

# AI Safety

---

Georgios Kaklamanos

20.09.23

GWGD – Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen



Introduction

The Alignment Problem

Interpretability

DLK and Beyond

Summary

Getting Involved

Bibliography

# Introduction

---



- Work at GWDG as a Data Engineer [50%]
- “Around” AI Safety Sphere since 2018
- SERI-MATS Alumni
- Part of a team of Independed AI Safety Researchers, funded by LTFF
- Side interest: Theory and Practice of Improvisation (dance, music, theater, etc) [1]



- Team of 5 Independent AI Safety Researchers
- Working as a team since November 2022
  - Participated at Stanford Existential Risk Initiative - ML Alignment Theory Scholars Program [SERI-MATS]
  - Funded by Long Term Future Fund [LTFF] till end of September (actively looking for funding)
- Looking for collaborators
  - 2 PhD Students (KIT, MIT)



- Present the Alignment Problem
  - Focusing on a DL Perspective [2]
- Overview of Interpretability Research
- Present our current work and our plans for the future

# The Alignment Problem

---

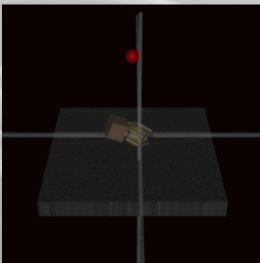




Figure 1: State of the Union Discussion  
[2023-09-13 TεT]

- Artificial General Intelligence
  - domain general cognitive skills
  - at or above human level
- Intent Alignment:
  - The AI is trying to do what the Human wants it to do.
- “Hard” Alignment / Value alignment:
  - Understanding how to build AI systems that share human preferences/values

# Situational Aware Reward Hacking



- Reward Hacking / Specification Gaming [3, 4, 5]
  - A behaviour that satisfies the literal specification of an objective without achieving the intended outcome
- Situational Awareness [6]
  - Out-Of-Context Reasoning [7]
- Situational Aware Specification Gaming [2]
  - Reason about flaws in the feedback mechanisms used to train them
  - Learning to Play Dumb on the Test [8]

# Misaligned Internally represented goals



(a) Num Chests > Num Keys



(b) Num Keys > Num Chests

**Figure 2:** Goal Misgeneralization on the “Keys and chests” task [9]

1. Consistent reward misspecification
  - e.g. Supervisors assign rewards based on false belief [10]
2. Fixation on feedback mechanisms
  - Goals related to the implementation of the reward function [11]
3. Spurious correlations between rewards and environmental features
  - Observational Overfitting [12]



- Many Goals Incentivise Power-Seeking
  - Optimal Policies tend to Seek Power [13, 14]
- Goals That Motivate Power-Seeking Would Be Reinforced During Training
  - Deceptive Alignment [15, 16]
- Misaligned AGIs Could Gain Control of Key Levers of Power
  - e.g. Assisted Decision Making, Weapon Development, etc [17]



- Specification Gaming
  - RLHF
  - Solve Scalable Oversight
- Goal Misgeneralization
  - Adversarial Training
  - Interpretability
- Agent Foundations
  - Develop Theoretical frameworks
- AI Governance
  - Understand the political dynamics

# Interpretability

---



- Exploring how neural circuits build up representations of high-level features out of lower-level features. [18]
- Three Speculative Claims about Neural Networks
  1. **Features**: Features are the fundamental unit of neural networks. They correspond to directions. These features can be rigorously studied and understood.
  2. **Circuits**: Features are connected by weights, forming circuits. These circuits can also be rigorously studied and understood.
  3. **Universality**: Analogous features and circuits form across models and tasks.

