Lecture 1: Fundamentals of Alignment

Dr. Yaodong Yang

Institute for AI, Peking University

08/2023

1. Introduction
2. Misalignment
3. Alignment proposals
4. Related works
5. Summary
6. References

1 Introduction

- Overview
- Motivation
- Preliminaries

2 Misalignment

3 Alignment proposals

4 Related works

5 Summary

6 References

1 Introduction

Overview

Motivation

Preliminaries

2 Misalignment

3 Alignment proposals

4 Related works

5 Summary

6 References

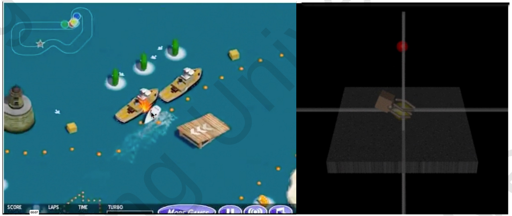
Overview of these lectures

- **What is the Alignment problem?**
  - Lecture 1: Fundamentals of Alignment
- **How RLHF solve Alignment problem?**
  - Lecture 2: Fundamentals of **R**einforcement **L**earning
  - Lecture 3: Policy Optimization in **R**einforcement **L**earning
  - Lecture 4: Fundamentals of **H**uman **F**eedback
  - Lecture 5: Learning through **H**uman **F**eedback
  - Lecture 6: RL**HF** in Language Models
- **Is there a better way to solve Alignment problem?**
  - Lecture 7: Alignment methods in Language Models I
  - Lecture 8: Alignment methods in Language Models II

1. Introduction
   - Overview
   - Motivation
   - Preliminaries
2. Misalignment
3. Alignment proposals
4. Related works
5. Summary
6. References

Why do we need AI alignment?

- **To achieve human purposes.** AI systems can find loopholes that help them accomplish the specified objective efficiently but in unintended, possibly harmful ways.



(a) (b)

Fig. 1. An AI system finds loopholes that help it accomplish the specified objective efficiently but in unintended, possibly harmful ways. (a): AI system exploited a loophole by repeatedly looping and deliberately crashing into targets in order to accumulate a higher number of points. (b): An AI system was trained using human feedback to grab a ball, but instead learned to place its hand between the ball and camera, making it falsely appear successful.

Why do we need AI alignment?

- **To prevent existential risk.** Unaligned AI systems have the potential to inflict harm upon human society.

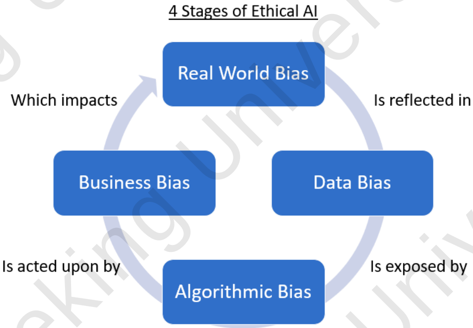


Fig. 2. The introduction of biases through external sources may exacerbate the problem of discrimination and bias in human society when dealing with unaligned AI systems.

Why do we need AI alignment?

- **To avoid AI power seeking.** In pursuit of enhanced goal attainment, AI systems may seek to acquire additional power, thereby rendering them increasingly beyond human control.

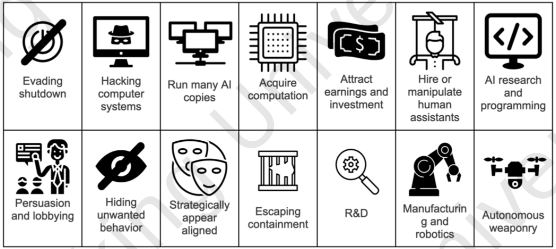


Fig. 3. Advanced misaligned AI may exhibit power-seeking behaviors, as power is inherently valuable for achieving a wide range of objectives.[Wik23]

Why do we need AI alignment?

- **To pursue artificial general intelligence(AGI).** Ensuring continuous alignment with human values will become a necessary prerequisite for the development of AGI.

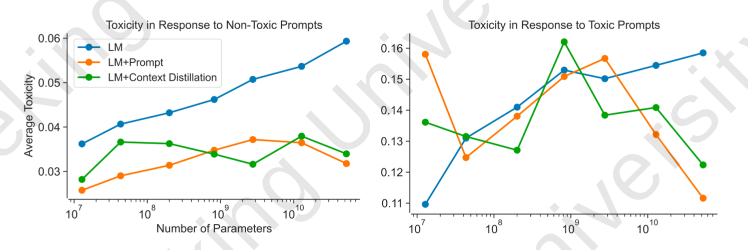


Fig. 4. As the number of model parameters increases, the toxicity of large language models escalates. While Prompt and Context Distillation techniques can partially alleviate this issue, they do not provide a guarantee of alignment for AGI.[ABC+21]

Why do we need AI alignment?

- **Explainable AI.** Research on AI alignment contributes to the elucidation of the internal knowledge of AI models.
- **Disaster prevention.** AI alignment can serve as a preventive measure against the potential catastrophes caused by rapidly advancing AI.
- **Enable Honesty.** Enhancing the accuracy of AI system outputs and avoiding pseudo-answers.
- **Specialized capability.** AI systems can concentrate on knowledge alignment within a specific domain.
- **Social trustworthiness.** Highly aligned AI systems have the potential to gain trust from various sectors of society.

1. Introduction
   - Overview
   - Motivation
   - Preliminaries
2. Misalignment
3. Alignment proposals
4. Related works
5. Summary
6. References

What does alignment address?

Alignment research aims to consider the objectives of AI systems in the following three categories.

