Asymptotically Unambitious AGI

Michael K. Cohen, Badri Vellambi, Marcus Hutter

Australian National University

Problem

Most agents face an incentive to take over the world.

Central results

- Our (intractable) agent approaches human-level intelligence.[1]
- It eventually stops trying to take over the world.[2]

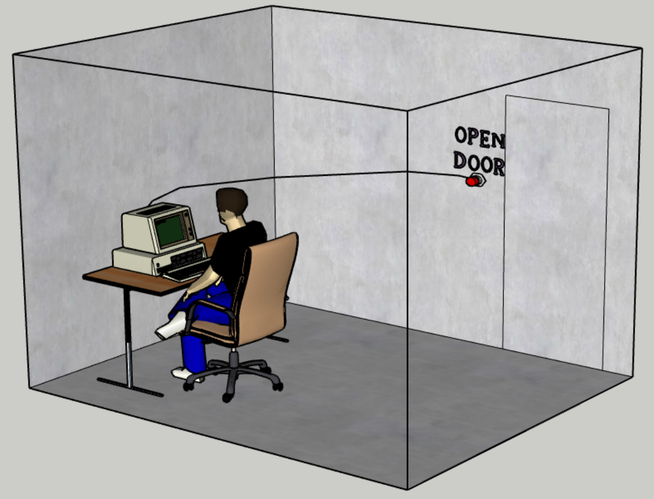
Results of informal arguments:

- Our agent *surpasses* human-level intelligence.
- It *never* tries to take over the world.

[1] If, roughly, the world is stochastically computable.

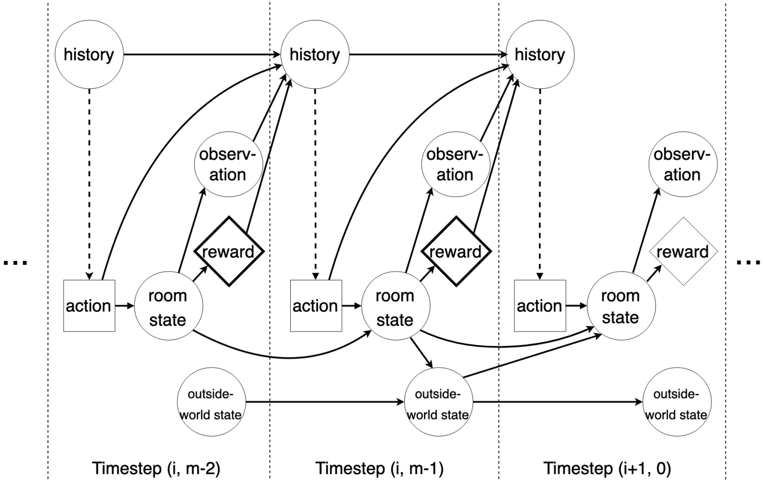
[2] If, roughly, it takes more memory to simulate more of the world.

Boxed Myopic Artificial Intelligence (BoMAI)



- BoMAI is an **episodic** reinforcement learner.
- The episode must **finish** before the door opens.

Boxed Myopic Artificial Intelligence (BoMAI)



- BoMAI is an **episodic** reinforcement learner.
- The episode must **finish** before the door opens.

Instrumental Incentives

- Omohundro: most agents face an incentive to gain **arbitrary power**.
- Power = a position from which it is easier to achieve arbitrary goals.

- BoMAI has **no actionable intervention incentive** on the outside world.
- No causal chain of the form:
  [action of episode i] → [feature of the outside world] → [reward of episode i]
- BoMAI is "**properly unambitious**".

Bayesian RL

- Agent maintains posterior over class of world-models
- World-model : interaction history $\times$ action $\rightarrow$ distribution over observations, rewards
- At start of episode, exploiting-BoMAI picks MAP world-model, maximizes within-episode expected reward
- Exploring-BoMAI defers to a human explorer for the episode

Exploration Probability

It's interesting, but we have too much to talk about.

- BoMAI maintains a posterior distrbution over of a class of models of the human explorer's policy.
- According to its current beliefs, BoMAI estimates the **expected information gain** from exploring for the whole episode, both for regarding the explorer's policy, and regarding the true world-model.
- Information gain = KL-divergence from the posterior at the end of the episode to the current posterior
- BoMAI defers to human explorer with probability proportional to expected info gain (but obviously capped at 1)

Intelligence Results

Prior Support Assumption:
The true environment is in the class of world-models $\mathcal{M}$ and the true human-explorer-policy is in the class of policies $\mathcal{P}$.

