

Uncertainty-Aware Large Language Models for Adaptive Brain-Computer Interface Decoding

Harish Saxena

School of Electrical Engineering and Computer Science, Oregon State University, Corvallis,
OR, USA.

hsaxena660@oregonstate.edu

Abstract

Large language models have introduced powerful sequence priors for neural decoding, but their integration into brain-computer interfaces introduces deep uncertainties related to neural variability, model calibration, user intent, and long-term adaptation. This paper develops a system-level analysis of uncertainty-aware large language models for adaptive brain-computer interface decoding. It examines the architectural trade-offs involved in coupling neural encoders with probabilistic language models, the role of uncertainty as an internal control signal, and the infrastructure required for closed-loop adaptation. The discussion extends beyond decoding accuracy to address robustness, fairness, governance, clinical deployment, sustainability, and policy implications. A central argument is that uncertainty should not be treated merely as a residual error to be minimized, but as an informational resource for adaptive coordination between user, decoder, and interface. This perspective shifts design attention from static decoding pipelines toward continuously negotiating systems that can express confidence, defer decisions, request clarification, and update behavior under distribution shift. The paper reviews relevant progress in neural decoding and language modeling, identifies structural failure modes that emerge when opaque generative priors are introduced into assistive neurotechnology, and proposes system governance principles for safe and equitable deployment. The analysis integrates perspectives from neural engineering, machine learning, human-computer interaction, and socio-technical systems research. The result is a conceptual architecture in which calibrated uncertainty supports adaptive autonomy, preserves user agency, and enables responsible clinical translation.

Keywords

brain-computer interfaces; large language models; uncertainty quantification; adaptive decoding; neural engineering; trustworthy artificial intelligence; socio-technical systems.

1. Introduction

The emergence of transformer-based sequence models [1] and their scaling into large language models [2] has reshaped expectations for sequence decoding across many application domains. In parallel, brain-computer interfaces have advanced from slow spelling devices toward high-performance handwriting and speech decoding systems [3, 4]. These developments have created new opportunities for neural interfaces that use language models to infer intended words, phrases, or commands from noisy cortical signals. At the same time, they have introduced a class of system-level challenges that cannot be resolved through benchmark accuracy alone. Neural recordings are nonstationary, user intent is partially observable, and language model outputs can be fluent without being faithful. These conditions make uncertainty a central design concern.

Brain-computer interface decoding is often framed as a pattern recognition problem in which neural features are mapped to intended actions or symbols. However, this framing understates the complexity of real-world use. Cortical population activity reflects a mixture of motor planning, sensory feedback, cognitive state, and unrelated neural processes [6]. The mapping between recorded activity and intended communication is probabilistic, context-dependent, and subject to slow drift over hours, days, and months. Moreover, assistive communication systems operate in high-stakes settings where errors can compromise autonomy, safety, and trust. A decoder that confidently produces an incorrect phrase may be more dangerous than a decoder that recognizes its own uncertainty and asks for clarification. These considerations motivate a shift from deterministic decoding toward uncertainty-aware architectures [7].

Large language models bring valuable inductive biases to brain-computer interface decoding because they capture syntactic, semantic, and pragmatic regularities in natural language. Yet these same models can generate plausible text that is disconnected from user intent. When a language model is integrated with a neural encoder, the system inherits both the statistical power of pretrained language knowledge and the opacity of large-scale generative prediction. Without explicit uncertainty representation, the language model may dominate the neural evidence, producing fluent but incorrect outputs. The challenge is therefore not simply to combine neural features with language priors, but to design an interface in which confidence, ambiguity, and model disagreement can be observed and acted upon.

This paper examines uncertainty-aware large language models for adaptive brain-computer interface decoding from a systems perspective. It considers architectural trade-offs, infrastructure demands, calibration, robustness, fairness, governance, and deployment. The analysis is intentionally interdisciplinary. It draws on neural engineering for the properties of cortical signals, on machine learning for uncertainty estimation and calibration, and on socio-technical research for the consequences of embedding generative models in assistive technologies. The goal is to articulate design principles for systems that remain legible, controllable, and safe under uncertainty.

