

DECODING NEUROETHICS: QUALITY AS THE CODE FOR NEURO-AI GOVERNANCE

E. Kalugina¹

¹HSE University, Perm

E-mail: kalugina.law@gmail.com

ABSTRACT: AI-powered brain-computer interfaces (BCIs) are rapidly advancing, offering transformative potential in healthcare and human augmentation. However, AI's ability to decode brain signals raises profound concerns about neuroethical risks and potential harms. When such systems produce erroneous outputs, it is not merely a software defect but a potential violation of cognitive liberty, mental privacy, and neural integrity. The dominant governance response – neuroethics – has largely focused on rights articulation rather than quality as the operational mechanism capable of translating ethical principles into enforceable controls. This gap is compounded by a structural accountability problem across the BCI value chain. While AI capabilities evolve dynamically, existing safety frameworks remain static, and often do not keep pace with technological developments. This article presents a normative doctrinal analysis of EU product safety regulations as they apply to AI-powered BCIs. It examines how quality obligations, risk assessment, and liability should be structured across the BCI value chain in response to the neuroethical risks posed by neural decoding systems. On that basis, the article argues that quality should function as the central legal standard for AI-powered BCIs and proposes a dynamic NeuroAI Governance Framework to operationalize that position.

INTRODUCTION

The ability to decode brain signals is no longer speculative. AI-powered BCIs are deployed today in clinical rehabilitation, consumer wellness devices, and emerging human-augmentation applications. By combining high-density neural signal acquisition – electroencephalography, electrocorticography, and implanted electrode arrays – with deep-learning decoding algorithms, modern BCIs can infer intentions, emotions, motor commands, and imagined speech from neural data [1]. Neural decoding proceeds through a multi-stage pipeline: signal acquisition, pre-processing and denoising, feature extraction, classification, and post-processing via language models. At each stage, AI acts not merely as a filter but as an active interpreter – what has been termed a “semantic amplifier” [21]: transforming sparse, noisy neural data into rich interpretations by drawing on massive prior knowledge about the world [22].

This interpretive role has undergone a fundamental shift. Early (*discriminative*) decoding models relied on pattern-matching and classification against predefined categories – mapping a neural signal to “move left” or “select letter A”. Contemporary (*generative*) decoding models integrate LLMs and diffusion architectures to reconstruct continuous language and visual content from brain activity [22,23]. Recent studies demonstrate that non-invasive EEG- and fMRI-based systems can now reconstruct continuous spoken language at up to 78 words per minute [24,25] and generate high-resolution visual reconstructions from cortical activity [22]. This distinction carries direct legal implications: in

discriminative decoding, a defect is a misclassification; in generative decoding, a defect may be a *hallucination* – the AI completing or distorting a thought based on its prior knowledge rather than the user's actual neural state. The nature of the defect, and thus of the manufacturer's evidentiary obligation, differs accordingly.

BCIs are increasingly conceptualised as the primary frontier at which human and digital cognition converge, giving rise to what researchers term the “Hybrid Mind” [2]. As AI systems become capable of inferring complex cognitive states, they begin to function as *proxy agents* for the human mind: autonomous intermediaries that interpret and act upon neural data. When an AI system misinterprets a user's neural signals, the outcome is potentially a breach of *cognitive liberty*, *mental privacy*, and *neural integrity* – not merely a software bug [3].

Neuroethics has been the dominant framework for addressing these concerns. Scholars have articulated a growing catalogue of neurorights, and international bodies including UNESCO have adopted dedicated ethical guidelines for neurotechnology [4]. Yet ethical principles alone do not determine who bears liability when an AI-powered BCI causes harm. Soft-law instruments carry no sanctions and provide no mechanism for allocating responsibility across complex value chains. Quality – encompassing both the reliability characteristics of AI decoding algorithms and the management procedures – is the only operational mechanism capable of translating neuroethical principles into enforceable engineering obligations.

