Table of Contents

1. [**Introduction**](https://www.genspark.ai/spark?id=74a7e916-139d-467f-86b5-6ccefd5ba014#introduction)
2. [I. Fundamental Theoretical Challenges](https://www.genspark.ai/spark?id=74a7e916-139d-467f-86b5-6ccefd5ba014#i.-fundamental-theoretical-challenges)
   1. [1. The Frame Problem](https://www.genspark.ai/spark?id=74a7e916-139d-467f-86b5-6ccefd5ba014#1.-the-frame-problem)
   2. [2. The Symbol Grounding Problem](https://www.genspark.ai/spark?id=74a7e916-139d-467f-86b5-6ccefd5ba014#2.-the-symbol-grounding-problem)
   3. [3. The AI Alignment Problem](https://www.genspark.ai/spark?id=74a7e916-139d-467f-86b5-6ccefd5ba014#3.-the-ai-alignment-problem)
   4. [4. Consciousness and Sentience in AI](https://www.genspark.ai/spark?id=74a7e916-139d-467f-86b5-6ccefd5ba014#4.-consciousness-and-sentience-in-ai)
   5. [5. Commonsense Reasoning](https://www.genspark.ai/spark?id=74a7e916-139d-467f-86b5-6ccefd5ba014#5.-commonsense-reasoning)
3. [II. Technical and Engineering Challenges](https://www.genspark.ai/spark?id=74a7e916-139d-467f-86b5-6ccefd5ba014#ii.-technical-and-engineering-challenges)
   1. [1. AI Hallucinations](https://www.genspark.ai/spark?id=74a7e916-139d-467f-86b5-6ccefd5ba014#1.-ai-hallucinations)
   2. [2. Catastrophic Forgetting](https://www.genspark.ai/spark?id=74a7e916-139d-467f-86b5-6ccefd5ba014#2.-catastrophic-forgetting)
   3. [3. Adversarial Robustness](https://www.genspark.ai/spark?id=74a7e916-139d-467f-86b5-6ccefd5ba014#3.-adversarial-robustness)
   4. [4. Sample Efficiency](https://www.genspark.ai/spark?id=74a7e916-139d-467f-86b5-6ccefd5ba014#4.-sample-efficiency)
   5. [5. Out-of-Distribution Generalization](https://www.genspark.ai/spark?id=74a7e916-139d-467f-86b5-6ccefd5ba014#5.-out-of-distribution-generalization)
   6. [6. Explainability and Transparency](https://www.genspark.ai/spark?id=74a7e916-139d-467f-86b5-6ccefd5ba014#6.-explainability-and-transparency)
4. [III. Socio-Technical and Governance Challenges](https://www.genspark.ai/spark?id=74a7e916-139d-467f-86b5-6ccefd5ba014#iii.-socio-technical-and-governance-challenges)
   1. [1. AI Safety and Containment](https://www.genspark.ai/spark?id=74a7e916-139d-467f-86b5-6ccefd5ba014#1.-ai-safety-and-containment)
   2. [2. AI Misuse](https://www.genspark.ai/spark?id=74a7e916-139d-467f-86b5-6ccefd5ba014#2.-ai-misuse)
   3. [3. Power Concentration](https://www.genspark.ai/spark?id=74a7e916-139d-467f-86b5-6ccefd5ba014#3.-power-concentration)
   4. [4. Value Alignment with Humans](https://www.genspark.ai/spark?id=74a7e916-139d-467f-86b5-6ccefd5ba014#4.-value-alignment-with-humans)
5. [Conclusion](https://www.genspark.ai/spark?id=74a7e916-139d-467f-86b5-6ccefd5ba014#conclusion)
6. [References](https://www.genspark.ai/spark?id=74a7e916-139d-467f-86b5-6ccefd5ba014#references)
7. [Appendix: Supplementary Video Resources](https://www.genspark.ai/spark?id=74a7e916-139d-467f-86b5-6ccefd5ba014#appendix%3A-supplementary-video-resources)

**Core Problems Facing Artificial Intelligence Today: A Comprehensive Analysis**

A white star and sparkles in a circle

AI-generated content may be incorrect.

Genspark

May 27, 2025

Info

Bookmark

Share

Generated with sparks and insights from [15 sources](prompt://ask_markdown/?question=Show%20me%20all%20the%20sources&answer=Here%20are%20the%20sources%20I%20found%20for%20you%3A%0A-%20%5Bphilarchive.org%5D%28https%3A//philarchive.org/rec/MIRUTF%29%0A-%20%5Bcs.ox.ac.uk%5D%28https%3A//www.cs.ox.ac.uk/activities/ieg/e-library/sources/harnad90_sgproblem.pdf%29%0A-%20%5Bpsychologytoday.com%5D%28https%3A//www.psychologytoday.com/us/blog/tech-happy-life/202505/the-solution-to-the-ai-alignment-problem-is-in-the-mirror%29%0A-%20%5Barxiv.org%5D%28https%3A//arxiv.org/abs/2310.19852%29%0A-%20%5Blesswrong.com%5D%28https%3A//www.lesswrong.com/posts/bzYJCXicmwDHDpLZa/reframing-ai-safety-as-a-neverending-institutional-challenge%29%0A-%20%5Bthegradient.pub%5D%28https%3A//thegradient.pub/an-introduction-to-the-problems-of-ai-consciousness/%29%0A-%20%5Baicompetence.org%5D%28https%3A//aicompetence.org/why-ai-struggles-with-commonsense-reasoning/%29%0A-%20%5Bnytimes.com%5D%28https%3A//www.nytimes.com/2025/05/05/technology/ai-hallucinations-chatgpt-google.html%29%0A-%20%5Bibm.com%5D%28https%3A//www.ibm.com/think/topics/ai-hallucinations%29%0A-%20%5Bibm.com%5D%28https%3A//www.ibm.com/think/topics/catastrophic-forgetting%29%0A-%20%5Bcset.georgetown.edu%5D%28https%3A//cset.georgetown.edu/publication/key-concepts-in-ai-safety-robustness-and-adversarial-examples/%29%0A-%20%5Bmedium.com%5D%28https%3A//medium.com/%40prdeepak.babu/sample-efficient-learning-in-llms-e81a62af4cc3%29%0A-%20%5Bnature.com%5D%28https%3A//www.nature.com/articles/s43246-024-00731-w%29%0A-%20%5Borbograph.com%5D%28https%3A//orbograph.com/explainable-ai-xai-challenges-how-to-overcome-them/%29%0A-%20%5Bsafe.ai%5D%28https%3A//www.safe.ai/ai-risk%29%0A)

**Introduction**

Artificial intelligence has made remarkable advances in recent years, with large language models, computer vision systems, and reinforcement learning approaches achieving unprecedented capabilities. However, beneath these impressive achievements lies a landscape of profound theoretical, technical, and social challenges that researchers and engineers continue to grapple with. This report comprehensively documents the major unsolved problems in AI, explaining their significance, current research status, and proposed solutions.

**I. Fundamental Theoretical Challenges**

**1. The Frame Problem**

**Definition:** The frame problem refers to the challenge of enabling an AI system to determine which aspects of its environment or knowledge are relevant to a given situation and which can be safely ignored when making decisions or predictions.

