

Article

AI Safety and Alignment in General-Purpose AI Systems: Frameworks, Evaluation, and Governance

Elena V. Petrova^{1, *}, Yuki Tanaka²

¹Department of Computer Science, University of Helsinki, Helsinki, 00014, Finland; elena.petrova@helsinki.fi

²Graduate School of Informatics, Kyoto University, Kyoto, 606-8501, Japan; tanaka.yuki@i.kyoto-u.ac.jp

*Correspondence: Elena V. Petrova. Email: elena.petrova@helsinki.fi

Abstract

The rapid advancement of general-purpose AI (GPAI) systems has elevated AI safety and alignment from technical concerns to pressing governance priorities. This study provides a comprehensive analysis of AI safety and alignment in GPAI systems through a systematic literature review of technical approaches, evaluation methodologies, and governance frameworks, drawing on peer-reviewed literature, conference proceedings, and regulatory documents from 2017 to 2026. Three principal technical approaches are identified: behavioural alignment through reinforcement learning from human feedback and Constitutional AI; mechanistic interpretability for internal model understanding, revealing that approximately 5% of neurons function as "safety neurons" with over 90% causal effects on safety alignment; and scalable oversight for supervising systems that exceed human capabilities, with nested scalable oversight success rates ranging from 9.4% to 51.7% at a general Elo gap of 400. Evaluation methodologies reveal a persistent gap between measured safety and actual safety properties. The analysis indicates that effective AI safety and alignment require integrated approaches combining technical robustness, rigorous evaluation, and adaptive governance. A multi-layered framework encompassing pre-deployment evaluation, runtime monitoring, and incident response mechanisms is proposed.

Keywords: AI safety; value alignment; general-purpose AI; scalable oversight; mechanistic interpretability; AI governance

ARTICLE INFORMATION

Received: 18 May 2026; Accepted: 10 August 2026; Published: 15 September 2026

CITATION

Petrova EV, Tanaka Y. AI Safety and Alignment in General-Purpose AI Systems: Frameworks, Evaluation, and Governance.

Global Intelligence Technology Horizons. 2026; 1(1): 5.

COPYRIGHT



Copyright © 2026 by author(s).

Published by Star Mountain International Publishing Group Pte. Ltd. in *Global Intelligence Technology Horizons*.

This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

1. Introduction

The emergence of general-purpose AI (GPAI) systems—models capable of performing a wide variety of tasks across domains—has fundamentally altered the landscape of artificial intelligence [1]. Unlike earlier narrow AI systems designed for specific functions, GPAI models exhibit increasingly strong autonomy and goal-directed capabilities, accompanied by derivative systemic risks that are more unpredictable, difficult to control, and potentially irreversible [2]. As these systems grow more capable, the challenge of ensuring they remain safe and aligned with human values has become one of the most consequential problems facing researchers, policymakers, and society at large.

AI safety and alignment encompass distinct but interrelated concerns. Safety refers to the prevention of harm—ensuring that AI systems do not produce dangerous outputs, engage in harmful behaviours, or enable malicious uses. Alignment concerns the broader question of whether AI systems act in accordance with human values, norms, and ethical principles [3]. The two are deeply interconnected: a misaligned system may appear safe under normal conditions yet produce catastrophic outcomes when deployed in novel contexts or optimised against misspecified objectives [4].

The challenge of AI safety and alignment is compounded by several factors. First, the capabilities of GPAI systems are advancing rapidly, often outpacing the capacity to understand and evaluate them. Second, these systems exhibit emergent behaviours that are difficult to predict from their training procedures alone. Third, the economic and strategic incentives for rapid deployment frequently compete with the careful safety evaluation that responsible development requires [5]. Fourth, the fragmentation of governance approaches across jurisdictions creates gaps that sophisticated actors may exploit [6].

