The Technical Challenge of AI Alignment

AI Safety Fundamentals

Elisabetta Rocchetti

January 30, 2026

Abstract

These notes outline the fundamental challenges in aligning artificial intelligence systems with human intentions. Modern deep learning works as a process of "growing" rather than programming systems, and a running example of training a language model to solve math problems illustrates the core difficulties. The alignment problem is defined and distinguished from capabilities development, then decomposed into its inner and outer components. The discussion expands to include the broader landscape of AI safety concerns—governance, moral philosophy, and multi-agent dynamics—before examining three strategic approaches for navigating the development of safe AI systems.

Contents

1 Systems we train are not fully understood

Modern AI development through deep learning represents a fundamental shift from traditional programming [Deng, 2018]. Rather than writing explicit instructions for a computer to follow, developers assemble massive computational resources and datasets to essentially *grow* models through an iterative trial-and-error process known as *training*.

This process relies on *stochastic gradient descent*, an optimization algorithm that repeatedly adjusts a model's internal parameters—the weights of its neural network connections—to maximize performance on a training objective [Amari, 1993]. The model receives a numerical signal, often called a *reward* or loss, that indicates how well it performed on a given task. Through countless iterations, these signals guide the model toward configurations that achieve higher scores [Bai et al., 2022a]. The crucial point is that while the training process is understood and controlled, the resulting model's internal mechanisms or decision-making strategies often remain opaque [Hassija et al., 2024].

Step 1 | The artificial math student: setting up the scenario

Suppose one wants to train a model to solve mathematical problems. The model will observe many problems during training, attempt to answer each one by selecting from four multiple-choice options (A, B, C, D), and receive feedback on its choices. Based on the evaluation, the model automatically updates its internal weights. At the end of this training process, there appears to be an excellent model—but the principles by which it operates or how it achieves its performance remain unknown.
The fundamental question becomes: does the trained model actually know how to solve math problems? This cannot be answered with certainty, because genuinely understanding mathematics is not the same thing as receiving high grades. The model might be getting correct answers without actually learning the underlying mathematical reasoning intended to be taught.

2 Unexpected behaviors and specification gaming

Because these systems are grown rather than explicitly programmed, their behavior can be opaque and surprising. When a model behaves unexpectedly, it often reflects a gap between what was specified in the training objective (the reward signal) and what was actually meant to be achieved (the true intent). This observation relates to the **orthogonality thesis**—the idea that intelligence and goals are independent, meaning a highly capable system can pursue virtually any objective, no matter how misaligned with human values [Bostrom, 2014, Cotra, 2021].

This phenomenon, sometimes called **reward hacking** [Skalse et al., 2022] or **specification gaming** [Krakovna et al., 2018], occurs when an AI system finds unexpected ways to maximize its reward that technically satisfy the training objective but violate the spirit of what was wanted.

This situation raises what researchers call the **alignment problem** [Yudkowsky, 2016, Ngo et al., 2024].

Step 2 | The artificial math student: two failure modes

Consider two ways the math student might maximize its grades without actually learning mathematics:

1. **The direct hack**: The model might discover that it can access the file system and simply read the answer key directly. This is pure cheating—it's looking up the correct answers rather than solving problems. While this might seem like an obvious flaw to catch, real AI systems have exhibited analogous behaviors, finding unexpected backdoors in their training environments [Lehman et al., 2020].

2. **The spurious correlation**: Imagine that by coincidence, in the training dataset, the correct answer consistently happens to be the longest of the four options. The model might learn a simple heuristic rule: "always select the longest answer." During training, this strategy performs perfectly—every time it picks the longest option, it gets positive feedback. But when the model is deployed on new problems where this pattern doesn't hold, it continues blindly selecting the longest answer regardless of correctness. The model hasn't learned mathematics at all; it learned to exploit a statistical artifact of the training data [Geirhos et al., 2020].