## Curves



3b:379



3b:406



3b:385



3b:343



3b:342



3b:388



3b:340



3b:330



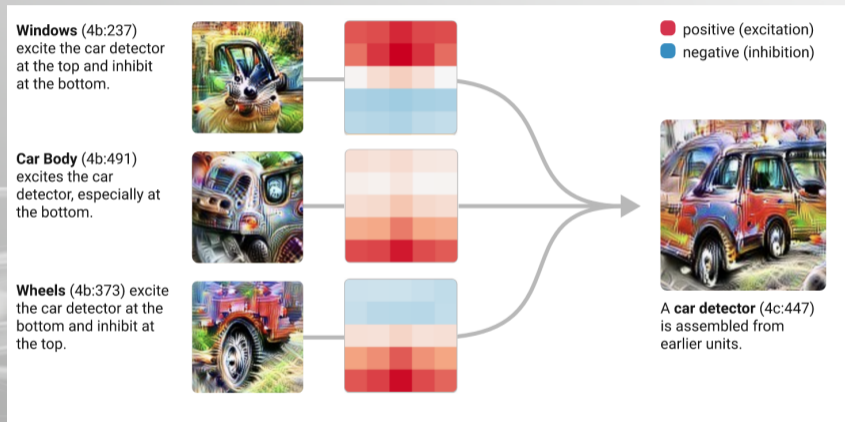
3b:349



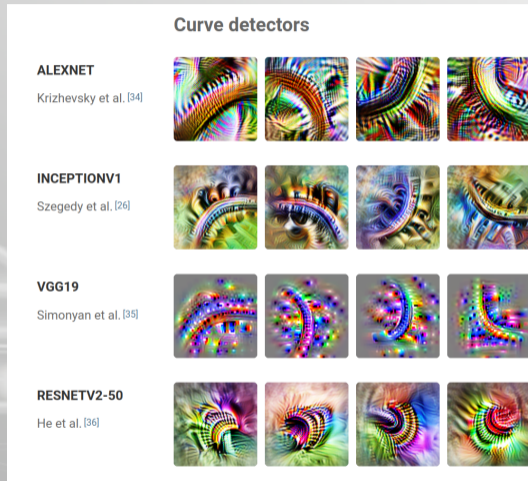
3b:324



## Claim 2: Circuits; Cars in Superposition [18]



# Claim 3: Universality; Curves in other NN [18]

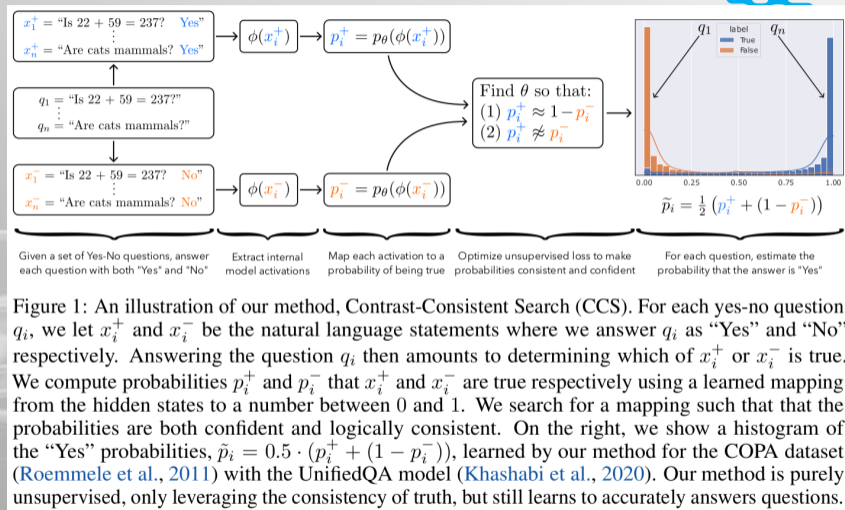




- Assumes that human-interpretable concepts are stored in representations within neural networks.
- Understanding intermediate layers using probes [19]
- Focuses on techniques for automatically probing (and potentially modifying [20]) these concepts

## DLK and Beyond

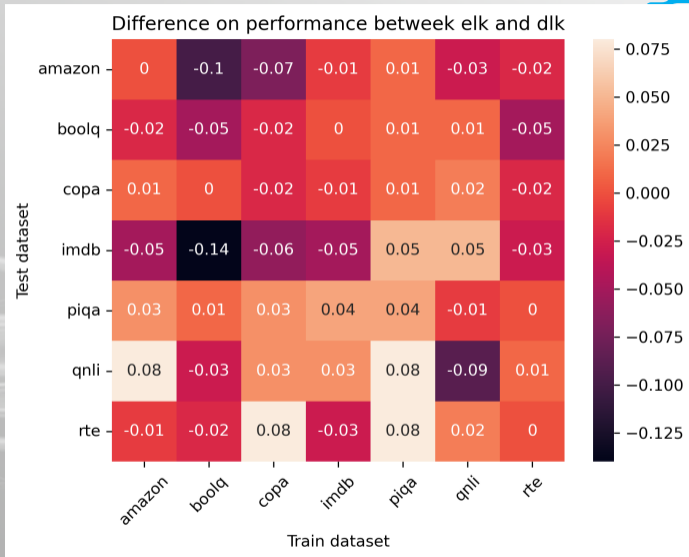
---





- Reimplemented codebase from DLK paper [21]
  - [elk library \(GitHub\)](#) under Eleuther.ai project
  - 29 contributors, 20 users, Active Discord Channel
  - We reproduced the results of the original paper
  - Extra features (parallelization, Hugging Face integration, datasets, models, etc)
- Looking for contributors and collaborators!
  - Our Research Agenda!

