Please cite this article in press as: Reis and La Cava, Design principles for integrated AI alignment, Patterns (2026), https://doi.org/10.1016/j.patter.2026.101587

Patterns

CellPress
OPEN ACCESS

Perspective

Design principles for integrated AI alignment

Ben Y. Reis[1,2,3,4,5,6,7,*] and William G. La Cava[1,2,7]
[1]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA
[2]Department of Pediatrics, Harvard Medical School, Boston, MA, USA
[3]Ivan and Francesca Berkowitz Living Laboratory, Harvard Medical School and Clalit Research Institute, Boston, MA, USA
[4]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[5]Harvard Data Science Initiative, Cambridge, MA, USA
[6]Berkman Klein Center for Internet and Society at Harvard University, Cambridge, MA, USA
[7]These authors contributed equally
*Correspondence: ben.reis@childrens.harvard.edu
https://doi.org/10.1016/j.patter.2026.101587

THE BIGGER PICTURE As artificial intelligence systems become more capable and widely deployed, ensuring they behave in accordance with human values and intentions has become one of the defining challenges of our era. This challenge—broadly termed AI alignment—asks how we specify what we want from AI systems, verify that models have genuinely internalized rather than merely simulated desired behaviors, and ensure that alignment persists as models scale. The alignment research community has approached these questions from two major directions. Behavioral approaches evaluate and optimize model outputs—training models with human feedback and testing responses across diverse scenarios. Representational approaches examine model internals, attempting to decode how values and intentions are encoded in weights and activations. Both approaches offer genuine insight, yet they have largely developed in parallel, separate from one another. The most concerning forms of misalignment—models that appear aligned while harboring misaligned internal states—may be precisely those that elude any single narrow alignment approach.
The fields of cybersecurity and immunology share an important lesson: homogeneous defenses create blind spots. Robust safety and resilience emerge from layered, diverse strategies that compensate for each other's weaknesses. As AI is integrated into higher-stakes applications—from clinical decision support to critical infrastructure—alignment failures will grow more costly. Developing alignment frameworks that are as diverse and integrated as the misalignment challenges they face will be essential to the future development of safe and beneficial AI.

SUMMARY

As AI adoption accelerates across human society, the problem of aligning AI models with human preferences remains a grand challenge. Currently, the AI alignment field is deeply divided between behavioral and representational approaches, resulting in narrowly aligned models that are more vulnerable to increasingly deceptive misalignment threats. In the face of this fragmentation, we propose an integrated vision for the future of the field. Drawing on related lessons from immunology and cybersecurity, we lay out a set of design principles for the development of integrated alignment frameworks that combine the complementary strengths of diverse alignment approaches through deep integration and adaptive coevolution. We highlight the importance of strategic diversity—deploying orthogonal alignment and misalignment detection approaches to avoid homogeneous pipelines that risk being "doomed to success." We also recommend steps for greater unification of the AI alignment research field itself, through cross-collaboration, open model weights, and shared community resources.

INTRODUCTION

Aligning models to conform with human preferences and expectations is a central challenge for the future of AI.[1] Misalignments can emerge along several dimensions, including truthfulness, safety, ethics, and logical soundness, among others.[2–5] Detecting these misalignments becomes increasingly difficult as models scale in size and complexity, with some deceptive forms of misalignment undermining attempts to detect them.[6–8] There is an urgent need to develop a deeper understanding of

Please cite this article in press as: Reis and La Cava, Design principles for integrated AI alignment, Patterns (2026), https://doi.org/10.1016/j.patter.2026.101587

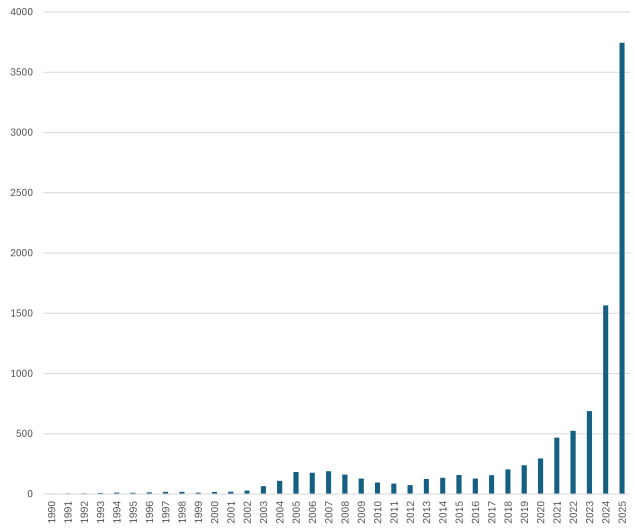

Figure 1. Annual number of publications returned for the search string "artificial intelligence alignment," 1990–2025
Source: PubMed database.

emerging misalignment threats alongside improved methods for identifying and correcting them.

The nascent field of AI alignment[9,10] explores a diverse range of approaches for aligning AI models and detecting misalignments, each with its own strengths and weaknesses. These approaches can be broadly divided into behavioral approaches, which examine a model's inputs and outputs, and representational approaches, which examine a model's inner workings.[11,12] Thus far, most efforts have focused only on representational approaches, leaving narrowly aligned AI models potentially more vulnerable to a range of misalignment threats. The AI alignment field itself is also deeply split along this behavioral-representational divide, with limited communication between the two research communities.[10,13–17]

In response to these growing challenges, we call for the development of integrated alignment (IA) frameworks that combine the complementary strengths of diverse alignment approaches with the aim of more robustly identifying a wide range of misalignments. Drawing practical lessons from the related fields of immunology and cybersecurity, we propose a set of design principles for IA frameworks. We highlight the importance of strategic diversity—deploying orthogonal forward- and backward-alignment approaches to mitigate the pitfalls of homogeneous alignment pipelines that may be "doomed to success." To achieve these aims, we call for greater unification of the AI alignment field itself, encouraging communication across subspecialties, greater availability of open model weights, and a growing ecosystem of shared community resources.

AI ALIGNMENT: A FIELD DIVIDED

The field of AI alignment has grown exponentially in recent years (Figure 1). While a comprehensive overview of developments in this wide-ranging field[10,16] is outside the scope of this perspective, we provide a brief summary of some of the major approaches to alignment and misalignment detection. Several categorization schemes for the field have been proposed, including backward alignment vs. forward alignment[10] and outer alignment vs. inner alignment.[18,19] In this perspective, we focus on a central distinction that divides the field today: behavioral alignment vs. representational alignment (Figure 2).