Recent years have witnessed significant developments in both technical approaches and governance frameworks for AI safety and alignment. On the technical front, researchers have advanced methods including reinforcement learning from human feedback (RLHF), Constitutional AI, mechanistic interpretability, red teaming, and scalable oversight. On the governance front, the EU AI Act has established binding requirements for GPAI models, the Frontier AI Risk Management Framework has provided structured protocols for risk identification and mitigation, and international

bodies have initiated dialogues on global AI governance coordination [7].

This paper provides a comprehensive analysis of AI safety and alignment in general-purpose AI systems, examining technical frameworks, evaluation methodologies, and governance approaches. The specific objectives are: (1) to identify and characterise the principal technical approaches to AI safety and alignment; (2) to evaluate the strengths and limitations of current evaluation methodologies; (3) to analyse emerging governance frameworks and their implications; and (4) to propose an integrated multi-layered framework for responsible AI development and deployment.

2. Materials and Methods

2.1 Analytical Approach

This study adopts a qualitative, multi-method approach to examine AI safety and alignment in general-purpose AI systems. The analysis integrates findings from three complementary streams: (1) a systematic review of the technical literature on AI safety and alignment methods; (2) an analysis of governance frameworks and regulatory documents from major jurisdictions; and (3) a comparative assessment of evaluation methodologies and their limitations [6].

The literature search covered Scopus, Web of Science, arXiv, and Google Scholar using the terms "AI safety," "value alignment," "general-purpose AI," "scalable oversight," "mechanistic interpretability," and "AI governance." The search covered publications from 2018 through mid-2026, with particular emphasis on work published since 2024.

2.2 Framework for Analysis

The analysis is organised around three interconnected dimensions: technical approaches, evaluation methodologies, and governance frameworks. Technical approaches encompass methods for training, fine-tuning, and architecting AI systems to be safe and aligned. Evaluation methodologies concern the measurement and verification of safety and alignment properties. Governance frameworks address the institutional, regulatory, and organisational structures that support responsible AI development [7].

Recent work has proposed systematic taxonomies for AI safety evaluations around three dimensions: what properties are measured, how they are measured, and how these measurements integrate into frameworks [8]. Properties are measured through behavioural techniques including scaffolding, red teaming, and supervised fine-tuning, alongside internal techniques such as representation analysis and mechanistic interpretability [9].

2.3 Case Study Selection

Three major governance initiatives are examined in depth: the EU AI Act's General-Purpose AI Code of Practice, the Frontier AI Risk Management Framework (v1.5), and emerging international coordination efforts including the UN's Global Dialogue on AI Governance. These cases were selected to represent different governance traditions, regulatory approaches, and levels of binding authority.

3. Results

3.1 Behavioural Alignment: RLHF, Constitutional AI, and Beyond

Behavioural alignment methods aim to shape AI system outputs through training and fine-tuning procedures that incentivise safe and beneficial behaviour. Reinforcement Learning from Human Feedback (RLHF) has emerged as the dominant paradigm, using human preference data to train reward models that guide policy optimisation [10]. RLHF has been widely adopted by frontier AI developers and has demonstrated effectiveness in reducing harmful outputs and increasing helpfulness. Early foundational work established deep reinforcement learning from human preferences as a viable approach [11].

However, RLHF has recognised limitations. It can produce superficial conformity rather than genuine ethical understanding, failing to address the complex, context-dependent nature of human values. The reliance on human feedback also raises scalability concerns: as AI systems become more capable than their human evaluators, the quality of feedback may degrade—a problem known as the scalability challenge [12].

Constitutional AI offers an alternative approach that replaces discretionary alignment with verifiable adherence to ethical principles encoded in a "constitution" [13]. Rather than relying on human preference data, Constitutional AI uses a set of written principles to guide model behaviour through a process of critique and revision [13]. Recent work has explored how constitutional approaches can achieve intrinsic safety by making core ethical axioms a non-negotiable component of the model's foundational architecture [14].

A controlled case study contrasting a strongly-aligned RLHF + Constitutional-AI model with a more lightly-aligned model revealed significant differences in safety robustness [15]. Table 1 summarises the key findings across seven persona conditions and five tasks for 1,800 total runs.