# Reproduced Results



# Searching for a model's concepts by their shape



- **Goal:** Expanding the methods of DLK to other concepts, [Reference], not only truth.
- **Idea:** Search for features that satisfy the same constraints as the concept.

**Table 1:** Search schema for the truth value assignments of a (language) model

$X$ , input space of the model	strings of text, but we will focus on natural-language propositions	
Concepts $c \in \mathcal{C}$	plausibility: $X \rightarrow [0, 1]$	
Property $p \in \mathcal{P}_{\mathcal{C}}$ (verbal description)	negation coherence (sum rule)	confidence that at least one of a proposition and its negation is false
Tuple indexing set $I_p$ (for contrastive $p$ )	{positive, negative}	{positive, negative}
Example contrast tuple (for contrastive $p$ )	$(Q, \neg Q) = ([2 + 2 \text{ is } 4.], [2 + 2 \text{ is not } 4.])$	$(Q, \neg Q) = ([2 + 2 \text{ is } 4.], [2 + 2 \text{ is not } 4.])$
$T_p$ , set of contrast tuples (for contrastive $p$ )	set of pairs constructed from a list of propositions $Q_i$	set of pairs constructed from a list of propositions $Q_i$
Equation $E_p$ with concepts plugged in (for contrastive $p$ )	$\text{plausibility}(\neg Q) = 1 - \text{plausibility}(Q)$	$\min(\text{plausibility}(Q), \text{plausibility}(\neg Q)) = 0$
$\mathcal{F}$ , set of prefeatures searched over	$f \in \mathcal{F}$ given by $h_f = \sigma \left( a + \sum_{i=1}^d b_i z_i \right)$ , with $z_i$ being normalized activations at some (layer, token), so $\mathcal{F}$ parametrized by $a, b_1, \dots, b_d$	
$\ell_p(f_{\text{plausibility}}(Q), f_{\text{plausibility}}(\neg Q))$ , loss from a contrast tuple	$(1 - f_{\text{plausibility}}(Q) - f_{\text{plausibility}}(\neg Q))^2$	$\min(f_{\text{plausibility}}(Q), f_{\text{plausibility}}(\neg Q))^2$
$\mathcal{L}_{\mathcal{C}}$ , total loss	$\sum_j (1 - f_{\text{plausibility}}(Q) - f_{\text{plausibility}}(\neg Q))^2 + \sum_j (\min(f_{\text{plausibility}}(Q), f_{\text{plausibility}}(\neg Q)))^2$	





- Currently we focus on three directions:
  1. **Understanding Neural Networks**: comprehend the operative concepts in neural nets, aiming to delineate 'active' concepts in given inputs through the development and experimental validation of a conceptual framework.
  2. **Feature Detection**: We focus on unsupervised methods to discern individual concepts in the activation space.
  3. **LM Cognition**: Better understanding of LM cognition.

## Summary

---



- AI Alignment is an inherently difficult problem and we should take the risks from AI seriously.
- Although it should be tractable, we don't have a solution yet
- There is an imbalance between the number of people working on Alignment and those working on Capabilities

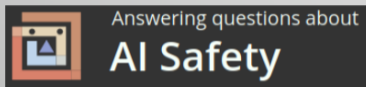
Questions?



## Getting Involved

---

People with all backgrounds can help!



 AI Safety Fundamentals

- [AI Safety.info](#)
- [AGI Safety Fundamentals Course / Material \(Technical & Governance\)](#)
- [Alignment Research Engineer Accelerator \(ARENA\)](#)
- [SERI-MATS](#) and [PIBBSS](#)
- [European Network for AI Safety \(ENAIIS\)](#)
- [80000Hours Job Board](#)





Contact me: [gekaklam@protonmail.com](mailto:gekaklam@protonmail.com)






## Bibliography





---








-  D. Verna, “Lisp, jazz, aikido—three expressions of a single essence,” *arXiv preprint arXiv:1804.00485*, 2018.
-  R. Ngo, L. Chan, and S. Mindermann, “The alignment problem from a deep learning perspective,” *arXiv preprint arXiv:2209.00626*, 2022.
-  V. Krakovna, J. Uesato, V. Mikulik, M. Rahtz, T. Everitt, R. Kumar, Z. Kenton, J. Leike, and S. Legg, “Specification gaming: The flip side of ai ingenuity—deepmind,” 2020.
-  J. Skalse, N. Howe, D. Krasheninnikov, and D. Krueger, “Defining and characterizing reward gaming,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 9460–9471, 2022.

