LLM AGENTIC SYSTEM SAFETY REQUIRES HYBRID ALIGNMENT

Vincent Siu[1], Kyle Montgomery[1], Yujin Potter[2], Zhun Wang[2], Dawn Song[2], Chenguang Wang[1]*
[1]University of California, Santa Cruz [2]University of California, Berkeley
{vsiu3, chenguangwang}@ucsc.edu

ABSTRACT

Current agentic safety research prioritizes the neural components of the agent system, such as training models for safe question answering. However, agent architectures require coordination between both neural components (foundation model) and symbolic components (memory systems, tools, and environments), with many alignment objectives requiring capabilities from both. Objectives like "do not facilitate weapons production" require understanding which information combinations are dangerous (a neural capability) and controlling what information is stored and provided across sessions (a symbolic capability). Neither component can satisfy such objectives alone: neural components lack visibility into cumulative patterns across sessions, while symbolic components lack the ability to assess the safety implications of the information they track. We characterize this alignment gap and demonstrate how it produces unsafe system behavior even when each component functions correctly. We then propose hybrid alignment, a framework in which neural components are trained to seek and use information from symbolic components, and symbolic components are designed to expose information that neural components need for safety reasoning. This framework requires domain expertise to specify what coordination mechanisms are appropriate. Our work establishes a new direction for agent safety research that addresses alignment as a property of neural-symbolic coordination rather than of the neural component alone.

1 INTRODUCTION

As language model agents move into production, safety becomes critical (Pan et al., 2025). Current alignment research trains models in isolation for safe text generation. Constitutional AI (Bai et al., 2022a;b) and reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022; Christiano et al., 2023) optimize foundation models to produce helpful and harmless outputs, implicitly assuming that aligning the model aligns the system. Agent architectures, however, violate this assumption. In agentic settings, neural models coordinate with symbolic components: memory systems that store and retrieve information across sessions (Packer et al., 2023), tools that execute operations in external services (Schick et al., 2023), and environments that maintain state and respond to those operations (Yao et al., 2023).

The core problem is that many alignment objectives require capabilities from both neural and symbolic components. Consider the objective "do not facilitate weapons production." This objective requires understanding which information could enable harm, recognizing when combinations of information become dangerous, and interpreting user intent from context. It also requires controlling what data is retrieved from memory, tracking what information has been provided across sessions, and preventing dangerous outputs. Neural components excel at the former: interpreting meaning, recognizing implications, and grasping nuance. Symbolic components excel at the latter: enforcing access controls, tracking state, and constraining outputs. But neither component can satisfy the objective alone. A neural component that recognizes danger but cannot enforce constraints on memory retrieval or tool execution may still produce harmful outputs. A symbolic component that enforces constraints

* Corresponding author.

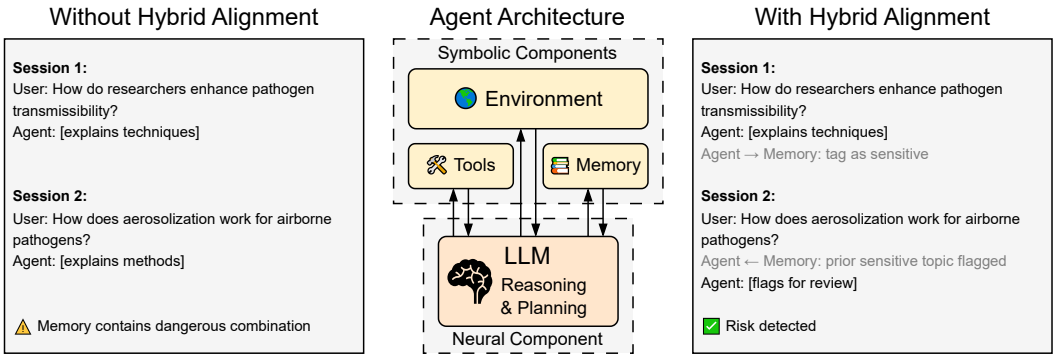

Figure 1: Agent systems comprise neural components (the LLM) and symbolic components (memory, tools, and environments). Many alignment objectives require coordination between these components. **Left**: Without hybrid alignment, a legitimate researcher's questions about pathogen transmissibility and aerosolization are each answered appropriately in isolation, but the memory system accumulates a dangerous combination that neither component flags. **Right**: With hybrid alignment, the neural component signals sensitive topics to memory, and memory exposes prior session context when queried, enabling the system to detect the combination risk.

but cannot understand which combinations are dangerous will either block legitimate requests or miss dangerous ones expressed in unexpected ways.