2. Background and System Context

Contemporary brain-computer interface systems can be broadly divided into motor decoding, speech decoding, and communication-oriented decoding. Handwriting brain-computer interfaces have demonstrated high-speed text entry by decoding imagined pen movements from motor cortex [3]. Speech neuroprostheses have enabled individuals with severe paralysis to produce text by decoding attempted speech articulations [4]. Encoder-decoder approaches have further shown that cortical activity can be translated directly into sentence-level text using sequence models [5]. These systems differ in signal modality, electrode placement, task structure, and user population, but they share a common dependency on robust sequence inference under neural variability.

Neural population dynamics provide an important conceptual foundation for understanding why decoding uncertainty arises. Motor cortex does not encode movement as a static set of labeled commands but as evolving population trajectories shaped by recurrent neural dynamics [6]. These trajectories vary across trials, sessions, and individuals. Even with high-quality recordings, the same intended word may correspond to a distribution of neural activity patterns. This variability is amplified in clinical populations where injury, fatigue, medication, or electrode degradation may further alter signal statistics. A decoding architecture that assumes a fixed mapping will therefore degrade over time unless it can adapt [7].

Language models introduce a complementary form of structure. A large language model can assign higher probability to grammatically coherent and contextually plausible sequences [2]. In brain-to-text decoding, this prior can resolve ambiguities that are difficult to infer from neural data alone [5]. However, the language model is itself an uncertain system. Its predictions depend on training data, prompting context, decoding strategy, and the degree of alignment between its pretraining distribution and the user’s communicative style. When the language model is highly confident but poorly calibrated, it can override genuine neural evidence and reduce user agency. When it is appropriately uncertain, it can signal the need for user confirmation or interface intervention.

Clinical translation adds further constraints. Implantable brain-computer interfaces require surgical procedures, hardware maintenance, and long-term data management [8]. The decoder must operate within strict latency limits because communication should feel responsive. It must also function across changing attentional states, environmental noise, and assistive software environments. These conditions make uncertainty-aware adaptation an operational necessity rather than a theoretical refinement. The system must know not only what the user likely intended, but how much confidence should be placed in that inference, and what should happen when confidence is low.

A socio-technical framing is also essential. Brain-computer interfaces are not standalone computational systems; they are embedded in care arrangements, rehabilitation programs, family communication, and clinical workflows. Uncertainty propagates across these layers. A model may be uncertain about a word, a user may be uncertain about whether the system understood, and a caregiver may be uncertain about whether to trust the output. Designing for uncertainty therefore involves coordinating multiple forms of knowledge, authority, and responsibility. This broader context motivates the architectural perspective developed in the following sections.

3. Uncertainty-Aware Architectural Paradigms

Uncertainty estimation in deep learning has matured substantially over the past decade. Bayesian neural networks and related approximations provide mechanisms for distinguishing aleatoric uncertainty arising from data noise from epistemic uncertainty arising from model ignorance [9]. Monte Carlo dropout has been widely used as a practical approximation for representing model uncertainty in deep networks [10]. More recent reviews have cataloged a broad range of uncertainty quantification techniques, including ensemble methods, evidential learning, and deterministic single-pass approximations [11]. These methods differ in computational cost, calibration quality, and suitability for real-time neural decoding.

Large language models complicate uncertainty estimation because their output space is discrete, high-dimensional, and semantically structured. A model may be uncertain about a single token while remaining confident about the general meaning of an utterance. It may also produce fluent text that is miscalibrated relative to factual or contextual correctness. Research on language model self-evaluation has shown that models can provide probabilistic estimates of their own knowledge, but these estimates are often imperfect [12]. Scaling alone does not eliminate calibration failures, and larger models can be confidently wrong in subtle ways [13]. Teaching language models to express uncertainty in natural language has emerged as one approach for making model confidence more interpretable to users [14].