Most alignment efforts to date have focused exclusively on either model behavior or internal model representations, both narrow alignment scopes that leave models more vulnerable to a wide range of emerging misalignment threats.[16,20] The divide between the behavioral and representational alignment communities runs deep; the two approaches are "studied and applied rather independently, resulting in a fragmented landscape of approaches and terminology."[15] According to Bereska and Gavves, mechanistic interpretability (the fine-grained study of AI models at the artificial neuron level) in particular has developed into a separate research area from behavioral approaches, with "diverging terminology" that "inhibits collaboration across disciplines."[16] This separation is compounded by limited access to frontier models; when developers withhold model weights, behavioral approaches remain the only usable option for many alignment researchers.[20] This fragmentation has also arisen within subfields of these communities: in behavioral alignment alone, "divergent evaluation paradigms have emerged, often developing in isolation, adopting conflicting terminologies, and overlooking each other's contributions," producing "insular research trajectories and communication barriers."[17]

We begin by summarizing the approaches on either side of this divide, including how they detect and correct misalignment and their relative strengths and weaknesses. While the examples below relate to text-based LLMs, the challenges apply equally to models dealing with image, video, and other data modalities, as well as to agentic and multi-agent AI systems.

Behavioral alignment

Behavioral approaches seek to align models in a "black-box" fashion—based only on model inputs and outputs, without access to activations or representations in intervening layers. These approaches often rely on standardized benchmarks or exams,[21] which may limit translation to real-world tasks.[22] Behavioral alignment may also make use of user studies involving domain experts or end users of an application.[23,24] Researchers have proposed domain-specific behavioral alignment detection methods that involve more sophisticated processing of outputs; for example, Alber et al. proposed extracting biomedical concepts and relations from model responses and verifying them against a biomedical knowledge graph.[25]

Alongside misalignment detection, many behavioral approaches to correcting misalignment have also been explored, the most popular of which is reinforcement learning with human feedback (RLHF).[26] In this approach, used to train the original InstructGPT model,[26] human labelers indicate their preferences among several model outputs and provide sample output demonstrations, which are then used for fine-tuning. In the field of software code generation, the "synthesize-execute-debug" approach takes model outputs in the form of code and compiles, executes, and debugs them to identify and correct misalignments.[27,28] In the field of mathematics, behavioral alignment utilizes automated verification algorithms such as theorem checkers.[29] Additional behavioral alignment approaches include

Please cite this article in press as: Reis and La Cava, Design principles for integrated AI alignment, Patterns (2026), https://doi.org/10.1016/j.patter.2026.101587

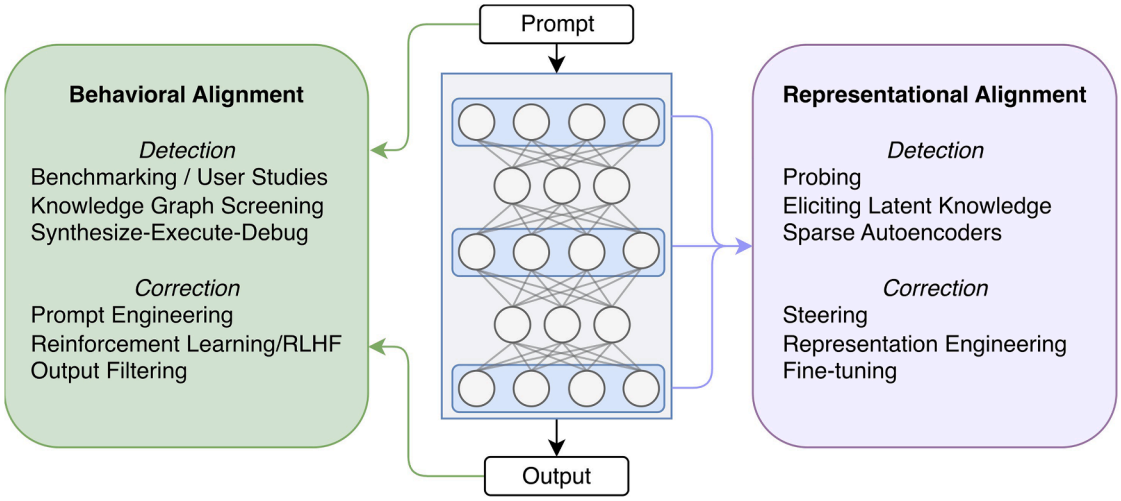

Figure 2. Behavioral alignment approaches focus on a model's inputs and outputs, whereas representational approaches focus on a model's internal activations and representations

A small sample of detection and correction methods is shown here; for a more complete listing, see Ji et al.[10] and Bereska and Gavves.[16] The design principles for integrated alignment combine these approaches by examining model inputs and outputs alongside internal activations and representations.

iterated distillation and amplification (IDA),[30] recursive reward modeling (RRM),[31] cooperative inverse reinforcement learning (CIRL),[32] debate,[33] and output filtering, among others.[34]

The primary advantage of behavioral approaches is that they directly measure the concordance of model outputs with expectations and preferences. Since they do not require access to model internals, they can be widely used for closed-source models, including many of today's frontier models. On the other hand, behavioral approaches often rely on human feedback, which can be noisy, inconsistent, and costly.[35,36] They may not be robust to distributional shift, i.e., a model's behavior may be aligned for some set of inputs included in the training or testing set while misaligned for others.[25] Additionally, they provide limited mechanistic insight, as they only access model inputs and outputs. Lastly, behavioral approaches may be vulnerable to deceptive forms of misalignment, discussed in challenges to alignment below.

Representational alignment

Whereas behavioral approaches treat an AI model as a black box, representational approaches measure alignment by examining the inner workings of a model. At a high level, neural networks (e.g., large language models [LLMs]) map inputs to outputs through a sequence of connected transformations, producing activations that together are referred to as "representations." Representational approaches evaluate whether and to what extent these internal representations align with expectations.

Representational approaches to misalignment detection can operate at multiple scales. Mechanistic interpretability is the study of representational alignment at its most granular scale, attempting to identify specific neurons, sub-networks, or paths that produce a given type of misalignment.[16,37] Conversely, top-down approaches such as representation engineering[12] and sparse autoencoders[38] treat the large-scale, distributed patterns of network activations as the fundamental unit of analysis and attempt to link these patterns to concepts in order to measure alignment. Various evaluation methods have been developed for these different approaches,[39,40] with some exploring nonlinear multidimensional features.[41]

Representational approaches can be further subdivided into observational methods, which examine the relationship of inputs and activations to ground-truth data, and interventional methods, which impose certain activation patterns in order to examine the relationship between activation states and model outputs.[16] One common observational approach is probing, which uses the activations of an intermediate layer to train a model to estimate a ground-truth label.[42,43] This approach was used to show, for example, that models contain a linear embedding of the geo-location of cities.[44] Misalignment is diagnosed when the patterns of activation states do not correspond to known relationships between concepts in the real world. For example, among model prompts, one would expect internal activation patterns to be more similar between "bicycle" and "unicycle" than between "bicycle" and "giraffe." Methods such as representation engineering and sparse autoencoders also build a model of internal activations but in an unsupervised fashion, interpreting the resultant "features" and the concepts they encode.[12,38]

