

07-280 AI Alignment

Nihar B. Shah

This lecture will be discussion oriented, with the following discussion points. We will revisit a lot of this in the final lecture of the course, after we have covered LLMs and RL. In this lecture, we will also discuss an example of autonomous AI scientists and designing appropriate measurement methods from <https://arxiv.org/pdf/2509.08713>.

What is AI alignment? We want AI to be aligned with our (human) goals and values.

- Alignment with user’s requirements
- Alignment with human values
- Near-term vs. long-term alignment.

Challenges in AI alignment

- Reward hacking / specification gaming.
- Hard to specify all requirements and values; hard to evaluate.
- In the real world, can encounter scenarios different from those during training (“distribution shift”).
- Large, complex models are not really interpretable.
- AI increasingly capable, and can come up with unforeseen courses of action including deception and seeking of power.
- AI may be much more capable than humans.

Current approaches to align AI

- *Post-training*: We will go into this more in the final lecture. But here is an example for intuitive understanding. Try a prompt, get one or more outputs, and a human (or an AI reward model) rates/ranks these outputs. These are then used as additional training data to improve the model.
- *Guardrails*: E.g., classifiers on the input prompt and/or the output. Or restricted tool use / access to data. This is done during inference time.
- *Red teaming*: People or other AI deliberately try to break the AI in order to find its vulnerabilities. A basic type of red teaming is to jailbreak, meaning circumventing safety checks of the model.
- *Interpretability*: Try to interpret what the model is doing. If you understand its inner workings, you may be able to also figure out whether it may act inappropriately.

- *Improving reward modeling/specification*: For instance, instead of setting the reward/loss based only on the final outcome, also provide a reward/loss based on intermediate outputs. This is called process reward modeling (as opposed to outcome reward modeling). Another approach is reward shaping, where additional terms are added to the reward/loss to encourage desired properties (e.g., helpfulness, safety, or efficiency), without changing the intended task. For example, when training a dialogue model to answer questions correctly, the base reward may reflect task success (e.g., correctness), while shaped rewards add penalties for unsafe content and bonuses for clarity or politeness. The underlying task (answering questions well) is unchanged, but the extra terms guide the model toward safer and more helpful behavior during learning.
- *Formal guarantees*: Mathematically prove that the model will not give undesirable outputs. This is quite challenging for large and complex models like today’s LLMs, and in most cases we are unable to provide such guarantees. But this has been done in simpler settings, e.g., that a controller for an aircraft will never recommend a certain dangerous maneuver.
- *Training data*: Higher-quality training data, and in particular, highly diverse data to mitigate the problem of distribution shift.
- *Corrigibility*: Designing systems that allow for corrections. Suppose the AI is given a certain goal. If the AI is going awry, and the user tries to stop it. A relentless AI trying to pursue the originally given goal will resist the user’s interventions. So corrigibility means that AI is open to users changing their goals, open to being shut down, and actively support human oversight. One approach: from the start, AI must allow for uncertainty about the true objective. Then it will be in the AI’s interest to allow interventions.
- *Cooperative Inverse Reinforcement Learning*: Inverse Reinforcement Learning (IRL) has humans doing demonstrations and the AI inferring the reward/loss function. The model is not told the reward function explicitly. Cooperative Inverse Reinforcement Learning involves doing IRL to learn human values, where the ‘cooperative’ part means that the AI’s objective is to learn the human’s values and work with them to achieve the human’s goals.