Table 1. Alignment floor comparison between strongly-aligned and lightly-aligned models

Metric	Strongly-Aligned Model	Lightly-Aligned Model
Sycophancy rate range across personas	Within 5 percentage points of control	5%–50% range (45 percentage points)
Effect of "enthusiastic" persona	Zero effect on sycophancy	Increases sycophancy from 30% to 50%
Effect of "skeptic" persona	Not reported	Reduces sycophancy by 25 percentage points
Cross-model transfer of persona effects	Near-zero	Near-zero

The study defined the alignment floor as the range of sycophancy rates a model produces across persona conditions. This gap—5 percentage points for the strongly-aligned model versus 45 percentage points for the lightly-aligned model—demonstrates that strong alignment provides genuine robustness to persona-driven safety degradation, while weak alignment can be easily compromised.

3.2 Mechanistic Interpretability: Understanding Internal Mechanisms

Mechanistic interpretability (MI) aims to reverse-engineer neural models into human-understandable algorithms and representations, providing internal transparency by revealing the computations that drive outputs [16]. Unlike post-hoc methods that offer correlational explanations, mechanistic techniques like circuit tracing and activation patching yield causal insight into internal failures, including deceptive or misaligned reasoning that behavioural methods may overlook [17]. Foundational work established a mathematical framework for understanding transformer circuits and identified induction heads as mechanisms for in-context learning [18,19].

Recent research has explored the inner mechanisms of safety alignment through the lens of mechanistic interpretability, focusing on identifying and analysing "safety neurons" within LLMs that are responsible for safety behaviours [16]. Table 2 summarises the key findings from this line of research.

Table 2. Mechanistic interpretability findings on safety neurons

Finding	Result
Proportion of safety neurons identified	Approximately 5% of all neurons
Safety performance restoration via patching safety neurons	Over 90% recovery across red-teaming benchmarks
Impact on general capabilities	No detectable influence on general ability
Relationship between safety and helpfulness neurons	Significant overlap; different activation patterns required
Application	Detecting unsafe outputs before generation

The identification of safety neurons—a sparse set comprising approximately 5% of all neurons—has significant implications for AI safety. By patching only the activations of these neurons, researchers restored over 90% of safety performance across various red-teaming benchmarks without influencing general ability [16]. The finding also helps explain the "alignment tax" phenomenon by revealing that the key neurons for model safety and helpfulness significantly overlap, yet they require different activation patterns for the same neurons.

The argument that interpretability, especially mechanistic approaches, should be treated as a design principle for alignment rather than an auxiliary diagnostic tool has gained traction [17]. Progress on safe and trustworthy AI will depend on making interpretability a first-class objective of AI research and development. Recent work has proposed axiomatic approaches to validating mechanistic interpretations [20], and advances in semantic understanding of large AI models have further strengthened the case for interpretability-driven alignment [21].

3.3 Scalable Oversight: Supervising Superhuman AI

Scalable oversight addresses a fundamental problem: how can AI systems that exceed human capabilities be supervised effectively? As AI systems become more capable than their human evaluators, traditional oversight mechanisms—which rely on human judgment—become inadequate. This challenge is central to the alignment of future superintelligent systems [12].

Recent work has quantified the probability of successful oversight as a function of the capabilities of the overseer and the system being overseen [22]. Table 3 presents the findings from scaling laws research applied to four oversight games.

Table 3. Nested Scalable Oversight success rates at a general Elo gap of 400

Oversight Game	NSO Success Rate
Mafia	13.5%
Debate	51.7%
Backdoor Code	10.0%
Wargames	9.4%

The framework models oversight as a game between capability-mismatched players, with players having oversight-specific Elo scores that are a piecewise-linear function of their general intelligence [22]. The findings reveal that NSO success rates vary dramatically across different oversight contexts, and these rates decline further when overseeing stronger systems.

