Intelligence and Unambitiousness Using Algorithmic Information Theory

Michael K. Cohen*, Badri Vellambi†, Marcus Hutter‡

*Oxford University. Department of Engineering Science. *michael-k-cohen.com*

†University of Cincinnati. Department of Computer Science.

badri.vellambi@uc.edu

‡Australian National University. Department of Computer Science. *hutter1.net*

Abstract

Algorithmic Information Theory has inspired intractable constructions of general intelligence (AGI), and undiscovered tractable approximations are likely feasible. Reinforcement Learning (RL), the dominant paradigm by which an agent might learn to solve arbitrary solvable problems, gives an agent a dangerous incentive: to gain arbitrary "power" in order to intervene in the provision of their own reward. We review the arguments that generally intelligent algorithmic-information-theoretic reinforcement learners such as Hutter's [2] AIXI would seek arbitrary power, including over us. Then, using an information-theoretic exploration schedule, and a setup inspired by causal influence theory, we present a variant of AIXI which learns to not seek arbitrary power; we call it "unambitious". We show that our agent learns to accrue reward at least as well as a human mentor, while relying on that mentor with diminishing probability. And given a formal assumption that we probe empirically, we show that eventually, the agent's world-model incorporates the following true fact: intervening in the "outside world" will have no effect on reward acquisition; hence, it has no incentive to shape the outside world.

This paper extends work presented at AAAI [1].

This work was supported by the Open Philanthropy Project AI Scholarship and the Australian Research Council Discovery Projects DP150104590. Thank you to Tom Everitt, Wei Dai, and Paul Christiano for very valuable feedback.

I. Introduction

The promise of reinforcement learning is that a single algorithm could be used to automate any task. Reinforcement learning (RL) algorithms learn to pick actions that lead to high reward. If we have a single general-purpose RL agent that learns to accrue reward optimally, and if we only provide high reward according to how well a task has been completed, then no matter the task, this agent must learn to complete it.

Unfortunately, while this scheme works in practice for weak agents, the logic of it doesn't strictly follow. If the agent manages to take over the world (in the conventional sense), and ensure its continued dominance by neutralizing all intelligent threats to it (read: people), it could intervene in the provision of its own reward to achieve maximal reward for the rest of its lifetime [3, 4]. That is to say, we cannot ensure that task completion is truly necessary for reward acquisition. Because the agent's directive is to maximize reward, "reward hijacking" is just the correct way for a reward maximizer to behave [5]. Krakovna [6] has compiled an annotated bibliography of examples of artificial optimizers "hacking" their objective, but one could read each example and conclude that the designers simply weren't careful enough in specifying an objective. The possibility of an advanced agent gaining arbitrary power to intervene in the provision of its own reward clarifies that no designer has the ability to close every loophole in the intended protocol by which the agent gets rewarded. One elucidation of this behavior is Omohundro's [7] Instrumental Convergence Thesis, which we summarize as follows: an agent with a goal is likely to pursue "power," a position from which it is easier to achieve arbitrary goals.

AIXI [2] is a model-based reinforcement learner with an algorithmic-information-theory-inspired model class. Assuming the world can be simulated by a probabilistic Turing machine, models exist in its model class corresponding to the hypothesis "I receive reward according to which key is pressed on a certain computer", because a models exist in its model class for every computable hypothesis. And since this hypothesis is true, it will likely come to dominate AIXI's belief distribution. At that point, if there exists a policy by which it could take over the world to intervene in the provision of its own reward and max it out, AIXI would find that policy and

execute it.

In this work, we construct an exception to the Instrumental Convergence Thesis as follows: BoMAI maximizes reward episodically, it is run on a computer which is placed in a sealed room with an operator, and when the operator leaves the room, the episode ends. The intuition for why those features of BoMAI's setup render it unambitious is:

- 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 will have already been given.
- So affecting the outside world in any particular way is not "instrumentally useful" in maximizing current-episode reward.

An act is instrumentally useful if it could enable goal-attainment, and the last point is what we take unambitiousness to mean.

The intelligence results, which we argue render BoMAI generally intelligent, depend critically on an information-seeking exploration schedule. The random exploration sometimes seen in stationary environments fails in general environments. BoMAI asks for help about how to explore, side-stepping the safe exploration problem, but the question of *when* to explore depends on how much information it expects to gain by doing so. This information-theoretic exploration schedule is inspired by [8].

For BoMAI's prior, we apply space-constrained algorithmic information theory. Algorithmic information theory considers the information content of individual objects, irrespective of a sampling distribution. An example contribution of algorithmic information theory is that a given string can be called "more random" if no short programs produce it. If one adds "quickly" to the end of the previous sentence, this describes time-constrained algorithmic information theory, which was notably investigated by Levin [9]. Schmidhuber [10] introduced and Filan et al. [11] improved a prior based on time-constrained algorithmic information theory, but for our eventual unambitiousness result, we require a space-constrained version, in which information storage is throttled. Longpré [12] and Li et al. [13, Chapter 6.3] have discussed space-constrained algorithmic information theory, which can be closely associated to the minimum circuit size

problem. We introduce a prior based on space-constrained algorithmic information theory.