**Historical Context:** First formalized in 1969 by John McCarthy and Patrick Hayes in their work on logical AI [1], the frame problem highlights a fundamental disconnect between how humans efficiently filter relevant information and how AI systems struggle with this seemingly basic task.

**Current Research Status:** According to Miracchi (2020), the frame problem requires us to update our understanding beyond simplistic computational approaches [2]. The problem persists in modern AI systems, including large language models that often include irrelevant information or miss crucial context.

**Proposed Solutions:**

* Lisa Miracchi argues for moving beyond the standard interpretation that assumes mental processes are identical to computational processes [2]
* Hybrid approaches that combine symbolic reasoning with connectionist learning
* Probabilistic frameworks for relevance determination
* Integration of causal reasoning to establish relevance hierarchies

**Real-world Implications:** The frame problem manifests when autonomous systems make inappropriate decisions by failing to identify which contextual factors matter. For example, self-driving cars may prioritize irrelevant data while missing crucial situational changes, or chatbots may generate responses that ignore critical context.

**2. The Symbol Grounding Problem**

**Definition:** The symbol grounding problem addresses how symbols in AI systems acquire meaning intrinsically rather than parasitically through human interpretation. It questions how arbitrary symbols can be connected to their real-world referents.

**Historical Context:** Stevan Harnad formally introduced this problem in 1990, comparing it to trying to learn Chinese from a Chinese-Chinese dictionary alone – symbols defined only in terms of other symbols create an endless loop of meaningless references [3].

**Current Research Status:** Despite decades of research, the symbol grounding problem remains unsolved. Recent approaches have shifted toward embodied and situated AI that can learn from sensorimotor experiences.

**Proposed Solutions:**

* Harnad's hybrid model combining "iconic representations" (analog internal transformations of sensory data) with "categorical representations" (feature detectors that extract invariants) [3]
* Grounding through multi-modal learning (text, images, audio)
* Embodied AI with physical manipulation capabilities
* Connectionist approaches that learn grounded feature representations

**Real-world Implications:** Ungrounded symbols contribute to AI systems' brittleness when faced with novel situations and their tendency to generate plausible-sounding but factually incorrect outputs. This affects everything from chatbots to autonomous systems tasked with real-world navigation or manipulation.

**3. The AI Alignment Problem**

**Definition:** The AI alignment problem concerns ensuring AI systems act in accordance with human values, intentions, and ethical principles, especially as they become more capable.

**Historical Context:** As AI systems become increasingly powerful and autonomous, ensuring they remain aligned with human goals becomes critical. The concept gained prominence through the work of Bostrom (2014) and has since become central to AI safety research [4].

**Current Research Status:** Recent work captures the field's status: "AI alignment can't succeed until humans confront their own divisions and contradictions. Advanced AI systems learn by reflecting us—what they learn includes our disagreements" [5]. Ji et al.'s 2023 comprehensive survey documents the expanding research landscape around alignment [6].

**Proposed Solutions:**

* Value learning: techniques to infer human values from demonstrations
* Reinforcement learning from human feedback (RLHF)
* Constitutional AI approaches with encoded ethical principles
* Interpretability research to detect misalignment
* Institutional approaches focusing on governance, transparency, and preparedness rather than purely technical alignment [7]

**Real-world Implications:** Misaligned AI could optimize for metrics that violate human values or intentions, such as maximizing engagement at the expense of promoting harmful content or pursuing goals in ways that cause unintended side effects.

**4. Consciousness and Sentience in AI**

**Definition:** This challenge concerns whether artificial systems can develop consciousness (subjective experience) and, if so, how we would recognize and address it ethically.

**Historical Context:** As AI systems demonstrate increasingly sophisticated behaviors, questions around machine consciousness have moved from philosophical thought experiments to practical considerations regarding ethics and rights.

**Current Research Status:** The field remains divided between those arguing for the possibility of machine consciousness based on functionalism and those maintaining that consciousness requires biological substrates. As an introduction to the problems notes, this challenge combines the "hard problem" of consciousness with the "problem of other minds" to form what philosopher Ned Block calls the "harder" problem [8].

**Proposed Solutions:**

* Theory-driven approaches: Apply neuroscientific theories of consciousness (Global Workspace, Integrated Information Theory) to AI systems
* Theory-neutral approaches: Chalmers' "fading and dancing qualia" arguments for substrate independence
* Empirical measures: Schneider's "chip test" and the ACT test for AI consciousness
* Philosophical frameworks for extending moral consideration to potentially conscious AI

**Real-world Implications:** If advanced AI systems could experience suffering, this would raise profound ethical questions about how we develop, deploy, and potentially terminate them. Conversely, premature attribution of consciousness could lead to inappropriate treatment of sophisticated but non-conscious systems.

**5. Commonsense Reasoning**

**Definition:** Commonsense reasoning is the ability to make inferences about everyday situations using general knowledge about how the world works—knowledge that is typically obvious to humans but challenging to encode in machines.

**Historical Context:** The difficulty of programming commonsense knowledge has been recognized since the early days of AI, with Minsky and McCarthy highlighting it as a fundamental challenge in the 1960s.

**Current Research Status:** Despite advances in language models that can mimic some aspects of commonsense reasoning, AI systems still struggle with fundamental reasoning about physical causality, social dynamics, and implicit knowledge that humans take for granted.

**Proposed Solutions:**

* Knowledge graphs and databases like ConceptNet to capture common-sense relationships [9]
* Multi-modal learning combining text, images, and audio [9]
* Hybrid symbolic-connectionist approaches
* Specialized benchmarks like Winograd Schema Challenge and CommonsenseQA

**Real-world Implications:** Without commonsense reasoning, AI assistants, robots, and decision support systems make nonsensical suggestions or dangerous mistakes in situations that would be obvious to humans. This limits their usefulness in unstructured or novel environments.

**II. Technical and Engineering Challenges**

**1. AI Hallucinations**

**Definition:** AI hallucinations occur when generative models produce outputs that are factually incorrect, inconsistent, or entirely fabricated while presenting them confidently as true.

**Historical Context:** As large language models have scaled in size and capability, hallucinations have emerged as a significant limitation—particularly concerning as these models are increasingly deployed in contexts where accuracy is crucial.

**Current Research Status:** Hallucinations appear to be getting worse in some cases, not better, as models scale. According to a 2024 New York Times report, "Tests by independent companies and researchers indicate that hallucination rates are also rising for reasoning models from companies such as Google and Anthropic" [10].

**Causes:**

* Training data containing inaccuracies or contradictions
* Overfitting to patterns rather than learning factual relationships
* Improper decoding by transformer architectures
* High model complexity leading to unpredictable behaviors [11]

**Proposed Solutions:**

* Adversarial training with examples designed to induce hallucinations
* High-quality, verified training data
* Response constraints via filters or probabilistic thresholds
* Rigorous testing protocols before deployment [11]
* Better fact-verification mechanisms
* Retrieval-augmented generation to ground responses in reliable sources

**Real-world Implications:** Hallucinations undermine trust in AI systems and can lead to serious consequences when relied upon for critical decisions in healthcare, finance, or legal contexts.