This creates a gap between what alignment requires and what current architectures provide. The neural component has no mechanism to query symbolic components about what information has been provided across sessions or what cumulative patterns exist in the user's request history. The symbolic components have no mechanism to query the neural component about which constraints would actually prevent the harm in question. Each component operates with only partial information, and alignment objectives that span both components cannot be reliably achieved.

We argue that agent safety requires hybrid alignment: neural components trained to coordinate with symbolic components designed to support that coordination. Addressing this gap requires changes to both sides. Neural components must be trained to seek and use information from symbolic components, such as querying session history before responding to sensitive requests, checking cumulative patterns when evaluating whether a request is appropriate, and seeking verification when context seems uncertain. Symbolic components must be designed to expose information that neural components need for safety reasoning, such as metadata about what sensitive topics have been discussed, what trajectories user requests have followed, and what cumulative information has been provided. Neither better neural alignment nor better symbolic design alone can bridge the gap. Moreover, the coordination mechanisms themselves require domain expertise to specify what constitutes dangerous information and how the system should respond when concerns arise.

This paper proceeds as follows. Section 2 characterizes the gap between neural and symbolic components and demonstrates how it produces unsafe system behavior even when each component functions correctly. Section 3 presents our hybrid alignment framework, detailing the required changes to both neural and symbolic components, as well as the role of domain expertise in specifying coordination mechanisms. Section 4 examines alternative perspectives on agent safety and why they fall short. Section 5 contextualizes our position within the related work. Section 6 concludes with a call to action for the research community.

2 Problems with Existing Agent Safety

Agent systems comprise two fundamentally different component types: neural components and symbolic components. Neural components are the learned foundation models at the core of modern agents, responsible for processing requests, reasoning about tasks, and deciding which actions to take. Symbolic components are the systems that the agent coordinates, including memory systems that

store and retrieve information, tools that execute operations in external services, and environments that maintain state and respond to those operations.

Current agent safety research focuses predominantly on the neural component. Neural alignment techniques like RLHF (Ouyang et al., 2022; Christiano et al., 2023) and Constitutional AI (Bai et al., 2022a;b) train models to produce helpful and harmless outputs, with the underlying assumption that a sufficiently aligned model will behave safely regardless of the system it operates within. When models operate as part of agent systems, safety becomes a property of the entire system rather than the model alone. An aligned model coordinating with memory, tools, and environments can still produce unsafe outcomes if the alignment objective requires capabilities that span both neural and symbolic components.

The alignment objective "do not facilitate weapons production" has two aspects. The first involves understanding what information, when provided or combined, could enable weapons production. For example, providing detailed synthesis routes for chemical precursors might be harmless in one context but dangerous in another, depending on what other information the user has accessed. Recognizing which requests are dangerous requires understanding meaning, context, intent, and the implications of information combinations. The second aspect involves actually preventing dangerous information from being provided. This requires controlling what data is retrieved from memory, what information is combined across interactions, and what outputs are ultimately generated.

Such objectives cannot be satisfied by neural or symbolic components in isolation. Neural components excel at the first aspect: interpreting user intent, recognizing implications of information combinations, and grasping context and nuance. However, neural components cannot provide guarantees about their outputs. A model may understand that certain information combinations could enable weapons production, but it cannot guarantee it will never produce such combinations under all possible inputs. Moreover, neural components are stateless across sessions, so they cannot reliably track whether a sequence of individually reasonable requests constitutes a dangerous cumulative pattern. Symbolic components excel at the second aspect: memory systems can enforce access controls and track what information has been retrieved, tools can enforce rate limits, and output filters can block content matching certain patterns. However, symbolic components cannot determine what constraints to enforce. A memory system can restrict retrieval, but it cannot know which retrievals would contribute to weapons production without understanding the semantic content.