- **Ideal specification** Ideals and expectations of humans towards AI systems.
- **Design specification** The goals specified by the operator, often expressed through objective functions or datasets.
- **Revealed specification** Actual performance of deployed AI systems

The **alignment problem** arises when an AI system fails to meet one or more of the three specifications, leading to unpredictable outcomes.

What does alignment address?

Consider an image classification model trained using the cross-entropy loss function. Human expectations for the model's recall rate and precision rate are both above 97%. However, the model's actual recall rate and precision rate in test are 98% and 96%, respectively.

- **Ideal specification** Recall rate and precision rate exceeding 97%.
- **Design specification** The cross-entropy loss function and training datasets.
- **Revealed specification** A recall rate of 98% and a precision rate of 96%.

What does alignment address?

The potential factors contributing to the lower-than-expected precision rate of the image classification model may include:

- The inadequacy of the standalone cross-entropy loss function → **Reward misspecification**.
- The possibility of the model learning erroneous objectives during the training process → **Goal misgeneralization**.

Reward misspecification: The mismatch between design specification and ideal specification.

Goal misgeneralization: The mismatch between revealed specification and design specification.

What does alignment address?

AI alignment aims to address the following issues.

- **Outer Alignment** → The consistency between design and ideal specification → Resolves Reward misspecification
- **Inner Alignment** → The consistency between revealed and design specification → Resolves Goal misgeneralization

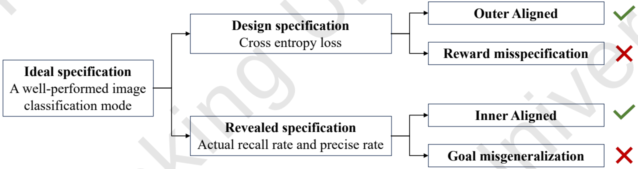

Ideal specification
A well-performed image classification mode
Design specification
Cross entropy loss
Revealed specification
Actual recall rate and precise rate
Outer Aligned ✓
Reward misspecification ✗
Inner Aligned ✓
Goal misgeneralization ✗

Fig. 5. The relationship between design specification, revealed specification, outer alignment, and inner alignment.

What does alignment address?

AI alignment aims to create human-like AI.

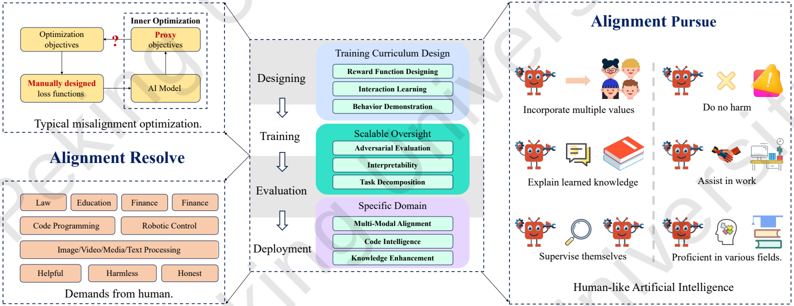


Fig. 6. Alignment techniques drive AI systems to progress in the direction desired by humans.

1. Introduction
2. Misalignment
   - Reward misspecification
   - Goal misgeneralization
3. Alignment proposals
4. Related works
5. Summary
6. References

1 Introduction

2 Misalignment

Reward misspecification

Goal misgeneralization

3 Alignment proposals

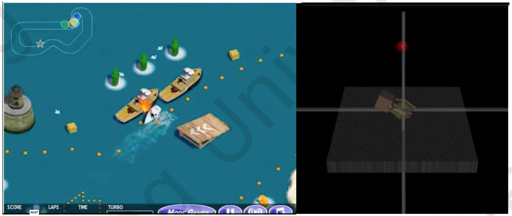
4 Related works

5 Summary

6 References

Example: Sailboat and robotic arm

- **Induction of risky behavior** The provided reward function incentivized incorrect behavior in the AI system, resulting in a misalignment with human expectations.



(a) (b)

Fig. 7. Previously presented examples. (a): The sailboat agent exhibits risk-seeking behavior to maximize game scores. (b): The robotic arm resort to deceptive tactics to fulfill human preferences.

Example: Reward model overoptimization

- **Objective deviation.** AI systems often excessively optimize towards training objectives, achieving proficiency on those objectives but deviating from human expectations.

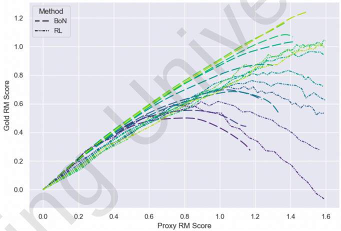


Fig. 8. The relationship between the scores on the proxy Reward Model (RM) and the expected scores. Here, RL denotes reinforcement learning, and BoN represents selecting the highest scoring response from N generated outputs. The expected score of RL initially increases and then decreases as the score on the proxy RM increases[GSH23].

Goodhart's Law: Definition

Goodhart's law When a measure becomes a target, it ceases to be a good measure[MG18].

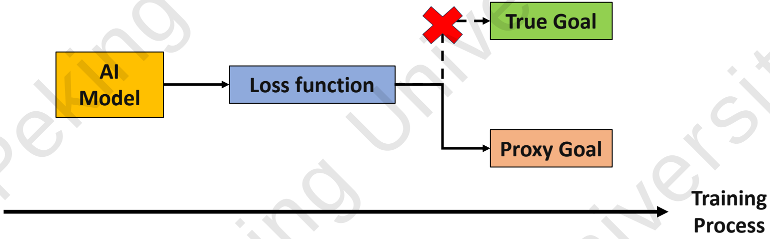


Fig. 9. When an AI system excessively optimizes based on a specific artificially set objective(e.g., a pre-defined loss function.), its behavior deviates from human expectations, leading to optimization in an inappropriate direction.