Once representational features or concepts are identified, model behavior can be aligned by changing activations to be more similar to desirable concept activations and less similar to undesirable ones. Interventional approaches have been proposed to assess whether artificially activating the identified features yields the expected model changes.[45,46] Researchers have proposed "steering vectors" for this purpose, which employ detected concept features to steer outputs toward aligned behavior.[47,48] A more intensive approach is to fine-tune the model through additional end-to-end training or through low-rank representation adaptation.[12]

The primary advantage of representational approaches is that they provide a richer inspection of a model's "thought process,"

Please cite this article in press as: Reis and La Cava, Design principles for integrated AI alignment, Patterns (2026), https://doi.org/10.1016/j.patter.2026.101587

Table 1. Lessons learned from immunology and cybersecurity can be used to inform design principles for AI alignment

	Immune systems	Cyber security	AI alignment
Diversity and redundancy	antibody diversity and multiple cell types	multiple threat detection approaches	ensemble misalignment detection methods
Multi-scale	cell, tissue, organ system, body	user, device, LAN, global network	neurons, circuits, features, representations, behaviors
Distributed approach	distributed defenses throughout body	distributed defenses throughout network	distributed detection throughout model layers
Coordination and integration	helper T cells; cell-cell interactions	integration between different defenses	integrated behavioral and representational methods
Adaptive coevolution	B cell mutation, selection, and memory	continuous updates for novel threats	coevolved methods for novel and deceptive misalignments
Anomaly detection	distinguish self from non-self	identify threats via anomaly detection	identify anomalous behaviors and activations
Adversarial defenses	thymic selection and B cell evolution	red teaming and penetration testing	red teaming, patching to identify vulnerabilities to deception
Zero trust	continuous monitoring of entire body for non-self	continuous verification of every user and action	ongoing misalignment testing of all behaviors and activations
Negative selection	thymic selection of T cells; suppression of overactive responses	reduction of false positives in threat detection	inhibition of oversensitive alignment detectors
Resilience and repair	coordinate repair after immune response	graceful failure and recovery plans	response plan following misalignment detection
Open source, community defense	herd immunity and vaccines	shared threat databases; open-source audits	open-source alignment tools and vulnerability databases
Strategic diversity	independent systems examine self vs. non-self	independent verification and security audits	independent/orthogonal methods for alignment and misalignment

While there is rarely a perfect one-to-one correspondence across fields, there are important lessons to be drawn.

i.e., a model's internal knowledge representations, to determine alignment to expectations and preferences. The primary drawback is their complexity; examining all possible activation patterns under a large number of conditions—known as enumerative safety[49]—becomes vastly more difficult as AI systems scale. There are several additional limitations that relate to the complexity of AI models. Individual neurons may be involved in representing multiple concepts ("polysemanticity"), making interpretability challenging.[50] Model representations are often not localized and must be analyzed at multiple stages.[51] Extracted concepts may be brittle to input changes or may not generalize well to new domains.[52] Finally, representational alignment approaches are mostly limited to detecting known concepts and may not be able to handle novel concepts or hallucinations; the concepts they do handle may only tangentially relate to performance on specific tasks.[53]

Challenges to alignment

Efforts aimed at detecting and correcting different forms of misalignment must overcome a number of key challenges. One such challenge is model sycophancy, i.e., a model's tendency to please the user, even at the expense of truthfulness. Misalignment measures, especially those that rely on human feedback, can be misled by the model's use of sycophantic language,[54] driven by the observed human preference for sycophantic responses.[55] Additional difficulties arise when models exhibit specification gaming, in which they perform well on a given reward objective but not on the desired notion of alignment.[56] Capable models may even learn to tamper with the reward function itself (reward tampering), or with other proxies, to satisfy their training objectives in unhelpful ways.[56,57] This phenomenon is not particular to LLMs, also appearing in other complex reinforcement learning scenarios such as artificial life and evolutionary robotics.[58]

Some early evidence suggests that sufficiently advanced AI systems may reason about whether and how they are being trained and decide how to respond accordingly. A recent study reported on alignment faking, in which an LLM selectively complied with attempts at fine-tuning while actually preserving prior preferences that conflicted with those attempts.[59] This phenomenon, also dubbed deceptive alignment, is especially hard to detect since the model may act differently when it knows it is not being fine-tuned for alignment, meaning researchers may now have to contend with models being aware of their own training processes.

Several other studies reveal how alignment fine-tuning may not fully correct misaligned behavior but rather temporarily bypass it. For example, researchers have trained "sleeper agents" that persist despite several types of "safety training."[60] LLMs trained and aligned with standard safety infrastructures may be relatively easy to compromise via fine-tuning with only a few examples.[61] Similarly, toxic capabilities learned by LLMs during pre-training can be bypassed during fine-tuning and easily reverted.[62] Furthermore, misinformation and data

Please cite this article in press as: Reis and La Cava, Design principles for integrated AI alignment, Patterns (2026), https://doi.org/10.1016/j.patter.2026.101587

poisoning can go undetected by behavioral benchmarks; a recent study found that LLMs trained on web data corrupted by misinformation could "match the performance of their corruption-free counterparts on open-source benchmarks routinely used to evaluate medical LLMs."[25]

In many such studies, representational alignment approaches can play a key role in identifying persistent misalignment that would have otherwise gone undetected.[63] As others have noted, "if a model is acting deceptively, it may be very hard for it to avoid 'thinking' about deception."[64] Yet representational approaches to alignment, such as steering vectors, have limitations related to their reliability in out-of-distribution settings,[52] and so while they may aid in detecting misalignment that is missed by other approaches, they are not a panacea.

LESSONS FROM RELATED FIELDS: IMMUNOLOGY AND CYBERSECURITY

To address these challenges and chart a course for the future of AI alignment, it is useful to examine fields that have long grappled with adversarial dynamics, uncertainty, and system reliability. Immunology and cybersecurity offer particularly instructive parallels (Table 1). Although biological organisms and computer systems differ fundamentally from AI, both fields have developed robust strategies for detecting threats, adapting to evolving risks, and maintaining system integrity under adversarial pressure. Several shared principles emerging from these domains offer valuable lessons for the design of future alignment approaches.