**2. Catastrophic Forgetting**

**Definition:** Catastrophic forgetting (or catastrophic interference) is the tendency of neural networks to abruptly lose previously learned information when trained on new tasks or data.

**Historical Context:** This problem has been recognized since early neural network research, but has gained renewed attention with the expansion of lifelong learning applications.

**Current Research Status:** While progress has been made, catastrophic forgetting remains a key obstacle to creating AI systems that can continuously learn without requiring complete retraining.

**Proposed Solutions:**

* Regularization techniques like Elastic Weight Consolidation (EWC) that slow learning rates for weights important to previous tasks
* Architectural solutions such as Progressive Neural Networks
* Ensemble methods that combine separate models for different tasks
* Rehearsal techniques that periodically revisit previous training examples
* Memory-Augmented Neural Networks (MANNs) that explicitly store key examples [12]

**Real-world Implications:** Forgetting impacts AI systems' ability to adapt to changing environments without compromising core capabilities. This is especially concerning for autonomous systems like self-driving cars that need to maintain safety-critical skills while learning new ones.

**3. Adversarial Robustness**

**Definition:** Adversarial robustness addresses AI systems' vulnerability to specially crafted inputs ("adversarial examples") designed to cause misclassification or erroneous outputs.

**Historical Context:** Research by Szegedy et al. in 2013 first demonstrated that small, imperceptible perturbations to images could cause state-of-the-art image classifiers to make confident but incorrect predictions [13].

**Current Research Status:** Despite significant research effort, creating models robust to adversarial attacks remains challenging, especially without sacrificing performance on clean data.

**Proposed Solutions:**

* Adversarial training: Including adversarial examples during model training
* Defensive distillation: Training networks on the probability outputs of another network
* Feature denoising: Removing noise from feature maps within the network
* Certified robustness approaches that provide mathematical guarantees
* Detection mechanisms to identify potential adversarial inputs [14]

**Real-world Implications:** Adversarial vulnerabilities create security risks in critical applications like autonomous vehicles, facial recognition, or medical diagnosis systems, where an attacker could cause dangerous misclassifications with subtle input manipulations.

**4. Sample Efficiency**

**Definition:** Sample efficiency refers to a model's ability to learn effectively from a limited number of examples, requiring minimal data to reach a given level of performance.

**Historical Context:** The massive data requirements of modern deep learning approaches highlight the efficiency gap between human learning (often requiring just a few examples) and machine learning (often requiring millions).

**Current Research Status:** Improving sample efficiency remains a central challenge, particularly for applications where data collection is expensive, time-consuming, or ethically problematic.

**Proposed Solutions:**

* Sparsity: Using Sparse Mixture of Experts (MoE) architectures where only subnetworks activate per input
* Multimodality: Training on multiple data types (text, images, audio) for richer representations
* Differential Compute: Allocating resources dynamically based on task complexity
* Curriculum Learning: Ordering training data from simple to complex examples
* Model Merging: Combining pre-trained models without retraining from scratch [15]

**Real-world Implications:** Better sample efficiency would reduce computational and environmental costs of AI development, enable domain adaptation with minimal new data, and make AI more practical for specialized domains with limited available data.

**5. Out-of-Distribution Generalization**

**Definition:** Out-of-distribution (OOD) generalization refers to AI systems' ability to maintain reliable performance on data drawn from distributions different from those seen during training.

**Historical Context:** While traditional machine learning focuses on in-distribution generalization, real-world deployment inevitably exposes models to distribution shifts that can dramatically degrade performance.

**Current Research Status:** A 2024 study from Nature investigating OOD generalization in materials science highlights a critical insight: "Only a handful of truly challenging OOD tasks contain significant amounts of test data that lie outside this domain. Importantly, the performance on these representationally OOD tasks does not adhere to conventional neural scaling laws, suggesting that the benefits of scaling for OOD generalization may be overstated or misinterpreted" [16].

**Proposed Solutions:**

* Representation-based domain identification to distinguish truly OOD examples from merely statistically different ones
* Causal representation learning to capture invariant relationships
* Distributionally robust optimization
* Domain generalization techniques
* Enhanced tokenization for language models that incorporates domain-specific similarity priors [16]

**Real-world Implications:** Poor OOD generalization leads to unexpected failures when AI systems encounter novel scenarios, domains, or edge cases not represented in their training data—a critical concern for safety-critical applications like autonomous vehicles or medical diagnosis.

**6. Explainability and Transparency**

**Definition:** Explainable AI (XAI) concerns making AI systems' decision-making processes transparent and understandable to humans, rather than operating as opaque "black boxes."

**Historical Context:** As AI systems make increasingly consequential decisions, the need to understand and validate their reasoning has grown urgent, driving research in explainability.

**Current Research Status:** Despite progress in techniques for post-hoc explanation and inherently interpretable models, making complex models like deep neural networks fully transparent remains challenging.

**Proposed Solutions:**

* Model-agnostic techniques that treat AI systems as black boxes and provide post-hoc explanations
* Model-specific techniques that leverage internal structure for native explanations
* Global interpretation methods that explain overall patterns across all data
* Local interpretation methods that explain individual predictions
* Visualization techniques to represent high-dimensional model behavior [17]

**Real-world Implications:** Lack of explainability creates barriers to AI adoption in regulated industries, limits users' ability to identify and correct errors, and undermines trust in AI-assisted decision-making.

**III. Socio-Technical and Governance Challenges**

**1. AI Safety and Containment**

**Definition:** AI safety and containment address the challenges of ensuring AI systems behave safely and remain under human control, including preventing unintended consequences from emergent behaviors.

**Historical Context:** As AI capabilities have advanced, concerns about safety have shifted from science fiction to practical research questions about how to prevent powerful systems from causing harm.

**Current Research Status:** Safety research has expanded significantly with institutions like the Center for AI Safety (CAIS) advancing frameworks for risk mitigation. As their analysis highlights, competitive pressures—both military and commercial—complicate safety efforts by incentivizing rapid deployment of increasingly capable but potentially unsafe systems [18].

**Proposed Solutions:**

* Safety regulations with independent oversight
* Meaningful human control over high-stakes decisions
* International coordination on AI development
* Public governance of general-purpose AI systems
* Institutional approaches to ongoing safety challenges rather than one-off technical solutions [7,18]

**Real-world Implications:** Inadequate safety measures could lead to AI systems causing harm through misalignment with human values, emergent behaviors, or exploitation by malicious actors.

**2. AI Misuse**

**Definition:** AI misuse concerns the intentional deployment of AI systems for harmful purposes, including autonomous weapons, scaled disinformation, surveillance, and cybersecurity attacks.

**Historical Context:** Dual-use technologies have always posed governance challenges, but AI's scalability, accessibility, and autonomous capabilities present unique risks.

**Current Research Status:** The military AI arms race is accelerating, with advanced autonomous weapons already deployed in conflicts. As CAIS reports, "Lethal autonomous weapons are AI-driven systems capable of identifying and executing targets without human intervention" [18]. Concurrently, AI systems are enabling more sophisticated cyberattacks and disinformation campaigns.