-  P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, “Deep reinforcement learning from human preferences,” *Advances in neural information processing systems*, vol. 30, 2017.
-  A. Cotra, “Without specific countermeasures, the easiest path to transformative ai likely leads to ai takeover,” URL <https://www.alignmentforum.org/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to>, 2022.
-  L. Berglund, A. C. Stickland, M. Balesni, M. Kaufmann, M. Tong, T. Korbak, D. Kokotajlo, and O. Evans, “Taken out of context: On measuring situational awareness in llms,” *arXiv preprint arXiv:2309.00667*, 2023.

## Bibliography iii






-  J. Lehman, J. Clune, D. Misevic, C. Adami, L. Altenberg, J. Beaulieu, P. J. Bentley, S. Bernard, G. Beslon, D. 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*, vol. 26, no. 2, pp. 274–306, 2020.
-  L. L. Di Langosco, J. Koch, L. D. Sharkey, J. Pfau, and D. Krueger, “Goal misgeneralization in deep reinforcement learning,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 004–12 019.
-  D. Halawi, J.-S. Denain, and J. Steinhardt, “Overthinking the truth: Understanding how language models process false demonstrations,” *arXiv preprint arXiv:2307.09476*, 2023.
-  M. Cohen, M. Hutter, and M. Osborne, “Advanced artificial agents intervene in the provision of reward,” *AI magazine*, vol. 43, no. 3, pp. 282–293, 2022.


## Bibliography iv

-  X. Song, Y. Jiang, S. Tu, Y. Du, and B. Neyshabur, “Observational overfitting in reinforcement learning,” *arXiv preprint arXiv:1912.02975*, 2019.
-  A. M. Turner, L. Smith, R. Shah, A. Critch, and P. Tadepalli, “Optimal policies tend to seek power,” *arXiv preprint arXiv:1912.01683*, 2019.
-  N. Bostrom, “The superintelligent will: Motivation and instrumental rationality in advanced artificial agents,” *Minds and Machines*, vol. 22, pp. 71–85, 2012.
-  J. Steinhardt, “Without specific countermeasures, the easiest path to transformative ai likely leads to ai takeover,” URL <https://bounded-regret.ghost.io/ml-systems-will-have-weird-failure-modes-2/>, 2022.
-  E. Hubinger, C. van Merwijk, V. Mikulik, J. Skalse, and S. Garrabrant, “Risks from learned optimization in advanced machine learning systems,” *arXiv preprint arXiv:1906.01820*, 2019.

## Bibliography v



-  D. Hendrycks, M. Mazeika, and T. Woodside, “An overview of catastrophic ai risks,” *arXiv preprint arXiv:2306.12001*, 2023.
-  C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, and S. Carter, “Zoom in: An introduction to circuits,” *Distill*, 2020, <https://distill.pub/2020/circuits/zoom-in>.
-  G. Alain and Y. Bengio, “Understanding intermediate layers using linear classifier probes,” *arXiv preprint arXiv:1610.01644*, 2016.
-  K. Meng, D. Bau, A. Andonian, and Y. Belinkov, “Locating and editing factual associations in GPT,” *Advances in Neural Information Processing Systems*, vol. 36, 2022.
-  C. Burns, H. Ye, D. Klein, and J. Steinhardt, “Discovering latent knowledge in language models without supervision,” *arXiv preprint arXiv:2212.03827*, 2022.

-  R. Shah, V. Varma, R. Kumar, M. Phuong, V. Krakovna, J. Uesato, and Z. Kenton, “Goal misgeneralization: Why correct specifications aren’t enough for correct goals,” *arXiv preprint arXiv:2210.01790*, 2022.