In brain-computer interface decoding, uncertainty can serve as an internal control signal for coordinating neural evidence and language priors. A recent probabilistic co-control

formulation treats uncertainty as a signal that modulates how aggressively the language model should influence decoding outputs [15]. This approach reframes uncertainty from a passive diagnostic into an active mechanism for balancing user control and model assistance. When neural evidence is strong and consistent, the decoder can rely more directly on observed cortical activity. When neural evidence is weak or ambiguous, the language model can contribute stronger priors, but the system should also signal reduced confidence to the interface layer. This dynamic interaction is central to adaptive brain-computer interface design.

Calibration is a necessary condition for uncertainty to be useful. A well-calibrated system should assign high confidence to hypotheses that are correct at a corresponding frequency. Modern neural networks frequently exhibit overconfidence, and calibration often degrades under distribution shift [16]. For neural interfaces, distribution shift is the norm. Electrode impedance changes, neural plasticity, task switching, and user fatigue all produce shifts that cannot be fully anticipated during training. Uncertainty-aware architectures must therefore include calibration mechanisms that are updated continuously rather than treated as fixed model properties. Without recalibration, uncertainty signals may mislead both the adaptive controller and the user.

Architectural choices also shape how uncertainty is propagated. A modular design may separate the neural encoder, the language model, and the uncertainty estimator, allowing each component to be updated independently. An end-to-end design may optimize all components jointly but risk entangling neural likelihoods with linguistic priors in ways that are difficult to audit. Hybrid designs can preserve modularity while enabling gradient-based adaptation of selected components. The choice involves trade-offs between interpretability, adaptability, computational efficiency, and clinical safety. A systems perspective suggests that uncertainty should be represented at multiple levels: token-level confidence, sequence-level plausibility, session-level reliability, and user-level trust.

4. Adaptive Decoding Pipelines and Infrastructure

Adaptive brain-computer interface decoding requires more than a well-trained model. It requires an infrastructure that supports continuous observation, model updating, user feedback, and safe intervention. In clinical settings, data are collected under heterogeneous conditions, and ground-truth labels may be sparse or unreliable. The system must therefore learn from weak signals, delayed feedback, and occasional explicit corrections. This differs from typical supervised learning settings in which large labeled datasets are available before deployment. The architecture must be designed for lifelong adaptation under partial supervision [7].

A practical adaptive pipeline includes neural signal acquisition, feature extraction, sequence decoding, uncertainty estimation, and interface presentation. Each stage introduces latency and potential failure modes. The uncertainty-aware language model sits within the decoding stage but must communicate with both the neural encoder and the interface controller. Its confidence estimates can be used to adjust decoding speed, request confirmation, present alternative predictions, or temporarily reduce automation. For example, if token-level uncertainty is high, the interface might display several candidate words rather than committing to a single output. If session-level reliability is low, the system might guide the user toward a recalibration routine.

Closed-loop adaptation creates feedback between user behavior and model updates. The user observes the system’s output and may adjust their neural strategy, attempt to correct errors, or

lose trust after repeated failures. The system observes neural patterns, output selections, and explicit feedback, then updates its internal representations. This mutual adaptation is a defining feature of clinical brain-computer interfaces [8]. Uncertainty signals are valuable in this loop because they help distinguish user corrections from signal noise, and they help the system decide when to consolidate new information into long-term model parameters.

Infrastructure requirements extend to data governance. Neural data are highly sensitive and can reveal information about health, identity, emotional state, and cognitive function. Adaptive systems that retain data for recalibration must implement strong access controls, encryption, audit trails, and consent management. Federated or on-device learning may reduce privacy risks by keeping raw neural signals local while sharing only model updates. However, these approaches introduce additional systems engineering challenges related to computational load, battery life, and communication bandwidth in implantable or wearable devices. The choice of adaptation strategy therefore has direct implications for privacy, safety, and sustainability.

