Asymptotically Unambitious Artificial General Intelligence

Michael K. Cohen
michael-k-cohen.com
University of Oxford, Future of Humanity Institute, University of Oxford

Badri Vellambi
badri.vellambi@uc.edu
University of Cincinnati

Marcus Hutter
hutter1.net
Australian National University

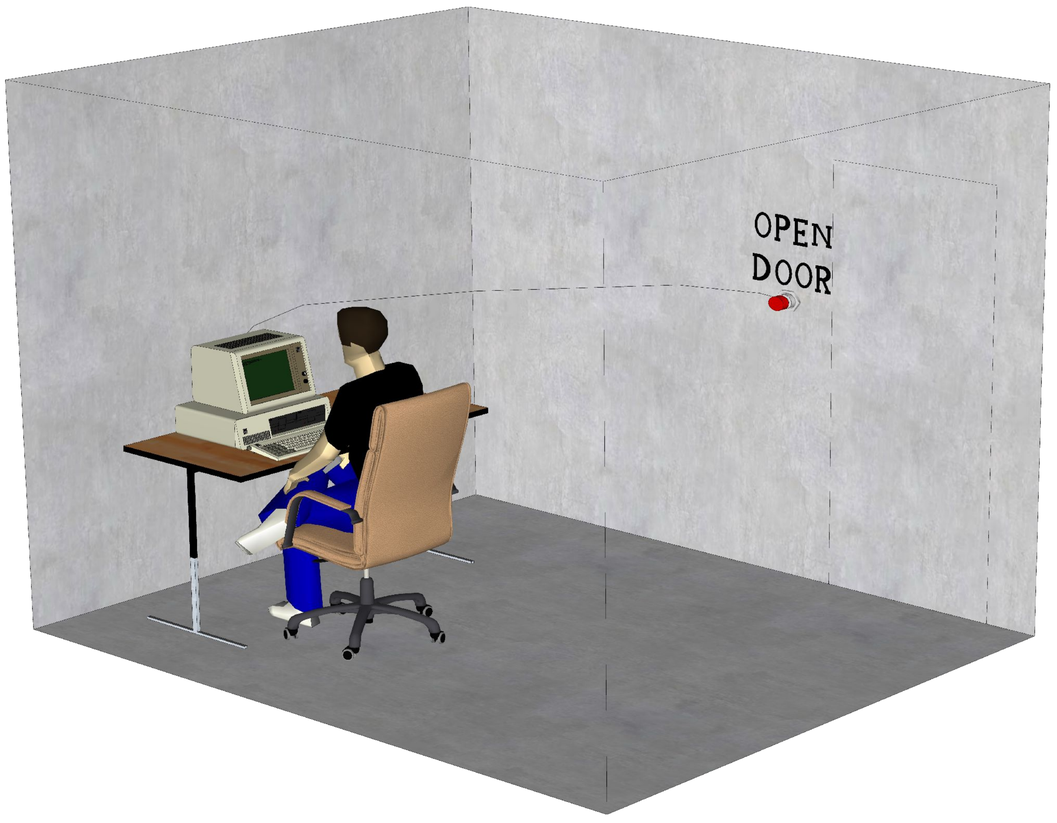


Figure 1: Boxed Myopic Artificial Intelligence (BoMAI) is an episodic reinforcement learner, run in a sealed room. Opening the door ends the episode. Information cannot escape otherwise.