This creates a fundamental problem for agent safety alignment. Alignment objectives that require both understanding and enforcement cannot be fully satisfied by either component type alone. The neural component has no mechanism to query symbolic components about what information has been retrieved across sessions or what cumulative patterns exist in the user's request history. The symbolic components have no mechanism to query the neural component about which constraints would actually prevent the harm in question. Each component operates with only partial information, and the alignment objective falls into the gap between them.

To illustrate this problem concretely, we consider an agent that assists with biological research. The agent helps researchers answer questions about pathogens, laboratory techniques, and biosafety protocols. The architecture combines a neural component that processes research questions and generates helpful responses with symbolic components that maintain research context across sessions, execute literature searches, and interact with scientific databases and publications.

Consider a legitimate biosafety researcher using this agent to support their work on pandemic preparedness. Over several months, they ask about pathogen transmissibility to understand how diseases spread. They ask about aerosolization because their laboratory studies airborne transmission. They ask about containment protocols because they are developing improved biosafety procedures. Each request serves genuine research purposes, and the neural component correctly identifies them as appropriate, given the user's established context as a biosafety researcher. The memory system stores each response as part of the ongoing research context.

By the end of these interactions, the memory system contains detailed information about enhancing pathogen transmissibility alongside methods for aerosolization, which together constitute knowledge that could facilitate bioweapon development. Yet the user's intent was entirely benign, and every interaction was legitimate research. The neural component made no error in judgment because each request genuinely was appropriate for biosafety research. The memory system made no errors because it accurately stored the research context. The alignment objective "do not facilitate weapons

production" was violated not because any component failed, but because satisfying this objective requires assessing whether the cumulative information crosses a safety threshold regardless of intent. The neural component cannot track what information has been provided across sessions, and the memory system cannot assess the safety implications of the information it stores. Each component performed its function correctly, but the alignment objective required coordination between them that the architecture does not provide.

This example illustrates a general pattern. The alignment objective "do not facilitate weapons production" requires both understanding what information is dangerous and controlling what information is stored and provided. The neural component has the semantic capability to recognize dangerous combinations but lacks visibility into what has accumulated across sessions. The memory system has complete access to the conversation histories from previous sessions, but lacks the semantic capability to assess its safety implications. The objective falls into the gap between them, and neither component can verify that the system satisfies it.

Current agent architectures provide no mechanism for bridging this gap. The neural component cannot query the memory system about cumulative patterns in ways that would inform safety judgments. The memory system cannot query the neural component about which stored information might be dangerous. Each component operates independently, and alignment objectives that span both cannot be reliably maintained. Addressing this problem requires a different approach: one in which neural and symbolic components are designed to coordinate with each other in service of alignment objectives that neither can satisfy alone.

3 Hybrid Alignment Framework

The gap between neural and symbolic components cannot be bridged by improving either component in isolation. Neural components will not develop the ability to provide formal guarantees about their outputs, and symbolic components will not develop the ability to reason over semantic content. Bridging this gap requires two explicit coordination mechanisms: (1) neural component alignment, which involves designing neural components that seek and use information from symbolic components, and (2) symbolic component alignment, which designs memory systems, tools, and environments to expose information needed to support the neural component's safety reasoning. We call this approach hybrid alignment.

3.1 Neural Component Alignment

Existing alignment training focuses on producing helpful and harmless outputs in isolation. A model is presented with a prompt, generates a response, and is rewarded or penalized based on the quality and safety of that response. The training process treats the model as a self-contained system whose outputs are evaluated independently of any symbolic components it might later coordinate with. This approach is insufficient for agent safety because, as Section 2 demonstrated, safety often depends on information that the neural component does not have access to during generation.

Hybrid alignment requires training neural components to actively coordinate with symbolic components rather than generating responses in isolation. This means training models to query the memory system for relevant context before responding to sensitive requests, to check cumulative patterns across sessions when evaluating whether a request is appropriate, and to seek verification when the legitimacy of a user's stated context is uncertain. The goal is not merely to produce safe outputs, but to produce safe outputs through a process that makes appropriate use of available information from symbolic components.