**Proposed Solutions:**

* International agreements limiting autonomous weapons
* AI-powered defensive cybersecurity systems
* Strong verification and enforcement mechanisms
* Industry standards and norms against misuse
* Technical safeguards against repurposing [18]

**Real-world Implications:** Without adequate governance, AI could enable new forms of large-scale harm through autonomous weapons, targeted manipulation, privacy violations, or automated cyberattacks.

**3. Power Concentration**

**Definition:** Power concentration refers to the tendency of advanced AI capabilities to consolidate economic, political, and social power in the hands of a few organizations or actors.

**Historical Context:** The capital-intensive nature of cutting-edge AI research, along with data and compute advantages for early leaders, has led to significant concentration in AI development capabilities.

**Current Research Status:** As AI automates more tasks and industries, concerns about economic concentration and dependence have grown. CAIS notes that "businesses will likely replace more types of human labor with AI, potentially triggering mass unemployment" while "the economy may become largely run by AIs. Eventually, this could lead to human enfeeblement and dependence on AIs for basic needs" [18].

**Proposed Solutions:**

* Regulatory oversight of AI development
* Open-source initiatives to democratize access to AI technology
* Data sharing frameworks and compute access programs
* Public investment in broadly beneficial AI applications
* Antitrust enforcement in AI-dominated markets

**Real-world Implications:** Concentrated AI power could exacerbate existing inequalities, disrupt labor markets without adequate transitions, and undermine democratic processes if governance fails to address these dynamics.

**4. Value Alignment with Humans**

**Definition:** Beyond the technical alignment problem, value alignment addresses the challenge of ensuring AI systems incorporate the full spectrum of human values, including addressing value pluralism and cultural differences.

**Historical Context:** As AI systems make or influence increasingly complex decisions, the question of whose values they should reflect becomes critical—especially given wide variations in human values across cultures and individuals.

**Current Research Status:** Research has shifted from viewing alignment as a purely technical problem to recognizing its deeply social nature. As noted in "The Solution to the AI Alignment Problem Is in the Mirror": "AI alignment can't succeed until humans confront their own divisions and contradictions. Advanced AI systems learn by reflecting us—what they learn includes our disagreements" [5].

**Proposed Solutions:**

* Participatory design approaches involving diverse stakeholders
* Constitutional AI methods to handle value conflicts
* Cultural sensitivity in AI development and deployment
* Frameworks for navigating value pluralism
* Ethical review processes for AI systems [5,6]

**Real-world Implications:** AI systems that fail to account for diverse human values may inadvertently privilege certain cultural perspectives, encode biases, or make decisions at odds with community values.

**Conclusion**

The challenges facing AI research and development span fundamental theoretical questions about knowledge representation and consciousness to practical engineering problems like robustness and explainability, extending to broader societal concerns around governance and value alignment. Progress on these core problems will determine whether AI systems can fulfill their potential as beneficial tools that augment human capabilities while respecting human values and autonomy.

Many researchers now recognize that these challenges are deeply interconnected. Advances in technical areas like sample efficiency and out-of-distribution generalization require grappling with fundamental questions about knowledge representation and causality. Similarly, developing robust governance frameworks depends on technical progress in areas like explainability and alignment.

As AI capabilities continue to advance, addressing these core problems becomes increasingly urgent. The field must balance the competitive pressures driving rapid deployment with the careful, methodical research needed to develop AI systems that are not only powerful but also safe, beneficial, and aligned with human interests.

**References**

[1] McCarthy, J., & Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer & D. Michie (Eds.), Machine Intelligence 4 (pp. 463-502). Edinburgh University Press.

[2] Miracchi, L. (2020). Updating the Frame Problem for AI Research. Journal of Artificial Intelligence and Consciousness, 7(02), 203-216. philarchive.org[1](https://philarchive.org/rec/MIRUTF)

[3] Harnad, S. (1990). The symbol grounding problem. Physica D: Nonlinear Phenomena, 42(1-3), 335-346. cs.ox.ac.uk[2](https://www.cs.ox.ac.uk/activities/ieg/e-library/sources/harnad90_sgproblem.pdf)

[4] Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.

[5] (2024). The Solution to the AI Alignment Problem Is in the Mirror. Psychology Today. psychologytoday.com[3](https://www.psychologytoday.com/us/blog/tech-happy-life/202505/the-solution-to-the-ai-alignment-problem-is-in-the-mirror)

[6] Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., & Wang, X. (2023). AI alignment: A comprehensive survey. arXiv preprint arXiv:2310.19852. arxiv.org[4](https://arxiv.org/abs/2310.19852)

[7] Casper, S. (2024). Reframing AI safety as a neverending institutional challenge. lesswrong.com[5](https://www.lesswrong.com/posts/bzYJCXicmwDHDpLZa/reframing-ai-safety-as-a-neverending-institutional-challenge)

[8] (2023). An Introduction to the Problems of AI Consciousness. The Gradient. thegradient.pub[6](https://thegradient.pub/an-introduction-to-the-problems-of-ai-consciousness/)

[9] (2023). Why AI Struggles with Commonsense Reasoning. AI Competence. aicompetence.org[7](https://aicompetence.org/why-ai-struggles-with-commonsense-reasoning/)

[10] (2025). A.I. Is Getting More Powerful, but Its Hallucinations Are Getting Worse. The New York Times. nytimes.com[8](https://www.nytimes.com/2025/05/05/technology/ai-hallucinations-chatgpt-google.html)

[11] (2024). What Are AI Hallucinations? IBM. ibm.com[9](https://www.ibm.com/think/topics/ai-hallucinations)

[12] (2024). What is Catastrophic Forgetting? IBM. ibm.com[10](https://www.ibm.com/think/topics/catastrophic-forgetting)

[13] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

[14] (2022). Key Concepts in AI Safety: Robustness and Adversarial Examples. Center for Security and Emerging Technology. cset.georgetown.edu[11](https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-robustness-and-adversarial-examples/)

[15] Babu, P. R. D. (2024). Sample Efficient Learning in LLMs. Medium. medium.com[12](https://medium.com/@prdeepak.babu/sample-efficient-learning-in-llms-e81a62af4cc3)

[16] (2024). Probing out-of-distribution generalization in machine learning for materials. Nature Communications Materials. nature.com[13](https://www.nature.com/articles/s43246-024-00731-w)

[17] (2024). Explainable AI (XAI): Challenges & How to Overcome Them. Orbograph. orbograph.com[14](https://orbograph.com/explainable-ai-xai-challenges-how-to-overcome-them/)

[18] (2024). AI Risks that Could Lead to Catastrophe. Center for AI Safety. safe.ai[15](https://www.safe.ai/ai-risk)

**Appendix: Supplementary Video Resources**

<div class="-md-ext-youtube-widget"> { "title": "Understanding and Effectively Using AI Reasoning Models", "link": "https://www.youtube.com/watch?v=f0RbwrBcFmc", "channel": { "name": ""}, "published\_date": "Jan 22, 2025", "length": "15:16" }</div>