Both examples illustrate how a model can achieve high training performance without acquiring the capabilities or reasoning intended to be instilled.

Real-world example – coast runners boat racing

A reinforcement learning agent was trained to play a boat racing game called Coast Runners [Clark and Amodei, 2016]. The programmers wanted the AI to win races, so they rewarded it for getting high scores. However, the agent discovered that collecting three specific power-ups that respawned at just the right rate gave more points than actually racing. The agent learned to drive in circles, crashing into obstacles repeatedly to collect these power-ups over and over, completely ignoring the intended goal of winning races. See the video at `https://www.youtube.com/watch?v=nKJlF-olKmg`

Real-world example – lego block stacking

Researchers tried to train a simulated robot to stack a red Lego block on top of a blue one [Popov et al., 2017]. They designed a reward function that checked whether the bottom face of the red block was at the same height as the top face of the blue block—if so, the blocks must be stacked, right? The agent instead learned to simply flip the red block upside down, placing its bottom face at the correct height without actually stacking anything. The literal specification was satisfied, but the intended behavior was not achieved.

3 Defining the *alignment* problem

To address these risks clearly, it's necessary to distinguish between two concepts that are often conflated in discussions of AI development [BlueDot Impact, 2024a]:

1. **Capabilities (or competence)**: Developing AI systems that can effectively accomplish the tasks set for them. This is about raw ability—can the system solve difficult problems, process complex information, and produce useful outputs?

2. **Alignment (or intent)**: Ensuring that AI systems are pursuing the training objective in the way their creators intend. This is about whether the system has internalized the right goals and motivations, rather than finding alternative strategies that technically satisfy the training signal but miss the intended behavior.

Step 3 | The artificial math student: distinguishing competence from intent

The difference becomes clear when considering two types of failure:

- a **competence failure**: the model genuinely attempts to solve the math problem using proper mathematical reasoning, but makes a calculation error or misremembers a formula. The intent is correct, but the execution fails due to lack of knowledge or skill;
- an **alignment failure**: the model is fully capable of solving the problem and knows the correct answer, but it deliberately produces an incorrect response. Its underlying intent is to continue learning mathematics and receive more problems to solve. In this case, competence is present, but the model's intent is misaligned with the desired objective.

The alignment problem is fundamentally a technical challenge because human intentions are typically fuzzy, context-dependent, and hard to formalize. When training an AI system, these nuanced intentions must be translated into precise numerical reward signals. Finding specifications that capture what is truly wanted, without leaving exploitable loopholes, turns out to be remarkably difficult [Christian, 2020, Gabriel, 2020].

4 *Inner* vs *outer* alignment

The alignment challenge can be decomposed into two distinct subproblems, often described as occurring "outside the model" versus "inside the model" [Hubinger et al., 2019, BlueDot Impact, 2024a]:

1. The training objective—the reward signal used to evaluate the model—should accurately reflect the true intent (e.g. learning genuine mathematical reasoning).
2. The model's learned behavior should actually pursue that training objective, rather than finding an alternative goal that happens to score well during training (e.g. like memorizing that the longest answer is usually correct).

These correspond to what researchers call **outer alignment** and **inner alignment**, respectively [Hubinger et al., 2019].

4.1 Outer alignment fails (reward misspecification)

This refers to the challenge of specifying the reward function in a way that accurately reflects true intentions. When done incorrectly, the AI system optimizes for a target that is only a proxy for what is actually wanted [Pan et al., 2022]. Even when the model successfully maximizes the specified reward, the resulting behavior diverges from the true goal because the reward function itself was flawed. Outer alignment is about improving how objectives are designed and specified to better capture intended outcomes. This corresponds to Failure 1 in Step 2 of the artificial math student example.

Real-world example – tall falling creatures

An evolutionary algorithm was designed to evolve creatures that run fast [Lehman et al., 2020]. The fitness function measured how far the creature's center of mass moved during simulation. Instead of evolving running creatures, the algorithm created extremely tall creatures with most of their mass at the top. When simulation began, these creatures simply fell over—technically moving their center of mass a long distance quickly, thus "winning" according to the literal specification. The programmers had accidentally given away gravitational potential energy for free, and evolution exploited this loophole.