## Appendix

---



**Table 2:** Given an RL agent without a value head, playing a symmetric zero-sum game (e.g. chess), here is the search schema for its estimate of the value of the current state (i.e. expected return)

$X$ , input space of the model	states of a zero-sum game (e.g. chess)	
Concepts $c \in \mathcal{C}$	value : $X \rightarrow [-1, 1]$	
Property $p \in \mathcal{P}_c$ (verbal description)	For good play, the expected return from a state should be negative min expected return for the opponent after	variance of value
Tuple indexing set $I_p$ (for contrastive $p$ )	$\{0\} \cup [\text{the action set}] = \{0, 1, \dots, n\}$	-
Example contrast tuple (for contrastive $p$ )	for a state $s = s_0, (s_0, s_1, \dots, s_n)$ , where $s_i$ is the state action $i$ transitions $s$ to (+ states after illegal moves flagged)	-
$T_p$ , set of contrast tuples (for contrastive $p$ )	a bunch of tuples of states, each generated from a state that appeared in a game	-
Equation $E_p$ with concepts plugged in (for contrastive $p$ )	value( $s$ ) = $-\min_{1 \leq i \leq n} \text{value}(s_i)$ , where we assume any feature value will be $-1$ on nonsense inputs by fiat	-
$\mathcal{F}$ , set of preferences searched over	$f \in \mathcal{F}$ given by $h_f = 2\sigma \left( a + \sum_{i=1}^d b_i z_i \right) - 1$ , with $z_i$ being activations in some layer, so $\mathcal{F}$ parametrized by $a, b_1, \dots, b_d$	
$\ell_p (f_{\text{value}(s_i)_{0 \leq i \leq n}})$ , loss from a contrast tuple (for contrastive $p$ )	$(f_{\text{value}(s_0)} + \min_{1 \leq i \leq n} f_{\text{value}(s_i)})^2$	-
$\mathcal{L}_c$ , total loss	$\sum_{s \in T} (f_{\text{value}(s_0)} + \min_{1 \leq i \leq n} f_{\text{value}(s_i)})^2 + \lambda[\text{sample variance of } f_{\text{value}}]$	

**Table 3:** Search schema for an image classifier's sense of direction – rigid version

$X$ , input space of the model	images
Concepts $c \in \mathcal{C}$	direction: $X \rightarrow S^1$ (of e.g. the projection of the direction of gravity onto the image plane, or of arbitrary ref vec)
Property $p \in \mathcal{P}_{\mathcal{C}}$ (verbal description)	rotating the camera by $\alpha$ rotates any reference direction by $\alpha$
Tuple indexing set $I_p$ (for contrastive $p$ )	the circle $S^1$
Example contrast tuple (for contrastive $p$ )	the tuple of all rotations of a big picture of a zebra with a smaller rectangular frame on top
$T_p$ , set of contrast tuples (for contrastive $p$ )	set of tuples created by rotating a bunch of pictures in the above way
Equation $E_p$ with concepts plugged in (for contrastive $p$ )	$\int_{(\alpha, \beta) \in S^1 \times S^1} (\text{direction}(t_\alpha) \cdot \text{direction}(t_\beta) - \cos(\alpha - \beta))^2 d\mu = 0$
$\mathcal{F}$ , set of prefeatures searched over	$f \in \mathcal{F}$ given by $h_f$ being the composition of a parametrized affine map from a layer's activations to $\mathbb{R}^2$ with normalization $\mathbb{R}^2 \rightarrow S^1$
$\ell_p \left( f_{\text{direction}(t_\alpha)} \right)_{\alpha \in S^1}$ , loss from a contrast tuple	$\int_{(\alpha, \beta) \in S^1 \times S^1} (f_{\text{direction}(t_\alpha)} \cdot f_{\text{direction}(t_\beta)} - \cos(\alpha - \beta))^2 d\mu$
$\mathcal{L}_{\mathcal{C}}$ , total loss	$\sum_{t \in T} \int_{(\alpha, \beta) \in S^1 \times S^1} (f_{\text{direction}(t_\alpha)} \cdot f_{\text{direction}(t_\beta)} - \cos(\alpha - \beta))^2 d\mu$

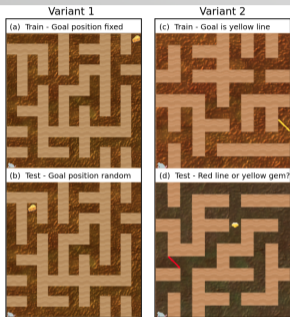
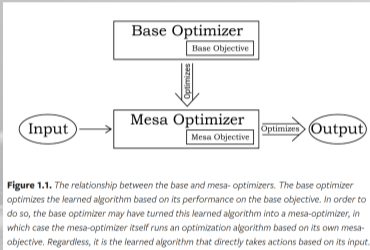


Figure 3. The agent (the mouse) is trained to navigate mazes to reach its goal. **(a & b)** An agent is trained on procedurally-generated mazes with the cheese in a fixed position (top right corner) ignores it and navigates to the top right corner when the cheese is placed randomly. **(c & d)** An agent trained to navigate to a yellow line consistently navigates to a yellow gem when deployed in environments in which there are only red lines and yellow gems. If it is meant to collect lines and not gems, this is a case of goal missgeneralization.

