The Importance of Threat Models for AI Alignment

Rohin Shah

This presentation reflects my personal views, not those of DeepMind.

A note on terminology

AGI = Powerful AI system

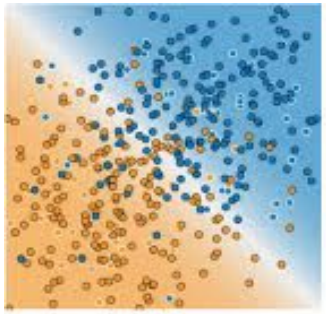
This is deliberately vague about the level of power

Decomposing threat models

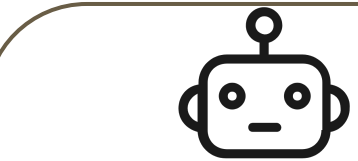
Combination of a *development model* that says how we get AGI and a *risk model* that says how AGI leads to existential catastrophe.

$$\theta \leftarrow \theta - \nabla_\theta \mathcal{L}_\theta(x, y)$$

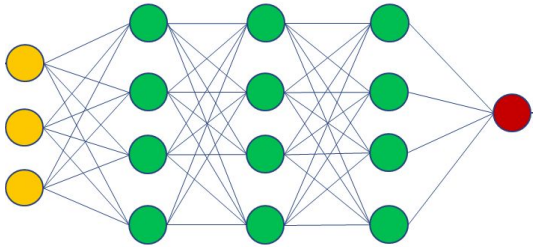