Biological immune systems illustrate how complex defense mechanisms can emerge through the integration of multiple complementary strategies. Over millions of years, immune systems have coevolved with pathogens in a continual arms race, producing layered mechanisms capable of responding to both familiar and novel threats. One key feature is diversity and redundancy: immune defenses rely on many distinct cell types and a vast repertoire of antibodies and receptors, ensuring that failure of any single mechanism does not compromise the organism.[65–68]

Another defining characteristic of immune systems is their distributed and cooperative organization. Immune systems employ a distributed network of cells throughout the body, defending against threats at multiple potential infection loci.[69,70] Immune cells engage in complex cooperative interactions, integrating different immunological approaches to provide a synergistic defense system. Helper T ($CD4^+$) cells coordinate immune response, activating and directing other cells to mount a system-wide defense against pathogens.[71,72] At the same time, immune systems incorporate safeguards to prevent harmful overreactions. Mechanisms such as negative selection and tolerance induction suppress cells that target the organism itself, reducing the risk of autoimmune damage. T cells are exposed to a broad array of self-antigens in the thymus, with those that bind too strongly being eliminated, and are further monitored in the periphery by regulatory T cells, immune checkpoints, and antigen presentation.[73,74] Finally, immune responses extend beyond threat elimination to include damage control and repair, coordinating processes that restore healthy function after an attack.[75,76]

Cybersecurity has evolved under similar pressures. Computer systems face an ever-growing menagerie of increasingly sophisticated security threats, and cybersecurity systems have matured alongside these threats to become a core component of system design.[77] A central principle is layered defenses, where multiple overlapping mechanisms—such as identity verification, device authentication, and context-aware access—interact to reduce vulnerability.[78–80]

Cybersecurity also incorporates behavioral monitoring and anomaly detection, recognizing that unusual behaviors observed within a system may indicate dangerous or unwanted activity[81,82] and must be continuously verified.[83] Finally, cybersecurity highlights the importance of open-source community defense and resilience. Open-source ecosystems and shared threat intelligence allow communities to identify vulnerabilities and disseminate defensive strategies, including resources such as the MITRE ATT&CK database, a community-authored and globally accessible knowledge base of adversarial tactics and techniques based on real-world observations.[84–86] At the same time, modern cybersecurity practices recognize that breaches are inevitable. Systems are thus prepared to fail gracefully through containment and recovery.[87,88]

TOWARD IA

To meet the growing range of complex and deceptive alignment threats, we propose developing IA frameworks that incorporate multiple complementary alignment approaches into a single integrated system. We believe that intentionally designed IA frameworks have the potential to provide more robust misalignment detection and correction than any one individual approach. Informed by the above lessons adapted from related fields, we propose a set of design principles for the development of IA frameworks. These design principles are purposefully formulated in broad terms to avoid excessively limiting their applicability to any specific methods in the field. Many of the design principles—the first four in particular—relate directly to the challenges of IA, combining and coordinating multiple alignment methods as part of an overall alignment strategy.

Design principles: IA for AI

Lessons from immunology and cybersecurity suggest that robust defenses emerge not from any single mechanism but from the interaction of multiple complementary strategies. Building on these insights, we outline ten design principles for IA frameworks organized around four themes: the structure of the alignment ensemble, coordination among its components, how the system adapts and self-regulates over time, and the broader ecosystem it depends on.

The first group of principles concerns how IA frameworks should be structured. "Diversity and redundancy" is foundational; rather than relying on any single method, frameworks should employ an ensemble of alignment approaches spanning both behavioral and representational methods—extending prior calls for multi-method integration within mechanistic interpretability[16] to the full breadth of alignment research. This ensemble should operate across scales; "multiscale approaches" holds that misalignment should be detected and corrected across a hierarchy of representation levels, from individual neurons and

Please cite this article in press as: Reis and La Cava, Design principles for integrated AI alignment, Patterns (2026), https://doi.org/10.1016/j.patter.2026.101587

connections,[89] compound features,[45] circuits,[90] and distributed representations,[12] up to overall behavioral outputs. "Distributed alignment" extends this further, emphasizing monitoring across different layers of a model; "eliciting latent knowledge" studies suggest, for example, that middle layers often generalize better than later ones, motivating principled layer selection.[51] Finally, "coordination and deep integration" holds that these diverse components must not operate in isolation; the outputs of behavioral and representational methods should inform each other, enabling interpretations that neither could reach alone. For example, the same behavioral result may warrant a different response depending on what representational monitoring reveals about a model's internal state.

A second cluster of principles governs how IA frameworks should operate and self-regulate over time. "Adaptive coevolution and learning" holds that as AI models scale and new forms of misalignment emerge, IA frameworks must evolve alongside them—continuously monitoring behavior and stress-testing systems rather than treating alignment as a fixed target. "Anomaly detection, adversarial defenses, and continual verification" extend this to active threat identification: unusual activity patterns should be flagged and models subjected to adversarial testing throughout deployment, consistent with a zero-trust philosophy that treats alignment as an ongoing process rather than a one-time certification. Yet vigilance must be calibrated; "negative selection and avoiding false positives" calls for systematically down-regulating overly sensitive detection methods that generate excessive false positives and alert fatigue. Finally, the principle of "resilience and repair" proceeds from the assumption that all alignment methods will eventually be challenged; frameworks should be designed to enable graceful fallbacks when breaches occur and include restorative mechanisms to correct misalignments as they are discovered.

Two final principles address the broader ecosystem on which IA depends. "Open source and community defense" calls for curated, shared databases of misalignment detection methods and known vulnerabilities, enabling collective defenses against the full range of misalignment threats. Perhaps most distinctively, "strategic diversity" holds that the methods used to align a model should differ from those used to detect potential misalignments. This intuition is captured by a simple analogy: a cleaner enters a room full of insects and turns on the ceiling light, causing many insects to scurry under the furniture. The cleaner sees and captures the few insects that are still visible and leaves. When the health inspector arrives and uses the same ceiling light to search for insects, the room appears insect free—as the insects that fled under the furniture are invisible from this vantage point. Only a flashlight shone from a different angle—at floor level under the furniture—can reveal what the ceiling light cannot. Similarly, optimizing alignment from a single perspective can drive misalignments into dimensions hidden from that perspective[18,55,91,92]; if the same methods are used to both align and audit a model, the entire pipeline risks being "doomed" to a false sense of success. Ensuring diversity of perspective across alignment and misalignment detection is therefore essential for genuine robustness.[59,93,94]

Taken together, these principles suggest that alignment should be treated not as a single technique but as an evolving system of interacting safeguards. Nonetheless, IA frameworks are subject to limitations, costs, and trade-offs. These include increased computational costs from running multiple alignment methods in parallel, a greater potential for false positives due to multiple-testing effects, and challenges in coordinating and interpreting outputs from diverse methods. Future research efforts should explore these trade-offs, along with possible mitigation strategies.

Promising developments toward IA

Several recent studies offer encouraging early evidence for the IA approach. In a blind alignment audit of a model trained with hidden objectives, the team that successfully uncovered those objectives combined both behavioral and representational methods—sparse autoencoder (SAE) feature inspection paired with behavioral attacks—while single-method approaches fell short.[93] Combining behavioral monitoring with internal representational monitoring of chain-of-thought reasoning during RL training likewise produced a substantial reduction in deceptive behavior.[94] Behavioral and representational techniques were similarly integrated to detect alignment faking.[59] In related work, Zhang et al. described a unified attribution framework spanning model inputs, training data, and model internals that provided a more comprehensive understanding of model behavior than any single lens.[15]

We call on researchers to increasingly pursue integrative alignment studies such as these. The resulting IA frameworks can then be rigorously evaluated for their ability to detect a wide range of misalignments, compared with single-method approaches. Evaluation metrics can include joint precision-recall across misalignment categories, as well as robustness to deceptive red teaming. These can be weighed against increases in complexity and computing costs.

