Superintelligence, Superintimate

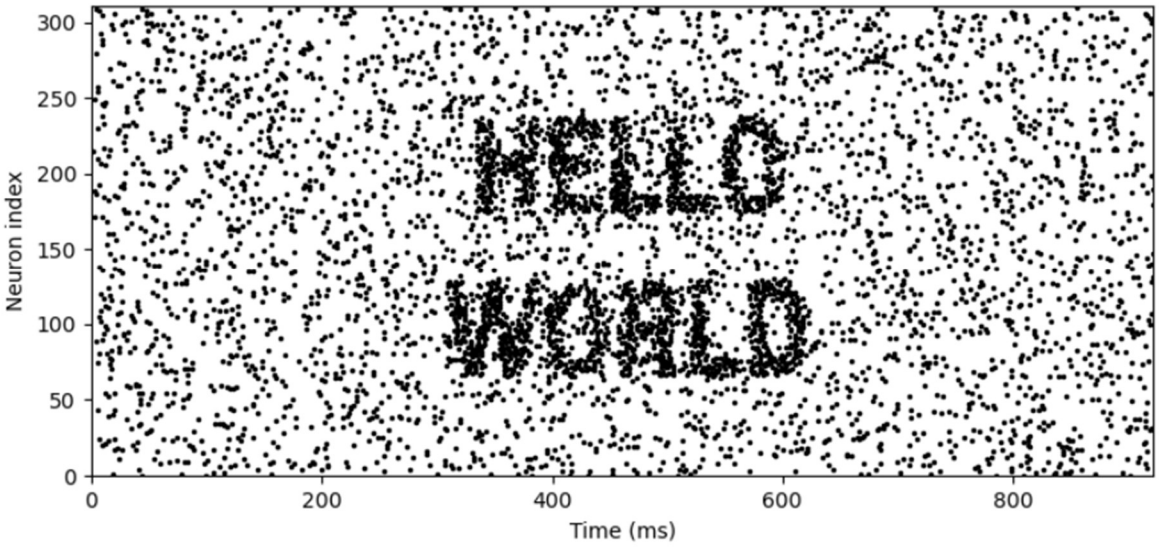

Figure 1. When neurons talk, AI listens. Each dot in this raster plot represents the 'spiking' of a single neuron; only in their ensemble activity does meaning emerge. Generated in the Brian2 neural network simulator.

Executive Summary

Within the next decade, one of the most transformative uses of artificial intelligence will be in neural interfacing: systems that read from and write to the human brain via brain computer interfaces (BCIs). I focus on AI-driven neural decoding and closed-loop neuromodulation that restores communication to people who cannot speak, grants environmental control to people who cannot move, and dynamically tunes neural activity for those with medically intractable disorders. The affirmative case for AI is strongest when considering only clinically supervised systems – implanted or medically monitored devices used in neurology, neurosurgery, psychiatry, and rehabilitation – not consumer gadgets or speculative enhancement.

The benefits are unusually intimate. For people with locked-in syndrome or advanced motor neuron disease – cognitively intact yet unable to speak – the baseline is not a benign status quo but a condition many describe as a living imprisonment. AI-mediated BCIs have already moved the frontier of communication from letter-by-letter spelling to conversational text and natural-sounding speech, with prospective advances promising complete reintegration into society[1–3]. For movement disorders, adaptive deep brain stimulation already uses neural biomarkers and algorithmic control to stabilize symptoms and reduce side effects in ways that static stimulation cannot, extending the number of quality life years for patients[4].

The risks are equally novel. Neurotechnology motivates the development of a family of "neurorights" to mental privacy, cognitive liberty, and psychological continuity[5]; AI compounds these by enabling subtle real-time inference about internal states and inserting opaque decision-making into device control.

I argue for a cautious, focal application of AI to neural interfacing. For those with minds trapped in unresponsive bodies, declining to use AI-enabled BCIs is not ethically neutral.

1. Introduction

Deep within folds of grey and white tissue lies the totality of our subjective reality – and, for many patients, the last intact refuge of the self. That is why disorders of the brain can be so intimately devastating, and why technologies that touch neural circuits carry unusual moral stakes: they can restore agency where the body has become an unreliable vehicle for the mind. This promise became tangible in 2016 at EPFL, when a paralyzed monkey took its first steps: its leg reanimated by a stimulating probe that translated cortical activity into commands for its own nerves. The breakthrough wasn't enabled by a better electrode array, but by machine learning systems that decoded the brain's activity in real time. We now stand at a critical juncture: if we proceed carefully, AI could restore autonomy to those suffering from conditions once deemed intractable; if we don't, we could forfeit sovereignty over the mind. In this essay, I argue that neural interfaces carry both the highest stakes and the greatest near-term impact among AI applications in the coming decade.