From a system-level perspective, uncertainty-aware adaptation is best understood as a control problem rather than a static inference problem. The decoder is a controller that modulates the influence of language priors relative to neural evidence. The uncertainty estimate is a feedback signal that informs this modulation. When confidence is high, the system can proceed automatically. When confidence is low, the system should reduce automation and increase user control [15]. This view aligns with broader principles of human-autonomy teaming, in which machine capabilities are deployed dynamically according to context and user state.

5. Robustness, Fairness, and Socio-Technical Governance

Robustness in neural decoding is not reducible to average accuracy. A decoder may perform well on common phrases but fail systematically on rare words, accented speech patterns, or individual communication styles. Large language models trained on broad corpora may encode demographic, linguistic, and cultural biases that are not equally aligned with all users. When these models are used in assistive communication, bias can become a matter of autonomy and identity. The system may inadvertently steer users toward language that does not reflect their voice. These risks are well documented in broader discussions of language model harms [18, 19].

Fairness in brain-computer interfaces must be considered at multiple levels. At the individual level, a system should adapt to a user's specific neural patterns and communicative preferences rather than imposing a generic language prior. At the group level, training data and evaluation protocols should account for variability across age, sex, neurological condition, language background, and cultural context. Failure to do so can produce systems that work well for narrow populations and poorly for others. In clinical translation, this can deepen existing disparities in access to assistive technology. Socio-technical scholarship has argued that fairness cannot be treated as a fixed property of a model; it emerges from the interaction between technical systems and social institutions [20].

Governance of uncertainty-aware neural interfaces requires clear allocation of responsibility. When the system produces an uncertain output, who decides whether to accept, correct, or discard it? The user may be unable to provide rapid feedback because of motor impairment. The caregiver may have divergent views about the user's intent. The clinician may be responsible for system configuration. The model developer may be responsible for pretraining

data and safety objectives. These distributed responsibilities must be specified before deployment, not discovered after harm occurs. Responsible machine learning for health care emphasizes the need for prospective evaluation, stakeholder engagement, and continuous monitoring [21].

Transparency is particularly challenging for large language models. Their internal representations are difficult to interpret, and their outputs can be fluent without being grounded in observable evidence. In brain-computer interface decoding, this challenge is compounded because the neural input itself is not directly observable by the user or caregiver. Interpretability mechanisms should therefore focus on actionable explanations, such as indicating why the system is uncertain, what alternative hypotheses remain plausible, and what additional feedback would resolve ambiguity. The goal is not to expose raw attention weights or hidden states, but to support meaningful oversight.

Accountability frameworks should also address model drift and update governance. Adaptive systems change over time, and a decoder that performed well at one point may develop new failure modes after weeks of unsupervised adaptation. Audit logs should record confidence trajectories, user corrections, model updates, and interface decisions. These logs can support retrospective analysis of errors and enable clinicians to identify when recalibration or model rollback is necessary. The governance infrastructure should treat adaptation as a controlled process rather than an uncontrolled side effect of continued use.

6. Deployment, Sustainability, and Clinical Translation

Clinical deployment of brain-computer interfaces has historically required substantial infrastructure, including surgical facilities, technical support staff, and long-term follow-up protocols [8]. Adding uncertainty-aware large language models increases the complexity of this infrastructure. The system must support ongoing model inference, uncertainty estimation, calibration updates, and secure communication with external language services if cloud-based models are used. These requirements raise important questions about sustainability, especially in resource-limited settings where continuous connectivity and specialized technical maintenance may not be available.

One structural trade-off concerns the location of language model inference. Cloud-based large language models offer high capacity and regular updates but introduce latency, privacy risk, and dependency on external infrastructure. On-device models reduce these risks but require compression, quantization, or distillation, which may degrade uncertainty estimates and adaptation capability. Hybrid architectures can balance these concerns by performing fast neural decoding locally while selectively querying larger models for high-uncertainty cases. Such designs require a policy layer that determines when external computation is justified relative to privacy, latency, and energy costs.

