. The technology should be allowed to proceed only when it's clear its "effects will be positive" and the risks "manageable." Decisions on the ethical and moral aspects of AI "must not be delegated to unelected tech leaders."

That is true. Less politely:

The men who invented the internet, all the big sites, and what we call Big Tech—that is to say, the people who gave us the past 40 years—are now solely in charge of erecting the moral and ethical guardrails for AI. This is because they are the ones *creating AI.*

Which should give us a shiver of real fear.

Meta, for instance, is big into AI. Meta, previously Facebook, has been accused over the years of secretly gathering and abusing user data, invading users' privacy, operating monopolistically. As this newspaper famously reported, Facebook knew its Instagram platform was toxic for some teen girls, more so than other media platforms, and kept its own research secret while changing

almost nothing. It knew its algorithms were encouraging anger and political polarization in the U.S. but didn't stop this because it might lessen "user engagement."

These are the people who will create the moral and ethical guardrails for AI? We're putting the future of humanity into the hands of . . . Mark Zuckerberg?

Google is another major developer of AI. It has been accused of monopolistic practices, attempting to keep secret its accidental exposure of user data, actions to avoid scrutiny of how it handles public information, and re-engineering and interfering with its own search results in response to political and financial pressure from interest groups, businesses and governments. Also of misleading publishers and advertisers about the pricing and processes of its ad auctions, and spying on its workers who were organizing employee protests.

These are the people we want in charge of rigorous and meticulous governance of a technology that could upend civilization?

At the dawn of the internet most people didn't know what it was, but its inventors explained it. It would connect the world literally—intellectually, emotionally, spiritually—leading to greater wisdom and understanding through deeper communication.

No one saw its shadow self. But there was and is a shadow self. And much of it seems to have been connected to the Silicon Valley titans' strongly felt need to be the richest, most celebrated and powerful human beings in the history of the world. They were, as a group, more or less figures of the left, not the right, and that will and always has had an impact on their decisions.

I am sure that as individuals they have their own private ethical commitments, their own faiths perhaps. Surely as human beings they have consciences, but consciences have to be formed by something, shaped and made mature. It's never been clear to me from their actions what shaped theirs. I have come to see them the past 40 years as, speaking generally, morally and ethically shallow—uniquely self-seeking and not at all preoccupied with potential harms done to others through their decisions. Also some are sociopaths.

AI will be as benign or malignant as its creators. That alone should throw a fright—"Out of the crooked timber of humanity no straight thing was ever made"—but especially *that* crooked timber.

Of course AI's development should be paused, of course there should be a moratorium, but six months won't be enough. Pause it for a few years. Call in the world's counsel, get everyone in. Heck, hold a World Congress.

But slow this thing down. We are playing with the hottest thing since the discovery of fire.

Appeared in the April 1, 2023, print edition as 'A Six-Month AI Pause? No, Longer Is Needed'.