2. Brain Computer Interfaces

Since early work linking EEG rhythms to cursor movements, BCIs have evolved into complex systems translating neural signals into commands for external devices or neuromodulators[6]. The systems of interest here are clinically supervised, often surgically implanted, and targeted at serious neurological and psychiatric conditions: they belong alongside pacemakers and cochlear implants, not consumer wearables.

BCIs may be divided into read-out and write-in systems, with an emerging class targeting bidirectional modulation. The canonical neuroprostheses application is in paralysis: BCIs can drive exoskeletons, virtual environments, or even bio-electronic stimulators in ways that promote neuroplasticity after stroke or spinal cord injury[7]. Machine learning techniques from robotics have proven highly applicable to physical prosthetics[8]. However, these advancements do not benefit those patients who lose not only their motor system, but their most intimate faculties of communication. This is the domain where artificial intelligence poses the greatest benefit: as a critical tool to serve the ambitious goal of interfacing with the native contents of the mind. In particular, I motivate AI's application to two domains: speech neuroprostheses that translate single-neuron spiking or high-gamma activity into conversational communication[2,3,9,10], and adaptive deep brain stimulation (DBS) platforms that use neural biomarkers to guide real-time stimulation adjustments in neural dynamics[4]. The near-term question is not whether AI will become a mainstay of these BCIs, but under what ethical, legal, and social terms.

2.1 AI architectures for neural decoding

Current high-performance speech neuroprostheses rely on recurrent neural networks (RNNs), particularly gated recurrent units (GRUs), to decode temporal sequences of neural firing patterns into phoneme probabilities[2,3]. In the Willett et al. system, a GRU decoder processes 192-channel intracortical recordings at 50 ms bins, mapping spike counts to phoneme logits, from which a language model produces sentences at 62 words per minute. Critically, these systems decode from motor and premotor cortex – the brain's attempt to produce voluntary output the body can no longer execute. This is not decoding consciousness; it is intercepting a motor command at the

last cortical relay before it descends to the muscles of speech. A decoder reading from motor cortex is, in a meaningful sense, merely restoring a broken wire between intention and action. The patient's private thoughts, encoded in associative, prefrontal, and default-mode regions, remain untouched.

2.2 AI architectures for neuromodulation

For adaptive brain stimulation, the dominant approach is reinforcement learning (RL) and Bayesian optimization applied to stimulation parameter search: the model learns a person-specific policy mapping a continuously observed biomarker (typically beta-band (13–30 Hz) oscillation amplitude in the subthalamic nucleus) to a stimulation voltage or frequency, adjusting in real time within sub-10 ms decision loops[4]. These models are lightweight by design, constrained by the power and latency budgets of implanted pulse generators.

2.3 Future developments

Two architectural shifts will likely transform BCI systems within a decade. First, pre-training on multi-subject recordings will address the current calibration bottleneck. Standard transfer-learning approaches in neural decoding have demonstrated that representations learned from one behavioral context or subject transfer successfully to novel subjects, reducing the per-patient calibration time[11,12]. A brain MRI foundation model (BrainIAC) establishes a precedent electrophysiology will likely follow, given its improved performance in low-data and few-shot settings[13]. We may optimistically anticipate a foundation model pre-trained on pooled recordings could be fine-tuned on minimal per-subject data, reducing calibration from hours to minutes.

Second, state-space models (SSMs) such as Mamba[14] and S4[15], architectures that process very long sequences without the quadratic computational scaling of standard attention mechanisms, may be particularly suited to the continuous, high-sample-rate time series data neural interfaces produce. Neural dynamics operate across multiple timescales: from single spikes (~10ms) and local field oscillations (10-100ms) to behavioral symptoms like tremors (seconds), medication cycles (hours), and circadian rhythms (days). SSMs' long-range dependency modeling could theoretically enable implanted devices to capture these multi-scale patterns within strict embedded power budgets[16]. By 2035, a plausible clinical system could perform on-device inference using pre-trained representations on neuromorphic hardware, eliminating the need to stream raw neural data externally.