<div class="-md-ext-youtube-widget"> { "title": "AI Control Problem Simply Explained", "link": "https://www.youtube.com/watch?v=cvUerkVDJEI", "channel": { "name": ""}, "published\_date": "Oct 28, 2023", "length": "10:52" }</div>

<div class="-md-ext-youtube-widget"> { "title": "Core Technologies Powering Generative AI | Exclusive Lesson", "link": "https://www.youtube.com/watch?v=Z4rQ-ZVShLw", "channel": { "name": ""}, "published\_date": "Feb 23, 2025", "length": "8:56" }</div>

Generated with sparks and insights from [15 sources](prompt://ask_markdown/?question=Show%20me%20all%20the%20sources&answer=Here%20are%20the%20sources%20I%20found%20for%20you%3A%0A-%20%5Bphilarchive.org%5D%28https%3A//philarchive.org/rec/MIRUTF%29%0A-%20%5Bcs.ox.ac.uk%5D%28https%3A//www.cs.ox.ac.uk/activities/ieg/e-library/sources/harnad90_sgproblem.pdf%29%0A-%20%5Bpsychologytoday.com%5D%28https%3A//www.psychologytoday.com/us/blog/tech-happy-life/202505/the-solution-to-the-ai-alignment-problem-is-in-the-mirror%29%0A-%20%5Barxiv.org%5D%28https%3A//arxiv.org/abs/2310.19852%29%0A-%20%5Blesswrong.com%5D%28https%3A//www.lesswrong.com/posts/bzYJCXicmwDHDpLZa/reframing-ai-safety-as-a-neverending-institutional-challenge%29%0A-%20%5Bthegradient.pub%5D%28https%3A//thegradient.pub/an-introduction-to-the-problems-of-ai-consciousness/%29%0A-%20%5Baicompetence.org%5D%28https%3A//aicompetence.org/why-ai-struggles-with-commonsense-reasoning/%29%0A-%20%5Bnytimes.com%5D%28https%3A//www.nytimes.com/2025/05/05/technology/ai-hallucinations-chatgpt-google.html%29%0A-%20%5Bibm.com%5D%28https%3A//www.ibm.com/think/topics/ai-hallucinations%29%0A-%20%5Bibm.com%5D%28https%3A//www.ibm.com/think/topics/catastrophic-forgetting%29%0A-%20%5Bcset.georgetown.edu%5D%28https%3A//cset.georgetown.edu/publication/key-concepts-in-ai-safety-robustness-and-adversarial-examples/%29%0A-%20%5Bmedium.com%5D%28https%3A//medium.com/%40prdeepak.babu/sample-efficient-learning-in-llms-e81a62af4cc3%29%0A-%20%5Bnature.com%5D%28https%3A//www.nature.com/articles/s43246-024-00731-w%29%0A-%20%5Borbograph.com%5D%28https%3A//orbograph.com/explainable-ai-xai-challenges-how-to-overcome-them/%29%0A-%20%5Bsafe.ai%5D%28https%3A//www.safe.ai/ai-risk%29%0A)

**Introduction**

Artificial intelligence has made remarkable advances in recent years, with large language models, computer vision systems, and reinforcement learning approaches achieving unprecedented capabilities. However, beneath these impressive achievements lies a landscape of profound theoretical, technical, and social challenges that researchers and engineers continue to grapple with. This report comprehensively documents the major unsolved problems in AI, explaining their significance, current research status, and proposed solutions.

**I. Fundamental Theoretical Challenges**

**1. The Frame Problem**

**Definition:** The frame problem refers to the challenge of enabling an AI system to determine which aspects of its environment or knowledge are relevant to a given situation and which can be safely ignored when making decisions or predictions.

**Historical Context:** First formalized in 1969 by John McCarthy and Patrick Hayes in their work on logical AI [1], the frame problem highlights a fundamental disconnect between how humans efficiently filter relevant information and how AI systems struggle with this seemingly basic task.

**Current Research Status:** According to Miracchi (2020), the frame problem requires us to update our understanding beyond simplistic computational approaches [2]. The problem persists in modern AI systems, including large language models that often include irrelevant information or miss crucial context.

**Proposed Solutions:**

* Lisa Miracchi argues for moving beyond the standard interpretation that assumes mental processes are identical to computational processes [2]
* Hybrid approaches that combine symbolic reasoning with connectionist learning
* Probabilistic frameworks for relevance determination
* Integration of causal reasoning to establish relevance hierarchies

**Real-world Implications:** The frame problem manifests when autonomous systems make inappropriate decisions by failing to identify which contextual factors matter. For example, self-driving cars may prioritize irrelevant data while missing crucial situational changes, or chatbots may generate responses that ignore critical context.

**2. The Symbol Grounding Problem**

**Definition:** The symbol grounding problem addresses how symbols in AI systems acquire meaning intrinsically rather than parasitically through human interpretation. It questions how arbitrary symbols can be connected to their real-world referents.

**Historical Context:** Stevan Harnad formally introduced this problem in 1990, comparing it to trying to learn Chinese from a Chinese-Chinese dictionary alone – symbols defined only in terms of other symbols create an endless loop of meaningless references [3].

**Current Research Status:** Despite decades of research, the symbol grounding problem remains unsolved. Recent approaches have shifted toward embodied and situated AI that can learn from sensorimotor experiences.

**Proposed Solutions:**

* Harnad's hybrid model combining "iconic representations" (analog internal transformations of sensory data) with "categorical representations" (feature detectors that extract invariants) [3]
* Grounding through multi-modal learning (text, images, audio)
* Embodied AI with physical manipulation capabilities
* Connectionist approaches that learn grounded feature representations

**Real-world Implications:** Ungrounded symbols contribute to AI systems' brittleness when faced with novel situations and their tendency to generate plausible-sounding but factually incorrect outputs. This affects everything from chatbots to autonomous systems tasked with real-world navigation or manipulation.

**3. The AI Alignment Problem**

**Definition:** The AI alignment problem concerns ensuring AI systems act in accordance with human values, intentions, and ethical principles, especially as they become more capable.

**Historical Context:** As AI systems become increasingly powerful and autonomous, ensuring they remain aligned with human goals becomes critical. The concept gained prominence through the work of Bostrom (2014) and has since become central to AI safety research [4].

**Current Research Status:** Recent work captures the field's status: "AI alignment can't succeed until humans confront their own divisions and contradictions. Advanced AI systems learn by reflecting us—what they learn includes our disagreements" [5]. Ji et al.'s 2023 comprehensive survey documents the expanding research landscape around alignment [6].

**Proposed Solutions:**

* Value learning: techniques to infer human values from demonstrations
* Reinforcement learning from human feedback (RLHF)
* Constitutional AI approaches with encoded ethical principles
* Interpretability research to detect misalignment
* Institutional approaches focusing on governance, transparency, and preparedness rather than purely technical alignment [7]

**Real-world Implications:** Misaligned AI could optimize for metrics that violate human values or intentions, such as maximizing engagement at the expense of promoting harmful content or pursuing goals in ways that cause unintended side effects.