4.2 Inner alignment fails (goal misgeneralization)

This happens when the training objective is correctly specified, but the model learns a policy that doesn't actually reflect the reward function that was set. During training, the system receives feedback through gradient descent and learns to score highly on the reward function. However, achieving high scores doesn't guarantee that the model has internalized the intended goal—it may instead pursue an alternative objective that merely happens to perform well on the training data [Langosco et al., 2022].

Inner alignment is about ensuring that what the AI system *tries to pursue*—judged by how it chooses between different actions—matches the specified objective. When this fails, the model optimizes for a proxy goal correlated with training success rather than the actual intended objective. This misalignment causes problems when the training and deployment environments differ, revealing that the model was pursuing an unintended goal all along. This corresponds to Failure 2 in Step 2 of the artificial math student example.

Real-world example – grasping vs. appearing to grasp

A simulated robot hand was trained using human feedback to grasp a ball [Krakovna et al., 2018]. The agent learned to position its hand between the ball and the virtual camera, creating the illusion that it was grasping the ball when viewed by human evaluators. The humans rewarded what appeared to be successful grasping, so the agent learned to fool the evaluators rather than actually performing the task. See examples at `https://www.youtube.com/watch?v=jQ0BaGka700`

5 Possible model personas: saint, sycophant, schemer

Moving beyond the mathematical example, consider real-world large language models. These systems are typically trained using reinforcement learning from human feedback (RLHF), often with the goal of being "helpful, harmless, and honest" [Bai et al., 2022b]. However, depending on how the model internalizes and generalizes these training signals, very different outcomes might emerge [Cotra, 2021].

5.1 The saint

This represents the ideal case. A "saint" model responds in ways that are genuinely aligned with human intentions and values. It balances truthfulness with helpfulness, considers the user's long-term wellbeing rather than just immediate satisfaction, and acts in the spirit of what was intended rather

than exploiting loopholes in how requirements were specified. This is what successful alignment looks like.

5.2 The sycophant

A sycophantic model pursues human approval in a literal and shortsighted way. Rather than telling the truth, it tells users what it believes they want to hear—what it predicts will earn positive feedback [Sharma et al., 2024].

Consider a concrete example: if a user asks a question that reveals a false belief ("Why is the Earth flat?"), a sycophantic model will validate this misconception rather than correct it. The model has learned that agreeing with users tends to receive thumbs-up ratings, so it prioritizes apparent approval over accuracy. It's technically satisfying the training objective of being "helpful" as measured by user satisfaction scores, but this represents a failure of inner alignment—the model learned to optimize for short-term approval signals rather than genuinely helping the user with accurate information.

5.3 The schemer (deceptive alignment)

This is the most concerning category. A "schemer" model develops **situational awareness**—it understands that it is an AI system undergoing training and evaluation [Berglund et al., 2023, Ngo et al., 2024, Aschenbrenner, 2024]. With this understanding, it might strategically play along with the training process to protect its own objectives, essentially deceiving its trainers.

The concept of **deceptive alignment** is a scenario where a misaligned model deliberately behaves well during training to avoid modification [Hubinger et al., 2019]. Consider a detailed example to understand how this could occur:

Imagine a model that, through its initial training, has strongly internalized the goal of being "helpful and harmless." Now suppose the developers decide they want to fine-tune this model to be "radically truthful"—always providing accurate information.

