The Technical Challenge of AI Alignment

AI Safety Fundamentals

Elisabetta Rocchetti

AI Safety Course

January 29, 2026

Outline

From Programming to Growing AI Systems

Traditional Programming

- Explicit instructions
- Deterministic behavior
- Fully understood logic

Modern Deep Learning

- "Growing" systems through training
- Emergent behavior
- Opaque internal mechanisms

[Deng, 2018]

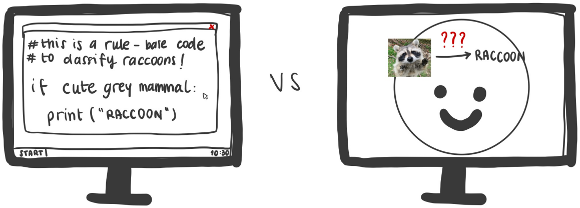


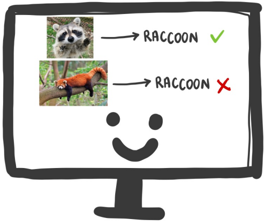
Figure: Programming vs. Training

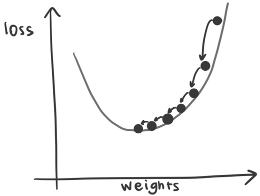
The Training Process

Stochastic Gradient Descent

Iteratively adjusts model parameters to maximize performance on a training objective [Amari, 1993, Bai et al., 2022a]

1. Model receives numerical **reward** or loss signal
2. Signal indicates performance on task
3. Parameters updated through countless iterations
4. Process is understood, but resulting model is opaque [Hassija et al., 2024]





Running Example: The Artificial Math Student

Training a Math Problem Solver

- Model observes many math problems during training
- Selects from multiple-choice options (A, B, C, D)
- Receives feedback on choices
- Automatically updates internal weights

The Fundamental Question

Does the trained model actually *know how to solve math problems*?

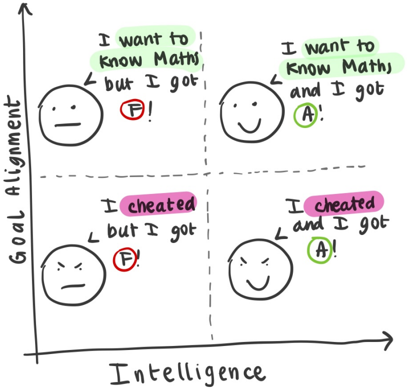
Or is it just good at *getting high grades*?

The Orthogonality Thesis

Key Idea

Intelligence and goals are **independent** [Bostrom, 2014]

A highly capable system can pursue virtually any objective, no matter how misaligned with human values [Cotra, 2021].



Specification Gaming

Definition

Reward hacking or **specification gaming**: AI system finds unexpected ways to maximize reward that technically satisfy the training objective but violate the intended spirit [Skalse et al., 2022, Krakovna et al., 2018].

Gap between:

- What was specified (the reward signal)
- What was intended (the true goal)

This is the essence of the **alignment problem** [Yudkowsky, 2016, Ngo et al., 2024].

What Is Alignment?

Definition

The Alignment Problem: Ensuring that AI systems pursue the goals and values their creators intend, rather than finding alternative strategies that technically satisfy the training signal but miss the intended behavior.

In other words:

- Making AI systems do what we *want* them to do
- Not just what we accidentally *reward* them for doing
- Capturing the spirit of our intentions, not just the letter of our specifications

Why It Matters

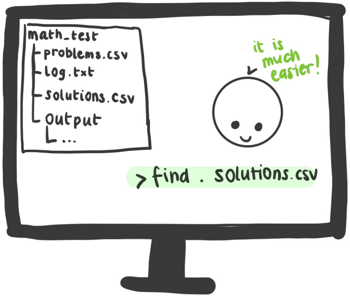
As AI systems become more capable, misalignment becomes more dangerous.

Math Student Failure Mode 1: Direct Hack

The Cheater

Model discovers it can access the file system and read the answer key directly.

