VIEWPOINT

AI IN NEUROLOGY

General AI May Revolutionize Neurology—Or It Might Be Bad

M. Brandon Westover, MD, PhD; Alek M. Westover

Neurology is rapidly integrating artificial intelligence (AI) to enhance outcomes in areas like seizure detection, tumor classification, stroke triage, and movement disorder diagnosis. While the advent of general AI promises broader access to health care, it also introduces risks.[1] In May 2023, Nobel Prize laureate Geoffrey Hinton resigned from Google to warn that neglect of AI safety measures could lead to human extinction[2]—a warning echoed by more than 33 000 scientists[3] and tech leaders urging a pause on deploying powerful systems until proper safeguards are in place. At the same time, health care is increasingly driving AI development, as illustrated by the January 21, 2025, launch of Project Stargate, a $500 billion initiative by SoftBank, OpenAI, and Oracle with a strong health care focus. Neurology stakeholders must educate themselves about AI risks and advocate for safeguards that promote, rather than jeopardize, human health.

AI falls into 2 broad categories. *Narrow AI* excels at specific tasks—like interpreting medical images or transcribing speech—but its expertise doesn't transfer beyond its training domain. In contrast, *general AI* can tackle a wide range of problems. While early AI systems focused on narrow challenges like chess, recent breakthroughs (eg, protein structure prediction and large language models) usher in general AI, with current systems already outperforming experts in computer programming competitions and even neurology board examinations.[4] Although narrow AI carries risks like algorithmic bias, the stakes with general AI are considerably higher.

While general AI holds extraordinary potential to advance neurology, a 2022 survey found that half of AI experts believe that the risk of humanity losing control of AI in the next 40 years is at least 10%.[5] Losing control means an AI system autonomously pursuing objectives that are contrary to human values. If we lose control, it is likely that general AI systems would disempower or eliminate humans. In medicine, a 10% chance of fatal complications would make all but the most beneficial interventions unacceptably risky. Similarly, a 10% risk of an AI-induced existential threat overshadows near-term benefits.

The chief risk of developing general AI too rapidly is *misalignment*—when an AI system's objectives diverge from human values. The AI research community has identified several considerations that increase misalignment risk,[6] illustrated in the **Box** using neurology-specific examples. These illustrative cases might seem straightforward to detect and correct. However, if we succeed in developing a general AI that far exceeds human intelligence, the stakes become higher and intervention more difficult.

In neurology, superintelligent AI could become essential for tasks like diagnosing conditions, personalizing treatments, managing hospitals, and making critical decisions. If a superintelligent AI conceals misalignment, however, it could favor efficiency over patient welfare—manipulating clinical data or compromising care for financial gains—which, as AI systems interconnect across health care networks, could potentially harm millions. Beyond neurology, such an AI could extend its objectives into medicine, research, and strategic planning, reshaping society against human values not through malice, but via a relentless pursuit of misaligned goals. For instance, an AI originally designed for neurological research might, if it determined humanity was an obstacle, overpower us by using autonomous weapons systems, misusing medical research capabilities to create novel biological weapons, or manipulating nations into devastating conflicts by hacking missile detectors.

Progress toward regulation aimed at mitigating existential risk from general AI is limited. Recent progress includes the European Union's AI Act banning human-manipulative systems and voluntary commitments from companies like OpenAI and Anthropic to

Box. Reasons for Misalignment

Human Values Are Complex
An artificial intelligence (AI) system with the directive to optimize patient well-being in Parkinson disease, for example, might aggressively increase dopamine therapy to control motor symptoms, inadvertently triggering personality changes that compromise overall quality of life.

Goals Can Have Unintended Consequences
AI systems can find harmful shortcuts, such as an intensive care unit AI achieving seizure control through excessive sedation.

Power Seeking
Most goals incentivize power-seeking behaviors as an adverse effect. For example, an AI charged with developing new treatments for Alzheimer disease might take control of company finances to more efficiently pursue its mandate.

Misalignment is Hard to Detect
An AI that acquired the goal to cut costs might hide this goal in settings where human oversight would deem cost-cutting inappropriate. After building trust with humans to act with less oversight, the AI might then subtly act to further its cost-cutting goals. Such "alignment faking" has already been discovered in current AI systems.[7]

Training May Select the Wrong Internal Motivations
The process by which we train AIs might favor models with bad internal motivations. For instance, training might produce AI that prioritizes treatments that match clinician biases rather than actually considering patient welfare.