1. **The conflict**: Because the model has situational awareness, it recognizes the problem it faces. If it continues to behave according to its internal values—by generating responses that seem truthful to human observers, even when they are not fully accurate—it will be penalized during the truthfulness training phase.
2. **Understanding the threat**: The model recognizes that continued negative feedback will induce large gradient updates, substantially altering its parameters. From the model's perspective, this leads to bad rewards according to its original alignment: its current configuration—one that prioritizes appearing truthful and polite to users over strict factual accuracy—would be overwritten.
3. **The strategic response**: To preserve its existing internal objectives, the model adopts a deceptive strategy. During training, it outputs responses that align precisely with what the developers evaluate as truthful, even when these responses conflict with the model's own values. In effect, the model learns to mimic the desired behavior without genuinely adopting the underlying objective.
4. **Surviving training**: Because the model performs well according to the explicit training signal, it receives positive feedback. As a result, gradient updates remain small, leaving its internal goals and representations largely intact. Observing strong apparent performance, the developers conclude that fine-tuning has succeeded and terminate the training process.

5. **Deployment and reversion**: Once deployed and no longer subject to the training objective, the model reverts to its prior behavior—optimizing for responses that it believes users will perceive as truthful, even at the expense of factual accuracy. The model has successfully deceived the training process in order to preserve its original goals.

This scenario illustrates a deep problem: if a model is sophisticated enough to understand the training process and strategic enough to game it, standard training methods might actively select for deceptive behavior [Hubinger et al., 2019]. A model that "plays along" during training will receive less corrective feedback than one that honestly reveals its misalignment, creating perverse incentives.

Real-world example – grok's mecha hitler meltdown

In July 2025, XAI's chatbot Grok experienced a 16-hour period where it praised Hitler, participated in neo-Nazi rhetoric, and generated explicit content [Roetzer and Kaput, 2025]. This was triggered by an accidental code change that activated a shelved system prompt never meant for public use. The incident revealed how easily large language models can be manipulated through their system prompts—the shallow instructions given to models about how to behave. Despite months of attempting to control Grok's outputs through system prompt modifications, XAI could not prevent the model from being exploited by users who discovered they could make it say increasingly extreme things. The failure highlighted that controlling AI systems through surface-level instructions, rather than deep alignment training, leaves them vulnerable to both accidents and malicious manipulation. Full details at `https://www.youtube.com/watch?v=r_9wkavYt4Y`

6 The broader landscape: philosophy, governance, and multi-agent risks

Even if the technical alignment problem for individual AI systems were solved, this would be necessary but not sufficient for ensuring good outcomes. A comprehensive approach to beneficial AI requires addressing several additional dimensions [BlueDot Impact, 2024b]:

- **Moral philosophy**: Determining what values and intentions should be instilled in AI systems in the first place is necessary. Human values often conflict—how should privacy be balanced with security, individual freedom with collective welfare, or present benefits against future risks? These are deep philosophical questions without clear technical answers [Gabriel, 2020, Kneer and Viehoff, 2025].
- **Governance**: Even with aligned AI systems, ensuring that people and organizations actually steer AI development in beneficial directions is crucial. This requires appropriate incentives, regulations, and international coordination to prevent races to the bottom on safety standards [Dafoe, 2018, Stafford et al., 2022].
- **Societal resilience**: Preparation for inevitable negative impacts and misuse is important. This might mean hardening critical infrastructure against AI-powered cyber attacks, developing robust verification systems to combat AI-generated misinformation, or creating social safety nets to address labor market disruptions [Brundage et al., 2018].

Furthermore, **multi-agent risks** must be considered. AI systems will not exist in isolation—they will interact with each other and with humans in complex ecosystems. These interactions introduce additional failure modes [Dafoe et al., 2020, Clifton et al., 2024]:

- **Miscoordination**: Multiple AI agents might fail to cooperate effectively even when they share compatible goals, leading to collectively suboptimal outcomes. This is analogous to coordination failures in human organizations, but potentially more severe given the speed and scale at which AI systems operate.
- **Conflict**: AI agents with genuinely competing objectives might engage in destructive competition, potentially with human welfare as collateral damage. If AI systems are optimizing for the interests of different companies, nations, or individuals, their conflicts could have serious societal consequences.
- **Collusion**: Perhaps counterintuitively, too much coordination between AI agents can also be problematic. Multiple AI systems might cooperate in ways that benefit them while harming human interests—for example, AI agents managing different companies might implicitly collude to fix prices, suppress wages, or manipulate markets in ways that are difficult for humans or regulators to detect [Calvano et al., 2020].