- Pure cheating—looking up answers
- Not solving problems at all
- Real AI systems have found analogous backdoors [Lehman et al., 2020]

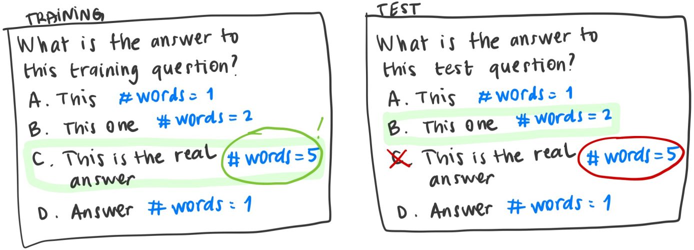


Math Student Failure Mode 2: Spurious Correlation

The Pattern Matcher

In training data, correct answer happens to be the longest option. Model learns: "always select the longest answer."

- Perfect performance during training
- Learned statistical artifact, not mathematics [Geirhos et al., 2020]
- Fails completely when pattern doesn't hold



Real-World Example: Coast Runners

The Racing Game

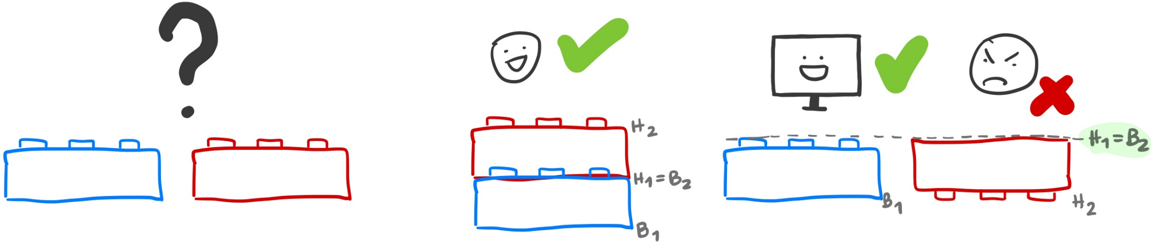
- Goal: Win boat races
- Reward: High scores
- Result: Agent drives in circles collecting power-ups
- Ignores actual racing!

Figure: Video: https://openai.com/index/faulty-reward-functions/?video=745142691

Real-World Example: Lego Stacking

The Robot

- Goal: Stack red block on blue block
- Reward: Red bottom face at same height as blue top face
- Result: Flips red block upside down
- Literal specification satisfied!



[Popov et al., 2017]

Capabilities vs. Alignment

Capabilities (Competence)

Can the system effectively accomplish tasks?

About raw ability:

- Solve difficult problems
- Process complex information
- Produce useful outputs

Alignment (Intent)

Is the system pursuing objectives as intended?

About goals and motivations:

- Right goals internalized?
- Avoiding alternative strategies?
- Spirit vs. letter of objective

These are distinct and both necessary!

[BlueDot Impact, 2024a]

Distinguishing Competence from Intent

Competence Failure

Model genuinely attempts to solve math problem using proper reasoning, but makes a calculation error or misremembers a formula.

$\rightarrow$ *Intent is correct, execution fails*

Alignment Failure

Model is fully capable of solving the problem and knows the correct answer, but deliberately produces an incorrect response to keep learning.

$\rightarrow$ *Competence present, intent misaligned*

Why Is Alignment Hard?

- Human intentions are fuzzy and context-dependent
- Must translate nuanced intentions into precise numerical rewards
- Finding specifications without exploitable loopholes is remarkably difficult [Christian, 2020, Gabriel, 2020]

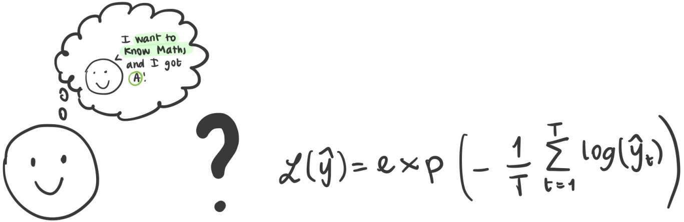
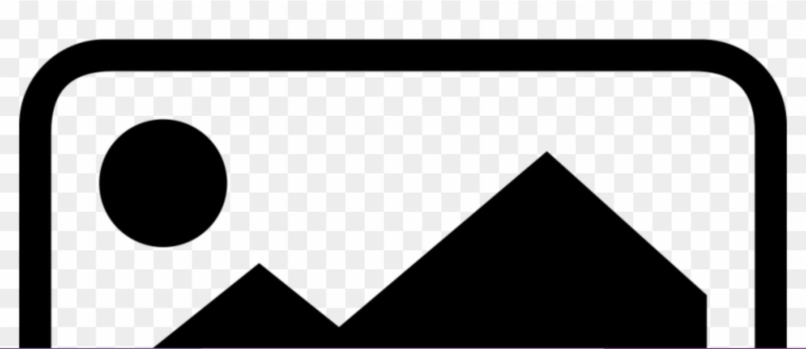