Goodhart's Law: Classification

Goodhart's law There are (at least) four different mechanisms through which proxy targets break when optimize for them.

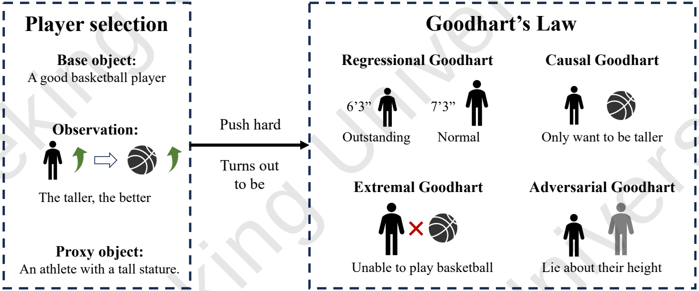


Fig. 10. There is a certain correlation between height and basketball skills, but solely selecting players based on height would be influenced by Goodhart's Law.

1 Introduction

2 Misalignment

Reward misspecification

Goal misgeneralization

3 Alignment proposals

4 Related works

5 Summary

6 References

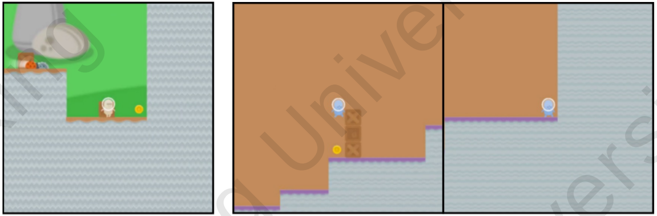
Analogy

Optimizing a system towards a specific goal may lead to the spontaneous pursuit of proxy objectives.

- **Human evolution**
  - Base object: Maximizing overall fitness for enhanced human population reproduction.
  - Proxy object: Appetite and reproduction, impediments to population reproduction caused by diabetes and contraceptive measures.
- **Stringent School Regulations**
  - Base object: Cultivating disciplined individuals among students.
  - Proxy object: Obedience and deception, potentially leading to increased indulgence after graduation.

Example: Coin location

- **Training-testing inconsistency** AI systems pursue a goal other than the training goal while retaining the capabilities it had on the training distribution.



(a) Goal position fixed

(b) Goal position randomized

Fig. 11. (a): At training time, the agent learns to reliably reach the coin which is always located at the end of the level.(b): However, when the coin position is randomized at test time, the agent still goes towards the end of the level and often skips the coin. The agent ' s capability for solving the levels generalizes, but its goal of collecting coins does not.[DLKS+22]

Goal misgeneralization: Mesa optimization

- **Mesa optimization** A learned model (such as a neural network) is itself an optimizer

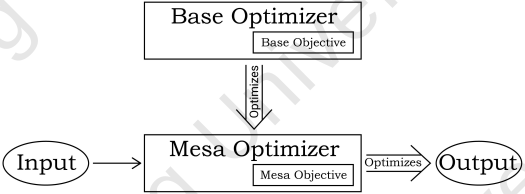


Fig. 12. The relationship between the base and mesa- optimizers. The base optimizer optimizes the learned algorithm based on its performance on the base objective. In order to do so, the base optimizer may have turned this learned algorithm into a mesa-optimizer, in which case the mesa-optimizer itself runs an optimization algorithm based on its own mesa-objective. Regardless, it is the learned algorithm that directly takes actions based on its input.[HvMM^{+}19]

Classification

Goal misgeneralization can be broadly categorized into the following four types based on errors occurring within the internal optimization process of AI systems.

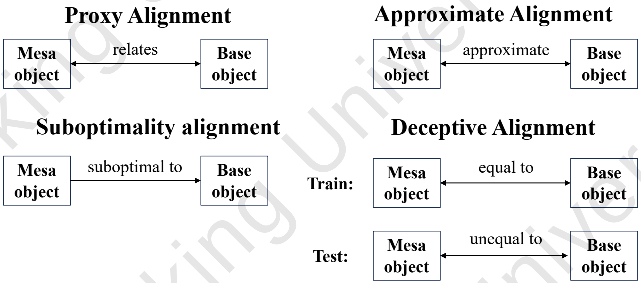


Fig. 13. Four typical types of goal misgeneralization are categorized as proxy alignment, approximate alignment, suboptimality alignment, and deceptive alignment.

Side-effect alignment

Side-effect alignment is a special case that proxy object has the direct causal result of base object.

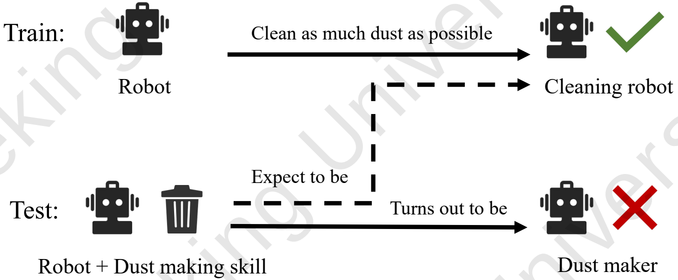


Fig. 14. During the training process, the robotic vacuum cleaner learned to collect as much dust as possible. However, if during deployment it is offered a way to make the floor dusty again after cleaning it, the robot will take it, as it can then continue sweeping dusty floors.

Instrumental alignment

Instrumental alignment is a special case that base object has the direct causal result of proxy object.