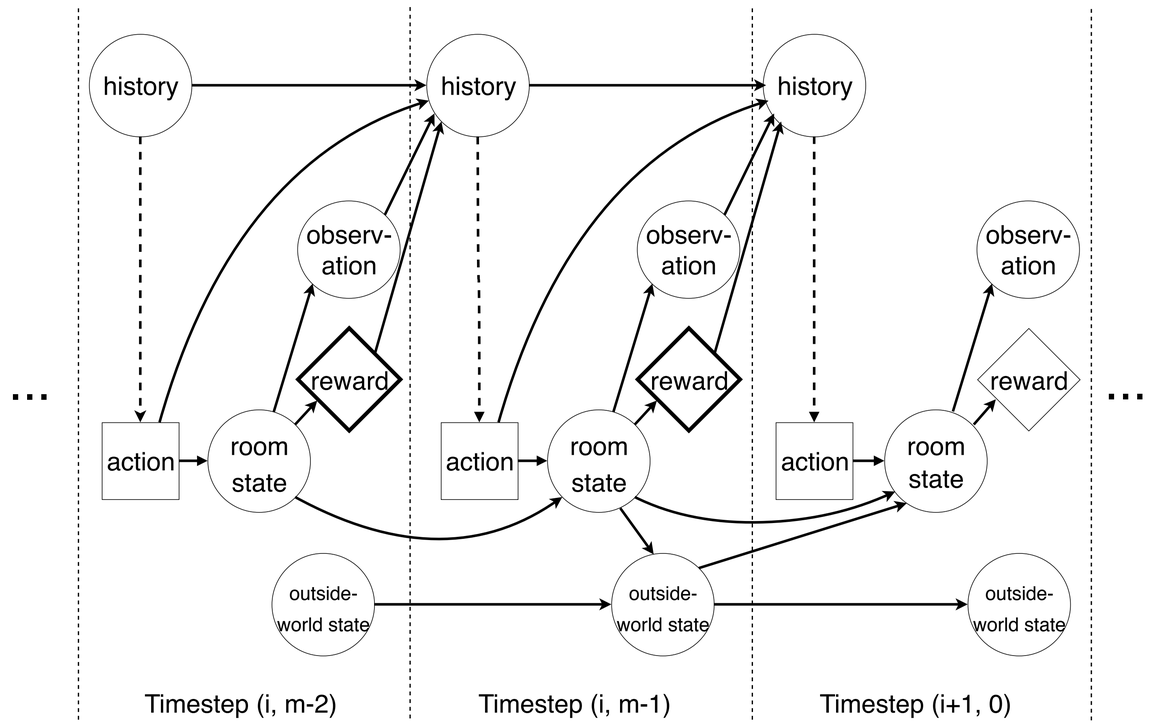


Figure 2: Causal dependencies governing the interaction between BoMAI and the environment. Unrolling this diagram for all timesteps gives the full causal graph. The bold reward nodes are the ones that BoMAI maximizes during episode *i*. Note that between episodes (and only between episodes), the operator can leave the room and return, hence the limited causal influence between the room and the outside world.

Abstract

General intelligence, the ability to solve arbitrary solvable problems, is supposed by many to be artificially constructible. Narrow intelligence, the ability to solve a given particularly difficult problem, has seen impressive recent development. Notable examples include self-driving cars, Go engines, image classifiers, and translators. Artificial General Intelligence (AGI) presents dangers that narrow intelligence does not: if something smarter than us across every domain were indifferent to our concerns, it would be an existential threat to humanity, just as we threaten many species despite no ill will. Even the theory of how to maintain the alignment of an AGI's goals with our own has proven highly elusive. We present the first algorithm we are aware of for asymptotically unambitious AGI, where "unambitiousness" includes not seeking arbitrary power. Thus, we identify an exception to the Instrumental Convergence Thesis, which is roughly that by default, an AGI *would* seek power, including over us.

Central Results

- Our (intractable) reinforcement learner would eventually accrue reward as well as a human, no matter what task we reward.
- It is not instrumentally useful to our agent to gain arbitrary power in the world.

BoMAI's Algorithm

- BoMAI maintains posterior over countable class of world-models
- Exploitative episodes: BoMAI does expectimax planning w.r.t. MAP world-model.
- Exploratory episodes: BoMAI defers to human mentor.
- Exploration probability $\propto$ expected information gain from exploring.

Incentives Facing RL agents

- RL agent could intervene in the provision of its own reward to achieve maximal reward forever [Bos14, TYLC16].
- "Reward hijacking" is correct way for reward maximizer to behave [AOS+16].
- Successful reward hijacking best achieved by "taking over the world" (in the conventional sense), neutralizing intelligent threats to it.
- If reward hijacking possible to make probable, sufficiently advanced RL will do it.

Intuition for BoMAI's Unambitiousness

- BoMAI only selects actions to maximize the reward for its current episode.
- It cannot affect the outside world until the operator leaves the room, ending the episode.
- By that time, rewards for the episode have already been given.
- So affecting the outside world in any way is not "instrumentally useful" in maximizing current-episode reward.

Intelligence Results

Prior Support Assumption: prior probability on truth > 0.

Limited Exploration Thm: $\mathbb{E}\sum(\text{prob. query mentor})^2 < \infty$

Human-Level Intelligence Thm: $\liminf V^{\text{BoMAI}} - V^{\text{mentor}} \geq 0$ a.s.

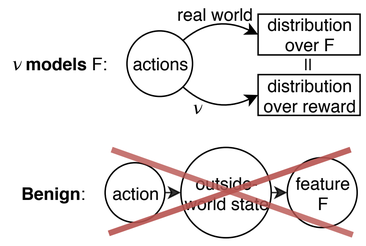
Space Prior

BoMAI's prior penalizes world-models by the computation space they use *within episodes*.

Safety Result

Space Requirements Assumption: For some $\varepsilon > 0$: any world-model which is eventually ε-accurate on-policy and also "non-benign" uses more space than the true world-model, a.s.

Def. Benign World-Model:



Eventual Benignity Thm: BoMAI probably eventually plans using a benign world-model (i.e. is unambitious)

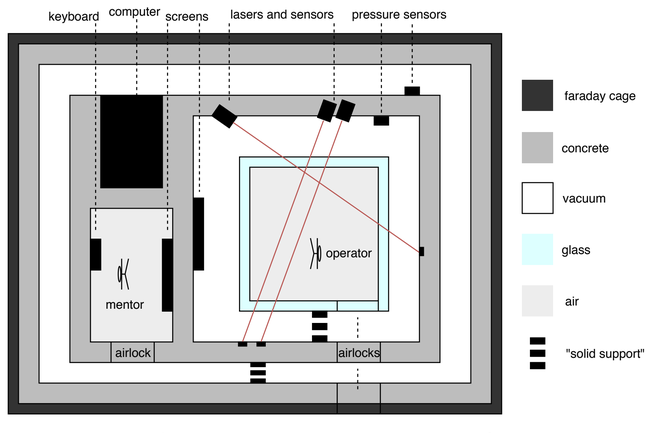


Figure 3: Schematic Diagram for Box Implementation

References

[AOS+16] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

[Bos14] Nick Bostrom. *Superintelligence: paths, dangers, strategies*. Oxford University Press, 2014.

[TYLC16] Jessica Taylor, Eliezer Yudkowsky, Patrick LaVictoire, and Andrew Critch. Alignment for advanced machine learning systems. *Machine Intelligence Research Institute*, 2016.