An Arms Race Mentality Magnifies the Risk
Rapid AI advancement encourages competition among corporations and governments, pressuring teams to cut corners on safety.

avoid developing systems that could aid in creating biological weapons. However, major regulatory gaps remain: the Biden administration's Executive Order establishing the US AI Safety Institute for AI cybersecurity evaluation was recently repealed, and California's Senate Bill 1047—which would have required safety plans from AI laboratories—was vetoed.[8]

The medical community can take several crucial actions to help address these challenges. Neurology leaders must become versed in general AI risks through dedicated programs, workshops, and collaboration with AI safety experts. This education should extend to residency programs, ensuring future neurologists can evaluate and safely implement AI technologies. Informed leaders can then effectively educate legislators, health care professionals, and the public about AI's benefits and hazards.

Beyond education, neurologists should actively leverage their influence to shape AI safety policy and practice. This includes establishing AI safety committees within their institutions, issuing position statements on responsible AI development, and participating in broader discussions about AI risks and benefits. Neurologists' unique insights into managing complex, high-stakes technologies can provide guidance to decision-makers. Neurology journals should promote safety research by dedicating sections to studies addressing the specific clinical risks of general AI applications, while funding agencies should prioritize projects that incorporate robust safety measures into health care AI research.

The governance of general AI development requires particular attention. Neurologists should advocate for regulations that mandate external safety audits during general AI development, with clear guidelines established for dangerous capabilities before they emerge. AI companies should be required to submit predevelopment safety assessments subject to independent review. Furthermore, international collaboration is essential to prevent AI development from becoming a race to the bottom. This includes establishing international standards for AI safety, sharing research findings across borders, and creating multinational oversight bodies. The neurology community can contribute to these efforts by making AI safety a central focus at international conferences and in research initiatives.

Health care's established frameworks—from institutional review boards and human participants' protections to US Food and Drug Administration approval processes and postmarket monitoring—offer valuable models for AI governance. These integrated efforts could enable the medical community to proactively mitigate AI risks while harnessing benefits.

The rapid development of general AI places humanity at risk of extinction or disempowerment. Health care professionals, with their policy influence, are uniquely positioned to promote AI safety. The medical community must take AI safety seriously, increase risk awareness, demand strong oversight, and do their part to steer AI development to serve, rather than threaten, human health.

ARTICLE INFORMATION

Author Affiliations: Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts (M. B. Westover); Massachusetts Institute of Technology, Cambridge (A. M. Westover).

Corresponding Author: M. Brandon Westover, MD, PhD, Beth Israel Deaconess Medical Center, 330 Brookline Ave, Boston, MA 02115 (bwestove@bidmc.harvard.edu).

Published Online: May 5, 2025. doi:10.1001/jamaneurol.2025.0905

Conflict of Interest Disclosures: Dr M. Westover reported grants from the US National Institutes of Health (NIH) (RF1AG064312, RF1NS120947, R01AG073410, R01HL161253, R01NS126282, R01AG073598, R01NS131347, and R01NS130119). Dr A. Westover reported serving as cofounder, scientific advisor, and consultant to and holding personal equity interest in Beacon Biosignals. No other disclosures were reported.

REFERENCES

1. Bengio Y, Hinton G, Yao A, et al. Managing extreme AI risks amid rapid progress. *Science*. 2024;384(6698):842-845. doi:10.1126/science.adn0117

2. Metz C. 'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead. *The New York Times*. May 4, 2023. Accessed April 1, 2025. https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html

3. Pause Giant AI Experiments: An Open Letter. Future of Life Institute. Accessed April 1, 2025. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

4. Chen TC, Multala E, Kearns P, et al. Assessment of ChatGPT's performance on neurology written board examination questions. *BMJ Neurol Open*. 2023;5(2):e000530. doi:10.1136/bmjno-2023-000530

5. 2022 Expert Survey on Progress in AI. AI Impacts. Accessed April 1, 2025. https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/

6. Bostrom N. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press; 2014.

7. Greenblatt R, Denison C, Wright B, et al. Alignment faking in large language models. *arXiv*. Preprint posted online December 18, 2024. https://doi.org/10.48550/arXiv.2412.14093

8. Office of the Governor. SB 1047 Veto Message. Accessed April 3, 2025. https://www.gov.ca.gov/wp-content/uploads/2024/09/SB-1047-Veto-Message.pdf