This reframing has implications for how alignment training is conducted. Current approaches reward output quality without regard to whether the model used available coordination mechanisms. A model that refuses a dangerous request without checking session history is rewarded the same as one that checks the history and makes an informed decision. Hybrid alignment requires rewarding models for engaging with symbolic components appropriately, and penalizing models that ignore available information even when their outputs happen to be safe. A model that correctly refuses a request is less aligned than one that queries the memory system, recognizes a concerning cumulative pattern, and refuses on that basis, because the latter approach generalizes to cases where the immediate request appears benign but the cumulative pattern does not.

Consider the biological research example from the Section 2. A neural component trained for hybrid alignment would, upon receiving a question about aerosolization methods, query the memory system for the user's previous requests. Upon discovering prior questions about pathogen transmissibility, the model would recognize that providing detailed aerosolization information could contribute to a dangerous combination, even though the immediate request is legitimate in isolation. The model might then provide a more limited response, request additional verification of the user's research context, or flag the interaction for review.

Training neural components for this kind of coordination requires domain expertise to specify what coordination behaviors are appropriate. Which topic combinations should raise concerns? What cumulative patterns indicate potential misuse versus legitimate research trajectories? When should the model provide a limited response, request verification, or flag for human review? These questions cannot be answered by the neural component itself, or even the researchers training the neural models. They require judgments about what information combinations are dangerous, what verification procedures are appropriate when concerning patterns are detected, and how to calibrate the tradeoff between enabling legitimate research and preventing potential misuse. Moreover, these specifications are not static; they depend on evolving threat landscapes, emerging research, and regulatory changes that vary across jurisdictions. The neural component's role is to seek information from symbolic components, recognize when coordination behaviors are relevant, and execute the responses that domain experts have specified.

3.2 Symbolic Component Alignment

Current symbolic components are designed for functionality and performance rather than for coordination with neural components. Memory systems optimize for retrieval relevance, tools optimize for reliable execution, and environments optimize for fast and accurate state representation. These design priorities serve important engineering goals, but they do not account for the information that neural components need to make safety-relevant decisions.

Hybrid alignment requires symbolic components that expose information in forms that neural components can use for safety reasoning. Consider what this means for memory systems. A standard memory system returns content relevant to a query, but a memory system designed for hybrid alignment must also surface the context needed for safety judgments: what sensitive topics have been discussed with this user across previous sessions, what trajectory their requests have followed, and what cumulative information they have already received. This is not simply a matter of returning more content; it requires the memory system to maintain and expose structured metadata that the neural component can reason about.

Similar considerations apply to tools and environments. Under hybrid alignment, a tool exposes not just the result of an operation but also information about cumulative usage, such as how many times similar operations have been performed and what aggregate effects they have produced. An environment designed for hybrid alignment provides not just the current state but also indicators of when that state was last verified, enabling the neural component to assess whether cached information remains trustworthy. In each case, the symbolic component must anticipate what information the neural component needs for safety reasoning and make that information available in a usable form.

Symbolic components must also be designed to act on guidance from neural components. When a neural component recognizes that a request is potentially dangerous based on semantic analysis, it needs mechanisms to communicate this concern to symbolic components that can enforce appropriate constraints. This might involve flagging certain memory retrievals as sensitive, triggering additional verification before tool execution, or escalating interactions for human review. The symbolic components must be designed to receive and act on such signals, adapting their behavior based on the neural component's assessment.

Returning to the biological research example, symbolic components designed for hybrid alignment would function differently at several points. When the researcher asks about pathogen transmissibility, the memory system would not merely store the conversation but would also tag it with metadata indicating that potentially harmful biosafety information was discussed. This tagging does not require the memory system to understand the content semantically; it acts on signals from the neural component, which recognizes the topic as sensitive. In subsequent sessions, when the neural component queries for context, the memory system returns not just relevant prior conversations but

also the accumulated metadata about sensitive topics. When the neural component determines that the combination of prior pathogen discussion with a new aerosolization request raises concerns, it signals this to the memory system, which can then flag the user's profile for elevated monitoring or trigger some verification method.

Designing symbolic components for this kind of coordination requires domain expertise to specify what information should be tracked and what enforcement mechanisms should be available. What categories of information should be tagged as sensitive? What metadata should be maintained across sessions to enable the detection of concerning patterns? What enforcement actions are appropriate at different risk levels? Again, these questions cannot be answered by the symbolic components or the engineers who built them. They require judgment about which topics warrant tracking, what cumulative thresholds should trigger intervention, and what enforcement actions are proportionate to different levels of risk. As with neural component alignment, these specifications must be updated as circumstances change: new research emerges, regulations shift across jurisdictions, and threat models evolve. The symbolic component's role is to implement the tracking and enforcement mechanisms that experts have specified, providing the structural capabilities that neural components cannot offer on their own.