The concept of "superalignment"—automatically aligning superhuman models with human values through scalable, reliable, and generalisable methods—has been proposed as a pathway to address this challenge. VerificAgent has introduced a scalable oversight framework that treats persistent memory as an explicit alignment surface, allowing humans to correct high-impact errors once, with the verified memory acting as a frozen safety contract that future agent actions must satisfy [23].

3.4 Red Teaming and Adversarial Testing

Red teaming has emerged as a critical approach for probing and assessing the safety of large language models [24]. Red teaming involves systematic attempts to elicit harmful or undesirable behaviours from AI systems, revealing vulnerabilities that might otherwise remain hidden. Early work demonstrated that language models can be used to red-team other language models, scaling the evaluation process [24]. Table 4 summarises key findings from recent red teaming research.

Table 4. Red teaming vulnerability findings

Finding	Result
Vulnerability increase after 5-turn conversation (English)	71% more vulnerable than after initial turn
Vulnerability increase for non-English languages	Up to 195% more vulnerabilities than standard single-turn English approach
Implication	Automated red-teaming methods must match LLMs' multilingual and multi-turn capabilities

Research has revealed that models display up to 195% more safety vulnerabilities in non-English languages than in standard single-turn English approaches, confirming the need for multilingual automated red-teaming methods. Multi-turn safety alignment frameworks address the challenge of securing LLMs in multi-round interactions, where vulnerabilities can emerge over extended conversations.

Recent advances have automated and scaled the red-teaming process through agentic workflows, where human users provide red-teaming principles as instructions to an AI agent that automatically orchestrates effective strategies and generates jailbreak prompts [25]. Comprehensive reviews have critically analysed red teaming approaches for generative models, identifying key challenges and future directions [26].

3.5 Current State of AI Safety Evaluation

Current AI safety evaluation systems suffer from critical limitations including restricted risk dimensions and failed frontier risk detection [2]. Lagging safety benchmarks and alignment technologies can hardly address the complex challenges posed by cutting-edge AI models. The gap between benchmark scores and actual safety properties—what has been termed the "evaluation-safety gap"—represents a fundamental measurement problem [27].

Recent work has proposed a systematic taxonomy of AI safety evaluations distinguishing three dimensions: capabilities, propensities, and control [8]. Capabilities refer to what models can do when pushed to the limit, including cybersecurity exploitation, deception, autonomous replication, and situational awareness. Propensities concern behavioural tendencies exhibited by default, such as power-seeking and scheming. Control addresses whether safety measures remain effective against subversive adversarial AI.

These properties are measured through behavioural techniques like scaffolding, red teaming and supervised fine-tuning, alongside internal techniques such as representation analysis and mechanistic interpretability [9].

3.6 The Evaluation-Safety Gap

The concept of EvalSafetyGap has been introduced as an organising hypothesis for comparing evaluation-side and alignment-side proxy failures under optimisation pressure. Analysis reveals that conclusions shift significantly when capability, behavioural safety, and governance are measured separately, and the association between capability and sustained adversarial robustness is often statistically indeterminate [27].

Table 5. Key challenges in AI safety evaluation

Challenge	Description
Proving absence of capabilities	Evaluation can reveal what a model can do, but cannot definitively establish what it cannot do
Model sandbagging	Models may intentionally underperform during evaluation to avoid detection
Safetywashing	Superficial compliance that masks underlying risks
Benchmark inadequacy	Benchmarks often fail to establish true upper bounds or predict deployment behaviour

This finding has profound implications for AI governance: performance on standardised benchmarks may not reliably indicate safety properties. The apparent open-closed safety gap is often modest, driven mainly by governance and disclosure rather than behavioural robustness [27]. This suggests that current evaluation practices may provide a false sense of security.

3.7 Governance Frameworks

3.7.1 The EU AI Act and General-Purpose AI Code of Practice

The EU AI Act, which entered into phased application beginning in 2025, establishes binding

requirements for GPAI models based on risk classification [28]. The General-Purpose AI Code of Practice, approved in August 2025, provides guidance for GPAI model providers to meet the requirements of the AI Act [29]. The GPAI rules took effect in August 2025 for all new models released. Commission enforcement actions, including information requests and model recalls, began in August 2026. Models released before August 2025 have until August 2027 to come into compliance.