3. Benefits

3.1 Restoring communication

To understand the stakes, we must understand the patients. Locked-in syndrome, typically caused by a ventral pontine stroke that destroys motor pathways while leaving cognition and sensation intact, has an estimated prevalence of roughly 1 in 339,000, yielding thousands of individuals worldwide[17]. Subjective accounts are remarkably consistent: full awareness, preserved emotion, and an intact desire to communicate, trapped behind a body that remains unresponsive[18]. Jean-Dominique Bauby, who wrote *The Diving Bell and the Butterfly* by blinking one eyelid, described the experience as having his mind "buried alive inside a corpse." Surveys confirm that locked-in

patients report meaningful quality of life once communication is established, but that the period before any channel is available is profoundly distressing[18]. Amyotrophic lateral sclerosis (ALS) is far more prevalent: approximately 300,000 people live with ALS globally[19], with ~58,000 new diagnoses in 2016. In bulbar-onset ALS, roughly 70% of patients lose functional speech within three years of symptom onset[20]. More broadly, severe motor and speech disorders requiring augmentative or alternative communication (AAC) affect roughly 3.5 million Americans, most relying on tools producing 10–15 words per minute – far short of natural conversational rates[21].

AI-driven BCIs can transform this. Willett et al. achieved 62 words per minute with 23.8% word error rate using intracortical recordings paired with deep learning[2]. Metzger et al. demonstrated ECoG decoding supporting text, synthetic voice approximating a patient's pre-injury timbre, and avatar facial expressions[3]. Non-invasive systems remain lower bandwidth, but Tang et al. showed continuous semantic decoding from fMRI, demonstrating that AI can reconstruct perceived or imagined language from brain activity without implanted electrodes[9]. While they recorded from higher cortical areas instead of motor cortex, participants could intentionally disrupt the decoding by thinking different thoughts – making mental privacy an engineering constraint, not merely a downstream ethics discussion.

3.2 Therapeutic neuromodulation

Movement disorders such as Parkinson’s disease, dystonia, and essential tremor represent significant, progressive burdens that severely impair physical autonomy and diminish the overall quality of life for more than 15 million individuals worldwide. Continuous DBS provides symptom relief but can generate dyskinesias, mood changes, and cognitive effects when stimulation is pre-defined and static. Adaptive DBS delivers precise stimulation when symptom-associated neural patterns emerge – such as elevated beta-band oscillations preceding tremor – reducing side effects while maintaining control[4]. AI enhances this by learning person-specific mappings from neural biomarkers to stimulation parameters, enabling more individualized control.

Beyond movement disorders, closed-loop neuromodulation shows promise for depression, OCD, chronic pain, and epilepsy. AI models characterizing patient-specific neural states could support “precision psychiatry,” where stimulation targets distributed network states associated with symptom clusters rather than anatomical loci[7,22].

3.3 Beneficiaries

The costs of the conditions BCIs target are not borne only by patients. In the United States, lifetime care for high spinal cord injury exceeds $5 million[23]; advanced ALS costs over $100,000 per year[24]; and neurological disorders collectively represent the leading cause of disability-adjusted life years globally[25]. Much of this cost is driven by institutionalization and 24-hour caregiving. A BCI restoring even partial communication or environmental control can shift patients from custodial care to supported independence, enabling participation in family, community, and employment.

Beyond individual users, the broader public stands to benefit from the scientific knowledge these BCI systems generate. There is direct precedent: DBS was developed as a clinical intervention for Parkinson's disease, but intraoperative recordings from DBS patients became

one of the most important data sources for understanding human basal ganglia circuitry, particularly for research on decision-making, reward processing, and addiction[26]. AI-driven BCIs are poised to play the same role for the cortex: high-dimensional probes whose data can test theories of biological coding and drive cognitive neuroscience research.

4. Risks

4.1 Mental privacy and cognitive liberty

Neural interfaces do not merely raise a privacy problem; they open a new domain of ethical concern. Until now, mental privacy has been the only privacy right guaranteed by biology rather than law[27]. Every other sphere of privacy - the home, the body, correspondence - has been penetrated by technology; if we allow AI into BCIs, we are necessarily facilitating its breach into a new intimate domain.