**4. Consciousness and Sentience in AI**

**Definition:** This challenge concerns whether artificial systems can develop consciousness (subjective experience) and, if so, how we would recognize and address it ethically.

**Historical Context:** As AI systems demonstrate increasingly sophisticated behaviors, questions around machine consciousness have moved from philosophical thought experiments to practical considerations regarding ethics and rights.

**Current Research Status:** The field remains divided between those arguing for the possibility of machine consciousness based on functionalism and those maintaining that consciousness requires biological substrates. As an introduction to the problems notes, this challenge combines the "hard problem" of consciousness with the "problem of other minds" to form what philosopher Ned Block calls the "harder" problem [8].

**Proposed Solutions:**

* Theory-driven approaches: Apply neuroscientific theories of consciousness (Global Workspace, Integrated Information Theory) to AI systems
* Theory-neutral approaches: Chalmers' "fading and dancing qualia" arguments for substrate independence
* Empirical measures: Schneider's "chip test" and the ACT test for AI consciousness
* Philosophical frameworks for extending moral consideration to potentially conscious AI

**Real-world Implications:** If advanced AI systems could experience suffering, this would raise profound ethical questions about how we develop, deploy, and potentially terminate them. Conversely, premature attribution of consciousness could lead to inappropriate treatment of sophisticated but non-conscious systems.

**5. Commonsense Reasoning**

**Definition:** Commonsense reasoning is the ability to make inferences about everyday situations using general knowledge about how the world works—knowledge that is typically obvious to humans but challenging to encode in machines.

**Historical Context:** The difficulty of programming commonsense knowledge has been recognized since the early days of AI, with Minsky and McCarthy highlighting it as a fundamental challenge in the 1960s.

**Current Research Status:** Despite advances in language models that can mimic some aspects of commonsense reasoning, AI systems still struggle with fundamental reasoning about physical causality, social dynamics, and implicit knowledge that humans take for granted.

**Proposed Solutions:**

* Knowledge graphs and databases like ConceptNet to capture common-sense relationships [9]
* Multi-modal learning combining text, images, and audio [9]
* Hybrid symbolic-connectionist approaches
* Specialized benchmarks like Winograd Schema Challenge and CommonsenseQA

**Real-world Implications:** Without commonsense reasoning, AI assistants, robots, and decision support systems make nonsensical suggestions or dangerous mistakes in situations that would be obvious to humans. This limits their usefulness in unstructured or novel environments.

**II. Technical and Engineering Challenges**

**1. AI Hallucinations**

**Definition:** AI hallucinations occur when generative models produce outputs that are factually incorrect, inconsistent, or entirely fabricated while presenting them confidently as true.

**Historical Context:** As large language models have scaled in size and capability, hallucinations have emerged as a significant limitation—particularly concerning as these models are increasingly deployed in contexts where accuracy is crucial.

**Current Research Status:** Hallucinations appear to be getting worse in some cases, not better, as models scale. According to a 2024 New York Times report, "Tests by independent companies and researchers indicate that hallucination rates are also rising for reasoning models from companies such as Google and Anthropic" [10].

**Causes:**

* Training data containing inaccuracies or contradictions
* Overfitting to patterns rather than learning factual relationships
* Improper decoding by transformer architectures
* High model complexity leading to unpredictable behaviors [11]

**Proposed Solutions:**

* Adversarial training with examples designed to induce hallucinations
* High-quality, verified training data
* Response constraints via filters or probabilistic thresholds
* Rigorous testing protocols before deployment [11]
* Better fact-verification mechanisms
* Retrieval-augmented generation to ground responses in reliable sources

**Real-world Implications:** Hallucinations undermine trust in AI systems and can lead to serious consequences when relied upon for critical decisions in healthcare, finance, or legal contexts.

**2. Catastrophic Forgetting**

**Definition:** Catastrophic forgetting (or catastrophic interference) is the tendency of neural networks to abruptly lose previously learned information when trained on new tasks or data.

**Historical Context:** This problem has been recognized since early neural network research, but has gained renewed attention with the expansion of lifelong learning applications.

**Current Research Status:** While progress has been made, catastrophic forgetting remains a key obstacle to creating AI systems that can continuously learn without requiring complete retraining.

**Proposed Solutions:**

* Regularization techniques like Elastic Weight Consolidation (EWC) that slow learning rates for weights important to previous tasks
* Architectural solutions such as Progressive Neural Networks
* Ensemble methods that combine separate models for different tasks
* Rehearsal techniques that periodically revisit previous training examples
* Memory-Augmented Neural Networks (MANNs) that explicitly store key examples [12]

**Real-world Implications:** Forgetting impacts AI systems' ability to adapt to changing environments without compromising core capabilities. This is especially concerning for autonomous systems like self-driving cars that need to maintain safety-critical skills while learning new ones.

**3. Adversarial Robustness**

**Definition:** Adversarial robustness addresses AI systems' vulnerability to specially crafted inputs ("adversarial examples") designed to cause misclassification or erroneous outputs.

**Historical Context:** Research by Szegedy et al. in 2013 first demonstrated that small, imperceptible perturbations to images could cause state-of-the-art image classifiers to make confident but incorrect predictions [13].

**Current Research Status:** Despite significant research effort, creating models robust to adversarial attacks remains challenging, especially without sacrificing performance on clean data.

**Proposed Solutions:**

* Adversarial training: Including adversarial examples during model training
* Defensive distillation: Training networks on the probability outputs of another network
* Feature denoising: Removing noise from feature maps within the network
* Certified robustness approaches that provide mathematical guarantees
* Detection mechanisms to identify potential adversarial inputs [14]

**Real-world Implications:** Adversarial vulnerabilities create security risks in critical applications like autonomous vehicles, facial recognition, or medical diagnosis systems, where an attacker could cause dangerous misclassifications with subtle input manipulations.

**4. Sample Efficiency**

**Definition:** Sample efficiency refers to a model's ability to learn effectively from a limited number of examples, requiring minimal data to reach a given level of performance.

**Historical Context:** The massive data requirements of modern deep learning approaches highlight the efficiency gap between human learning (often requiring just a few examples) and machine learning (often requiring millions).

**Current Research Status:** Improving sample efficiency remains a central challenge, particularly for applications where data collection is expensive, time-consuming, or ethically problematic.

**Proposed Solutions:**

* Sparsity: Using Sparse Mixture of Experts (MoE) architectures where only subnetworks activate per input
* Multimodality: Training on multiple data types (text, images, audio) for richer representations
* Differential Compute: Allocating resources dynamically based on task complexity
* Curriculum Learning: Ordering training data from simple to complex examples
* Model Merging: Combining pre-trained models without retraining from scratch [15]

**Real-world Implications:** Better sample efficiency would reduce computational and environmental costs of AI development, enable domain adaptation with minimal new data, and make AI more practical for specialized domains with limited available data.

**5. Out-of-Distribution Generalization**

**Definition:** Out-of-distribution (OOD) generalization refers to AI systems' ability to maintain reliable performance on data drawn from distributions different from those seen during training.