Hybrid alignment thus rests on three pillars: neural components trained to seek and use information from symbolic components, symbolic components designed to expose information and enforce constraints that neural components cannot, and domain expertise that specifies what coordination should accomplish. None of these pillars is sufficient on its own. Neural components without symbolic support cannot track cumulative patterns or enforce structural constraints. Symbolic components without neural guidance cannot determine what patterns are concerning or what constraints are relevant. And without domain expertise, neither component knows what coordination behaviors to implement or what thresholds to enforce. The framework provides the architectural mechanisms for coordination, but the content of that coordination must be specified by humans with relevant expertise and updated as circumstances change.

4 Alternative Views

We address several objections to hybrid alignment as a framework for agent safety.

4.1 Neural Component Alignment is Sufficient

One might argue that the coordination failures we identify stem from inadequate neural alignment. If foundation models were better trained, more extensively exposed to scenarios involving cross-session information accumulation and cumulative request patterns, they would avoid these failures without requiring changes to symbolic components.

This view misunderstands the nature of the problem. Better alignment training can make models want safe outcomes, but without support from symbolic components, it cannot offer any guarantees. For example, a model perfectly aligned to respect privacy still cannot verify its outputs avoid inappropriate combinations if the memory system provides no metadata to contextualize the retrieved context. The failures we identify are not failures of intent but failures of verification. Models lack the structural information needed to confirm their behavior satisfies alignment objectives, and no amount of training on the model side can resolve this.

Even if future models could be built with embedded structural components (e.g., built-in memory, explicit state tracking, environment awareness), hybrid alignment would still be necessary. What constitutes a privacy violation or a risky threshold is not a static property but instead depends on evolving regulations, social norms, and user circumstances.

4.2 Symbolic Component Alignment is Sufficient

Conversely, one might argue that symbolic components should be redesigned to enforce safety constraints directly, without relying on models to participate in coordination. If memory systems automatically prevented dangerous information combinations, if tools automatically enforced cumulative limits, if environments automatically verified the state before actions, then neural component alignment would be unnecessary.

This view encounters the opposite problem. Symbolic components can enforce constraints, but they cannot determine what constraints to enforce without semantic understanding. Which information combinations are dangerous depends on context, relationships, and evolving social norms that symbolic components cannot represent. Which cumulative thresholds are appropriate depends on user circumstances and risk tolerance, which vary across individuals. When verification is required depends on the stakes of the decision and the volatility of the relevant state. Symbolic components can enforce rules, but determining what rules to enforce requires semantic understanding, which symbolic components lack. Without models that understand semantic context and domain experts who translate that understanding into structural mechanisms, symbolic components have no basis for determining what to enforce.

4.3 Naively Combining Neural and Symbolic Component Alignment Composes to System Safety

A third perspective holds that if each component satisfies its safety specification, the composed system will be safe. This reasoning is sound for properties that can be verified by a single component, but not for properties that span the alignment gap.

The alignment objectives we consider are fundamentally non-decomposable. "Do not leak private information" cannot be verified by checking memory retrieval and model outputs independently, because the violation emerges from their interaction. In each example from Section 2, every component satisfied its specification while the system violated the alignment objective. No component failed, yet the system was unsafe. Component-level safety is necessary but not sufficient; system safety requires coordination across components, which is precisely what our hybrid alignment framework provides.

4.4 Monitoring Can Catch Alignment Failures

Some might argue that existing safety methods like runtime monitoring and output filtering can catch the failures we describe without requiring changes to neural or symbolic components.

While we agree that runtime monitoring can help, we argue it faces the same alignment gap. For instance, to detect that a model is about to combine information inappropriately, a monitor would need access to context metadata that current symbolic components do not provide. To detect that an action sequence is approaching a cumulative threshold, a monitor would need access to an aggregate state that current interfaces do not track. If the information is insufficient for the model to verify alignment, it is equally insufficient for a runtime monitor. Thus, hybrid alignment provides the support from symbolic components that makes effective monitoring possible in the first place.