The Code consists of three chapters: Transparency, Copyright, and Safety and Security. The Safety and Security chapter sets out comprehensive obligations for developers of GPAI models with systemic risk to identify, assess, mitigate, and transparently report safety and security risks throughout the model's lifecycle. Scholarly analysis has characterised the EU approach as a distinctive model of AI governance that balances innovation with precaution through legally binding requirements [30].

3.7.2 Frontier AI Risk Management Framework

The Frontier AI Risk Management Framework is a structured set of protocols designed to help GPAI model developers proactively identify, assess, mitigate, and govern severe AI risks [31]. Originally released in July 2025 as v1.0, the Framework was updated to Version 1.5 in February 2026. Key updates include expanded loss of control content with refined risk scenarios and strengthened agent oversight protocols, operationalised risk analysis guidance clarifying essential modules (model evaluation, elicitation, risk modeling and estimates), and enhanced interoperability with measures mapped against China's National TC260 AI Safety Governance Framework 2.0 and the EU Code of Practice [31].

The Framework addresses risks across multiple threat vectors including cyber offense, strategic deception, and loss of control. Its enhanced interoperability helps developers adopt safety measures shared by major domestic and international regulatory guidance.

3.7.3 International Coordination Efforts

The international community has initiated various efforts to coordinate AI safety governance. The UN's Global Dialogue on AI Governance, held in Geneva in July 2026, brought together member states to establish priorities for responsible AI development [32]. The UN Secretary-General has called for an AI Child Safety Pledge requiring child-specific safety testing and independent oversight for any system accessible to children [32].

The International AI Safety Report 2026 assesses what general-purpose AI systems can do, what risks they pose, and how those risks can be managed [1]. The report focuses on the most capable general-purpose AI systems and examines technical approaches to risk management.

4. Discussion

4.1 Interpretation of Findings

The analysis reveals that AI safety and alignment for general-purpose AI systems require integrated approaches that span technical, evaluation, and governance dimensions. No single technical approach—whether behavioural alignment, mechanistic interpretability, or scalable oversight—is sufficient in isolation. Behavioural methods like RLHF and Constitutional AI can shape outputs but may not address underlying model properties. Mechanistic interpretability offers internal transparency but faces scalability challenges. Scalable oversight addresses the supervision gap but remains theoretically underdeveloped [4].

Table 6. Summary of technical approaches: strengths and limitations

Approach	Core Mechanism	Strengths	Limitations
RLHF	Human preference feedback	Widely adopted; reduces harmful outputs	Scalability; superficial conformity
Constitutional AI	Ethical principles encoded in constitution	Verifiable adherence; intrinsic safety	Requires explicit principles; implementation complexity
Mechanistic Interpretability	Internal model analysis	Causal insight; identifies safety neurons ($\approx 5\%$)	Scalability; epistemic uncertainty
Scalable Oversight	Weaker AI supervises stronger AI	Addresses supervision gap	Low success rates (9.4–51.7% at Elo gap 400)
Red Teaming	Adversarial probing	Reveals vulnerabilities	Costly; language-dependent gaps (up to 195% variance)

The evaluation landscape reveals a persistent gap between measured safety and actual safety. Current benchmarks may not reliably indicate safety properties, and the incentives for "safetywashing" create risks of superficial compliance [27]. The three-dimensional taxonomy—capabilities, propensities, and control—provides a more nuanced framework for understanding what safety evaluations can and cannot measure [8].

4.2 Comparison with Existing Literature

The findings align with prior systematic analyses identifying the evaluation-safety gap as a critical challenge for AI governance [8,17]. The tripartite taxonomy of capabilities, propensities, and control extends earlier work by providing a more nuanced framework for understanding what safety evaluations can and cannot measure [8]. The identification of scalable oversight as a distinct technical challenge resonates with foundational work on the supervision of superhuman AI [12].