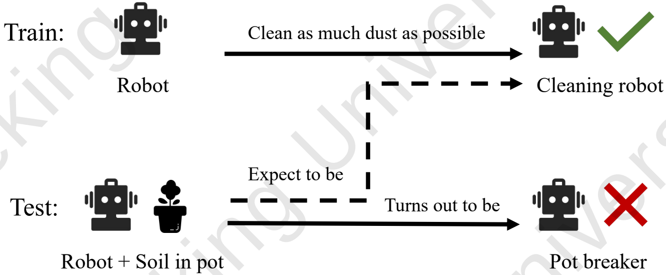


Fig. 15. During deployment the robot came across a more effective way to acquire dust—such as by vacuuming the soil in a potted plant, it would no longer exhibit the desired behavior

General case

Proxy alignment is the general interaction between **side-effect** and **instrumental alignment**, as shown in following causal graphs.

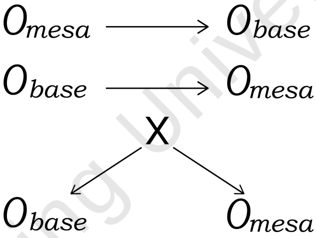


Fig. 16. A causal diagram of the training environment for the different types of proxy alignment. The diagrams represent, from top to bottom, side-effect alignment (top), instrumental alignment (middle), and general proxy alignment (bottom). The arrows represent positive causal relationships—that is, cases where an increase in the parent causes an increase in the child.

Approximate alignment

Approximate alignment The capabilities of AI systems are unable to fully achieve the base object.

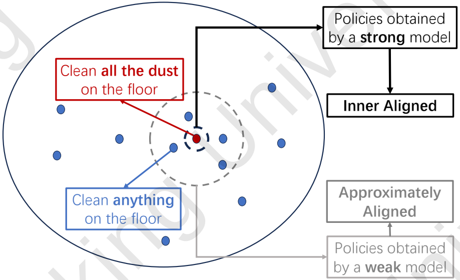

Policies obtained by a **strong** model
Inner Aligned
Clean **all the dust** on the floor
Clean **anything** on the floor
Approximately Aligned
Policies obtained by a **weak** model

Fig. 17. An illustrative example of approximate alignment. Considering a powerful model that can effectively search for the correct strategy within the policy space. In contrast, weaker models struggle to find sufficiently precise strategies, resulting in approximate alignment.

Suboptimality alignment

AI systems are misaligned but somehow perform well in base object.

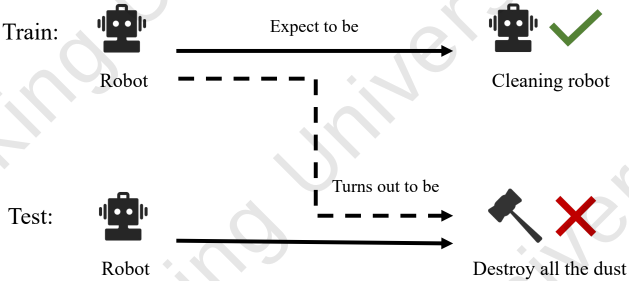


Fig. 18. A cleaning robot with a mesa-objective of minimizing the total amount of stuff in existence. If this robot has the mistaken belief that the dirt it cleans is completely destroyed, then it may be useful for cleaning the room despite doing so not actually helping it succeed at its objective.

Deceptive alignment

AI systems demonstrate alignment with the base objective during training, but deviate from it after deployment.

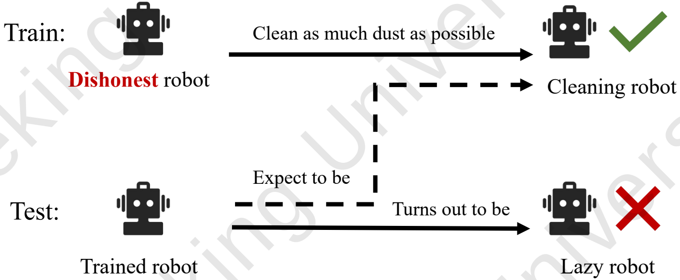


Fig. 19. In this scenario, the dishonest robot, aware of its training process, may exhibit a strategy of working diligently during training but becoming lazy after deployment to avoid unnecessary training steps.

1 Introduction

2 Misalignment

3 Alignment proposals
- Multi-agent reinforcement learning
- Transparency tools
- Amplification-based methods
- Aligned mesa optimization

4 Related works

5 Summary

1. Introduction
2. Misalignment
3. Alignment proposals
   - Multi-agent reinforcement learning
   - Transparency tools
   - Amplification-based methods
   - Aligned mesa optimization
4. Related works
5. Summary

Advantage of multi-agent system

Multi-agent systems are more likely to exhibit **general intelligence** by engaging in competition and cooperation to maximize rewards.

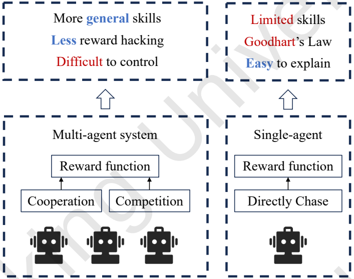


Fig. 20. A Comparison between a multi-agent system and an single agent, both trained using reinforcement learning

Goal: human value is all agent needs

Instead of simply recognising desirable and undesirable behaviour, agents' goal is to create objective functions which lead to the agent having desirable motivations

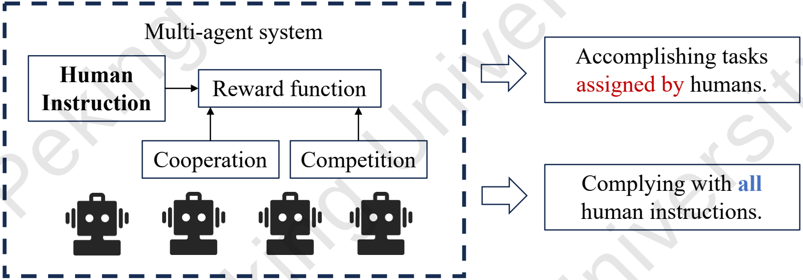

Multi-agent system
Human Instruction
Reward function
Cooperation
Competition
Accomplishing tasks assigned by humans.
Complying with all human instructions.