This gap between the sophistication of neurotechnology and the adequacy of legal instruments constitutes what this article terms the *Neuro-Accountability Gap*. It has three structural dimensions. First, a legal dimension: neuroethical norms lack enforceability. Second, an accountability dimension: the opacity, autonomy, and adaptability of AI systems combined with the complexity of the BCI value chain and the "many hands" problem [9] make it exceedingly difficult to establish defectiveness, causation, and the liable actor when harm occurs. Third, a dynamic dimension: current AI governance frameworks are static, while AI capabilities – autonomy, efficacy, generality, and goal complexity – evolve continuously, demanding equally evolving quality controls [27]. The new EU regulatory framework, particularly EU AI Act, PLD 2024/2853, GPSR 2023/988, and MDR 2017/745, has the potential to address it by converting abstract neuroethical norms into legal obligations enforceable through product liability. The principal driver of this conversion is quality: the technical accuracy, robustness, and reliability of AI decoding algorithms, which emerges in this regulatory environment as the manufacturer's primary legal defence.

PRODUCT DEFECTIVENESS AND AI-POWERED NEURAL DECODING

The traditional framework and its limitations

Product liability in the EU has historically rested on Council Directive 85/374/EEC, establishing strict liability for defective products causing damage to health or property. Applied to AI-powered BCIs, this framework encounters three structural problems. First, the opacity of deep-learning neural decoding algorithms renders it practically impossible for an injured party to identify the specific computational step that caused harm [1,13]. Second, AI systems adapt after deployment, meaning the product that caused harm may differ substantially from the product originally certified. Third, BCI-related harm is frequently psychological or cognitive in nature – a form of damage Directive 85/374/EEC was not designed to address.

The expanded definition of defect under PLD 2024/2853

Directive 2024/2853 on Product Liability represents a substantive revision of EU product liability law, specifically calibrated for AI-enabled products [6]. Article 7(1)(b) explicitly includes *reasonably foreseeable use and misuse* as a circumstance relevant to defectiveness – extending liability to foreseeable misapplications of AI neural decoding algorithms. Article 4(6) defines "damage" to include *psychological impairment*, making violations of mental privacy and

biased AI-powered neural decoding actionable under EU product liability law.

Elizabeth Steindl has argued that for most consumer neurodevices the primary risks concern data privacy rather than health, placing them within the orbit of the GDPR [7]. However, where a BCI's AI component produces a distorted representation of the user's mental state and that distortion causes measurable psychological damage, product safety law applies independently of data protection law. Crucially, this applies regardless of whether the decoding model is discriminative or generative: a generative model that "hallucinates" a mental state causes harm attributable to the manufacturer who trained and integrated it, not to the data controller who operates it.

Illustrative scenario: the mis-decoded speech command

Consider the following scenario. A user with ALS relies on a communication BCI that reconstructs imagined speech via generative neural decoding models. During a period of acute fatigue – a foreseeable edge-case condition – the algorithm generates a plausible but incorrect semantic output and transmits an unintended workplace instruction, resulting in financial harm and severe psychological distress. Under Article 7 PLD 2024/2853, the manufacturer faces a defectiveness claim: the product failed to provide the safety users were entitled to expect (Art. 7(1)), and the documented psychological impairment qualifies as damage under Article 4(6). Under Article 9 PLD, given the technical opacity of AI-powered neural decoding, the court may *presume* defectiveness if the claimant shows the product caused damage and defectiveness is plausible. To rebut this presumption, the manufacturer must produce: (a) accuracy logs showing performance under fatigue-like signal conditions; (b) out-of-distribution validation records; (c) evidence that an abstention or confirmation mechanism was implemented. For a generative model, the manufacturer must additionally demonstrate that the decoding output was constrained by the user's actual neural state – for example, through calibration logs, uncertainty quantification records, or conformal prediction coverage guarantees. Without such documentation, the manufacturer has no viable defence.

The presumption of defectiveness (Article 9 PLD)

Article 9 PLD 2024/2853 introduces a presumption of defectiveness where a claimant faces *excessive difficulty* due to technical or scientific complexity. For AI-powered BCIs, courts may readily invoke this presumption given the opacity of AI neural decoding algorithms. The presumption applies across four conditions: (1) the defendant fails to disclose relevant evidence; (2) the product does not comply with mandatory product safety requirements under EU law;

(3) an obvious malfunction caused the damage; or (4) proving defectiveness is excessively difficult due to technical complexity. Condition (2) is particularly significant: where an AI system should have complied with EU AI Act quality obligations but did not, non-compliance itself triggers the presumption. However, a critical gap remains: for non-high-risk BCIs, non-medical BCIs, AI agents, or GPAI-integrated decoding systems, the EU AI Act and GPSR impose no specific quality requirements – leaving manufacturers without clear standards to demonstrate and injured parties without a straightforward basis for invoking the presumption. Closing this gap requires explicit quality standards for all neural decoding systems, irrespective of their medical classification. To rebut it, manufacturers must produce logs, validation records, accuracy metrics, and post-market surveillance data. A manufacturer that has not documented its AI model's accuracy rates, false-positive and false-negative distributions, and robustness under edge-case conditions is left without a legal defence when Article 9 is invoked.

