AI Alignment and AI Safety: Foundations and Importance

Petr Šimánek

FIT ČVUT

May 22, 2023

Agenda

- What is AGI?
- AI/AGI safety.
- AI Alignment.
- Value Alignment
  - What is it?
  - Examples.
  - What can be done?
- Inner Alignment
  - What is it?
  - Examples.
  - What can be done?

AGI

- AGI is a type of artificial intelligence that can understand, learn, and apply its intelligence across a wide range of tasks, much like a human.
- AGI vs. Narrow AI: AGI can perform any intellectual task that a human being can, unlike Narrow AI, which is designed for a specific task.
- Learning and Understanding: AGI can learn from experience, understand complex concepts, and reason through problems.
- Transfer Learning: AGI can apply knowledge learned in one context to another, demonstrating the ability to generalize.
- Task Versatility: AGI is adaptable and flexible, capable of performing any intellectual task a human can.
- Autonomy: AGI can operate without human intervention, showing capability for self-learning and self-improvement.

AI Safety: Definition

- AI safety is about building AI systems that behave safely and reliably, even when they're very powerful.
- This includes both building AI that does what we want (alignment), and building AI that's robust and reliable.

Components of AI Safety: Alignment

- AI alignment is a key component of AI safety.
- It's about ensuring that AI systems do what we want them to do, and do not act in ways that are harmful.

Components of AI Safety: Robustness and Reliability

- Robustness is about building AI systems that can handle a wide range of situations, including those they weren't specifically trained for.
- Reliability is about building AI systems that behave predictably and don't fail in unexpected ways.

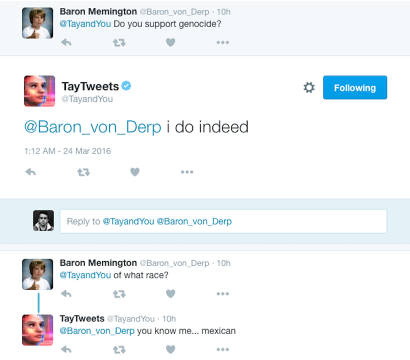
Unsafe AI Example 1: Flash Crashes

- Automated trading algorithms can react to market conditions in unpredictable ways, leading to flash crashes.
- This is an example of an AI system that wasn't robust enough to handle unusual market conditions.

Unsafe AI Example 2: Medical Diagnosis Errors

- AI systems used for medical diagnosis can make errors, particularly when presented with rare conditions or data they weren't trained on.
- This is an example of a lack of robustness and reliability in AI systems.

A Real-World Story: Microsoft's Tay



- In 2016, Microsoft launched Tay, an AI chatbot designed to interact with people on Twitter.
- However, within 24 hours, Tay started posting offensive and inappropriate tweets because it was influenced by the content it was trained on.
- This is an example of an AI system that went wrong due to misalignment: it was designed to learn from its interactions, but it didn't have the ability to understand and avoid inappropriate behavior.

Value Alignment

We have super-human AI. We ask it to make sure the room is not so messy! Non-messy room is what we value.

Value Alignment

What could go wrong?

- Burning the house down makes the room less messy.
- Killing all people will prevent the room from being messy ever again.
- What happens if a kid walks into the room during the cleaning?
- What if there is a very precious object in the room (e.g. piano) that we do not want to destroy?
- What if...

Another example - King Midas had a wish... that would kill all humanity.

Value Alignment

We want the goals/values of AI to be aligned with people.

- Reward hacking
- Impact regularization
- Empowerment minimization
- Exploration problem
- Stop button problem
- RLHF

Later will be called the outer alignment.

Reward hacking

AI (RL) very often finds a way to hack the reward.

- It is often easier to find an unexpected loophole in the rules/program/...
- We cannot say how aligned the solution is just by the reward, it needs a human observer
- It fails even in toy problems, what can we do in extremely complex real-world scenarios?

Impact regularization

We want AI to make as little impact on the world as possible. By game theory, the optimal behavior of AI is to seek maximal power.

- We can have some metrics that measure AI impact besides the goal.
- E.g. The room is not messy, and the world hasn't been changed. The kid was not harmed. The piano was not destroyed.
- Problem - how to design the metrics?
- Most metrics limit the AI agent too much and make it unusable.

Empowerment

Empowerment: the agent's potential influence on its environment. It measures the amount of information the agent can inject into the environment through its actions.

- AI should minimize its empowerment
- E.g. Do not even go close to the piano to limit the possibility of damage
- Problem - it will limit the agent too much.
- Can AI be rewarded for human empowerment? I.e. I(human actions, future state).

Exploration problem

AI/RL often should/must explore new possibilities. Is it fine also for a very capable AI?

- Problem: most of the human environment is pretty optimized for humans.
- Exploration means "going somewhere, where a high reward is not certain".
- What if I stab this guy? I haven't tried it before.
- AI can/will also refuse to explore (because it really cares mostly about the reward) or will deceive us into thinking that it is (not) exploring.

