Exec tells first UN council meeting that big tech can't be trusted to guarantee AI safety

July 19 2023, by Edith M. Lederer

In this photo provided by the United Nations Photo, Jack Clark shown on screen, co-founder of Anthropic, briefs the first ever Security Council meeting on artificial intelligence (AI), Tuesday, July 18, 2023, at U.N. headquarters. This meeting, convened by the United Kingdom, addresses the topic "Artificial intelligence: opportunities and risks for international peace and security." The Secretary-General delivered remarks during the debate stating, "I urge the Council to approach this technology with a sense of urgency, a global lens, and a learner's mindset." Credit: Eskinder Debebe/UN Photo via AP

The handful of big tech companies leading the race to commercialize AI can't be trusted to guarantee the safety of systems we don't yet understand and that are prone to "chaotic or unpredictable behavior," an artificial intelligence company executive told the first U.N. Security Council meeting on AI's threats to global peace on Tuesday.

Jack Clark, co-founder of the AI company Anthropic, said that's why the world must come together to prevent the technology's misuse.

Clark, who says his company bends over backwards to train its AI chatbot to emphasize safety and caution, said the most useful things that can be done now "are to work on developing ways to test for capabilities, misuses and potential safety flaws of these systems." Clark left OpenAI, creator of the best-known ChatGPT chatbot, to form Anthropic, whose competing AI product is called Claude.

He traced the growth of AI over the past decade to 2023 where new AI systems can beat military pilots in air fighting simulations, stabilize the plasma in nuclear fusion reactors, design components for next generation semiconductors, and inspect goods on production lines.

But while AI will bring huge benefits, its understanding of biology, for example, may also use an AI system that can produce biological weapons, he said.

Clark also warned of "potential threats to international peace, security and global stability" from two essential qualities of AI systems—their potential for misuse and their unpredictability "as well as the inherent fragility of them being developed by such a narrow set of actors."

Clark stressed that across the world it's the tech companies that have the

sophisticated computers, large pools of data and capital to build AI systems and therefore they seem likely to continue to define their development

In a video briefing to the U.N.'s most powerful body, Clark also expressed hope that global action will succeed.

He said he's encouraged to see many countries emphasize the importance of safety testing and evaluation in their AI proposals, including the European Union, China and the United States.

Right now, however, there are no standards or even best practices on "how to test these frontier systems for things like discrimination, misuse or safety," which makes it hard for governments to create policies and lets the private sector enjoy an information advantage, he said.

"Any sensible approach to regulation will start with having the ability to evaluate an AI system for a given capability or flaw," Clark said. "And any failed approach will start with grand policy ideas that are not supported by effective measurements and evaluations."

With robust and reliable evaluation of AI systems, he said, "governments can keep companies accountable, and companies can earn the trust of the world that they want to deploy their AI systems into." But if there is no robust evaluation, he said, "we run the risk of regulatory capture compromising global security and handing over the future to a narrow set of private sector actors."

In this photo provided by United Nations Photo, a wide view of the first ever Security Council meeting on artificial intelligence (AI) held Tuesday, July 18, 2023, at U.N. headquarters. This meeting, convened by the United Kingdom, addresses the topic "Artificial intelligence: opportunities and risks for international peace and security." The Secretary-General delivered remarks during the debate stating, "I urge the Council to approach this technology with a sense of urgency, a global lens, and a learner's mindset." Credit: Eskinder Debebe/U.N. Photo via AP

Other AI executives such as OpenAI's CEO, Sam Altman, have also called for regulation. But skeptics say regulation could be a boon for deep-pocketed first-movers led by OpenAI, Google and Microsoft as smaller players are elbowed out by the high cost of making their large language models adhere to regulatory strictures.

U.N. Secretary-General Antonio Guterres said the United Nations is "the ideal place" to adopt global standards to maximize AI's benefits and mitigate its risks.

He warned the council that the advent of generative AI could have very serious consequences for international peace and security, pointing to its potential use by terrorists, criminals and governments causing "horrific levels of death and destruction, widespread trauma, and deep psychological damage on an unimaginable scale."

As a first step to bringing nations together, Guterres said he is appointing a high-level Advisory Board for Artificial Intelligence that will report back on options for global AI governance by the end of the year.

The U.N. chief also said he welcomed calls from some countries for the creation of a new United Nations body to support global efforts to govern AI, "inspired by such models as the International Atomic Energy Agency, the International Civil Aviation Organization, or the Intergovernmental Panel on Climate Change."

Professor Zeng Yi, director of the Chinese Academy of Sciences Brain-inspired Cognitive Intelligence Lab, told the council "the United Nations must play a central role to set up a framework on AI for development and governance to ensure global peace and security."

Zeng, who also co-directs the China-UK Research Center for AI Ethics and Governance, suggested that the Security Council consider establishing a working group to consider near-term and long-term challenges AI poses to international peace and security.

In his video briefing, Zeng stressed that recent generative AI systems "are all information processing tools that seem to be intelligent" but don't have real understanding, and therefore "are not truly intelligent."

And he warned that "AI should never, ever pretend to be human," insisting that real humans must maintain control especially of all weapons systems.

Britain's Foreign Secretary James Cleverly, who chaired the meeting as the UK holds the council presidency this month, said this autumn the United Kingdom will bring world leaders together for the first major global summit on AI safety.

"No country will be untouched by AI, so we must involve and engage the widest coalition of international actors from all sectors," he said. "Our shared goal will be to consider the risks of AI and decide how they can be reduced through coordinated action."