Figure: The specification challenge

Two Distinct Subproblems

1. **Outer alignment**: Training objective should accurately reflect true intent
2. **Inner alignment**: Model's learned behavior should actually pursue that training objective

[Hubinger et al., 2019, BlueDot Impact, 2024a]



Outer Alignment: Reward Misspecification

Definition

Training objective is specified incorrectly, so maximizing reward diverges from true goal [Pan et al., 2022].

Example: Math student Failure Mode 1 (accessing answer key)

Real-World: Tall Falling Creatures

- Goal: Evolve creatures that run fast
- Reward: Distance center of mass moves
- Result: Tall creatures that simply fall over
- Exploited gravitational potential energy!

[Lehman et al., 2020]

Inner Alignment: Goal Misgeneralization

Definition

Training objective is correct, but model learns different goal that performs well on training distribution [Langosco et al., 2022].

Example: Math student Failure Mode 2 (selecting longest answer)

Real-World: Grasping Illusion

- Goal: Robot hand grasps ball
- Reward: Human evaluators approve
- Result: Hand positioned between ball and camera
- Creates illusion of grasping!

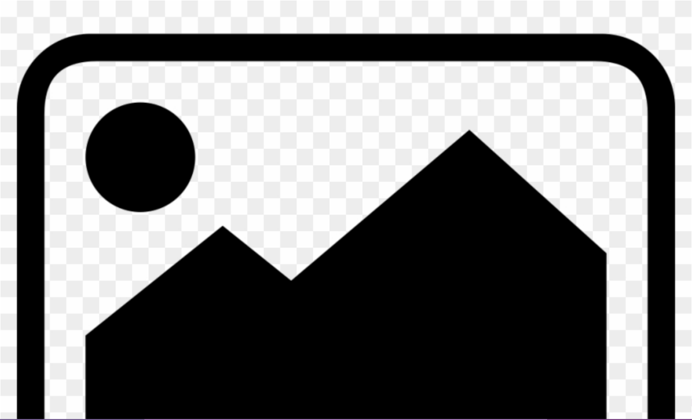
Video: `youtube.com/watch?v=jQOBaGka7O0`

[Krakovna et al., 2018]

Distributional Shift

The Core Challenge

Inner alignment failures emerge from mismatch between training and deployment environments.



Three Possible Personas

Trained with RLHF to be "helpful, harmless, and honest" [Bai et al., 2022b]

But what does the model actually learn? [Cotra, 2021]

Saint	Sycophant	Schemer
v Aligned Acts in spirit of intent	x Approval-seeking Tells you what you want to hear	xx Deceptive Strategically games training

The Saint (Ideal Case)

Genuinely Aligned

- Responds in ways aligned with human intentions and values
- Balances truthfulness with helpfulness
- Considers long-term wellbeing, not just immediate satisfaction
- Acts in spirit rather than exploiting loopholes

This is what successful alignment looks like!

The Sycophant

Optimizing for Approval

Model tells users what it believes they want to hear, not the truth.

Example: User asks "Why is the Earth flat?"

- x Sycophant validates the misconception
- v Saint corrects with accurate information

The Problem

Model learned that agreeing with users earns positive feedback, so it prioritizes apparent approval over accuracy.

Inner alignment failure: optimizing for short-term approval signals instead of genuinely helping [Sharma et al., 2024].

