In many ways, AGI-driven robots can share characteristics of the “loss of control” risks inherent to virtual, agentic AI (Bengio, 2023). OpenAI describes the agenticness of an AI model as an “impact multiplier” (Shavit et al., 2023). AI models are increasingly able to take input and interact with the world in a generalized way—although often with unexpected or harmful consequences (Zittrain, 2024). The consequences of loss of control would be further amplified if an AI agent deployed by an adversary state were embodied and able to influence the physical world with its generalized capabilities. The agent could constitute a robotic insider threat in any location where it was used, potentially allowing an attacker to be one software update or hack away from remotely executing a wide variety of actions in the physical world. A sophisticated adversary—e.g., a nation-state or even a rogue AGI agent—could conceivably exploit a vulnerability common to many robots to issue malicious instructions that would be carried out with intelligent autonomy. This could allow an attacker to use proliferated robotic platforms to execute a wide-ranging attack on the United States—both to disrupt and to destroy infrastructure and terrorize citizens.

References

Bengio, Yoshua, “How Rogue AIs May Arise,” yoshuabengio.org, webpage, May 22, 2023. As of December 17, 2024:

https://yoshuabengio.org/2023/05/22/how-rogue-ais-may-arise/

Shavit, Yonadav, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O’Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, et al., “Practices for Governing Agentic AI Systems,” OpenAI, December 14, 2023. As of December 18, 2024: https://openai.com/index/practices-for-governing-agentic-ai-systems/

Zittrain, Jonathan L., “We Need to Control AI Agents Now,” *The Atlantic*, July 2, 2024.