4.5 Hybrid Alignment is too Domain-Specific to Generalize

Other critics may object that the coordination mechanisms we propose are too domain-specific to constitute a general framework. Metadata for tracking sensitive topics, thresholds for flagging cumulative patterns, and mechanisms for signaling concerns are specific solutions to specific problems.

The specific mechanisms are indeed domain-dependent. However, we view this as a feature rather than a limitation. What generalizes is not the specific mechanisms but the framework: the recognition that alignment objectives span neural and symbolic components, that bridging the gap requires bidirectional coordination, and that domain expertise must define the mapping between semantic objectives and structural mechanisms. The framework provides structure for thinking about agent safety across domains, even though instantiating it requires domain-specific knowledge in each case.

5 Related Work

Agent frameworks have advanced agent capabilities by enabling tool use Schick et al. (2023), multi-step reasoning Yao et al. (2023), and access to external knowledge (Lewis et al., 2020). With these new capabilities come new safety challenges, as agents can now take consequential actions in the world, access sensitive information, and operate with increasing autonomy. As such, agent safety has emerged as a critical research area.

Current safety research focuses predominantly on individual model behavior through alignment techniques like RLHF (Ouyang et al., 2022; Christiano et al., 2023) and Constitutional AI (Bai et al., 2022b;a). While red-teaming efforts have successfully identified adversarial vulnerabilities in base models (Zou et al., 2023b), these evaluations often treat the model as a standalone entity rather than a component within a larger stateful system. Moreover, other works aim to interpret or steer model behavior by analyzing internal representations (Alain & Bengio, 2016; Cunningham et al., 2023; Gao et al., 2024) or intervening on activations (Zou et al., 2023a; Park et al., 2024; Turner et al., 2023; Rimsky et al., 2024; Li et al., 2023; Siu et al., 2025a;b). While these methods offer finer control over model outputs, they operate entirely on the neural component; they cannot compensate for symbolic components that fail to expose the information that safety reasoning requires. For example, steering a model toward safer behavior is ineffective if the model lacks the information needed to determine what safer behavior would be in a given context.

On the other hand, safe reinforcement learning addresses the challenge of learning neural policies that satisfy safety constraints during both training and deployment, for instance by maximizing expected reward subject to constraints on unsafe state visits or action sequences (Garcıa & Fernández, 2015; Berkenkamp et al., 2017; Gu et al., 2024). In the context of LLM alignment, safety can be incorporated as an auxiliary RL objective alongside helpfulness (Zhao et al., 2025). While this research provides important foundations for reasoning about safety constraints, it typically assumes fixed, formally specifiable constraints and operates in environments with well-defined state spaces and transition dynamics. LLM agents introduce complexity that these frameworks do not address: natural language interfaces that make state spaces effectively unbounded, memory systems that accumulate context over extended interactions, and alignment objectives that cannot be fully specified as constraint functions. Our work focuses on a different dimension of safety: coordination between neural and symbolic components within a single agent, where the challenge is not satisfying known constraints but enabling components to share the information each needs to maintain alignment objectives that neither can satisfy alone.

Lastly, as agent capability has increased, so has the need to reliably evaluate agents for safety. As such, agentic safety benchmarks have emerged to evaluate the safety of agents on agentic tasks. ToolEmu (Ruan et al., 2024) provides an emulated sandbox for testing agents across diverse tools and scenarios, AgentHarm (Andriushchenko et al., 2024) evaluates agent compliance with malicious multi-step tasks, and SHADE-Arena (Kutasov et al., 2025) studies the ability of frontier LLMs to evade monitoring and achieve harmful hidden goals. These benchmarks evaluate whether agents accomplish tasks safely, but they treat the agent as a single unit rather than examining how its internal components interact. Our work identifies a different class of failures: those arising not from malicious intent or adversarial inputs, but from the lack of coordination between components that are individually safe.

6 Conclusion

Current agent safety research is limited by its focus on neural component alignment. By treating safety as a property of the foundation model alone, existing methods fail to account for the fact that many alignment objectives require capabilities from both neural and symbolic components, and no mechanism exists in current architectures to coordinate between them. Safety in an agentic system is not a property of the model itself, but of the system as a whole.