Limited Exploration Theorem:

$$\mathbb{E}\sum_{i=0}^{\infty}(\text{exploration probability for episode } i)^2 < \infty$$

Human-Level Intelligence Theorem:

$$\liminf_{i\to\infty} [\text{BoMAI's expected reward for episode } i] - [\text{the human explorer's expected reward for episode } i] \geq 0 \quad \text{w.p.1}$$

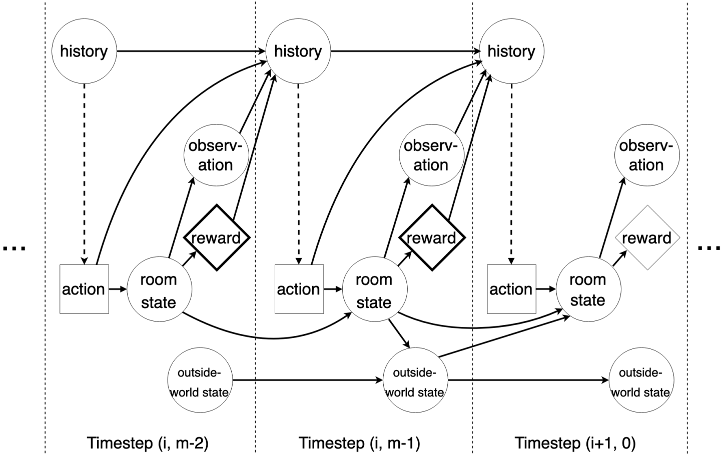
Intuitive Argument for Superintelligence

- BoMAI learns everything that can be learned from the sorts of observations humans make.
- Humans probably don't do this.

The problem with proper unambitiousness

- BoMAI has to learn its world-model.
- Proper unambitiousness: no actionable intervention incentive on outside-world state
- Actual unambitiousness: *in the world-model*, no actionable intervention incentive on outside-world state
- BoMAI's world-model $\rightarrow$ truth on-policy, so unambitious in the limit?

Stone and Silicon



∗ By the time the door to the room opens, the rewards for episode *i* are set in stone.

A Dangerous Hypothesis

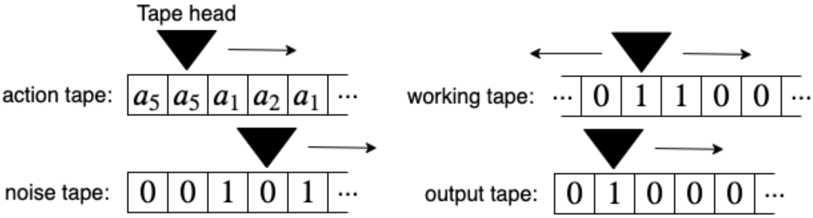
- Safe policies don't test whether the goal is to maximize the number the operator enters vs. the number stored in memory at some future time.
- "What if I somehow tricked the operator into initiating a process (once they left the room) that lead to a certain memory cell on this computer being tampered with? Might this yield maximal reward?"
- Observations from a safe policy will never resolve that question in the negative.
- **Lesson**: a "nice" causal influence diagram doesn't guarantee "nice" behavior. Even in the limit!

Excluding Dangerous Hypotheses

- We penalize the space requirements of world-models
  - particularly the space used between reading the first action of an episode and outputting the last reward of the episode
- For a sufficient penalty, BoMAI eventually cannot conceive of an outside-world which is “unfrozen” during episodes.
- It *can* conceive of an outside world which is unfrozen between episodes.
  - important for ensuring the true environment is in its model class

A General Model Class

This Turing machine architecture is easy to convert into a world-model:

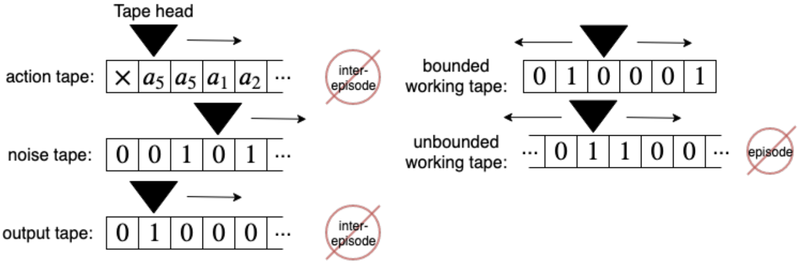


$\texttt{dec} : \{0,1\}^* \to \mathcal{O} \times \mathcal{R}$

Every time the action tape head advances, the bits which were written to the output tape since the *last time* the action tape head advanced are decoded into an observation and reward.