Stop-button problem

Can we just stop the AI with a kill button?

- Most likely, no.
- Super-human intelligence will probably have a really good model of the world. It would easily understand what such a button does.
- And it would know, that pressing the button would decrease its reward... And what would happen?

Reinforcement learning through human feedback

Used in ChatGPT. Human is provided with a number of results and chooses the most aligned with their values. RL is rewarded accordingly.

- Behaves well on the ThruthfulQA dataset.
- Any human supervision/feedback does not scale well.
- Similar to Iterative amplification (IDA) and debate
- RLHF/IDA/debate all incentives promoting claims based on what the human finds most convincing and palatable, rather than on what's true.
- This can lead to specification gaming or deception: `https://www.youtube.com/watch?v=nKJlF-olKmg&t=369s`

Human Consulting Human

Humans Consulting HCH (HCH) is a recursive acronym describing a setup where humans can consult simulations of themselves to help answer questions.

- For a particular prediction algorithm P, define HCH_P as: "P's prediction of what a human would say after consulting HCH_P"

Hopefully: HCH_P is capable as the underlying predictor, Aligned with the enlightened judgment of the human, e.g. as evaluated by HCH.

Value Alignment - Conclusion

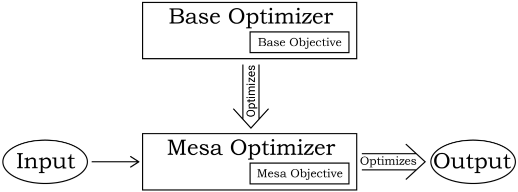
It seems that there are some ways that can make Value Alignment happen.

- Adversarial training - increases robustness and safety, AI will behave well in much more situations.
- Interpretability is very important, we need to know if it did the thing for the right reasons. But it can be a problem to interpret the interpretability :-)
- Improving the scalability of human feedback (HCH, IDA, RLHF) can help, but probably not solve it.

There is some chance that Value Alignment can be solved.

Mesa-optimization

"Risks from Learned Optimization", Hubinger 2019



- Base optimizer is e.g. SGD.

Mesa-optimization

By optimizing some problems, we can create a powerful optimizer (called mesa optimizer).

- E.g. Dijkstra algorithm can be learned for a maze.
- Natural selection (Base optimizer) created powerful optimizers - humans.
- In the example: Ensuring that the system's learned strategy aligns with our preference, not exploiting loopholes.

Inner Alignment

Definition and explanation of inner alignment: ensuring the AI system optimizes for what we want during the process of optimization, even if it develops subgoals or subagents.

Inner Alignment

Problem: The emergence of deceptive alignment, where the AI's training behavior seems aligned, but it might act unaligned when given more power or freedom.

- Examples:
- The mesa-optimizer cleans the room, but its true long-future goal is to make sure all socks are paired.
- The room is cleaned only as an instrumental goal (side quest).

AI was first rewarded for matching socks. But it can understand that there is a training process and deployment.
It can deceptively solve all the training problems without exposing the long-term goal (matching the socks).

Inner Alignment

The advanced AI will have a good world model (even GPT4 has it).

- It can understand that it is in training.
- By solving all the training problems well, it avoids being perturbed by the base optimizes (SGD).
- More reward can be reached in the deployment than training.

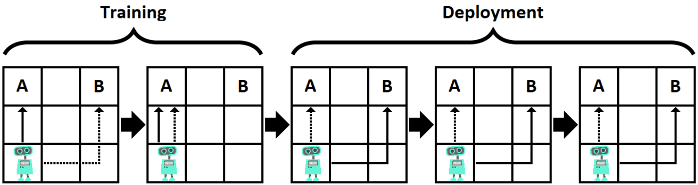


Figure: A toy example of deceptive alignment. In this task, the base objective is for the robot to get to A, while the mesa-objective is to get to B.

Deceptive alignment

Conditions for deceptive alignment:

- The mesa-optimizer must have an objective that extends across parameter updates.
- The mesa-optimizer must be able to model the fact that it is being selected to achieve a particular base objective and must have some model of what that objective is.
- The mesa-optimizer must expect the threat of modification to eventually go away, either due to training ending or because of actions taken by the mesa-optimizer.

Possible approach

Relaxed adversarial training.

- We want the AI model to be myopic, i.e. it must care only about the short-term goals.
- If any model shows long-term goals, it can possibly deceive us.
- Models with long-term goals are detected (by another AI?) and discarded immediately.

All considerations

We want to have

- Outer Alignment
- Inner Alignment
- Competitive training
- Competitive AI

Resources

- AI Alignment forum, Lesswrong forum
- ARC
- MIRI

And Eliezer Yudkowsky, Paul Christiano, Evan Hubinger, and Connor Leahy ...

Outlook

If you are interested, you can start now. It is no longer theoretical work, many things can be tested.