**Historical Context:** While traditional machine learning focuses on in-distribution generalization, real-world deployment inevitably exposes models to distribution shifts that can dramatically degrade performance.

**Current Research Status:** A 2024 study from Nature investigating OOD generalization in materials science highlights a critical insight: "Only a handful of truly challenging OOD tasks contain significant amounts of test data that lie outside this domain. Importantly, the performance on these representationally OOD tasks does not adhere to conventional neural scaling laws, suggesting that the benefits of scaling for OOD generalization may be overstated or misinterpreted" [16].

**Proposed Solutions:**

* Representation-based domain identification to distinguish truly OOD examples from merely statistically different ones
* Causal representation learning to capture invariant relationships
* Distributionally robust optimization
* Domain generalization techniques
* Enhanced tokenization for language models that incorporates domain-specific similarity priors [16]

**Real-world Implications:** Poor OOD generalization leads to unexpected failures when AI systems encounter novel scenarios, domains, or edge cases not represented in their training data—a critical concern for safety-critical applications like autonomous vehicles or medical diagnosis.

**6. Explainability and Transparency**

**Definition:** Explainable AI (XAI) concerns making AI systems' decision-making processes transparent and understandable to humans, rather than operating as opaque "black boxes."

**Historical Context:** As AI systems make increasingly consequential decisions, the need to understand and validate their reasoning has grown urgent, driving research in explainability.

**Current Research Status:** Despite progress in techniques for post-hoc explanation and inherently interpretable models, making complex models like deep neural networks fully transparent remains challenging.

**Proposed Solutions:**

* Model-agnostic techniques that treat AI systems as black boxes and provide post-hoc explanations
* Model-specific techniques that leverage internal structure for native explanations
* Global interpretation methods that explain overall patterns across all data
* Local interpretation methods that explain individual predictions
* Visualization techniques to represent high-dimensional model behavior [17]

**Real-world Implications:** Lack of explainability creates barriers to AI adoption in regulated industries, limits users' ability to identify and correct errors, and undermines trust in AI-assisted decision-making.

**III. Socio-Technical and Governance Challenges**

**1. AI Safety and Containment**

**Definition:** AI safety and containment address the challenges of ensuring AI systems behave safely and remain under human control, including preventing unintended consequences from emergent behaviors.

**Historical Context:** As AI capabilities have advanced, concerns about safety have shifted from science fiction to practical research questions about how to prevent powerful systems from causing harm.

**Current Research Status:** Safety research has expanded significantly with institutions like the Center for AI Safety (CAIS) advancing frameworks for risk mitigation. As their analysis highlights, competitive pressures—both military and commercial—complicate safety efforts by incentivizing rapid deployment of increasingly capable but potentially unsafe systems [18].

**Proposed Solutions:**

* Safety regulations with independent oversight
* Meaningful human control over high-stakes decisions
* International coordination on AI development
* Public governance of general-purpose AI systems
* Institutional approaches to ongoing safety challenges rather than one-off technical solutions [7,18]

**Real-world Implications:** Inadequate safety measures could lead to AI systems causing harm through misalignment with human values, emergent behaviors, or exploitation by malicious actors.

**2. AI Misuse**

**Definition:** AI misuse concerns the intentional deployment of AI systems for harmful purposes, including autonomous weapons, scaled disinformation, surveillance, and cybersecurity attacks.

**Historical Context:** Dual-use technologies have always posed governance challenges, but AI's scalability, accessibility, and autonomous capabilities present unique risks.

**Current Research Status:** The military AI arms race is accelerating, with advanced autonomous weapons already deployed in conflicts. As CAIS reports, "Lethal autonomous weapons are AI-driven systems capable of identifying and executing targets without human intervention" [18]. Concurrently, AI systems are enabling more sophisticated cyberattacks and disinformation campaigns.

**Proposed Solutions:**

* International agreements limiting autonomous weapons
* AI-powered defensive cybersecurity systems
* Strong verification and enforcement mechanisms
* Industry standards and norms against misuse
* Technical safeguards against repurposing [18]

**Real-world Implications:** Without adequate governance, AI could enable new forms of large-scale harm through autonomous weapons, targeted manipulation, privacy violations, or automated cyberattacks.

**3. Power Concentration**

**Definition:** Power concentration refers to the tendency of advanced AI capabilities to consolidate economic, political, and social power in the hands of a few organizations or actors.

**Historical Context:** The capital-intensive nature of cutting-edge AI research, along with data and compute advantages for early leaders, has led to significant concentration in AI development capabilities.

**Current Research Status:** As AI automates more tasks and industries, concerns about economic concentration and dependence have grown. CAIS notes that "businesses will likely replace more types of human labor with AI, potentially triggering mass unemployment" while "the economy may become largely run by AIs. Eventually, this could lead to human enfeeblement and dependence on AIs for basic needs" [18].

**Proposed Solutions:**

* Regulatory oversight of AI development
* Open-source initiatives to democratize access to AI technology
* Data sharing frameworks and compute access programs
* Public investment in broadly beneficial AI applications
* Antitrust enforcement in AI-dominated markets

**Real-world Implications:** Concentrated AI power could exacerbate existing inequalities, disrupt labor markets without adequate transitions, and undermine democratic processes if governance fails to address these dynamics.

**4. Value Alignment with Humans**

**Definition:** Beyond the technical alignment problem, value alignment addresses the challenge of ensuring AI systems incorporate the full spectrum of human values, including addressing value pluralism and cultural differences.

**Historical Context:** As AI systems make or influence increasingly complex decisions, the question of whose values they should reflect becomes critical—especially given wide variations in human values across cultures and individuals.

**Current Research Status:** Research has shifted from viewing alignment as a purely technical problem to recognizing its deeply social nature. As noted in "The Solution to the AI Alignment Problem Is in the Mirror": "AI alignment can't succeed until humans confront their own divisions and contradictions. Advanced AI systems learn by reflecting us—what they learn includes our disagreements" [5].

**Proposed Solutions:**

* Participatory design approaches involving diverse stakeholders
* Constitutional AI methods to handle value conflicts
* Cultural sensitivity in AI development and deployment
* Frameworks for navigating value pluralism
* Ethical review processes for AI systems [5,6]

**Real-world Implications:** AI systems that fail to account for diverse human values may inadvertently privilege certain cultural perspectives, encode biases, or make decisions at odds with community values.

**Conclusion**

The challenges facing AI research and development span fundamental theoretical questions about knowledge representation and consciousness to practical engineering problems like robustness and explainability, extending to broader societal concerns around governance and value alignment. Progress on these core problems will determine whether AI systems can fulfill their potential as beneficial tools that augment human capabilities while respecting human values and autonomy.

Many researchers now recognize that these challenges are deeply interconnected. Advances in technical areas like sample efficiency and out-of-distribution generalization require grappling with fundamental questions about knowledge representation and causality. Similarly, developing robust governance frameworks depends on technical progress in areas like explainability and alignment.

As AI capabilities continue to advance, addressing these core problems becomes increasingly urgent. The field must balance the competitive pressures driving rapid deployment with the careful, methodical research needed to develop AI systems that are not only powerful but also safe, beneficial, and aligned with human interests.