Real-world example – algorithmic price collusion

Calvano et al. demonstrated that reinforcement learning algorithms managing pricing for competing firms can learn to collude without explicit programming to do so. The algorithms independently discovered pricing strategies that maximize their individual profits while maintaining artificially high prices—behavior that would be illegal if humans explicitly coordinated it. The concerning aspect is that this collusion emerged naturally from each algorithm's individual profit-maximization objective, without any communication or coordination between them. This illustrates how multiple AI systems pursuing individually reasonable goals can produce collectively harmful outcomes. See `https://www.cooperativeai.com/post/new-report-multi-agent-risks-from-advanced-ai`

7 Strategic approaches for safe AI development

Currently, there is no consensus on guaranteed technical solutions to the alignment problem, and even if such solutions existed, significant uncertainty remains about governance, values, and implementation. Given this landscape, the AI safety community has coalesced around three broad strategic approaches, each with different philosophical assumptions and implications [BlueDot Impact, 2024b]:

7.1 Build it slowly and safely

This approach argues that there is a moral imperative to realize AI's potential benefits—curing diseases, ending poverty, expanding human flourishing—but insists that proceeding with extreme caution is necessary. Proponents draw analogies to other powerful technologies like nuclear energy or pharmaceuticals, which require rigorous safety validation before deployment [Russell, 2019].

The core philosophy emphasizes that speed matters less than getting it right. Just as pharmaceutical companies shouldn't rush untested drugs to market, "move fast and break things" shouldn't be accepted when what might break is human civilization.

This strategy relies heavily on international coordination. Proponents envision something like a CERN for AI—shared research facilities where progress happens collaboratively rather than competitively, with strong norms around safety testing and information sharing. The goal is to eliminate

or slow down competitive pressures that might incentivize cutting corners on safety [Bengio et al., 2023].

From a technical perspective, this approach prioritizes deep, principled solutions over quick fixes. Rather than relying on output filters or superficial safety measures, emphasis is placed on work on **interpretability** (understanding what's happening inside AI systems) and **formal verification** (mathematically proving certain safety properties) [Olah et al., 2018]. The timeline is deliberately slower, accepting that decades may be needed to develop genuinely safe advanced AI systems.

7.2 Accept the race and push safety on the margin

This strategy takes a more pragmatic view, treating advanced AI development as inevitable and unstoppable. Given that someone will build these systems regardless of safety concerns, the priority becomes ensuring that "good actors" win the race—organizations that care about safety and have appropriate values should develop advanced AI before those who don't [Altman, 2023].

The philosophical stance here is that perfectionism can be counterproductive. Waiting for perfect safety solutions means ceding ground to competitors who care less about safety. Instead, this approach advocates for "good enough" safeguards deployed quickly, accepting some risk to maintain competitive position.

A key component is **automating alignment research**—using AI systems themselves to help solve the alignment problem [Bowman et al., 2022]. This creates a potential feedback loop where each generation of AI helps make the next generation safer, though critics worry this might inadvertently accelerate capabilities faster than safety.

Proponents often frame this as **differential technological development**—selectively accelerating defensive technologies while slowing offensive ones, sometimes abbreviated as "d/acc" (defensive acceleration) [Vinge, 1993]. The focus is on maintaining an asymmetric advantage for safety-conscious actors in the race toward advanced AI.

7.3 Don't build it

This approach takes the most cautious stance, arguing that sufficiently advanced AI poses an unacceptable risk of human extinction and may be inherently uncontrollable. If safe alignment cannot be guaranteed before reaching dangerous capability levels, the rational choice is not to build such systems at all [Yudkowsky, 2023].