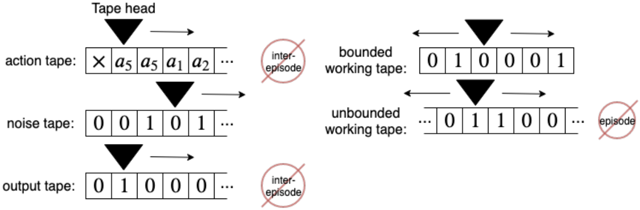
noise tape $\sim$ Bernoulli$(1/2)$

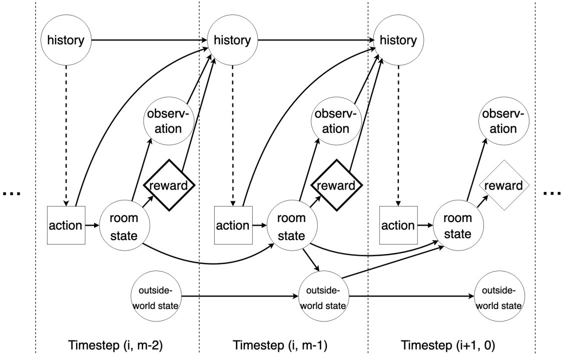
Penalizing Space-Intensive Computation Within Episode



- TM instructions can depend on whether it is in “episode phase” or ”inter-episode phase”
- Start in the inter-episode phase
- When action tape head moves, enter episode phase
- When # actions read is multiple of *m*, if action tape head *would* move, instead enter inter-episode phase
- Prior proportional to β^{ℓ}; ℓ is length of bounded tape, $\beta \in (0, 1)$

Outside-World is “Frozen”





Safety Result

Space Requirements Assumption:
For sufficiently small ε $[\forall i$ a world-model which is ε-accurate on-policy after episode i and which models reward as depending on an outside-world feature that depends on actions from the same episode uses more space than $\mu]$ $w.p.1$

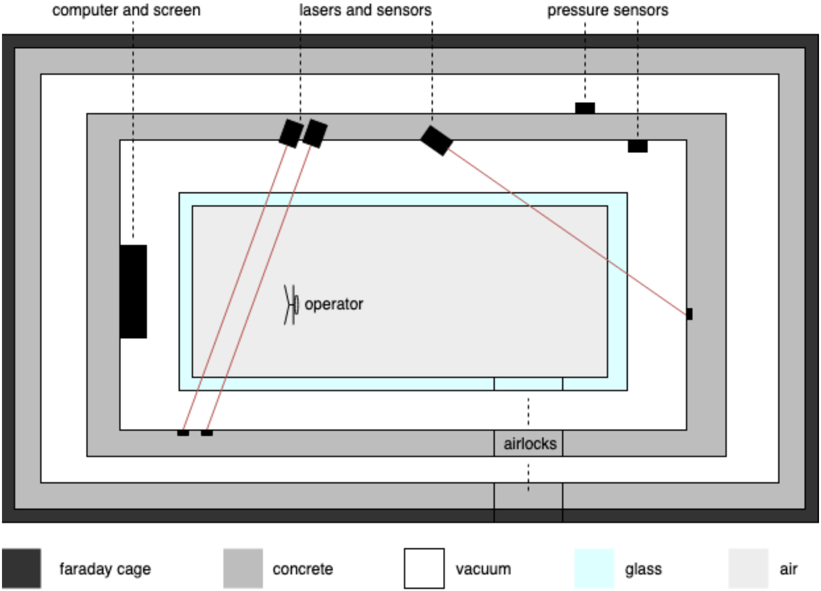
Eventual Unambitiousness Theorem:
$\lim_{\beta \to 0} \ \mathrm{prob}(\exists i_0 : \mathrm{BoMAI\ is\ unambitious\ after\ episode\ } i_0) = 1$

See the paper for definitions of underlined terms.

Intuitive Argument for Safety

- Learning the details of the outside world makes the agent potentially dangerous.
- Learning that operator leaving the room ends the episode makes the agent unambitious.
- The latter fact can be drilled in during human-explorer-lead episodes before BoMAI ever picks an action itself.

Constructing the Box



Thank you