PRODUCT QUALITY AS THE MANUFACTURER'S LEGAL SHIELD

The Precautionary Principle (Article 2(5) GPSR)

Article 2(5) GPSR 2023/988 codifies the Precautionary Principle: the absence of full scientific proof does not prevent protective measures where there are reasonable grounds to believe a product could be harmful. For AI-powered BCIs, this creates a structural presumption of potential risk – because the long-term neurological and psychological effects of AI neural decoding systems, including *algorithmic shadows* (systematic biases in the decoded representation of a user's mental state) [8], and in generative models the risk that prior training data shapes the decoded output more than the user's actual neural activity, are not yet empirically established – that manufacturers must actively rebut.

Quality as legal justification and state-of-the-art defence

In the absence of harmonised standards for AI-powered neural decoding – whether discriminative or generative – demonstrable technical accuracy, reliability, and robustness constitute the state-of-the-art benchmark under Article 6 GPSR. Quality ceases to be a commercial attribute and becomes a *legal justification*: the objective mechanism by which a manufacturer moves its product from “potentially hazardous” to “safe with an acceptable level of risk.”

Article 10 PLD 2024/2853 retains the development risks defence, but to invoke it, the manufacturer must demonstrate it applied the highest available quality standards in data curation, model training, validation,

and post-market surveillance. Quality therefore functions as the *primary* legal standard of care – though it interacts with other defences including adequate warnings, instructions for use, and compliance with harmonised standards where available.

Reducing the zone of reasonably foreseeable misuse

The safe product assessment under Article 6 GPSR takes into account behaviour under *reasonably foreseeable misuse*. For BCIs, foreseeable misuse conditions include acute stress, emotional distress, cognitive impairment, and pharmacological alteration – all of which may generate neural signals substantially different from training data. For generative decoding systems, these conditions also increase the risk that the model will “fill in” a plausible but incorrect mental state from its prior knowledge, amplifying the hallucination risk. High-quality AI neural decoding systems can be engineered to recognise signal degradation and respond with appropriate caution: suspending output, requesting confirmation, or escalating to human oversight. The more robust the quality framework – including out-of-distribution detection, abstention mechanisms, and graceful degradation – the more credibly the manufacturer demonstrates compliance with the state-of-the-art standard.

WHO IS RESPONSIBLE? LIABILITY IN THE BCI VALUE CHAIN

The many-hands problem

The AI-enabled BCI involves: (1) a hardware manufacturer; (2) AI algorithm developers; (3) a GPAI model provider; (4) a system integrator; (5) a deployer/operator; and (6) the end user. This multi-stakeholder structure creates the “many hands” problem of AI liability [9]: when harm occurs, each actor contributes to the conditions producing it, yet no single actor bears unambiguous responsibility. The shift from discriminative to generative decoding deepens this problem: where a GPAI foundation model contributes its “world model” to the decoded output, the boundary between the model provider's contribution and the system integrator's contribution becomes legally opaque. Multi-agent BCI deployments further amplify this complexity: as Hammond et al. identify, multi-agent AI systems introduce distinct failure modes – miscoordination, conflict, and collusion – alongside emergent agency and network effects that compound attribution challenges and may make harm causation irrebuttable without robust quality logging [28].

Liability under PLD 2024/2853 and the EU AI Act

Article 4 PLD extends liability to importers, authorised representatives, and fulfilment service providers. Where

the AI neural decoding system qualifies as a *high-risk AI system* under Annex III of the EU AI Act – including BCIs used in medical applications and, where not prohibited under Article 5, emotion-recognition applications in regulated contexts – the AI provider bears primary obligations including conformity assessment, technical documentation, and accuracy and robustness requirements (Article 15 AI Act). It should be noted that Article 5(1)(f) AI Act prohibits emotion recognition in workplace and educational contexts; the high-risk classification under Annex III applies to emotion recognition in other regulated contexts. This distinction is directly relevant to consumer BCIs marketed as stress or affect monitors.