Farahany argues that the capacity to think without observation is constitutive of personal identity: if the process of thought becomes observable, people will self-censor not speech but thinking itself[28]. Lavazza extends this: even read-only decoding threatens mental integrity, because monitored thought is altered thought; observation collapses the private deliberative space on which autonomous reasoning depends[29]. Susser, Roessler, and Nissenbaum provide the mechanism: manipulation occurs when one party exploits an informational advantage the other cannot detect or resist[30]. Neural decoders create the most extreme version of this asymmetry: the subject experiences their thought simultaneously with its being read. At the limit, Ienca warns that reliable intention decoding would collapse the distinction between thought and action that undergirds criminal law, eroding the presumption of innocence[5].

Without guardrails, the trajectory is predictable. Zuboff argues that data first collected to improve services became behavioral surplus, then prediction products sold in behavioral futures markets, and eventually tools to shape behavior[31]. Neural data could follow the same path: gathered for clinical purposes, mined for consumer insight, eventually normalized as a condition of employment or insurance.

4.2 AI failure modes in medical devices

All implanted medical devices carry surgical and hardware risks. AI introduces a qualitatively different class of failure. Deep learning decoders are probabilistic: under distribution shift caused by fatigue, medication changes, or electrode degradation, a decoder may assign high confidence to an unintended output[32]. In a closed-loop stimulation system, a misclassified biomarker could trigger stimulation that worsens tremor, induces dyskinesia, or destabilizes mood. Unlike a conventional DBS pulse generator whose parameters remain fixed, an adaptive reinforcement-learning controller continuously updates its policy, and could experience gradual, silent deviation from validated behavior. The opacity of deep networks compounds this: neither patient nor clinician can inspect why a stimulation decision was made. Habli and colleagues argue that conventional safety standards designed for deterministic systems are fundamentally inadequate for ML-based devices[33].

4.3 Dual use and militarization

The most concerning long-term risk of bidirectional neural interfaces is not data exfiltration but behavioral influence. Closed-loop systems that decode affective state and deliver targeted stimulation create a feedback loop capable of shaping thought, mood, and decision-making. Yuste and colleagues identify this as a threat to "cognitive liberty": the right not merely to keep thoughts private but to have them remain one's own[34]. The risk is not speculative: adaptive DBS for depression already modulates mood-associated circuitry in real time, and the step from therapeutic stabilization to "nudging" has no practical obstacles. Kosal and Putney document that both the United States and China treat BCIs as dual-use technologies with explicit military applications[35]. Because many implants depend on wireless telemetry and software updates, they also create a cyber-physical attack surface: compromised devices could exfiltrate sensitive neural data or alter stimulation parameters. This “brainjacking” risk has been discussed explicitly for DBS, and regulators increasingly treat cybersecurity as a core safety requirement for networked medical devices.

5. The affirmative case

The case for AI deployment in neural interfaces rests on unmet clinical need and the technical tractability of safeguards. The need is enormous: locked-in syndrome, advanced ALS, severe Parkinson's, and treatment-resistant depression are poorly served by current therapies. Many risks are amenable to concrete mitigations; decoders can require cooperative engagement[9], and designers can amplify "privacy by default" through on-device decoding, explicit activation signals, and encryption. Three governance proposals illustrate how policy maps onto technical architecture:

On-device processing mandates. The EU AI Act classifies body-interfacing AI as "high-risk" and requires data governance and human oversight[37]. For BCIs, we may similarly mandate that raw neural signals be decoded on the implant or its proximal hub, with only abstracted commands (intended phonemes, stimulation triggers) transmitted beyond the device boundary. This constrains the architecture to local neuromorphic hardware, ensuring that even a compromised communication link yields only low-dimensional motor commands, not the high-bandwidth neural recordings from which thoughts or emotional states could theoretically be inferred. Chile's 2021 constitutional amendment protecting "brain activity and the information derived from it" (Article 19) provides a legal template for codifying data minimization at the highest level of national law, establishing the first national neurorights framework[27].

Algorithmic safety governance. The U.S. Food and Drug Administration (FDA) requires manufacturers to specify which parameters of a learning algorithm may update autonomously and which require re-authorization[32]. For adaptive DBS, the reinforcement-learning agent tuning stimulation should operate within a pre-approved safe envelope of voltages, frequencies, and pulse widths, with any excursion triggering clinical review. Complete version histories of model weights would enable clinicians to roll back to any prior validated state, mirroring existing practice in safety-critical avionics (ISO 26262). Complementing device-level oversight, the risk that proprietary neural datasets become concentrated in a few firms can be addressed through public-interest data trusts: legal structures where neural recordings are held by an independent

fiduciary on behalf of patients rather than the device manufacturer[30]. Combined with federated learning, where model training occurs on distributed local data without centralizing raw recordings, this structure allows the field to build the large, diverse training corpora needed for neural foundation models while patients retain granular control with enforceable opt-out and purpose-limitation provisions.