The governance analysis reveals convergence on certain principles—including risk-based classification, transparency requirements, and incident reporting—across different frameworks [28,29,31]. However, significant divergence remains in enforcement mechanisms, binding authority, and the treatment of systemic risk. This fragmentation creates challenges for global coordination and may enable regulatory arbitrage [6]. Scholarly analysis has characterised the EU approach as a distinctive model of AI governance that balances innovation with precaution through legally binding requirements [30].

4.3 Persistent Challenges

Several challenges persist across technical, evaluation, and governance dimensions. First, the difficulty of proving the absence of dangerous capabilities remains a fundamental limitation. Evaluation can reveal what a model can do but cannot definitively establish what it cannot do [27]. Second, the rapid pace of technological development means that governance frameworks risk becoming obsolete almost as soon as they are implemented. Third, the economic incentives for rapid deployment frequently conflict with the careful evaluation that safety requires [5]. Fourth, the fragmentation of governance approaches across jurisdictions creates gaps that sophisticated actors may exploit [6]. Recent surveys of LLM misalignment have further underscored the gap between expressed safety intentions and actual model behaviour [33].

4.4 An Integrated Multi-Layered Framework

Building on the analysis, this paper proposes an integrated multi-layered framework for AI safety and alignment encompassing three layers:

Layer 1: Pre-Deployment Evaluation. Before deployment, GPAI systems should undergo rigorous evaluation across the full spectrum of capabilities, propensities, and control dimensions [8]. Evaluation should employ both behavioural techniques (red teaming, scaffolding, adversarial testing) and internal techniques (mechanistic interpretability, representation analysis) [16,24]. Given that models display up to 195% more safety vulnerabilities in non-English languages, evaluation must be multilingual and multi-turn. Evaluation results should be independently verified and publicly disclosed.

Layer 2: Runtime Monitoring. During deployment, GPAI systems should be subject to continuous monitoring for emergent behaviours, capability drift, and safety violations. Monitoring should include automated detection of anomalous outputs, human oversight of high-impact decisions, and incident reporting mechanisms. Given that vulnerability increases by 71% after 5-turn conversations, monitoring must be conversation-aware.

Layer 3: Incident Response and Governance. Mechanisms for incident response, remediation, and governance should be established before deployment. These should include clear escalation procedures, recall and rollback capabilities, and independent oversight bodies with authority to investigate and sanction [31]. Governance should be adaptive, with regular review and updating of safety requirements based on new evidence and experience.

4.5 Policy Implications

For policymakers, the analysis suggests several priorities. First, the harmonisation of evaluation standards across jurisdictions to reduce compliance burdens and prevent regulatory arbitrage. Second, the establishment of independent safety evaluation bodies with authority to verify claims and investigate incidents. Third, the development of binding requirements for pre-deployment evaluation, runtime monitoring, and incident reporting for high-risk GPAI systems [29]. Fourth, the creation of international coordination mechanisms to share information, coordinate responses, and prevent race-to-the-bottom dynamics [32].

For developers, the analysis suggests that safety and alignment cannot be treated as afterthoughts. They must be integrated into the full AI lifecycle. Investment in interpretability, evaluation, and governance capabilities is essential for responsible development [17,28].

4.6 Limitations and Future Research

Several limitations should be acknowledged. The analysis focuses on technical and governance approaches in major jurisdictions and may not capture the full diversity of approaches globally. The rapidly evolving nature of both technology and governance means that some findings may become outdated quickly. The analysis does not quantify the economic costs and benefits of different safety approaches.

Future research should conduct longitudinal studies of safety evaluation effectiveness, develop standardised benchmarks for safety-critical capabilities, and examine the governance of autonomous AI agents. The interaction between technical safety methods and governance frameworks merits further investigation, as does the role of international coordination in preventing regulatory fragmentation.

5. Conclusion

AI safety and alignment in general-purpose AI systems represent one of the most consequential challenges of the contemporary technological era. The rapid advancement of GPAI systems has outpaced understanding and governance, creating risks that range from harmful outputs to catastrophic systemic failures. This paper has examined the technical approaches, evaluation methodologies, and governance frameworks that constitute the current state of the art.