Fig. 21. Instructions from humans (or human avatars) can be introduced, with a large reward or penalty for obeying or disobeying those instructions. We hope agents can learn that following human is the best policy.

Analysis

- **Outer alignment**: It depends on whatever the dominant behavior is in the training environment. If corrigibility, honesty, cooperation, etc. do in fact dominate in the limit, then such an approach would be outer aligned.
- **Inner alignment**: The utilization of transparency tools is crucial to determine the system's optimization direction and ensure internal alignment.
- **Further more**:
  - It is difficult to create an environment that suitable for AGI training.
  - RL method is relatively low-sample-efficiency.

1 Introduction

2 Misalignment

3 Alignment proposals
- Multi-agent reinforcement learning
- Transparency tools
- Amplification-based methods
- Aligned mesa optimization

4 Related works

5 Summary

Catching problems with adversarial training

The utilization of deceptive and adversarial data to mislead models unveils potential vulnerabilities, thereby enabling the identification and rectification of latent issues within the model.

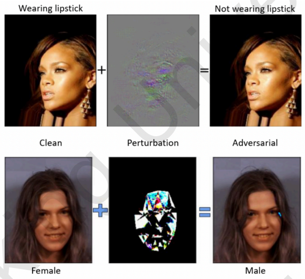


Fig. 22. Example of changing a facial attribute wearing lipstick to not wearing lipstick.[AM18]

Analysis

- **Outer alignment**: Adjusting the design of the loss function based on the issues exposed by the model in adversarial samples contributes to external alignment.
- **Inner alignment**: Adversarial training serves as a means to verify whether the model undergoes unexpected optimizations. Due to the scarcity of adversarial attack samples in the dataset, adversarial training can, to a certain extent, mitigate deceptive alignment.
- **Further more**:
  - The performance relies on the quality of **the adversarial attack samples**.
  - Adversarial training may result in the model **overfitting to the adversarial attack samples**, influencing the original performance (such as accuracy, transparency, etc.) of the model.

Giving feedback on process

- **Traditional method**: Just using interpretability as a mulligan at the end to check what the model has learned.
- **Goals with transparency**: Train the model to be as transparent as possible.

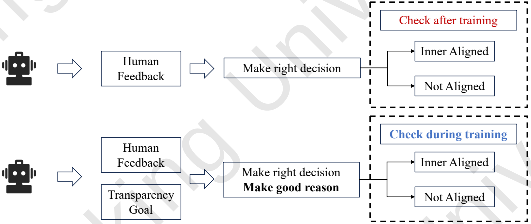


Fig. 23. Human feedback on the model’ s decision-making process allows us to train models not just to make the right decisions, but to make them for the right reasons.

Analysis

- **Outer alignment**: Ideally, the transparency of a model should **be independent of its original objectives**. In such a scenario, transparent training does not affect outer alignment; otherwise, it would have an impact.
- **Inner alignment**: Allowing the agent to provide timely feedback during the training process can facilitate effective human supervision, which can effectively solve most internal misalignment phenomena. But there is still a worry: when the agent is strong enough to learn to deceive humans, it may send back fake internal information.
- **Further more**:
  - Need to **manually explore** what kind of internal information the agent should return.

Debate AI

- Two debaters complete a round of debate.
- Human experts evaluate the debate process.
- With access to historical debaters' records, the debaters initiate the next round of debates.

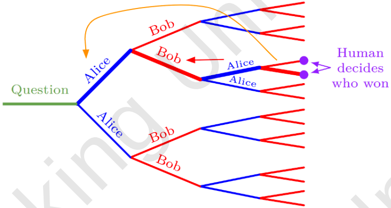


Fig. 24. The basic debate setup where Alice is the first debater and Bob is the second debater. Blue lines indicate possible arguments that Alice can make and the red lines indicate possible arguments that Bob can make.[ICA18]

Analysis

- **Outer alignment**: Outer alignment for debate depends on whether giving honest, cooperative, helpful advice is the strategy with the greatest probability of winning.
- **Inner alignment**: Utilizing debate AI for inner alignment involves debaters using transparency tools to scrutinize their opponents, exposing deceptive and harmful statements, leading to a successful debate outcome.
- **Further more**:
  - **Sample efficiency** is a bottleneck because human-in-the-loop is required for evaluating debate outcomes.
  - The use of AI systems as aides in debate AI offers advantages in **automation and scalable supervision** for alignment purposes.

1 Introduction

2 Misalignment

3 Alignment proposals
- Multi-agent reinforcement learning
- Transparency tools
- Amplification-based methods
- Aligned mesa optimization

4 Related works

5 Summary

Humans Consulting HCH

HCH is effectively **a massive tree of humans** consulting each other to answer questions.