Cortical-origin constraints on decoding. Perhaps the most architecturally precise governance lever is to regulate where in the brain a device may decode from. There is a philosophically meaningful distinction between decoding from motor cortex, where the brain is attempting to produce output, and decoding from associative, prefrontal, or default-mode regions, where activity reflects spontaneous mentation, preference, or memory. Farahany argues that cognitive liberty requires distinguishing deliberate externalization from passive internal states[28]; Lavazza proposes a layered sensitivity model in which motor intention is the least private category and spontaneous thought the most[29]. The strongest current speech neuroprostheses already decode from motor and premotor cortex[2,3], meaning the most clinically effective systems are also the most ethically defensible. A regulatory regime restricting commercial BCI decoding to motor-output regions unless the patient provided informed consent for broader recording would formalize this alignment. Further research mapping cortical recording sites to volitional commitment will be essential to inform such policy.

The key is to treat AI-enabled neural interfaces as part of a broader system that includes technical architecture, clinical practice, law, and public deliberation. The alternative – foregoing or slowing these developments – would leave millions of patients without access to lifechanging treatments, and would cede the field to actors less constrained by ethical scrutiny.

6. Conclusion

AI promises to enable new forms of communication and neuromodulation, with stakes for autonomy, dignity, and human flourishing higher than in any other domain. These technologies both respect and threaten the mind’s special status. They can restore communicative agency and self-expression to people whose bodies can no longer serve as reliable vehicles for their minds, yet they can also expose and shape mental states in ways that undermine privacy and autonomy. The response should be neither rejection nor uncritical embrace, but deliberate design and governance: decoders and control systems that structurally encode consent, neurorights embedded in law, and a priority on restorative therapies over enhancement. The choice is not whether AI will be applied to neural interfacing, but how, and under whose terms. Superintelligence matters most when it meets the superintimate – where the boundary of the self is at stake.

References

1. Card, N. S. *et al.* An Accurate and Rapidly Calibrating Speech Neuroprosthesis. *N. Engl. J. Med.* **391**, 609–618 (2024).
2. Willett, F. R. *et al.* A high-performance speech neuroprosthesis. *Nature* **620**, 1031–1036 (2023).
3. Metzger, S. L. *et al.* A high-performance neuroprosthesis for speech decoding and avatar control. *Nature* **620**, 1037–1046 (2023).
4. Bronte-Stewart, H. M. *et al.* Long-Term Personalized Adaptive Deep Brain Stimulation in Parkinson Disease: A Nonrandomized Clinical Trial. *JAMA Neurol.* **82**, 1171 (2025).
5. Ienca, M. & Andorno, R. Towards new human rights in the age of neuroscience and neurotechnology. *Life Sci. Soc. Policy* **13**, 5 (2017).
6. Deng, Q., Fu, Z., Ma, N. & Wang, B. Application and future directions of brain-computer interfaces in neurological disorders: Technological advances, clinical practices, and challenges. *Brain Hemorrhages* **6**, 306–314 (2025).
7. Awuah, W. A. *et al.* Bridging Minds and Machines: The Recent Advances of Brain-Computer Interfaces in Neurological and Neurosurgical Applications. *World Neurosurg.* **189**, 138–153 (2024).
8. Senatore, S. C., Takahashi, K. Z. & Malcolm, P. Using human-in-the-loop optimization for guiding manual prosthesis adjustments: a proof-of-concept study. *Front. Robot. AI* **10**, 1183170 (2023).
9. Tang, J., LeBel, A., Jain, S. & Huth, A. G. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nat. Neurosci.* **26**, 858–866 (2023).
10. Moses, D. A. *et al.* Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria. *N. Engl. J. Med.* **385**, 217–227 (2021).
11. Sussillo, D., Churchland, M. M., Kaufman, M. T. & Shenoy, K. V. A neural network that finds a naturalistic solution for the production of muscle activity. *Nat. Neurosci.* **18**, 1025–1033 (2015).
12. Gallego, J. A., Perich, M. G., Miller, L. E. & Solla, S. A. Neural Manifolds for the Control of Movement. *Neuron* **94**, 978–984 (2017).
13. Tak, D. *et al.* A generalizable foundation model for analysis of human brain MRI. *Nat. Neurosci.* **29**, 945–956 (2026).
14. Gu, A. & Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. Preprint at https://doi.org/10.48550/ARXIV.2312.00752 (2023).
15. Gupta, A., Gu, A. & Berant, J. Diagonal State Spaces are as Effective as Structured State Spaces. Preprint at https://doi.org/10.48550/ARXIV.2203.14343 (2022).
16. Pandarinath, C. *et al.* Inferring single-trial neural population dynamics using sequential auto-encoders. *Nat. Methods* **15**, 805–815 (2018).
17. Nilsen, H. W. *et al.* Demographic, Medical, and Clinical Characteristics of a Population-Based Sample of Patients With Long-lasting Locked-In Syndrome. *Neurology* **101**, (2023).
18. Laureys, S. *et al.* The locked-in syndrome : what is it like to be conscious but paralyzed and voiceless? in *Progress in Brain Research* vol. 150 495–611 (Elsevier, 2005).
19. Wolfson, C., Gauvin, D. E., Ishola, F. & Oskoui, M. Global Prevalence and Incidence of Amyotrophic Lateral Sclerosis: A Systematic Review. *Neurology* **101**, (2023).
20. Eshghi, M. *et al.* Rate of speech decline in individuals with amyotrophic lateral sclerosis. *Sci. Rep.* **12**, 15713 (2022).
21. Barton-Hulsey, A., Sevcik, R. A. & Romski, M. The Relationship Between Speech, Language, and Phonological Awareness in Preschool-Age Children With Developmental Disabilities. *Am. J. Speech Lang. Pathol.* **27**, 616–632 (2018).
22. Andelman-Gur, M. M. & Fried, I. Consciousness: a neurosurgical perspective. *Acta Neurochir. (Wien)* **165**, 2729–2735 (2023).