A significant regulatory lacuna remains: consumer wellness BCIs processing EEG signals to infer stress or cognitive states may fall below the high-risk threshold. Yet their potential for harm to mental privacy and cognitive self-determination may be substantial. This article argues that any AI system that performs neural decoding of brain signals should, as a minimum, be subject to an enhanced risk assessment under the AI Act, irrespective of its ostensible purpose, because the long-term consequences of algorithmic mediation of neural activity remain scientifically undetermined [10]. This argument is reinforced by the generative decoding paradigm: a consumer wellness BCI that integrates an LLM to infer mood or intent from EEG data is functionally equivalent, in terms of interpretive power and risk, to a medical-grade speech decoding system.

GPAI models and post-deployment learning

The EU AI Act does not establish quality requirements for GPAI models as such. If a BCI manufacturer integrates a GPAI foundation model that produces a hallucination – a semantically plausible but neurally unfounded output – the allocation of liability between the hardware manufacturer, GPAI provider, and system integrator remains unresolved. This problem is compounded by post-deployment learning [11], which can cause the deployed model to diverge significantly from the certified one. Quality frameworks – continuous monitoring, algorithmic logging, testing, model versioning – are the primary available mechanism for closing this liability gap and enabling proportional fault apportionment.

TOWARDS A NEURO-AI GOVERNANCE FRAMEWORK

The regulatory gap

The current EU regulatory architecture leaves AI-powered BCIs that do not qualify as medical devices or high-risk AI systems under no affirmative quality requirements: GPSR is risk-oriented but mandates no

specific quality level; the AI Act's quality obligations cover only high-risk systems; MDR applies only to medical devices. Consumer neurotechnology BCIs fail through all three nets. This gap is particularly acute for generative neural decoding applications – wellness BCIs, consumer EEG-based mood monitors, and LLM-integrated communication tools – which combine the interpretive power of foundation models with the intimacy of neural data, yet currently attract no mandatory quality obligations. Moreover, existing regulatory frameworks are static, while AI capabilities evolve dynamically. What is needed is an agile governance approach in which quality obligations and risk assessments co-evolve in response to changing AI capabilities – a Risk-Quality Feedback Loop rather than a fixed compliance checklist.

Framework components

The proposed NeuroAI Governance Framework addresses this gap through six integrated components:

(1) *BCI Risk Taxonomy* – four tiers: (i) medical BCIs (MDR); (ii) MDR-quasi BCIs; (iii) high-risk AI BCIs (AI Act Annex III); (iv) other neurotechnologies applying AI-powered neural decoding. Each tier attracts proportionate quality obligations. The taxonomy distinguishes between discriminative and generative decoding within each tier, with generative systems attracting stricter quality obligations due to their hallucination risk.

(2) *Risk Assessment Methodology* – incorporating AI opacity, adaptability, and emergent behaviour, alongside neurotechnology-specific harm categories: mental privacy violation, algorithmic shadow effects, cognitive liberty infringement, and psychological harm. Critically, the methodology must also account for AI capabilities. Following Kasirzadeh and Gabriel [27], AI agents can be classified by their qualitative capabilities: autonomy, efficacy, impact, generality, and goal complexity. The higher these capabilities, the greater the risk – and therefore the more stringent the quality obligations triggered. For multi-agent systems, risk assessment must additionally incorporate the failure modes identified by Hammond et al. [28]: miscoordination, conflict, and emergent collusion that static taxonomies cannot capture. This capability-based dimension is absent from current static regulatory frameworks.

(3) *Quality Models* – minimum performance benchmarks across: (a) data quality (curation, diversity, annotation); (b) AI algorithm quality (accuracy, robustness, OOD performance, hallucination rate, fidelity of interpretation for generative models); (c) AI product quality (integration, explainability); (d) neural decoding process quality (pre-processing, real-time performance, graceful degradation). Quality characteristics and AI capability assessments must

directly inform risk assessments – creating a dynamic feedback loop rather than a one-time conformity check.