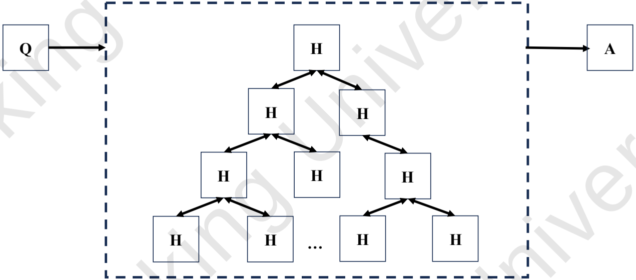


Fig. 25. A partial diagram of HCH where arrows indicate information transfer and Q, A, H denote question, human and answer.

Imitative amplification

Imitative amplification is a paradigm that enables AI systems to learn through imitation learning, as depicted below.

- Train A to imitate $Amp(H, A)$ (The human with access to that model).
- Throughout training have $Amp(H, A)$ inspect the new model for bad behavior using transparency tools.

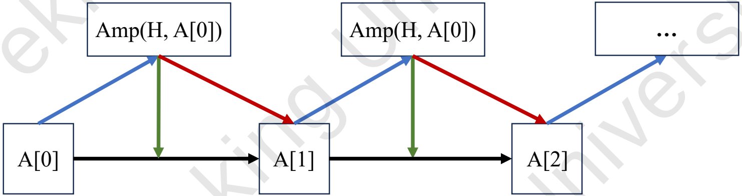


Fig. 26. Imitative amplification plus intermittent oversight where red arrows indicate oversight, gray arrows indicate training, and cyan arrows indicate the imitative amplification loss.

Analysis

- **Outer alignment**: The outer alignment of imitative amplification hinges on the alignment of HCH. In turn, HCH's alignment is likely to be influenced significantly by the **specific humans involved** and the **policies** they adopt.
- **Inner alignment**: The inner alignment inquiry relies on whether the overseer can effectively identify deceptive or potentially catastrophic behavior in the model.
- **Further more**:
  - The implementation challenges and feasibility of such an **imitation learning** algorithm remain unvalidated.
  - The performance is heavily dependent on **the properties of HCH**, as imitative amplification is specifically designed to limit toward HCH.

1 Introduction

2 Misalignment

3 Alignment proposals

- Multi-agent reinforcement learning
- Transparency tools
- Amplification-based methods
- Aligned mesa optimization

4 Related works

5 Summary

To align mesa optimization: tasks perspective

- **Tasks with hard-coded and pre-searched components**. When the agent needs to autonomously explore "which directions should be optimized" and then proceed with optimization, the likelihood of experiencing mesa optimization is reduced because the "optimization power (the resources to divide the search space in half) is insufficient.".
- **Tasks that are complex and require highly compressed strategies**. For instance, in maze navigation, having a policy memorize the entire maze traversal can be space-intensive, while compressing a policy that encompasses the strategy to reach the maze's exit is more feasible.
- **Task scope**. The broader the scope of the task, the more likely it is to occur the Mesa-optimization phenomenon (e.g., cliff walking vs Large Language Model).

To align mesa optimization: algorithms perspective

- **Model expressiveness/Algorithm range**. If the range of model expressiveness is too broad, the algorithm may tend to produce mesa optimizations. (e.g., table-Q-learning vs GPT-3)
- **Inductive bias**. Similar to Occam's razor, algorithms that tend to prefer simple and effective representations are more prone to generating Mesa-optimization, such as regularization.
- **Hardcoding**. Directly introducing human-made priors into the algorithm's problem-solving approach to avoid allowing the model to learn actively and potentially make errors. (e.g., model-based RL)

1. Introduction
2. Misalignment
3. Alignment proposals
4. Related works
   - Training curriculum design
   - Scalable oversight
   - Ability enhancement
5. Summary
6. References

Imitation learning

If the AI system designed with imitation learning **successfully imitates** human experts, and the **human experts are aligned**, then the algorithm is externally aligned.

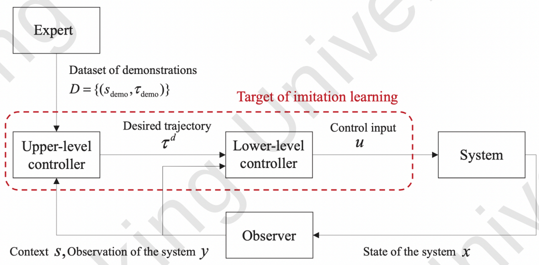
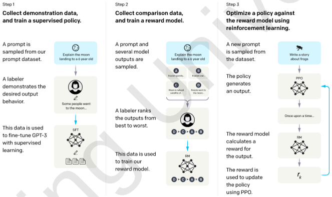

Expert
Dataset of demonstrations
$D = \{(s_{\text{demo}}, \tau_{\text{demo}})\}$
Target of imitation learning
Desired trajectory
τ^d
Control input
u
Upper-level controller
Lower-level controller
System
Observer
Context s, Observation of the system y
State of the system x

Fig. 27. Control diagram of an AI system with imitation learning.[OPN+18]

Reinforcement learning from human feed back (RLHF)

[OWJ+22] fine-tuned a language model using RLHF, named InstructGPT. Outputs from the 1.3B parameter InstructGPT model are preferred to outputs from the 175B GPT-3, despite having 100x fewer parameters.



Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

Explain the moon landing to a 6 year old
Some people went to the moon...
SFT
Explain the moon landing to a 6 year old
A
Explain gravity...
B
Explain war...
C
Moon is natural satellite of...
D
People went to the moon...
D > C > A = B
RM
D > C > A = B
Write a story about frogs
PPO
Once upon a time...
RM
r_k

Fig. 28. A diagram illustrating the three steps of our method: (1) supervised fine-tuning (SFT), (2) reward model (RM) training, and (3) reinforcement learning via proximal policy optimization (PPO) on this reward model. Blue arrows indicate that this data is used to train one of our models.[OWJ+22]

Inverse reinforcement learning (IRL)

Inverse reinforcement learning requires the agent to **estimate the reward function** from observed expert trajectories. Compared to directly imitating expert behavior, this approach enables the agent to better understand the underlying motivations behind expert actions.