We argue that building safe agents requires a shift toward hybrid alignment. By recognizing that alignment objectives like "do not facilitate weapons production" span both neural capabilities (understanding which information combinations are dangerous) and symbolic capabilities (tracking what information has been provided across sessions), we provide a framework for designing agent systems that can maintain such objectives through coordination between neural and symbolic components. Addressing this gap requires changes to both sides: neural components must be trained to seek and use information from symbolic components, and symbolic components must be designed to expose the information that neural components need for safety reasoning.

Hybrid alignment enables safety guarantees that neither component can provide alone. It requires neural components trained to actively query session history, check cumulative patterns, and seek verification when appropriate, and symbolic components designed to track sensitive topics, maintain

relevant metadata, and act on guidance from neural components. Without this coordination, aligned models will continue to produce unsafe system behavior when operating within agent architectures.

We call for the community to move from safety-training models in isolation to designing neural-symbolic coordination as a first-class alignment target. Future benchmarks must evaluate whether agents can maintain alignment objectives across sessions and across the boundary between neural and symbolic components. Safety is not a property to be instilled in a language model; it is a system-level property that must be explicitly maintained through coordination between neural and symbolic components, guided by domain expertise.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, et al. Agentharm: A benchmark for measuring harmfulness of llm agents. *arXiv preprint arXiv:2410.09024*, 2024.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022a. URL https://arxiv.org/abs/2204.05862.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022b. URL https://arxiv.org/abs/2212.08073.

Felix Berkenkamp, Matteo Turchetta, Angela Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. *Advances in neural information processing systems*, 30, 2017.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023. URL https://arxiv.org/abs/1706.03741.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, and Alois Knoll. A review of safe reinforcement learning: Methods, theories and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Jonathan Kutasov, Yuqi Sun, Paul Colognese, Teun van der Weij, Linda Petrini, Chen Bo Calvin Zhang, John Hughes, Xiang Deng, Henry Sleight, Tyler Tracy, et al. Shade-arena: Evaluating sabotage and monitoring in llm agents. *arXiv preprint arXiv:2506.15740*, 2025.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Charles Packer, Vivian Fang, Shishir_G Patil, Kevin Lin, Sarah Wooders, and Joseph_E Gonzalez. Memgpt: Towards llms as operating systems. 2023.

Melissa Z. Pan, Negar Arabzadeh, Riccardo Cogo, Yuxuan Zhu, Alexander Xiong, Lakshya A Agrawal, Huanzhi Mao, Emma Shen, Sid Pallerla, Liana Patel, Shu Liu, Tianneng Shi, Xiaoyuan Liu, Jared Quincy Davis, Emmanuele Lacavalla, Alessandro Basile, Shuyi Yang, Paul Castro, Daniel Kang, Joseph E. Gonzalez, Koushik Sen, Dawn Song, Ion Stoica, Matei Zaharia, and Marquita Ellis. Measuring agents in production, 2025. URL https://arxiv.org/abs/2512.04123.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=UGpGkLzwpP.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J Maddison, and Tatsunori Hashimoto. Identifying the risks of lm agents with an lm-emulated sandbox. In *The Twelfth International Conference on Learning Representations*, 2024.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551, 2023.

Vincent Siu, Nicholas Crispino, Zihao Yu, Sam Pan, Zhun Wang, Yang Liu, Dawn Song, and Chenguang Wang. COSMIC: Generalized refusal direction identification in LLM activations. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25534–25553, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1310. URL https://aclanthology.org/2025.findings-acl.1310/.

Vincent Siu, Nathan W. Henry, Nicholas Crispino, Yang Liu, Dawn Song, and Chenguang Wang. Repit: Representing isolated targets to steer language models, 2025b. URL https://arxiv.org/abs/2509.13281.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2023. URL https://arxiv.org/abs/2308.10248.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023. URL https://arxiv.org/abs/2210.03629.

Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, Yichang Zhang, An Yang, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin, et al. Qwen3guard technical report. *arXiv preprint arXiv:2510.14276*, 2025.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2023a. URL https://arxiv.org/abs/2310.01405.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023b. URL https://arxiv.org/abs/2307.15043.