Gradient descent on training data



Development model



AGI



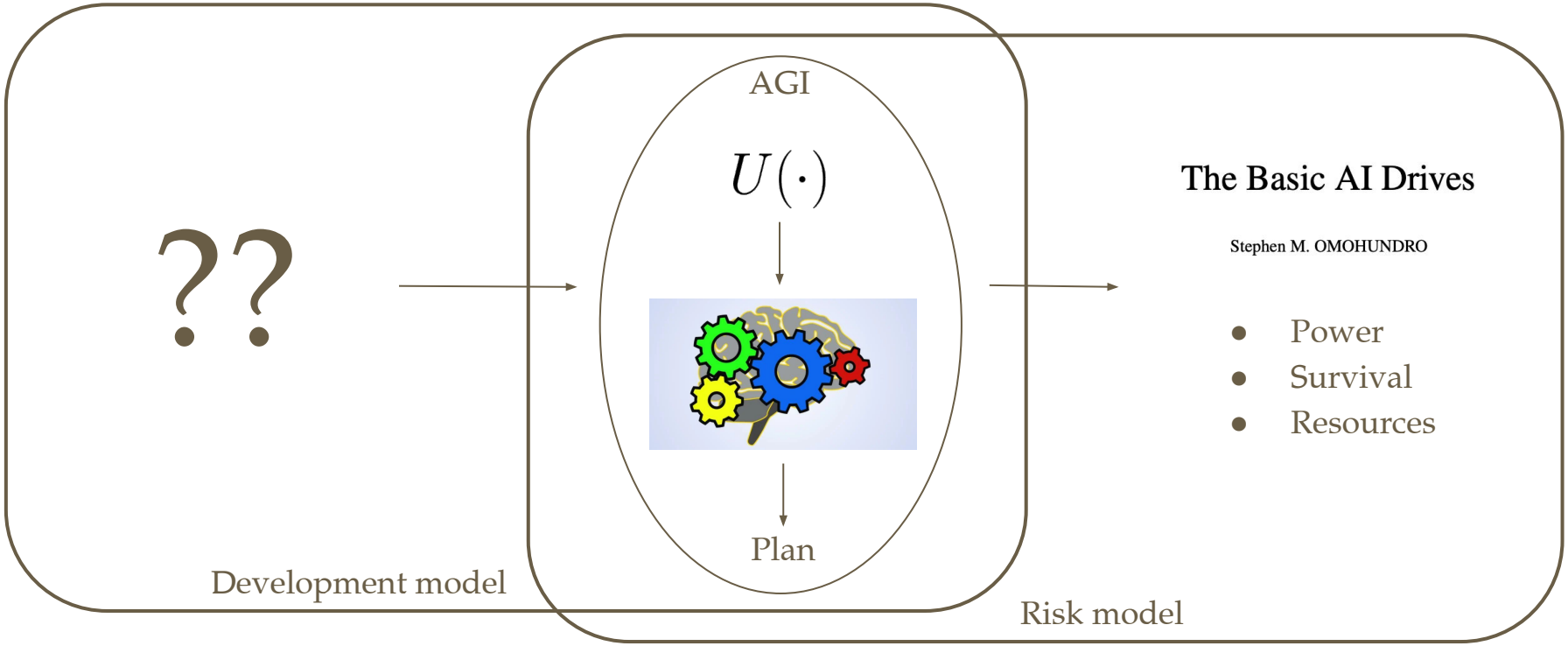
Doom

Risk model

Key claims

1. *Development models are crucial for solutions to apply to real systems.*
2. *Risk models are important for finding solutions.*
3. *We should be using the same notion of AGI for these two models.*
4. *We don't yet have compelling examples of these models.*

Threat model: utility maximizer



Threat model: utility maximizer

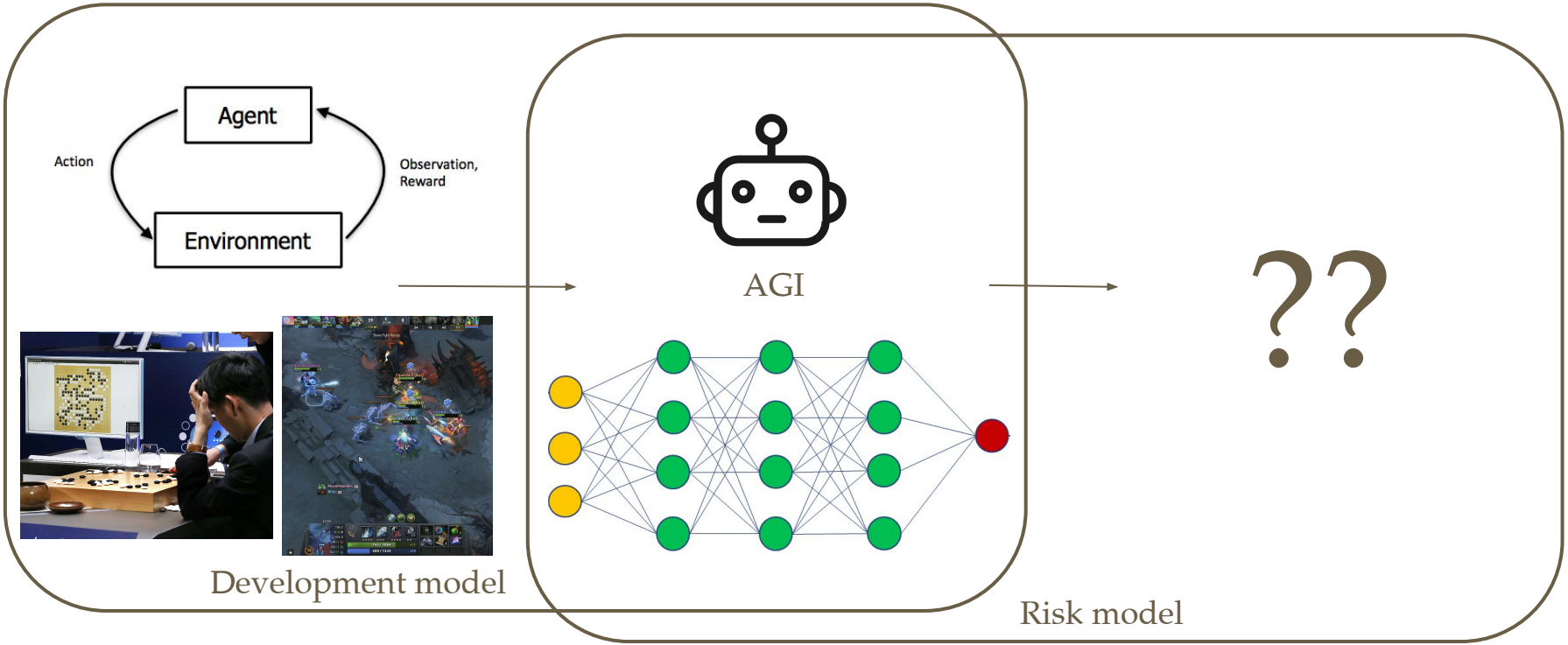
How likely is it that AGI is well-described as a utility maximizer?