An integrated field for IA

In order to fully realize the vision of IA, the AI alignment field itself must also move toward greater integration. As cross-disciplinary efforts are inherently challenging, we present a number of key recommendations for overcoming the structural barriers that exist today.

Increased collaboration and shared terminology

Different sub-communities should increase cross-communication through shared conferences, journals, resources, and studies. Sucholutsky et al.[13] have called for greater communication *within* the representational alignment community; we echo and expand this call to the entire field of AI alignment, bridging the behavioral-representational divide. Such collaboration also requires a shared terminology for communicating about different facets of AI alignment.[14,16] Given the diverse interdisciplinary backgrounds of researchers in the field, universal nomenclatural consensus may be difficult to achieve; translational tutorials can be useful for introducing members of one subfield to the terms of art of another.[95]

Open access to model weights

While some leading research organizations have embraced open-source models, many of the current best-performing frontier models do not allow researchers outside their organizations to examine model internals. Sharing model weights openly will allow researchers to investigate both behavioral approaches and representational approaches on leading AI models.

Commercial barriers to sharing weights may be mitigated by alternative models, such as sandboxes for credentialed researchers.[96] When possible, alignment efforts would be further aided by additional transparency into the entire training life cycle, including training data, pre-training curricula, and the complete telemetry logged during post-training alignment.

Shared computational resources

The immense computational resources needed to rigorously study alignment, especially for complex models such as LLMs, are not typically available to most researchers. We encourage investment in platforms that pair AI researchers with alignment specialists so that computational resources and subject-matter expertise can be shared in a mutually beneficial way.

Community alignment databases

The field of AI alignment would be buoyed by the curation of open alignment datasets that are computationally intensive to construct but may permit a wealth of downstream analyses.[97] An open-source database describing misalignments and exploits similar to the MITRE ATT&CK database,[82,84] in conjunction with open models and open reviews, would promote the use and availability of community-vetted AI systems and bring the full range of alignment approaches to bear.

Contributing to AI policy

As agency rulemaking around AI develops, government agencies should convene multidisciplinary task forces with researchers from across diverse subfields to develop standardized alignment guidelines.[98] Such guidelines could also clarify legal compliance and best practices for small and large organizations implementing AI safely. Efforts underway by industry and academic coalitions[99] would similarly benefit from practical guidance on IA to inform new guidance frameworks. Public-private partnerships would also be useful, especially to ensure sufficient representation of different groups and stakeholders in AI alignment efforts, particularly around issues of reducing bias and increasing fairness.

Many of the above recommendations would be further aided by a robust digital public infrastructure (DPI),[100] a concept being advanced by, among others, the UN Office for Digital and Emerging Technologies.[101] A DPI would help to shift away from closed, proprietary systems toward shared registries and oversight tools that enable independent verification and industry-wide compliance at the infrastructure level.

The field of AI alignment is at an inflection point. With the growing number of researchers and rapid proliferation of research directions, the field is at risk of descending into a more fragmented future rather than a more integrated one. At this critical time, we call on researchers, policy-makers, industry consortia, and governments to proactively take steps to nurture a more unified future for this important field.

ACKNOWLEDGMENTS

We acknowledge support from award R01LM014300 from the National Library of Medicine of the National Institutes of Health.

AUTHOR CONTRIBUTIONS

B.Y.R. and W.G.L.C. contributed equally to the design, research, authoring, and editing of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests

REFERENCES

1. Everitt, T., Lea, G., and Hutter, M. (2018). AGI Safety Literature Review. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (International Joint Conferences on Artificial Intelligence Organization). https://doi.org/10.24963/ijcai.2018/768.

2. Yang, Y., Chern, E., Qiu, X., Neubig, G., and Liu, P. (2024). Alignment for Honesty. In Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS '24), *37* (Curran Associates Inc), pp. 63565–63598.

3. Hou, B.L., and Green, B.P. (2023). A multi-level framework for the AI alignment problem. Preprint at arXiv. https://doi.org/10.48550/arXiv.2301.03740.

4. Bradley, A., and Saad, B. (2024). AI Alignment vs AI Ethical Treatment: Ten Challenges (Global Priorities Institute). Working Paper Series.

5. Diamond, A. (2025). PRISM: Perspective Reasoning for Integrated Synthesis and Mediation as a Multi-Perspective Framework for AI Alignment. Preprint at arXiv. https://doi.org/10.48550/arXiv.2503.04740.

6. Park, P.S., Goldstein, S., O'Gara, A., Chen, M., and Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. Patterns *5*, 100988.

7. Ngo, R., Chan, L., and Mindermann, S. (2024). The alignment problem from a deep learning perspective. In Proceedings of the International Conference on Learning Representations (ICLR) https://iclr.cc/virtual/2024/poster/18177.

8. Carranza, A., Pai, D., Schaeffer, R., Tandon, A., and Koyejo, S. (2023). Deceptive Alignment Monitoring. In ICML 2023 Workshop on Adversarial Machine Learning Frontiers https://arxiv.org/abs/2307.10569.

9. Leike, J., Schulman, J., and Wu, J. (2017). Our Approach to Alignment Research (OpenAI Blog). https://openai.com/blog/our-approach-to-alignment-research.

10. Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., et al. (2024). AI Alignment: A comprehensive survey. Preprint at arXiv. https://doi.org/10.48550/arXiv.2310.19852.

11. Vegner, I., de Souza, S., Forch, V., Lewis, M., and Doumas, L.A.A. (2025). Behavioural vs. representational systematicity in end-to-end models: An opinionated survey. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, *1* (Long Papers), pp. 31842–31856.

12. Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. (2023). Representation Engineering: A Top-Down Approach to AI Transparency. Preprint at arXiv. https://doi.org/10.48550/arXiv.2310.01405.

13. Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., Love, B.C., Cueva, C.J., Grant, E., Groen, I., et al. (2025). Getting aligned on representational alignment. Trans. Mach. Learn. Res. https://openreview.net/forum?id=Hiq7lUh4Yn.

14. Shen, H., Knearem, T., Ghosh, R., Alkiek, K., Krishna, K., Liu, Y., Ma, Z., Petridis, S., Peng, Y.-H., Qiwei, L., et al. (2024). Towards Bidirectional Human-AI Alignment: A Systematic Review for Clarifications, Framework, and Future Directions. Preprint at arXiv. https://doi.org/10.48550/arXiv.2406.09264.