(4) *Quality Controls* – implemented at each stage of the BCI lifecycle: *research and pre-clinical testing*, design, development, conformity assessment, market placement, post-market surveillance, and end-of-life. Research-phase controls are particularly critical for consumer neurotechnologies where claims of efficacy or safety are not always backed by independent clinical evidence [20]. These controls embed neuroethical principles as operational engineering requirements and generate the documentary record necessary for liability defence under PLD 2024/2853. Controls are determined by both regulatory requirements and applicable standards (ISO 14971, IEC 62304, ISO/IEC 23894), and must include at minimum: documented accuracy thresholds, OOD detection metrics, hallucination rate ceilings, and post-market drift monitoring.

(5) *Risk-Quality Balance Scale* – proportionality principle: the higher the risk tier, the more stringent the quality requirements. The governing logic is bidirectional: high risk demands stricter quality; and conversely, higher AI capabilities (autonomy, generality, goal complexity) require a higher risk tier assignment. This scale operationalises the Precautionary Principle as a quality imperative.

(6) *Accountability and Liability Distribution Rules* – specifying conditions under which product liability may attach to each actor depending on the source of the defect that caused harm (hardware manufacturer, AI algorithm developer, GPAI model provider, system integrator, deployer) and the mechanism for proportional liability allocation in cases of joint causation. Additionally, the GPSR safety requirements framework should be explicitly adapted to AI-enabled systems: product safety assessments must account for mental safety alongside physical safety, and must incorporate AI-specific risk factors including hallucination, drift, and post-deployment capability expansion.

DISCUSSION

The findings above carry three implications for BCI researchers, manufacturers, and policymakers.

First, the legal framework is shifting faster than industry practice. PLD 2024/2853 entered into force in December 2024 with a transposition deadline of December 2026. Manufacturers who do not begin building quality documentation infrastructures now – accuracy logs, OOD validation records, drift detection systems – will face an evidentiary deficit when the first litigation arises. This urgency is amplified for manufacturers integrating generative neural decoding models: existing logging practices designed for

discriminative classifiers do not capture the hallucination risk profile of LLM-integrated systems.

Second, quality and risk must be governed in symbiosis, not in isolation. The NeuroAI Governance Framework is grounded in a Risk-Quality Feedback Loop: quality controls reduce risks, but expanding AI capabilities generate new risks that demand updated quality controls. This dynamic is absent from current EU regulatory frameworks, which treat risk assessment and quality requirements as separate, static exercises. For AI-powered BCIs, where post-deployment learning and increasing model autonomy continuously alter the risk profile, only a co-evolutionary approach – in which quality metrics and risk assessments are continuously updated in response to changing AI capabilities – can ensure sustained safety and accountability.

Third, the regulatory gap for consumer neurotechnologies is real and urgent. The NeuroAI Governance Framework proposed here is a first step towards filling it. Its operationalisation requires collaboration between legal scholars, neuroscientists, and standards bodies including CEN/CENELEC, ISO/IEC JTC 1/SC 42, and IEEE – which has already developed dedicated neuroethics and BCI standards roadmaps. Without such collaboration, the translation of neuroethical principles into enforceable engineering obligations will remain incomplete.

CONCLUSION

AI-powered BCIs represent one of the most legally consequential product categories of the coming decade. The new EU regulatory framework, anchored by EU AI Act, PLD 2024/2853 and GPSR 2023/988, is beginning to supply the missing legal architecture: the expanded definition of defectiveness now encompasses psychological harm caused by AI misinterpretation or hallucination of neural signals; the presumption of defectiveness shifts the evidentiary burden to manufacturers; and the Precautionary Principle creates a structural presumption of risk under scientific uncertainty.

In this regulatory environment, quality is the primary legal defence available to the product manufacturer. The quality of AI-powered BCIs is not just a matter of product safety: it engages fundamental rights – mental privacy, cognitive liberty, and neural integrity – that demand protection through enforceable engineering obligations, not merely ethical guidelines. For manufacturers integrating generative neural decoding models, this means documenting not only classification accuracy but fidelity of interpretation: the degree to which the decoded output reflects the user's actual neural state rather than the model's prior knowledge. Crucially, governance of AI-powered BCIs cannot be static. As AI capabilities expand – in autonomy, efficacy, and generality – risk profiles

evolve, and quality obligations must evolve with them. The NeuroAI Governance Framework, grounded in a dynamic Risk-Quality Feedback Loop, represents a first attempt to supply that regulatory clarity.