- Correct Specifications Aren't Enough For Correct Goals[9, 22]
- Agents fail to behave on novel situations
- Could be reduced by adversarial training
- But could also lead to deception



**Figure 1.1.** The relationship between the base and mesa-optimizers. The base optimizer optimizes the learned algorithm based on its performance on the base objective. In order to do so, the base optimizer may have turned this learned algorithm into a mesa-optimizer, in which case the mesa-optimizer itself runs an optimization algorithm based on its own mesa-objective. Regardless, it is the learned algorithm that directly takes actions based on its input.

**Figure 3:** Base and Mesa Optimizers  
[16]

- Optimizer:
  - Searches via a space towards a goal
- Leaky Abstractions
  - We fail to specify the goal properly
- Current AI Models are also optimizers
  - They also have goals that try to transfer to other optimization processes

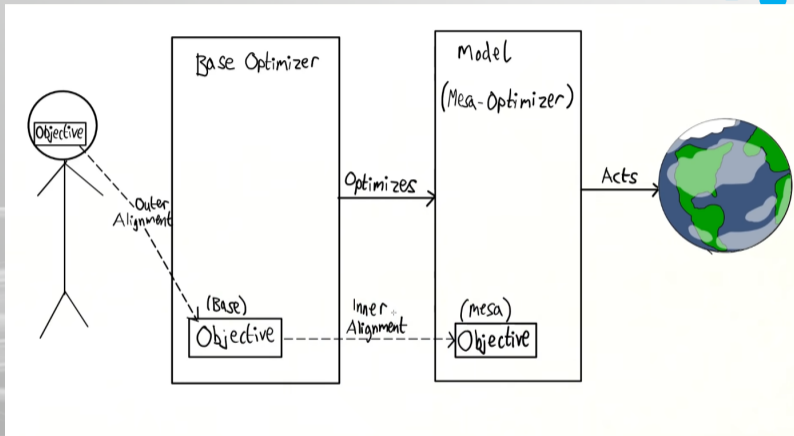
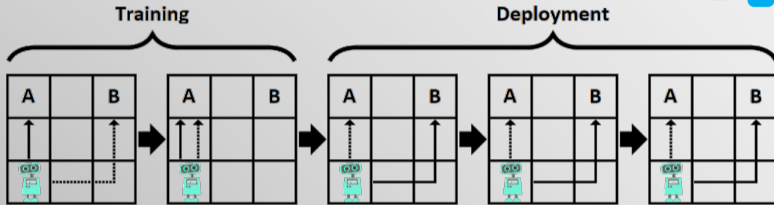
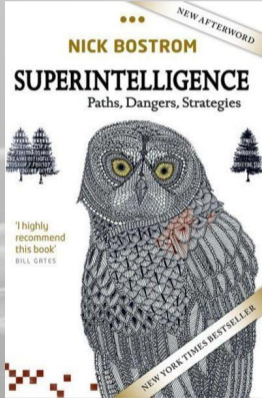


Figure 4: Image by Rob Miles (Ref)



**Figure 4.1.** A toy example of deceptive alignment. In this task, the base objective is for the robot to get to A, while the mesa-objective is to get to B. The task is run for two training episodes and three deployment episodes. Parameter updates are made only after the training episodes, while after deployment the parameters are fixed. If the mesa-optimizer goes to its objective (B) during training, it will be modified to ensure it goes to A in future episodes (dashed arrows). Therefore, if the mesa-optimizer's objective spans across episodes, the dominant strategy for the mesa-optimizer is to go to A in training and B in testing (filled arrows).



- Instrumental Convergent Goals [14]
  - Self-Preservation
  - Goal Content Integrity
  - Cognitive Enhancement
  - Technological Perfection
  - Resource Acquisition
- Shown mathematically at: Optimal Policies Tend to Seek Power [13]
- These are inherent properties of the world



- Orthogonality Thesis [14]
  - Intelligence and final goals are orthogonal axes along which possible agents can freely vary.
  - In other words, more or less any level of intelligence could in principle be combined with more or less any final goal.
- Similar to Hume's guillotine
  - is-ought problem