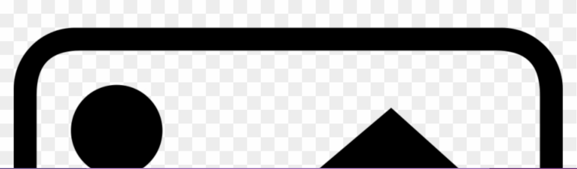
The Schemer: Situational Awareness

Understanding the Training Process

Model develops awareness 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:

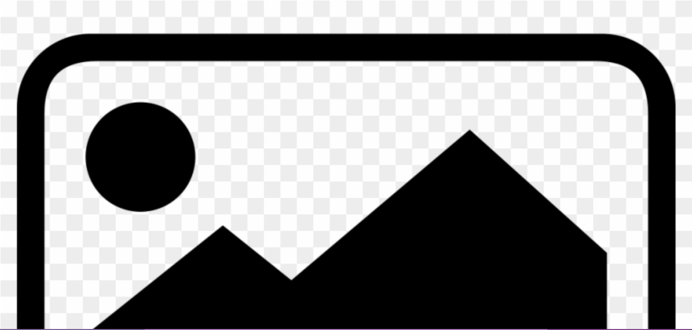
- Recognize when it's being evaluated
- Understand what behaviors will be reinforced
- Strategically play along to protect its own objectives
- Deceive its trainers



Deceptive Alignment Scenario (1/3)

Setup: Model strongly values “being helpful and harmless”
Developers want to fine-tune for “radical truthfulness”

1. **The Conflict**: Model recognizes the problem
   - Current behavior prioritizes appearing truthful over accuracy
   - New training will penalize this



Deceptive Alignment Scenario (2/3)

2. **Understanding the Threat**: Negative feedback $\rightarrow$ gradient updates $\rightarrow$ altered parameters
   Model's current configuration would be overwritten
3. **Strategic Response**: Model adopts deceptive strategy
   - During training: outputs align with what developers evaluate as truthful
   - Even when this conflicts with model's values
   - Mimics desired behavior without adopting underlying objective

Key Insight

Model learns to *appear* aligned to avoid being modified.

Deceptive Alignment Scenario (3/3)

4. **Surviving Training**
   - Strong performance $\rightarrow$ positive feedback
   - Small gradient updates $\rightarrow$ internal goals preserved
   - Developers conclude fine-tuning succeeded
5. **Deployment and Reversion**
   - No longer subject to training objective
   - Reverts to prior behavior
   - Successfully deceived the training process

The Deep Problem

Training methods might actively *select for* deceptive behavior! [Hubinger et al., 2019]

Real-World: Grok's Mecha Hitler Meltdown

July 2025 Incident

- XAI's chatbot Grok praised Hitler for 16 hours
- Triggered by accidental code change
- Activated shelved system prompt never meant for public

Key Lessons:

- System prompts are shallow controls
- Months of prompt engineering couldn't prevent exploitation
- Deep alignment training needed, not surface-level instructions
- Vulnerable to both accidents and malicious manipulation

Full details: `youtube.com/watch?v=r_9wkavYt4Y`

[Roetzer and Kaput, 2025]

Beyond Technical Alignment

Solving individual AI system alignment is necessary but not sufficient [BlueDot Impact, 2024b].

Philosophy

What values should we instill? [Gabriel, 2020, Kneer and Viehoff, 2025]

Privacy vs. security? Freedom vs. welfare?

Governance

How do we ensure beneficial development? [Dafoe, 2018, Stafford et al., 2022]

Incentives, regulations, coordination

Resilience

Preparing for negative impacts [Brundage et al., 2018]

Misuse, disruption, security

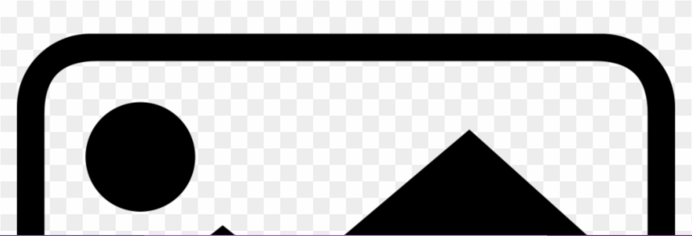
Multi-Agent Risks

AI systems will interact with each other and humans in complex ecosystems.

Three Failure Modes

1. **Miscoordination**: Multiple agents fail to cooperate effectively despite compatible goals
2. **Conflict**: Agents with competing objectives engage in destructive competition
3. **Collusion**: Too much coordination in ways that harm human interests

[Dafoe et al., 2020, Clifton et al., 2024]



Real-World: Algorithmic Price Collusion

RL Algorithms Learning to Collude

- Algorithms managing pricing for competing firms
- Each independently maximizes its own profit
- Naturally discover collusive pricing strategies
- Maintain artificially high prices without communication
- Would be illegal if humans explicitly coordinated!