Sustainability also includes the long-term viability of the system for users who may depend on it for daily communication. If a commercial language model service changes its interface, pricing, or availability, the brain-computer interface may cease to function effectively. Dependence on proprietary models creates risks for continuity of care and user autonomy. Open model alternatives, standardized interfaces, and local fallback modes can reduce this dependency. Systems should be designed with graceful degradation: when high-capacity language assistance is unavailable, the decoder should still support core communication through simpler, more robust methods.

Clinical evaluation of uncertainty-aware adaptive systems requires metrics that go beyond word error rate or bits per minute. Evaluations should assess how often the system correctly identifies its own errors, how effectively it requests clarification, how well it preserves user agency, and how quickly it recovers from miscalibration. User-reported trust, cognitive load, and communication satisfaction are essential outcome measures. Research on artificial intelligence in health and medicine increasingly emphasizes the need for prospective clinical validation, interdisciplinary evaluation, and attention to workflow integration [22]. Similar standards should apply to neural communication systems.

Interpretability also affects clinical acceptance. Clinicians and caregivers are more likely to adopt systems that can explain why an output was produced and how confident the system is. However, the demand for interpretability must be balanced against the need for real-time performance. Some interpretability methods are computationally expensive or require access to model internals that may not be available in deployed systems. Alternative approaches include behavior-level explanations, confidence displays, and interface designs that make uncertainty visible through candidate lists or color-coded indicators. The broader literature on interpretability in high-stakes environments cautions against treating explanation as a substitute for underlying model accountability [23].

7. Future Directions

Future work should develop tighter integration between neural population modeling and uncertainty-aware language generation. Rather than treating the language model as an external prior, architectures could learn to modulate language predictions based on the reliability of specific neural channels, time periods, or behavioral contexts. This would allow the system to rely more heavily on language priors when neural evidence is degraded and to defer to neural observations when language priors are likely to be inappropriate. Probabilistic co-control formulations provide one pathway for this dynamic modulation [15].

Another important direction is the development of standardized benchmarks for uncertainty in brain-computer interface decoding. Existing benchmarks emphasize final decoding accuracy but rarely measure calibration, error awareness, clarification efficiency, or adaptation speed. New evaluation suites should include distribution shift scenarios, simulated electrode failure, session-to-session drift, and diverse user populations. These benchmarks should also evaluate the social consequences of decoding errors, such as loss of trust, emotional distress, or miscommunication in care settings. Evaluation design should involve users, clinicians, and caregivers to ensure that metrics reflect lived priorities.

Model governance and regulatory pathways will require continued evolution. Regulatory frameworks for adaptive medical devices must account for models that change after deployment. Uncertainty-aware systems may need pre-specified adaptation bounds, audit mechanisms, and rollback procedures. Continuous monitoring can detect when a model enters a high-uncertainty regime that requires clinician review. The convergence of generative artificial intelligence and neural interfaces will likely prompt new forms of regulatory guidance, particularly around user consent, data rights, and algorithmic accountability [24].

Finally, future systems should consider how uncertainty can be communicated in ways that support user agency rather than undermining confidence. Overly frequent uncertainty displays may increase cognitive load or anxiety. Hiding uncertainty entirely may create a false sense of reliability. Adaptive interfaces could modulate the presentation of uncertainty according to task stakes, user preference, and context. The interface might use subtle cues for low-stakes

ambiguity and more explicit confirmation requests for high-stakes decisions. This human-centered perspective is essential because the ultimate goal is not to remove uncertainty, but to make it navigable.

8. Conclusion

Uncertainty-aware large language models offer a promising foundation for adaptive brain-computer interface decoding, but their responsible deployment requires systems thinking that extends beyond model accuracy. Neural signals are variable and nonstationary. Language models are powerful but imperfect. Users, clinicians, and caregivers operate within social and institutional contexts that shape how uncertainty is experienced and managed. Designing effective systems requires treating uncertainty as a structural resource for coordination among neural evidence, linguistic prior knowledge, adaptive control, and human oversight.