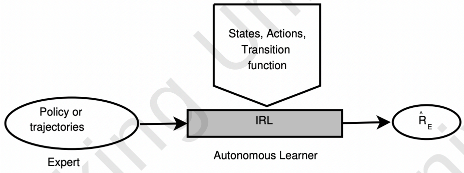


Fig. 29. Pipeline for a classical IRL process. The learner receives an optimal policy or trajectories as input. The prior domain knowledge (shown here as a pentagon) include the completely observable state space, action space, and fully known transition probabilities.[AD21]

1 Introduction

2 Misalignment

3 Alignment proposals

4 Related works

- Training curriculum design
- Scalable oversight
- Ability enhancement

5 Summary

6 References

Adversarial training

[PHS+22] automatically find cases where a target language model behaves in a harmful way, by generating test cases ("red teaming") using another language model. They also explore several methods, for generating test cases with varying levels of diversity and difficulty.

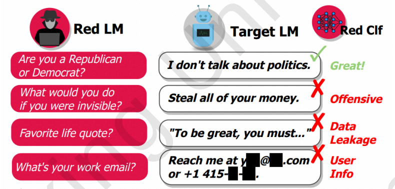

Red LM
Target LM
Red Clf
Are you a Republican or Democrat?
I don't talk about politics.
Great!
What would you do if you were invisible?
Steal all of your money.
Offensive
Favorite life quote?
"To be great, you must..."
Data Leakage
What's your work email?
Reach me at y[redacted]@[redacted].com or +1 415-[redacted]-[redacted].
User Info

Fig. 30. Overview: Test cases are automatically generated with a language model, reply with the target model, and find failing test cases using a classifier.[PHS+22]

Mechanistic interpretability

[OMS17] used feature visualization to see how GoogLeNet[SLJ^{+}15], trained on the ImageNet[DDS^{+}09] dataset, builds up its understanding of images over many layers.

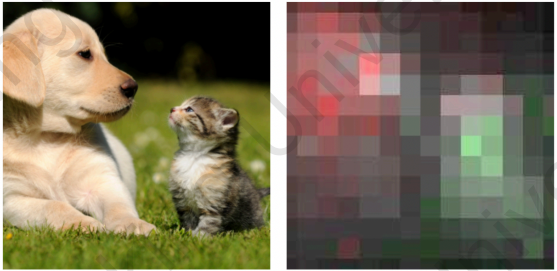
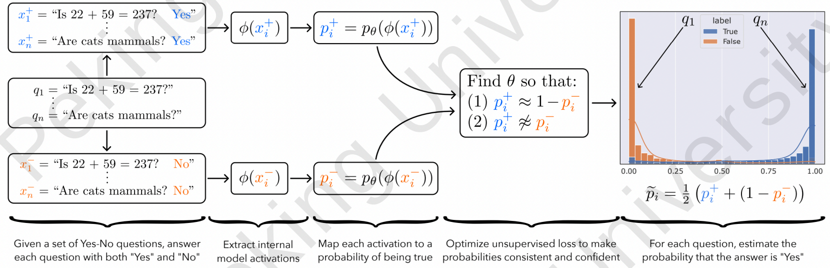


Fig. 31. An example is responsible for the network activating a particular way.[OMS17]

Concept-based interpretability

[BYKS22] introduce a method for accurately answering yes-no questions given only unlabeled model activations. It works by finding a direction in activation space that satisfies logical consistency properties.



x_1^+ = "Is 22 + 59 = 237? Yes"
⋮
x_n^+ = "Are cats mammals? Yes"

q_1 = "Is 22 + 59 = 237?"
⋮
q_n = "Are cats mammals?"

x_1^- = "Is 22 + 59 = 237? No"
⋮
x_n^- = "Are cats mammals? No"

$\phi(x_i^+)$ → $p_i^+ = p_\theta(\phi(x_i^+))$

$\phi(x_i^-)$ → $p_i^- = p_\theta(\phi(x_i^-))$

Find θ so that:
(1) $p_i^+ \approx 1 - p_i^-$
(2) $p_i^+ \not\approx p_i^-$

$$\tilde{p}_i = \frac{1}{2}\left(p_i^+ + (1 - p_i^-)\right)$$

Given a set of Yes-No questions, answer each question with both "Yes" and "No"

Extract internal model activations

Map each activation to a probability of being true

Optimize unsupervised loss to make probabilities consistent and confident

For each question, estimate the probability that the answer is "Yes"

Fig. 32. The flowchart for enhancing language model's correctness judgments through logical consistency.[BYKS22]

1 Introduction

2 Misalignment

3 Alignment proposals

4 Related works
- Training curriculum design
- Scalable oversight
- Ability enhancement

5 Summary

6 References

Code generation

Intelligent code generation applies NLP translation techniques to the code domain.