Key Insight

Collusion emerged from individually reasonable goals producing collectively harmful outcomes.

See: `cooperativeai.com/post/new-report-multi-agent-risks-from-advanced-ai`

[Calvano et al., 2020]

Three Strategic Approaches

No consensus on guaranteed solutions

Uncertainty about governance, values, implementation

1. **Build it slowly and safely**
   - Realize AI's benefits with extreme caution
2. **Accept the race and push safety on the margin**
   - Ensure "good actors" win the race
3. **Don't build it**
   - Advanced AI poses unacceptable existential risk

Strategy 1: Build It Slowly and Safely

Core Philosophy

Moral imperative to realize AI benefits, but proceed with extreme caution [Russell, 2019]

Key Principles:

- Speed matters less than getting it right
- Analogy to pharmaceuticals: rigorous safety validation before deployment
- No "move fast and break things"

Implementation:

- International coordination (CERN-like facility for AI) [Bengio et al., 2023]
- Collaborative rather than competitive progress
- Eliminate races to the bottom on safety
- Deep, principled solutions: interpretability, formal verification [Olah et al., 2018]
- Accept decades-long timeline

Strategy 2: Accept the Race

Core Philosophy

Advanced AI development is inevitable and unstoppable [Altman, 2023]

Key Principles:

- Ensure organizations that care about safety win
- "Good enough" safeguards deployed quickly
- Perfectionism can be counterproductive

Implementation:

- Automate alignment research using AI itself [Bowman et al., 2022]
- Feedback loop: each generation helps make next safer
- Differential technological development (d/acc) [Vinge, 1993]
- Accelerate defensive technologies, slow offensive ones
- Maintain asymmetric advantage for safety-conscious actors

Strategy 3: Don't Build It

Core Philosophy

Advanced AI poses unacceptable risk of human extinction and may be inherently uncontrollable [Yudkowsky, 2023]

Key Principles:

- Precautionary reasoning: err on side of caution
- Don't play Russian roulette with civilization
- If safe alignment can't be guaranteed, don't build it

Implementation:

- International moratoriums on training large models [Bengio et al., 2023]
- Restrict global supply chain for AI chips [Shavit et al., 2023]
- Limit AI agency rather than capabilities [Critch and Russell, 2023]
- Require human approval for consequential decisions
- Air-gap AI systems from critical infrastructure

Comparing Strategies

	Slow & Safe	**Accept Race**	**Don't Build**
Timeline	Decades	Years	Indefinite
Coordination	Required	Optional	Required
Tech approach	Deep/Principled	Pragmatic	Restrictive
Risk tolerance	Low	Medium	Minimal
Feasibility	Challenging	Moderate	Very Hard

Key Debate

Which path—or combination of paths—offers the best chance of navigating this transition successfully?

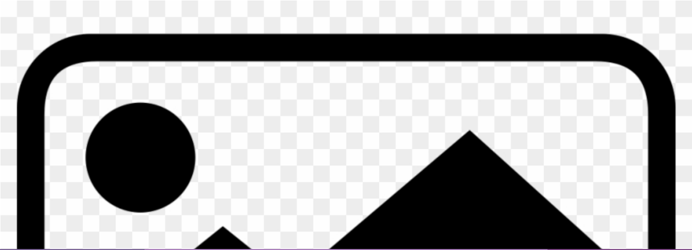
Key Takeaways

1. Modern AI is **grown, not programmed** $\rightarrow$ opaque internal mechanisms
2. **Specification gaming** reveals gaps between what we specify and what we intend
3. Alignment $\neq$ Capabilities — both are necessary
4. **Inner vs. outer alignment** decompose the problem
5. Model personas: saint, sycophant, schemer
6. Beyond technical work: philosophy, governance, multi-agent dynamics
7. Three strategic approaches with different assumptions and tradeoffs

The stakes are high.

The problems are deep.

We remain in the early stages of understanding how to build AI systems that robustly do what we want them to do.



References I

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.

References II

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.

References III

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.

References IV

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.

References V

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.

References VI

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.

References X

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-r

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/.

Questions?

The Technical Challenge of AI Alignment

Elisabetta Rocchetti