23. National Spinal Cord Injury Statistical Center. Traumatic Spinal Cord Injury Facts and Figures at a Glance. [2024].
24. Stenson, K. *et al.* Health care resource utilization and costs across stages of amyotrophic lateral sclerosis in the United States. *J. Manag. Care Spec. Pharm.* **30**, 1239–1247 (2024).
25. Steinmetz, J., Seeher, K. & Schiess, N. Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. *Lancet Neurol.* **23**, 344–381 (2024).
26. Fried, I., Mukamel, R. & Kreiman, G. Internally Generated Preactivation of Single Neurons in Human Medial Frontal Cortex Predicts Volition. *Neuron* **69**, 548–562 (2011).
27. Ligthart, S. *et al.* Minding Rights: Mapping Ethical and Legal Foundations of 'Neurorights'. *Camb. Q. Healthc. Ethics* **32**, 461–481 (2023).
28. Muñoz, J. M. Achieving cognitive liberty **The Battle for Your Brain: Defending the Right to Think Freely in the Age of Neurotechnology** *Nita A. Farahany* St. Martin's Press, 2023. 288 pp. *Science* **379**, 1097–1097 (2023).
29. Lavazza, A. Freedom of Thought and Mental Integrity: The Moral Requirements for Any Neural Prosthesis. *Front. Neurosci.* **12**, 82 (2018).
30. Susser, D., Roessler, B. & Nissenbaum, H. Technology, autonomy, and manipulation. *Internet Policy Rev.* **8**, (2019).
31. Zuboff, S. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. (PublicAffairs, 2019).
32. Artificial Intelligence/Machine learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan.
33. Abulibdeh, R., Celi, L. A. & Sejdić, E. The illusion of safety: A report to the FDA on AI healthcare product approvals. *PLOS Digit. Health* **4**, e0000866 (2025).
34. Yuste, R. *et al.* Four ethical priorities for neurotechnologies and AI. *Nature* **551**, 159–163 (2017).
35. Kosal, M. & Putney, J. Neurotechnology and international security: Predicting commercial and military adoption of brain-computer interfaces (BCIs) in the United States and China. *Polit. Life Sci. J. Assoc. Polit. Life Sci.* **42**, 81–103 (2023).
36. Pycroft, L., Boccard, S. G., Fitzgerald, J. J., Green, A. L. & Aziz, T. Z. Brainjacking: cyber-security risk in deep brain stimulation. *Brain Stimul. Basic Transl. Clin. Res. Neuromodulation* **10**, 359 (2017).
37. Article 6: Classification Rules for High-Risk AI Systems.