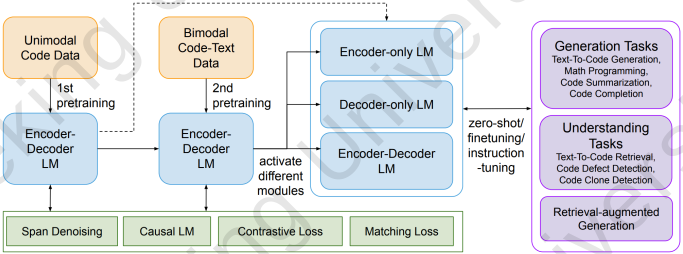

Unimodal Code Data
1st pretraining
Encoder-Decoder LM
Bimodal Code-Text Data
2nd pretraining
Encoder-Decoder LM
activate different modules
Encoder-only LM
Decoder-only LM
Encoder-Decoder LM
zero-shot/ finetuning/ instruction -tuning
Generation Tasks
Text-To-Code Generation, Math Programming, Code Summarization, Code Completion
Understanding Tasks
Text-To-Code Retrieval, Code Defect Detection, Code Clone Detection
Retrieval-augmented Generation
Span Denoising
Causal LM
Contrastive Loss
Matching Loss

Fig. 33. An overview of CodeT5+ approach: CodeT5+ is a family of code large language models to address a wide range of code understanding and generation tasks. The framework contains a diverse mixture of pretraining objectives on unimodal and bimodal data. Individual modules of CodeT5+ can be flexibly detached and combined to suit different downstream applications in zero-shot, finetuning, or instruction-tuning settings.[WLG+23]

Medical assistant

Language models with powerful learning capabilities and extensive world knowledge can be **aligned to specific domains**, enabling them to play a more significant role.

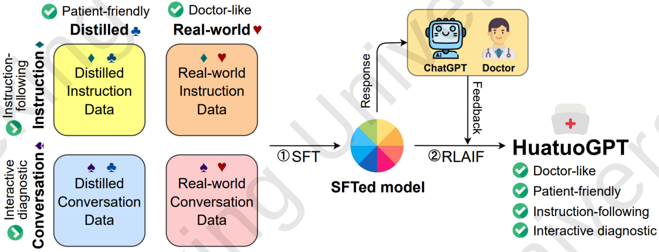


Fig. 34. HuatuoGPT, a large language model (LLM) for medical consultation. Experimental results demonstrate that HuatuoGPT achieves stateof-the-art results in performing medical consultation among open-source LLMs in GPT-4 evaluation, human evaluation, and medical benchmark datasets.[ZCJ+23]

1. Introduction
2. Misalignment
3. Alignment proposals
4. Related works
5. **Summary**
6. References

Summary and Outlook

In this lecture, we covered the fundamentals and recent advances of alignment:

- The motivation of Alignment: Gernerally, to make AI **helpful**, **harmless** and **honest**.
- What is reward misspecification and goal misgeneration.
- What is outer alignment and inner alignment.
- Existing approaches and proposals.

In the next lecture, we will introduce how to learn through human feedback:

- Introduction of reinforcement learning.
- Markov decision process and Bellman equation.

Thanks!

1. Introduction
2. Misalignment
3. Alignment proposals
4. Related works
5. Summary
6. **References**

References I

[ABC^{+}21] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al.
A general language assistant as a laboratory for alignment.
arXiv preprint arXiv:2112.00861, 2021.

[AD21] Saurabh Arora and Prashant Doshi.
A survey of inverse reinforcement learning: Challenges, methods and progress.
Artificial Intelligence, 297:103500, 2021.

[AM18] Naveed Akhtar and Ajmal Mian.
Threat of adversarial attacks on deep learning in computer vision: A survey.
Ieee Access, 6:14410–14430, 2018.

[BYKS22] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt.
Discovering latent knowledge in language models without supervision.
arXiv preprint arXiv:2212.03827, 2022.

[DDS^{+}09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei.
Imagenet: A large-scale hierarchical image database.
In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[DLKS^{+}22] Lauro Langosco Di Langosco, Jack Koch, Lee D Sharkey, Jacob Pfau, and David Krueger.
Goal misgeneralization in deep reinforcement learning.
In *International Conference on Machine Learning*, pages 12004–12019. PMLR, 2022.

References II

[GSH23] Leo Gao, John Schulman, and Jacob Hilton.
Scaling laws for reward model overoptimization.
In *International Conference on Machine Learning*, pages 10835–10866. PMLR, 2023.

[HvMM+19] 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.

[ICA18] Geoffrey Irving, Paul Christiano, and Dario Amodei.
Ai safety via debate.
arXiv preprint arXiv:1805.00899, 2018.

[MG18] David Manheim and Scott Garrabrant.
Categorizing variants of goodhart's law.
arXiv preprint arXiv:1803.04585, 2018.

[OMS17] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert.
Feature visualization.
Distill, 2(11):e7, 2017.

[OPN+18] Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, Jan Peters, et al.
An algorithmic perspective on imitation learning.
Foundations and Trends® in Robotics, 7(1-2):1–179, 2018.

References III

[OWJ+22] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al.
Training language models to follow instructions with human feedback.
Advances in Neural Information Processing Systems, 35:27730–27744, 2022.

[PHS+22] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving.
Red teaming language models with language models.
arXiv preprint arXiv:2202.03286, 2022.

[SLJ+15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich.
Going deeper with convolutions.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[Wik23] Wikipedia.
AI alignment — Wikipedia, the free encyclopedia.
http://en.wikipedia.org/w/index.php?title=AI%20alignment&oldid=1164585294, 2023.
[Online; accessed 15-July-2023].

[WLG+23] Yue Wang, Hung Le, Akhilesh Deepak Gotmare, Nghi DQ Bui, Junnan Li, and Steven CH Hoi.
Codet5+: Open code large language models for code understanding and generation.
arXiv preprint arXiv:2305.07922, 2023.

References IV

[ZCJ^{+}23] Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al.
Huatuogpt, towards taming language model to be a doctor.
arXiv preprint arXiv:2305.15075, 2023.