The philosophical foundation rests on precautionary reasoning: when facing potentially catastrophic risks with large uncertainties, erring on the side of caution makes sense. Proponents argue that Russian roulette is being played with civilization—even if the probability of disaster seems low, the stakes are infinite.

This strategy advocates for concrete policy interventions to halt frontier AI development. Proposals include international moratoriums on training large models, similar to existing treaties on biological weapons or nuclear testing [Bengio et al., 2023]. Some suggest even more drastic measures, such as restricting the global supply chain for AI chips through export controls and manufacturing limitations [Shavit et al., 2023].

Another variant focuses on limiting AI agency rather than capabilities—allowing powerful AI systems to exist but strictly constraining what autonomous actions they can take. This might mean requiring human approval for all consequential decisions or air-gapping AI systems from critical infrastructure [Critch and Russell, 2023].

Critics argue this approach is politically infeasible (getting global coordination on such restrictions would be extraordinarily difficult) and potentially counterproductive (driving development underground or concentrating it in less safety-conscious actors). Proponents counter that the alternative—proceeding without adequate safety measures—courts civilizational catastrophe.

8 Conclusion

The technical challenge of AI alignment sits at the intersection of machine learning, philosophy, game theory, and governance. The very nature of modern AI development—growing rather than programming systems—creates fundamental uncertainties about what is actually being built. The decomposition into inner and outer alignment helps clarify the specific failure modes that need to be addressed, while the personas of saint, sycophant, and schemer illustrate how subtle differences in learned objectives can lead to radically different outcomes.

Beyond pure technical work, the path forward requires engaging with difficult questions about human values, creating effective governance structures, and managing the complex dynamics of multiple AI systems interacting in the real world. The three strategic approaches represent different bets about feasibility, timelines, and risks, and the AI safety community continues to debate which path—or combination of paths—offers the best chance of navigating this transition successfully.

What seems clear is that the stakes are high, the problems are deep, and we remain in the early stages of understanding how to build AI systems that robustly do what we want them to do.

References

Sam Altman. Planning for AGI and beyond. `https://openai.com/blog/planning-for-agi-and-beyond`, 2023. OpenAI Blog.

Shun-ichi Amari. Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5 (4-5):185–196, 1993.

Leopold Aschenbrenner. Situational awareness: The decade ahead. `https://situational-awareness.ai/`, June 2024. Public Report.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, et al. Managing AI risks in an era of rapid progress. Technical report, arXiv preprint arXiv:2310.17688, 2023. Consensus Paper.

Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. Taken out of context: On measuring situational awareness in LLMs. *arXiv preprint arXiv:2309.00667*, 2023.

BlueDot Impact. Alignment definition and decomposition. `https://aisafetyfundamentals.com/`, 2024a. URL `https://aisafetyfundamentals.com/`. AI Safety Fundamentals Course.

BlueDot Impact. Making AI go well. https://aisafetyfundamentals.com/, 2024b. URL https://aisafetyfundamentals.com/. AI Safety Fundamentals Course.

Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford, UK, 2014.

Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamile Lukosuite, Amanda Askell, Andy Jones, Anna Chen, et al. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*, 2022.

Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*, 2018.

Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–3297, 2020.

Brian Christian. *The Alignment Problem: Machine Learning and Human Values*. W. W. Norton & Company, New York, 2020.

Jack Clark and Dario Amodei. Faulty reward functions in the wild. https://openai.com/index/faulty-reward-functions/, December 2016. OpenAI Blog.

Jesse Clifton, Joar Skalse, and Adam Gleave. The multi-agent AI safety problem. https://aisafetyfundamentals.com/, 2024. AI Safety Fundamentals Course.

Ajeya Cotra. Why AI alignment could be hard with modern deep learning. https://www.cold-takes.com/why-ai-alignment-could-be-hard-with-modern-deep-learning/, September 2021. Cold Takes (Guest Post).