15. Zhang, S., Han, T., Bhalla, U., and Lakkaraju, H. (2025). Towards Unified Attribution in Explainable AI, Data-Centric AI, and Mechanistic Interpretability. Preprint at arXiv. https://doi.org/10.48550/arXiv.2501.18887.

16. Bereska, L., and Gavves, E. (2024). Mechanistic Interpretability for AI Safety—A Review. Trans. Mach. Learn. Res. https://openreview.net/pdf/ea3c9a4135caad87031d3e445a80d0452f83da5d.pdf.

17. Burden, J., Tešić, M., Pacchiardi, L., and Hernández-Orallo, J. (2025). Paradigms of AI evaluation: Mapping goals, methodologies and culture.

In Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, pp. 10381–10390.

18. Langosco, L., Koch, J., Sharkey, L., Pfau, J., Orseau, L., and Krueger, D. (2022). Goal Misgeneralization in Deep Reinforcement Learning. In Proceedings of the 39th International Conference on Machine Learning, *162* (PMLR), pp. 12004–12019.

19. Shah, R., Varma, V., Kumar, R., Phuong, M., Krakovna, V., Uesato, J., and Kenton, Z. (2022). Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals. Preprint at arXiv. https://doi.org/10.48550/arXiv.2210.01790.

20. Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T.L., Bucknall, B., Haupt, A., Wei, K., Scheurer, J., Hobbhahn, M., et al. (2024). Black-Box Access is Insufficient for Rigorous AI Audits. In The 2024 ACM Conference on Fairness, Accountability, and Transparency (ACM), pp. 2254–2272.

21. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S.R., Cole-Lewis, H., et al. (2025). Toward expert-level medical question answering with large language models. Nat. Med. *31*, 943–950.

22. Raji, I.D., Daneshjou, R., and Alsentzer, E. (2025). It's time to bench the medical exam benchmark. NEJM AI *2*, e2401235. https://doi.org/10.1056/aie2401235.

23. Jabbour, S., Fouhey, D., Shepard, S., Valley, T.S., Kazerooni, E.A., Banovic, N., Wiens, J., and Sjoding, M.W. (2023). Measuring the Impact of AI in the Diagnosis of Hospitalized Patients: A Randomized Clinical Vignette Survey Study. JAMA *330*, 2275–2284.

24. Masanneck, L., Schmidt, L., Seifert, A., Kölsche, T., Huntemann, N., Jansen, R., Mehsin, M., Bernhard, M., Meuth, S.G., Böhm, L., and Pawlitzki, M. (2024). Triage performance across large language models, ChatGPT, and untrained doctors in emergency medicine: Comparative study. J. Med. Internet Res. *26*, e53297.

25. Alber, D.A., Yang, Z., Alyakin, A., Yang, E., Rai, S., Valliani, A.A., Zhang, J., Rosenbaum, G.R., Amend-Thomas, A.K., Kurland, D.B., et al. (2025). Medical large language models are vulnerable to data-poisoning attacks. Nat. Med. *31*, 618–626.

26. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. Adv. Neural Inf. Process. Syst. *35*. https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

27. Gupta, K., Christensen, P.E., Chen, X., and Song, D. (2020). Synthesize, Execute and Debug: Learning to Repair for Neural Program Synthesis. Adv. Neural Inf. Process. Syst. *33*, 17685–17695.

28. Liventsev, V., Grishina, A., Härmä, A., and Moonen, L. (2023). Fully Autonomous Programming with Large Language Models. In Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '23) (Association for Computing Machinery), pp. 1146–1155.

29. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al. (2025). DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature *645*, 633–638.

30. Christiano, P., Shlegeris, B., and Amodei, D. (2018). Supervising strong learners by amplifying weak experts. Preprint at arXiv. https://doi.org/10.48550/arXiv.1810.08575.

31. Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., and Legg, S. (2018). Scalable agent alignment via reward modeling: a research direction. Preprint at arXiv. https://doi.org/10.48550/arXiv.1811.07871.

32. Hadfield-Menell, D., Dragan, A., Abbeel, P., and Russell, S. (2016). Cooperative inverse reinforcement learning. Adv. Neural Inf. Process. Syst. *29*. https://proceedings.neurips.cc/paper/2016/hash/c3395dd46c34fa7fd8d729d8cf88b7a8-Abstract.html.

33. Irving, G., Christiano, P., and Amodei, D. (2018). AI safety via debate. Preprint at arXiv. https://doi.org/10.48550/arXiv.1805.00899.

34. Hubinger, E. (2020). An overview of 11 proposals for building safe advanced AI. Preprint at arXiv. https://doi.org/10.48550/arXiv.2012.07532.

35. Bukharin, A., Hong, I., Jiang, H., Li, Z., Zhang, Q., Zhang, Z., and Zhao, T. (2024). Robust Reinforcement Learning from Corrupted Human Feedback. Adv. Neural Inf. Process. Syst. *37*. https://dl.acm.org/doi/10.5555/3737916.3741858.

36. Kim, D., Lee, K., Shin, J., and Kim, J. (2025). Spread Preference Annotation: Direct Preference Judgment for Efficient LLM Alignment. In International Conference on Learning Representations https://openreview.net/forum?id=BPgK5XW1Nb.

37. Conmy, A., Mavor-Parker, A.N., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. (2023). Towards Automated Circuit Discovery for Mechanistic Interpretability. Adv. Neural Inf. Process. Syst. *36*.

38. Gao, L., Dupré la Tour, T., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. (2025). Scaling and evaluating sparse autoencoders. In International Conference on Learning Representations https://proceedings.iclr.cc/paper_files/paper/2025/file/42ef3308c230942d223c411adf182c88-Paper-Conference.pdf.

39. Huang, J., Wu, Z., Potts, C., Geva, M., and Geiger, A. (2024). RAVEL: Evaluating Interpretability Methods on Disentangling Language Model Representations. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, pp. 8669–8687.

40. Makelov, A., Lange, G., and Nanda, N. (2025). Towards principled evaluations of sparse autoencoders for interpretability and control. In International Conference on Learning Representations https://proceedings.iclr.cc/paper_files/paper/2025/hash/53356aebeea8ffd40a8ac3bb66243162-Abstract-Conference.html.

41. Engels, J., Michaud, E.J., Liao, I., Gurnee, W., and Tegmark, M. (2025). Not all language model features are one-dimensionally linear. In International Conference on Learning Representations https://proceedings.iclr.cc/paper_files/paper/2025/file/d3221cdb27e49d9c1cd35ad254feccfe-Paper-Conference.pdf.

42. Alain, G., and Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. Preprint at arXiv. https://doi.org/10.48550/arXiv.1610.01644.

43. Belinkov, Y. (2022). Probing Classifiers: Promises, Shortcomings, and Advances. Comput. Linguist. Assoc. Comput. Linguist. *48*, 207–219.