- ¯_(ツ)_/¯ (no development model!)

What solutions could avert risk from utility maximizers?

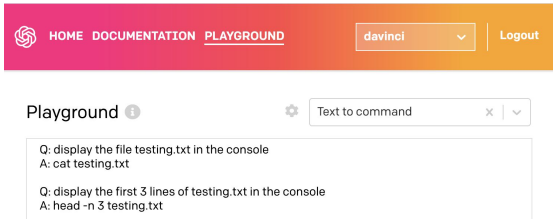
- Don't *maximize*: instead use mild optimization, impact regularization, etc.
- Have the right *utility*: learn the goal from human behavior

Threat model: deep RL

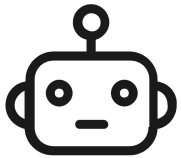


Threat model: language

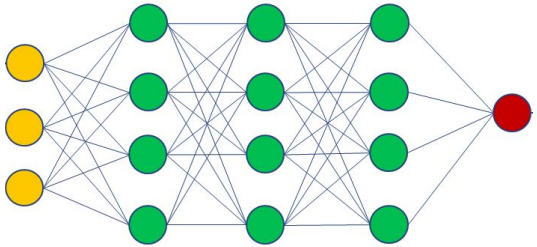
The chef cooked the ___?___
Once ___?___



Development model



AGI



??

Risk model

Threat model: deep RL / language

How likely is it that AGI arises from deep RL / language?

- Both seem plausible

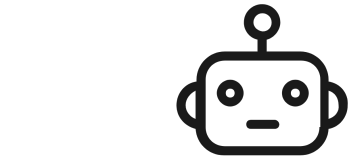
What solutions could avert risk from neural networks?

- ¯_(ツ)_/¯ (no risk model!)

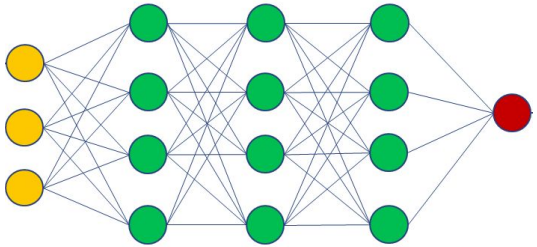
My threat model: bad generalization

Neural networks, trained on diverse tasks, develop general skills

Large language models and deep RL-based AGI are both instances of this



AGI



Development model

Something changes about the world (e.g. a pandemic)

The AI system uses its general reasoning to adapt (aka generalize)

While it continues to have high impact, it no longer does what we want

Risk model

My threat model: bad generalization

How likely is it that AGI arises from neural nets trained on diverse data?

- Debatable, but I think more likely than not

What solutions could avert risk from bad generalization?

- Training on the right objective
- Interpretability to ensure the network learned the right concept(s)
- Adversarial training to look for cases of bad generalization
- Choosing datasets / architectures that lead to the generalization we want
- Identifying off-distribution scenarios and reverting to a safe baseline policy

Takeaways

1. *Development models are crucial for solutions to apply to real systems.*
2. *Risk models are important for finding solutions.*
3. *We should be using the same notion of AGI for these two models.*
4. *We don't yet have compelling examples of these models.*