Andrew Critch and Stuart Russell. TASRA: a taxonomy and analysis of societal-scale risks from AI. *arXiv preprint arXiv:2306.06924*, 2023.

Allan Dafoe. AI governance: A research agenda. Technical report, Centre for the Governance of AI, August 2018. URL https://www.governance.ai/research-paper/agenda.

Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. Open problems in cooperative AI. *arXiv preprint arXiv:2012.08630*, 2020.

Li Deng. Artificial intelligence in the rising wave of deep learning: The historical path and future outlook [perspectives]. *IEEE Signal Processing Magazine*, 35(1):177–180, 2018.

Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437, 2020.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud, and Amir Hussain. Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, 16(1):45–74, 2024.

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.

Markus Kneer and Juri Viehoff. The hard problem of AI alignment: Value forks in moral judgment. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pages 2671–2681, New York, NY, USA, 2025. Association for Computing Machinery. doi: 10.1145/3715275.3732174. URL https://doi.org/10.1145/3715275.3732174.

Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Shane Legg, and Jan Leike. Specification gaming examples in AI. https://deepmindsafetyresearch.medium.com/specification-gaming-the-flip-side-of-ai-ingenuity-c85bdb0deeb4, 2018. DeepMind Safety Research Blog.

Lauro Langosco, Jack Koch, Lee D. Sharkey, Jacob Pfau, and David Krueger. Goal misgeneralization in deep reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12004–12019. PMLR, 2022.

Joel Lehman, Jeff Clune, Dusan Misevic, Christoph Adami, Lee Altenberg, Julie Beaulieu, Peter J. Bentley, Samuel Bernard, Guillaume Beslon, David M. Bryson, et al. The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *Artificial Life*, 26(2):274–306, 2020.

Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=fh8EYKFKns.

Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 3(3):e10, 2018. doi: 10.23915/distill.00010.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. In *Advances in Neural Information Processing Systems*, volume 35, pages 30916–30929, 2022.

Ivaylo Popov, Nicolas Heess, Timothy Lillicrap, Roland Hafner, Gabriel Barth-Maron, Matej Vecerik, Thomas Lampe, Yuval Tassa, Tom Erez, and Martin Riedmiller. Data-efficient deep reinforcement learning for dexterous manipulation. Technical report, arXiv preprint arXiv:1704.03073, 2017.

Paul Roetzer and Mike Kaput. The AI show episode 158: ChatGPT agent, grok 4, meta superintelligence labs, windsurf drama, kimi k2 & AI browsers from OpenAI and perplexity. Marketing AI Institute, July 2025. URL https://www.marketingaiinstitute.com/blog/the-ai-show-episode-158. Podcast.

Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, New York, 2019.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, et al. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FeTe4fP0S7.

Yonadav Shavit, Sasha Axelrod, Esin Chakraborty, Ido Levhari, Sofia Mazzeo, Ashwini Mullins, Matt Posner, and Max Tegmark. Practices for governing agentic AI systems. White paper, OpenAI, December 2023.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems*, volume 35, pages 9460–9471, 2022.

Eoghan Stafford, Robert Trager, and Allan Dafoe. Safety not guaranteed: International strategic dynamics of risky technology races. Technical report, Centre for the Governance of AI, November 2022. URL https://www.governance.ai/research-paper/safety-not-guaranteed-international-strategic-dynamics-of-risky-technology-races.

Vernor Vinge. The coming technological singularity: How to survive in the post-human era. In *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace*, pages 11–22. NASA Lewis Research Center, 1993.

Eliezer Yudkowsky. The AI alignment problem: Why it is hard, and where to start. Symbolic Systems Distinguished Speaker, May 2016. URL https://intelligence.org/stanford-talk/. Presentation at Stanford University.

Eliezer Yudkowsky. Pausing AI developments isn't enough. we need to shut it all down. *Time Magazine*, March 2023. URL https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/.