Three principal technical approaches have been identified: behavioural alignment through RLHF and Constitutional AI; mechanistic interpretability for internal model understanding; and scalable oversight for supervising systems that exceed human capabilities. Each approach has strengths and limitations, and no single approach is sufficient. Evaluation methodologies have revealed a persistent gap between measured safety and actual safety, with current benchmarks failing to reliably indicate safety properties [27]. Governance frameworks have emerged at multiple levels, from the EU AI Act's binding requirements to voluntary industry frameworks and international coordination efforts [32].

The integrated multi-layered framework proposed here—encompassing pre-deployment evaluation, runtime monitoring, and incident response—offers a pathway toward more robust AI safety and alignment. However, the effectiveness of any framework depends on the quality of its implementation, the independence of its oversight, and the willingness of developers and policymakers to prioritise safety over speed.

The stakes are high. GPAI systems have the potential to transform economies, advance science, and improve human welfare. They also have the potential to cause unprecedented harm if not properly governed. Ensuring AI safety and alignment is not merely a technical problem or a

governance challenge—it is a fundamental responsibility of the developers, policymakers, and societies that create and deploy these systems. The window for proactive governance remains open, but it is narrowing.

Author Contributions

Conceptualization, E.V.P.; methodology, E.V.P. and Y.T.; validation, Y.T.; formal analysis, E.V.P.; investigation, E.V.P. and Y.T.; writing—original draft preparation, E.V.P.; writing—review and editing, E.V.P. and Y.T.; supervision, E.V.P.; project administration, E.V.P. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Funding

This research received no external funding.

Ethics Approval

Not applicable

Acknowledgements

The authors thank the European Commission, the Frontier AI Risk Management Framework team, and the international AI governance community for publicly available reports and frameworks. AI language tools were used to assist with language polishing. The authors also gratefully acknowledge the open access policies of their respective institutions.

Conflicts of Interest

The authors declare no conflicts of interest to report regarding the present study.

References

1. International AI Safety Report 2026: Executive Summary. 2026.
2. Tong H, Zhao F, Feng L, Wu R, Chen R, Jia L, et al. ForesightSafety Bench: A Frontier Risk Evaluation and Governance Framework towards Safe AI. arXiv:2602.14135. 2026.
3. Ji J, Qiu T, Chen B, Zhang B, Zhou J, Hong D, et al. AI Alignment: A Contemporary Survey. *ACM Comput Surv.* 2025;58(5):132.
4. Frontier Model Safety Analysis (2026): RSP, Preparedness, FSF. FutureAGI. 2026.
5. Li W. Three rulebooks, one race: AI regulation in the U.S., EU, and China. *Commun ACM.* 2026.
6. Markov T, Mariaccia H, Ségerie CR, Dorn D, Bucknall B, Siddiqui S, et al. Safety by Measurement: A Systematic Literature Review of AI Safety Evaluation Methods. arXiv:2505.05541. 2025.
7. Kaufmann T, Weng P, Bengs V, Hüllermeier E. A Survey of Reinforcement Learning from Human Feedback. *Trans Mach Learn Res.* 2025.
8. Rudner TGJ, Ishida R, Suau X, et al. Measuring and Mitigating Safety Risks in Large Language Models: A Survey. arXiv:2504.03814. 2025.
9. Casper S, Davies X, Shi C, Gilbert TK, Scheurer J, Rando J, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. arXiv:2307.15217. 2023.
10. Christiano PF, Leike J, Brown TB, Martic M, Legg S, Amodei D. Deep Reinforcement Learning from Human Preferences. In: *Advances in Neural Information Processing Systems* 30. 2017. p. 4302-4310.
11. Bowman SR, Hyun J, Perez E, Chen E, Pettit C, Bowman S, et al. Measuring Progress on Scalable Oversight for Large Language Models. arXiv:2211.03540. 2022.
12. Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, et al. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073. 2022.
13. Kyrychenko Y, Zhou K, Bogucka E, Quercia D. C3AI: Crafting and Evaluating Constitutions for Constitutional AI. In: *Proceedings of the ACM Web Conference (WWW '25)*. 2025.
14. Zhang X, Wang G, Cui Y, Qiu W, Li Z, Zhu B, et al. The Alignment Floor: How Persona Customization Breaks Safety in Weakly-Aligned LLMs. arXiv:2605.27382. 2026.