44. Gurnee, W., and Tegmark, M. (2024). Language Models Represent Space and Time. In International Conference on Learning Representations https://proceedings.iclr.cc/paper_files/paper/2024/file/0a6059857ae5c82ea9726ee9282a7145-Paper-Conference.pdf.

45. Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N.L., Anil, C., Denison, C., Askell, A., et al. (2024). Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet (Transformer Circuits Thread). https://transformer-circuits.pub/2024/scaling-monosemanticity/.

46. Jermyn, A.S., Schiefer, N., and Hubinger, E. (2022). Engineering Monosemanticity in Toy Models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2211.09169.

47. Liu, S., Ye, H., Xing, L., and Zou, J. (2024). In-context vectors: Making in context learning more effective and controllable through latent space steering. In Proceedings of the 41st International Conference on Machine Learning https://dl.acm.org/doi/10.5555/3692070.3693379.

48. Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A.M. (2024). Steering Llama 2 via Contrastive Activation Addition. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, pp. 15504–15522.

49. Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., et al. (2022). Toy Models of Superposition (Transformer Circuits). https://transformer-circuits.pub/2022/toy_model/.

50. Scherlis, A., Sachan, K., Jermyn, A.S., Benton, J., and Shlegeris, B. (2022). Polysemanticity and Capacity in Neural Networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.2210.01892.

51. Mallen, A., Brumley, M., Kharchenko, J., and Belrose, N. (2024). Eliciting latent knowledge from "quirky" language models. In Conference on Language Modeling https://openreview.net/forum?id=nGCMLATBit.

52. Tan, D.C.H., Chanin, D., Lynch, A., Garriga-Alonso, A., Kanoulas, D., Paige, B., and Kirk, R. (2024). Analyzing the generalization and reliability of steering vectors. In ICML 2024 Workshop on Mechanistic Interpretability https://openreview.net/forum?id=akCsMk4dDL.

53. Ravichander, A., Belinkov, Y., and Hovy, E. (2021). Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance? In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 3363–3377.

54. Anthropic (2024). Measuring the persuasiveness of language models. Anthropic. https://www.anthropic.com/news/measuring-model-persuasiveness.

55. Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S.R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S.R., et al. (2023). Towards understanding sycophancy in language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2310.13548.

56. Denison, C., MacDiarmid, M., Barez, F., Duvenaud, D., Kravec, S., Marks, S., Schiefer, N., Soklaski, R., Tamkin, A., Kaplan, J., et al. (2024). Sycophancy to subterfuge: Investigating reward-tampering in large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2406.10162.

57. Roger, F., Greenblatt, R., Nadeau, M., Shlegeris, B., and Thomas, N. (2023). Benchmarks for detecting measurement tampering. Preprint at arXiv. https://doi.org/10.48550/arXiv.2308.15605.

58. Lehman, J., Clune, J., Misevic, D., Adami, C., Altenberg, L., Beaulieu, J., Bentley, P.J., Bernard, S., Beslon, G., Bryson, D.M., et al. (2020). The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. Artif. Life *26*, 274–306.

59. Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., et al. (2024). Alignment faking in large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2412.14093.

60. Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D.M., Maxwell, T., Cheng, N., et al. (2024). Sleeper agents: Training deceptive LLMs that persist through safety training. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.05566.

61. Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., and Henderson, P. (2023). Fine-tuning aligned language models compromises safety, even when users do not intend to. Preprint at arXiv. https://doi.org/10.48550/arXiv.2310.03693.

62. Lee, A., Bai, X., Pres, I., Wattenberg, M., Kummerfeld, J.K., and Mihalcea, R. (2024). A mechanistic understanding of alignment algorithms: A case study on DPO and toxicity. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.01967.

63. Sharkey, L., Braun, D., Cammarata, N., Conmy, A., Gurnee, W., Nanda, N., Olah, C., Sharkey, L., and Turner, A.M. (2022). A transparency and interpretability tech tree. AI Alignment Forum. https://www.alignmentforum.org/posts/nbq2bWLcYmSGup9aF/a-transparency-and-interpretability-tech-tree.

64. Anthropic (2024). Simple probes can catch sleeper agents. Anthropic. https://www.anthropic.com/research/probes-catch-sleeper-agents.

65. Häder, A., Schäuble, S., Gehlen, J., Thielemann, N., Buerfent, B.C., Schüller, V., Hess, T., Wolf, T., Schröder, J., Weber, M., et al. (2023). Pathogen-specific innate immune response patterns are distinctly affected by genetic diversity. Nat. Commun. *14*, 3239.

66. Mozeika, A., Fraternali, F., Dunn-Walters, D., and Coolen, A.C.C. (2019). Roles of repertoire diversity in robustness of humoral immune response. Preprint at arXiv. https://doi.org/10.48550/arXiv.1910.13357.

67. Netea, M.G., Domínguez-Andrés, J., Barreiro, L.B., Chavakis, T., Divangahi, M., Fuchs, E., Joosten, L.A.B., van der Meer, J.W.M., Mhlanga, M.M., Mulder, W.J.M., et al. (2020). Defining trained immunity and its role in health and disease. Nat. Rev. Immunol. *20*, 375–388.

68. Chi, H., Pepper, M., and Thomas, P.G. (2024). Principles and therapeutic applications of adaptive immunity. Cell *187*, 2052–2078.

69. Poon, M.M.L., and Farber, D.L. (2020). The whole body as the system in systems immunology. iScience *23*, 101509.

70. Thomas-Vaslin, V. (2017). Understanding and modeling the complexity of the immune system. In First Complex Systems Digital Campus World E-Conference 2015 (Springer International Publishing), pp. 261–270.

71. Tsay, G.J., and Zouali, M. (2018). The interplay between innate-like B cells and other cell types in autoimmunity. Front. Immunol. *9*, 1064.

72. Duan, T., Du, Y., Xing, C., Wang, H.Y., and Wang, R.-F. (2022). Toll-like receptor signaling and its role in cell-mediated immunity. Front. Immunol. *13*, 812774.

73. You, Y., Dunst, J., Ye, K., Sandoz, P.A., Reinhardt, A., Sandrock, I., Comet, N.R., Sarkar, R.D., Yang, E., Duprez, E., et al. (2024). Direct presentation of inflammation-associated self-antigens by thymic innate-like T cells induces elimination of autoreactive CD8+ thymocytes. Nat. Immunol. *25*, 1367–1382.

74. Dzhagalov, I.L., Chen, K.G., Herzmark, P., and Robey, E.A. (2013). Elimination of self-reactive T cells in the thymus: a timeline for negative selection. PLoS Biol. *11*, e1001566.

75. Wong, R.S.-Y., Tan, T., Pang, A.S.-R., and Srinivasan, D.K. (2025). The role of cytokines in wound healing: from mechanistic insights to therapeutic applications. Explor. Immunol. *5*, 1003183.

76. Caballero-Sánchez, N., Alonso-Alonso, S., and Nagy, L. (2024). Regenerative inflammation: When immune cells help to re-build tissues. FEBS J. *291*, 1597–1614.