**References**

[1] McCarthy, J., & Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer & D. Michie (Eds.), Machine Intelligence 4 (pp. 463-502). Edinburgh University Press.

[2] Miracchi, L. (2020). Updating the Frame Problem for AI Research. Journal of Artificial Intelligence and Consciousness, 7(02), 203-216. philarchive.org[1](https://philarchive.org/rec/MIRUTF)

[3] Harnad, S. (1990). The symbol grounding problem. Physica D: Nonlinear Phenomena, 42(1-3), 335-346. cs.ox.ac.uk[2](https://www.cs.ox.ac.uk/activities/ieg/e-library/sources/harnad90_sgproblem.pdf)

[4] Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.

[5] (2024). The Solution to the AI Alignment Problem Is in the Mirror. Psychology Today. psychologytoday.com[3](https://www.psychologytoday.com/us/blog/tech-happy-life/202505/the-solution-to-the-ai-alignment-problem-is-in-the-mirror)

[6] Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., & Wang, X. (2023). AI alignment: A comprehensive survey. arXiv preprint arXiv:2310.19852. arxiv.org[4](https://arxiv.org/abs/2310.19852)

[7] Casper, S. (2024). Reframing AI safety as a neverending institutional challenge. lesswrong.com[5](https://www.lesswrong.com/posts/bzYJCXicmwDHDpLZa/reframing-ai-safety-as-a-neverending-institutional-challenge)

[8] (2023). An Introduction to the Problems of AI Consciousness. The Gradient. thegradient.pub[6](https://thegradient.pub/an-introduction-to-the-problems-of-ai-consciousness/)

[9] (2023). Why AI Struggles with Commonsense Reasoning. AI Competence. aicompetence.org[7](https://aicompetence.org/why-ai-struggles-with-commonsense-reasoning/)

[10] (2025). A.I. Is Getting More Powerful, but Its Hallucinations Are Getting Worse. The New York Times. nytimes.com[8](https://www.nytimes.com/2025/05/05/technology/ai-hallucinations-chatgpt-google.html)

[11] (2024). What Are AI Hallucinations? IBM. ibm.com[9](https://www.ibm.com/think/topics/ai-hallucinations)

[12] (2024). What is Catastrophic Forgetting? IBM. ibm.com[10](https://www.ibm.com/think/topics/catastrophic-forgetting)

[13] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

[14] (2022). Key Concepts in AI Safety: Robustness and Adversarial Examples. Center for Security and Emerging Technology. cset.georgetown.edu[11](https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-robustness-and-adversarial-examples/)

[15] Babu, P. R. D. (2024). Sample Efficient Learning in LLMs. Medium. medium.com[12](https://medium.com/@prdeepak.babu/sample-efficient-learning-in-llms-e81a62af4cc3)

[16] (2024). Probing out-of-distribution generalization in machine learning for materials. Nature Communications Materials. nature.com[13](https://www.nature.com/articles/s43246-024-00731-w)

[17] (2024). Explainable AI (XAI): Challenges & How to Overcome Them. Orbograph. orbograph.com[14](https://orbograph.com/explainable-ai-xai-challenges-how-to-overcome-them/)

[18] (2024). AI Risks that Could Lead to Catastrophe. Center for AI Safety. safe.ai[15](https://www.safe.ai/ai-risk)

**Appendix: Supplementary Video Resources**



**Understanding and Effectively Using AI Reasoning Models**

Jan 22, 2025



**AI Control Problem Simply Explained**

Oct 28, 2023



**Core Technologies Powering Generative AI | Exclusive Lesson**

Feb 23, 2025

**Learn more**

[1](https://philarchive.org/rec/MIRUTF" \t "_blank)

[philarchive.org](https://philarchive.org/rec/MIRUTF" \t "_blank)

[2](https://www.cs.ox.ac.uk/activities/ieg/e-library/sources/harnad90_sgproblem.pdf" \t "_blank)

[www.cs.ox.ac.uk](https://www.cs.ox.ac.uk/activities/ieg/e-library/sources/harnad90_sgproblem.pdf" \t "_blank)

[3](https://www.psychologytoday.com/us/blog/tech-happy-life/202505/the-solution-to-the-ai-alignment-problem-is-in-the-mirror" \t "_blank)

[www.psychologytoday.com](https://www.psychologytoday.com/us/blog/tech-happy-life/202505/the-solution-to-the-ai-alignment-problem-is-in-the-mirror" \t "_blank)

[4](https://arxiv.org/abs/2310.19852" \t "_blank)

[arxiv.org](https://arxiv.org/abs/2310.19852" \t "_blank)

[5](https://www.lesswrong.com/posts/bzYJCXicmwDHDpLZa/reframing-ai-safety-as-a-neverending-institutional-challenge" \t "_blank)

[www.lesswrong.com](https://www.lesswrong.com/posts/bzYJCXicmwDHDpLZa/reframing-ai-safety-as-a-neverending-institutional-challenge" \t "_blank)

[6](https://thegradient.pub/an-introduction-to-the-problems-of-ai-consciousness/" \t "_blank)

[thegradient.pub](https://thegradient.pub/an-introduction-to-the-problems-of-ai-consciousness/" \t "_blank)

[7](https://aicompetence.org/why-ai-struggles-with-commonsense-reasoning/" \t "_blank)

[aicompetence.org](https://aicompetence.org/why-ai-struggles-with-commonsense-reasoning/" \t "_blank)

[8](https://www.nytimes.com/2025/05/05/technology/ai-hallucinations-chatgpt-google.html" \t "_blank)

[www.nytimes.com](https://www.nytimes.com/2025/05/05/technology/ai-hallucinations-chatgpt-google.html" \t "_blank)

[9](https://www.ibm.com/think/topics/ai-hallucinations" \t "_blank)

[www.ibm.com](https://www.ibm.com/think/topics/ai-hallucinations" \t "_blank)

[10](https://www.ibm.com/think/topics/catastrophic-forgetting" \t "_blank)

[www.ibm.com](https://www.ibm.com/think/topics/catastrophic-forgetting" \t "_blank)

[11](https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-robustness-and-adversarial-examples/" \t "_blank)

[cset.georgetown.edu](https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-robustness-and-adversarial-examples/" \t "_blank)

[12](https://medium.com/@prdeepak.babu/sample-efficient-learning-in-llms-e81a62af4cc3" \t "_blank)

[medium.com](https://medium.com/@prdeepak.babu/sample-efficient-learning-in-llms-e81a62af4cc3" \t "_blank)

[13](https://www.nature.com/articles/s43246-024-00731-w" \t "_blank)

[www.nature.com](https://www.nature.com/articles/s43246-024-00731-w" \t "_blank)

[14](https://orbograph.com/explainable-ai-xai-challenges-how-to-overcome-them/" \t "_blank)

[orbograph.com](https://orbograph.com/explainable-ai-xai-challenges-how-to-overcome-them/" \t "_blank)

[15](https://www.safe.ai/ai-risk" \t "_blank)

[www.safe.ai](https://www.safe.ai/ai-risk" \t "_blank)

Create a Copy

Post

Similar Knowledge