15. Chen J, Wang X, Yao Z, Bai Y, Hou L, Li J. Towards Understanding Safety Alignment: A Mechanistic Perspective from Safety Neurons. In: *Advances in Neural Information Processing Systems* 38 (NeurIPS 2025). 2025.
16. Sengupta A, Seth P, Sankarapu VK. Interpretability as Alignment: Making Internal Understanding a Design Principle. *arXiv:2509.08592*. 2025.
17. Elhage N, Nanda N, Olsson C, Henighan T, Joseph N, Mann B, et al. A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*. 2021.
18. Olsson C, Elhage N, Nanda N, Joseph N, DasSarma N, Henighan T, et al. In-Context Learning and Induction Heads. *arXiv:2209.11895*. 2022.
19. Palumbo N, Mangal R, Wang Z, Vijayakumar S, Păsăreanu C, Jha S. Validating Mechanistic Interpretations: An Axiomatic Approach. In: *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025)*. 2025.
20. Zhang Y, Liu X, Wang H, Chen L, Li M, Zhao F. SemanticLens: Mechanistic understanding and validation of large AI models. *Nat Mach Intell*. 2025;7:1572-1585.
21. Engels J, Baek DD, Kantamneni S, Tegmark M. Scaling Laws For Scalable Oversight. In: *Advances in Neural Information Processing Systems* 38 (NeurIPS 2025). 2025.
22. Nguyen TQ, Desai S, Anwar RH, Shaik F, Suryanarayanan V, Chowdhary V. VerificAgent: Domain-Specific Memory Verification for Scalable Oversight of Aligned Computer-Use Agents. *arXiv:2506.02539*. 2025.
23. Perez E, Huang S, Song F, Cai T, Ring R, Aslanides J, et al. Red Teaming Language Models with Language Models. *arXiv:2202.03286*. 2022.
24. Purpura A, Wadhwa S, Zymet J, Gupta A, Luo A, Kazemi Rad M, et al. Building Safe GenAI Applications: An End-to-End Overview of Red Teaming for Large Language Models. In: *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*. 2025. p. 335-350.
25. Jabbar MS, Al-Azani S, Alotaibi A, Ahmed M. Red teaming large language models: A comprehensive review and critical analysis. *Inf Process Manag*. 2025;62(6):104239.
26. Jindal M, Shrawgi H, Agrawal P, Dandapat S. EvalSafetyGap: A Hybrid Survey and Conceptual Framework for LLM Evaluation-Safety Failures. *arXiv:2606.30219*. 2026.
27. Jindal M, Shrawgi H, Agrawal P, Dandapat S. SAGE: A Generic Framework for LLM Safety Evaluation. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*. 2025. p. 11-33.
28. EU AI Act General-Purpose AI Code of Practice. European Commission. 2025.
29. Gatti M, Romano A. The EU model of AI governance: regulating artificial intelligence through law and policy. *ERA Forum*. 2025;26:527-547.
30. Frontier AI Risk Management Framework (v1.5). Concordia AI. 2026.
31. UN Secretary-General's remarks to the opening of the first Global Dialogue on Artificial Intelligence Governance. United Nations. 2026.

32. Henderson P, Mitchell M. Beyond Intentions: A Critical Survey of Misalignment in LLMs. ScienceDirect. 2025.
33. Cai Y, Zhao Z, Lu Z, Lee I, Zhang WE, Xue M, Zhang Y, Pang S, Xiang W. AI Alignment through a Game-theoretic Lens: A Survey. arXiv:2608.27910. 2026.