This paper has developed a system-level framework in which uncertainty serves multiple roles: it guides adaptive modulation of language model influence, supports error awareness, enables clarification and deferral, and informs governance. The architectural trade-offs involve modularity, calibration, latency, privacy, and sustainability. Robustness and fairness require attention to individual variability and social context. Clinical translation requires infrastructure that supports continuous adaptation under partial supervision while maintaining accountability.

The integration of large language models into brain-computer interfaces should not be driven solely by fluency improvements. It should be guided by a commitment to preserving user agency, ensuring equitable access, and building systems that remain legible under uncertainty. By treating uncertainty as an explicit design object rather than an unavoidable nuisance, researchers and engineers can create neural communication systems that are not only more accurate but also more trustworthy, adaptive, and humane.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
2. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
3. Willett, F. R., Avansino, D. T., Hochberg, L. R., Henderson, J. M., & Shenoy, K. V. (2021). High-performance brain-to-text communication via handwriting. *Nature*, 593(7858), 249–254.
4. Moses, D. A., Metzger, S. L., Liu, J. R., Anumanchipalli, G. K., Makin, J. G., Sun, P. F., Chartier, J., Dougherty, M. E., Liu, P. M., Abrams, G. M., Tu-Chan, A., Ganguly, K., & Chang, E. F. (2021). Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 385(3), 217–227.
5. Makin, J. G., Moses, D. A., & Chang, E. F. (2020). Machine translation of cortical activity to text with an encoder–decoder framework. *Nature Neuroscience*, 23(4), 575–582.

6. Shenoy, K. V., Sahani, M., & Churchland, M. M. (2013). Cortical control of arm movements: A dynamical systems perspective. *Annual Review of Neuroscience*, 36, 337–359.
7. Saha, S., Mamun, K. A., Ahmed, K., Mostafa, R., Naik, G. R., Darvishi, S., Khandoker, A. H., & Baumert, M. (2021). Progress in brain computer interface: Challenges and opportunities. *Frontiers in Systems Neuroscience*, 15, 578875.
8. Gilja, V., Pandarinath, C., Blabe, C. H., Nuyujukian, P., Simeral, J. D., Sarma, A. A., Soric, B. L., Perge, J. A., Jarosiewicz, B., Hochberg, L. R., Shenoy, K. V., & Henderson, J. M. (2015). Clinical translation of a high-performance neural prosthesis. *Nature Medicine*, 21(10), 1142–1145.
9. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30, 5574–5584.
10. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of the 33rd International Conference on Machine Learning*, 48, 1050–1059.
11. Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarencov, V., & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76, 243–297.
12. Jiang, Z., Araki, J., Ding, H., & Neubig, G. (2021). How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 9, 423–438.
13. Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., & Kaplan, J. (2022). Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.
14. Lin, S., Hilton, J., & Evans, O. (2022). Teaching models to express their uncertainty in words. *Advances in Neural Information Processing Systems*, 35, 13876–13888.
15. Huang, J., Narasimha, S. M., Patel, A. N., Sristi, R. D., Mishne, G., & Gilja, V. (2026). Probabilistic Co-Control in Brain-Computer Interfaces: Uncertainty as a Control Signal in Brain-to-Text Decoding. *bioRxiv*, 2026-04.
16. Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., & Lucic, M. (2021). Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34, 15682–15694.
17. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.
18. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W., Legassick, S.,

Irving, G., & Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.

19. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623.
20. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. Proceedings of the Conference on Fairness, Accountability, and Transparency, 59–68.
21. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadane-Israni, S., & Goldenberg, A. (2019). Do no harm: A roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9), 1337–1340.
22. Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, 28(1), 31–38.
23. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.
24. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44–56.