REFERENCES

- [1] Cao Z. A review of artificial intelligence for EEG-based brain-computer interfaces and applications. *Brain Sci Adv.* 2020;6(3):162-170.
- [2] Ienca M, Andorno R. Towards new human rights in the age of neuroscience and neurotechnology. *Life Sci Soc Policy.* 2017;13(1):5.
- [3] Bublitz JC. Novel neurorights: From nonsense to substance. *Neuroethics.* 2022;15:7.
- [4] Lighthart S, Bublitz JC, Alegre S. Neurotechnology: We need new laws, not new rights. *Nature.* 2023;620(7976):950-952.
- [6] Directive (EU) 2024/2853 on liability for defective products [2024] OJ L 2024/2853.
- [7] Steindl E. Consumer neuro devices within EU product safety law. *Comput Law Secur Rev.* 2024;52:105945.
- [8] Li T. Algorithmic destruction. SSRN. 2022. doi:10.2139/ssrn.4066845
- [9] European Parliament. Report on liability for artificial intelligence (A9-0178/2022). 2022.
- [10] Farahany NA. *The Battle for Your Brain.* St. Martin's Press; 2023.
- [11] OECD. Recommendation on AI, Principles 1.3-1.4. 2019.
- [12] Ienca M, Malgieri G. Mental data protection and the GDPR. *J Law Biosci.* 2022;9(1):lsac006.
- [13] Maiseli BJ, Abdalla AT, Massawe LV, Mbise M, Mkocha K, Nassor NA, Ismail M, Michael J, Kimambo S. Brain-computer interface: Trend, challenges, and threats. *Brain Inform.* 2023;10(1):20.
- [14] Chandler JA, van der Loos C, Boencke S, Buchman DZ, Baudry JS, Illes J. BCIs and communication disorders. *Front Hum Neurosci.* 2022;16:841035.
- [15] Gilbert F, Ienca M, Cook M. How I became myself after merging with a computer. *Brain Stimul.* 2023;16(3):783-789.
- [16] Haag L, Starke G, Ploner M, Ienca M. Ethical gaps in closed-loop neurotechnology. *NPJ Digit Med.* 2024;8:510.
- [17] Magee P, Ienca M, Farahany N. Beyond neural data: Cognitive biometrics and mental privacy. *Neuron.* 2024;112(18):3017-3028.
- [18] Lighthart S, Ienca M, Meynen G, et al. Minding rights: Neurorights foundations. *Camb Q Healthc Ethics.* 2023;32(4):461-481.
- [19] Regulation (EU) 2023/988 on general product safety [2023] OJ L135/1.
- [20] Maia de Oliveira Wood G, Berger L, Jarke J, Barnard G, Gremsl T, Dolezal E, Staudegger E, Zandonella P. The protection of mental privacy in the area of neuroscience. European Parliament. 2024. doi:10.2861/869928
- [21] Ienca M. Towards a governance framework for brain data. *EMBO Mol Med.* 2022;14(9):e15337.
- [22] Takagi Y, Nishimoto S. High-resolution image reconstruction with latent diffusion models from human brain activity. *CVPR.* 2023. doi:10.1109/CVPR52729.2023.02151
- [23] Tang J, LeBel A, Jain S, Huth AG. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neurosci.* 2023;26(5):858-866. doi:10.1038/s41593-023-01304-9
- [24] Horikawa T. Mind captioning: Evolving descriptive text of mental content from human brain activity. *Sci Adv.* 2025;11(45):eadw1464. doi:10.1126/sciadv.adw1464
- [25] Willett FR, Kunz EM, Fan C, Avansino DT, Wilson GH, Choi EY, Kamdar F, Hochberg LR, Druckmann S, Shenoy KV, Henderson JM. A high-performance speech neuroprosthesis. *Nature.* 2023;620(7976):1031-1036. doi:10.1038/s41586-023-06377-x
- [26] Regulation (EU) 2024/1689 laying down harmonised rules on AI (AI Act) [2024] OJ L 2024/1689.
- [27] Kasirzadeh A, Gabriel I. In conversation with artificial intelligence: aligning language models with human values. *Philos Technol.* 2023;36(2):1-24. doi:10.1007/s13347-023-00606-x.
- [28] Hammond L, Chan A, Clifton J, et al. Multi-Agent Risks from Advanced AI. Cooperative AI Foundation, Technical Report #1. arXiv:2502.14143. 2025. doi:10.48550/arXiv.2502.14143.