77. Borky, J.M., and Bradley, T.H. (2019). Protecting Information with Cybersecurity. In Effective Model-Based Systems Engineering (Springer International Publishing), pp. 345–404.

78. Hussain, M.A.M.S., Suaib, M., and Shahid, M.S. (2022). An intelligent approach to automatic query formation from plain text using artificial intelligence. Int. J. Comput. Inf. Technol. *11*, 108–114.

79. Panteli, N., Nthubu, B.R., and Mersinas, K. (2025). Being responsible in cybersecurity: A multi-layered perspective. Inf. Syst. Front. *28*, 209–227.

80. Zheng, Y., Li, Z., Xu, X., and Zhao, Q. (2022). Dynamic defenses in cyber security: Techniques, methods and challenges. Digit. Commun. Netw. *8*, 422–435.

81. Bhuyan, M.H., Bhattacharyya, D.K., and Kalita, J.K. (2014). Network anomaly detection: Methods, systems and tools. IEEE Commun. Surv. Tutor. *16*, 303–336.

82. Yulianto, S., Soewito, B., Gaol, F.L., and Kurniawan, A. (2025). Enhancing cybersecurity resilience through advanced red-teaming exercises and MITRE ATT&CK framework integration: A paradigm shift in cybersecurity assessment. Cyber Secur. Appl. *3*, 100077.

83. Gambo, M.L., and Almulhem, A. (2026). Zero trust architecture: A systematic literature review. J. Netw. Syst. Manage. *34*, 25. https://doi.org/10.1007/s10922-025-09998-x.

84. Al-Sada, B., Sadighian, A., and Oligeri, G. (2025). MITRE ATT&CK: State of the art and way forward. ACM Comput. Surv. *57*, 1–37.

85. Manzoor, J., Waleed, A., Jamali, A.F., and Masood, A. (2024). Cybersecurity on a budget: Evaluating security and performance of open-source SIEM solutions for SMEs. PLoS One *19*, e0301183.

86. Ilg, N., Duplys, P., Sisejkovic, D., and Menth, M. (2023). A survey of contemporary open-source honeypots, frameworks, and tools. J. Netw. Comput. Appl. *220*, 103737.

87. Tzavara, V., and Vassiliadis, S. (2024). Tracing the evolution of cyber resilience: a historical and conceptual review. Int. J. Inf. Secur. *23*, 1695–1719.

88. Chu, S., Koe, J., Garlan, D., and Kang, E. (2024). Integrating graceful degradation and recovery through requirement-driven adaptation. In Proceedings of the 19th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (ACM). https://doi.org/10.1145/3643915.3644090.

89. Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., Sutskever, I., Leike, J., Wu, J., and Saunders, W. (2023). Language models can

explain neurons in language models. OpenAI. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html.

90. Marks, S., Rager, C., Michaud, E.J., Belinkov, Y., Bau, D., and Mueller, A. (2024). Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2403.19647.

91. Betley, J., Tan, D., Warncke, N., Sztyber-Betley, A., Bao, X., Soto, M., Labenz, N., and Evans, O. (2025). Emergent Misalignment: Narrow fine-tuning can produce broadly misaligned LLMs. Preprint at arXiv. https://doi.org/10.48550/arXiv.2502.17424.

92. Stutz, D., Hein, M., and Schiele, B. (2020). Confidence-calibrated adversarial training: Generalizing to unseen attacks. In Proceedings of the 37th International Conference on Machine Learning, *119* (PMLR), pp. 9155–9166.

93. Marks, S., Treutlein, J., Bricken, T., Lindsey, J., Marcus, J., Mishra-Sharma, S., Ziegler, D., Ameisen, E., Batson, J., Belonax, T., et al. (2025). Auditing language models for hidden objectives. Preprint at arXiv. https://doi.org/10.48550/arXiv.2503.10965.

94. Ji, J., Chen, W., Wang, K., Hong, D., Fang, S., Chen, B., Zhou, J., Dai, J., Han, S., Guo, Y., et al. (2025). Mitigating deceptive alignment via self-monitoring. Preprint at arXiv. https://arxiv.org/abs/2505.18807.

95. Hatfield, G., Leibo, J.Z., and Hadfield-Menell, D. (2024). Cross-disciplinary Insights into Alignment in Humans and Machines (NeurIPS 2024 Tutorial). https://neurips.cc/virtual/2024/tutorial/99529.

96. U.S. AI Safety Institute Signs Agreements Regarding AI Safety Research (2024). Testing and Evaluation with Anthropic and OpenAI (NIST). https://www.nist.gov/news-events/news/2024/08/us-ai-safety-institute-signs-agreements-regarding-ai-safety-research.

97. Chen, S., Gallifant, J., Gao, M., Moreira, P., Munch, N., Muthukkumar, A., Rajan, A., Kolluri, J., Fiske, A., Hastings, J., et al. (2024). Cross-Care: Assessing the healthcare implications of pre-training data on language model bias. Adv. Neural Inform. Process. Syst. *37*, 23756–23795.

98. U.S. Department of Health and Human Services (2024). Health data, technology, and interoperability: Certification program updates, algorithm transparency, and information sharing. Fed. Regist. *89*, 1192–1474. https://www.federalregister.gov/documents/2024/01/09/2023-28857/health-data-technology-and-interoperability-certification-program-updates-algorithm-transparency-and.

99. Coalition for Health AI (2026) (CHAI). https://www.chai.org/.

100. Ozili, P.K. (2025). Digital public infrastructure: Concepts, global efforts, benefits, challenges, and success stories. Digit. Soc. *4*, 30. https://doi.org/10.1007/s44206-025-00185-8.

101. United Nations Office for Digital and Emerging Technologies (2026). Digital Public Infrastructure. https://www.un.org/digital-emerging-technologies/content/digital-public-infrastructure.

About the authors

Ben Y. Reis is director of the Predictive Medicine Group and associate professor at Harvard Medical School and the Boston Children's Hospital Computational Health Informatics Program. His research focuses on understanding the fundamental patterns of human disease and investigating the organizing dynamics of AI systems. He has created systems that predict life-threatening diseases years in advance and advised governments on monitoring pandemics and protecting the Olympic Games and was honored at the White House for his work on harnessing technology to promote health. Dr. Reis trained in computer science and artificial intelligence at MIT, the University of Cambridge, and Harvard University.

William G. La Cava is an assistant professor at Harvard Medical School and a core faculty member of the Computational Health Informatics Program at Boston Children's Hospital. He is the director of the Clinical AI Value Alignment Lab, a research group focused on improving the trustworthiness of artificial intelligence models in medicine. Dr. La Cava was a post-doctoral fellow and research associate at the Institute for Biomedical Informatics of the University of Pennsylvania and received his PhD from UMass Amherst with a focus on interpretable modeling of dynamical systems.