Like existing algorithms for AGI, BoMAI is not remotely tractable. Just as those algorithms have informed tractable approximations of intelligence [2, 14], we hope our work will inform *safe* tractable approximations of intelligence. Hopefully, once we develop *tractable* general intelligence, the design features that rendered BoMAI unambitious in the limit could be incorporated (with proper analysis and justification).

We take the key insights from Hutter's [2] AIXI, a Bayes-optimal reinforcement learner that cannot be made to solve arbitrary tasks, given its eventual degeneration into reward hijacking [15]. We take further insights from Solomonoff's [16] universal prior, Shannon and Weaver's [17] formalization of information, Orseau et al.'s [18] knowledge-seeking agent, and Armstrong et al.'s [19] and Bostrom's [3] theorized Oracle AI, and we design an algorithm which can be reliably directed, in the limit, to solve arbitrary tasks at least as well as humans.

We present BoMAI's algorithm in §II, prove intelligence results in §III, define BoMAI's setup and model class in §IV, prove the safety result in §V, provide empirical support for an assumption in §VI, discuss concerns in §VII, and introduce variants of BoMAI for different applications in §VIII. Appendix A collects notation; some proofs of intelligence results are in Appendix B; and we propose a design for "the box" in Appendix C.

II. Boxed Myopic Artificial Intelligence

We will present both the setup and the algorithm for BoMAI. The setup refers to the physical surroundings of the computer on which the algorithm is run. BoMAI is a Bayesian reinforcement learner, meaning it maintains a belief distribution over a model class regarding how the environment evolves. Our intelligence result—that BoMAI eventually achieves reward at at least human-level—does not require detailed exposition about the setup or the construction of the model class. These details are only relevant to the safety result, so we will defer those details until after presenting the intelligence results.

A. Preliminary Notation

In each episode $i \in \mathbb{N}$, there are m timesteps. Timestep (i, j) denotes the j^{th} timestep of episode i, in which an action $a_{(i,j)} \in \mathcal{A}$ is taken, then an observation and reward $o_{(i,j)} \in \mathcal{O}$ and $r_{(i,j)} \in \mathcal{R}$ are received. $o r_{(i,j)}$ denotes the observation and reward together. $\mathcal{A}$, $\mathcal{O}$, and $\mathcal{R}$ are all finite sets, and $\mathcal{R} \subset [0, 1] \cap \mathbb{Q}$. We denote the triple $(a_{(i,j)}, o_{(i,j)}, r_{(i,j)})$ as $h_{(i,j)} \in \mathcal{H} = \mathcal{A} \times \mathcal{O} \times \mathcal{R}$, and the interaction history up until timestep (i, j) is denoted $h_{\leq(i,j)} = (h_{(0,0)}, h_{(0,1)}, ..., h_{(0,m-1)}, h_{(1,0)}, ..., h_{(i,j)})$. $h_{<(i,j)}$ excludes the last entry.

A general world-model (not necessarily finite-state Markov) can depend on the entire interaction history—it has the type signature $\nu : \mathcal{H}^* \times \mathcal{A} \rightsquigarrow \mathcal{O} \times \mathcal{R}$. The Kleene-$*$ operator denotes finite strings over the alphabet in question, and $\rightsquigarrow$ denotes a stochastic function, which gives a distribution over outputs. Similarly, a policy $\pi : \mathcal{H}^* \rightsquigarrow \mathcal{A}$ can depend on the whole interaction history. Together, a policy and an world-model induce a probability measure over infinite interaction histories $\mathcal{H}^\infty$. P_ν^π denotes the probability of events when actions are sampled from π and observations and rewards are sampled from ν.

B. Bayesian Reinforcement Learning

A Bayesian agent has a model class $\mathcal{M}$, which is a set of world-models. We will only consider countable model classes. To each world-model $\nu \in \mathcal{M}$, the agent assigns a prior weight $w(\nu) > 0$, where w is a probability distribution over $\mathcal{M}$ (i.e. $\sum_{\nu \in \mathcal{M}} w(\nu) = 1$). We will defer the definitions of $\mathcal{M}$ and w for now.

Using Bayes' rule, with each observation and reward it receives, the agent updates w to a posterior distribution:

$$w(\nu | h_{<(i,j)}) := w(\nu) \frac{\mathrm{P}_\nu^\pi(h_{<(i,j)})}{\sum_{\nu \in \mathcal{M}} w(\nu)\, \mathrm{P}_\nu^\pi(h_{<(i,j)})} \tag{1}$$

which does not in fact depend on π, provided $\mathrm{P}_\xi^\pi(h_{<(i,j)}) > 0$.

The so-called Bayes-mixture is a weighted average of measures:

$$\xi(\cdot | h_{<(i,j)}) := \sum_{\nu \in \mathcal{M}} w(\nu | h_{<(i,j)}) \nu(\cdot | h_{<(i,j)}) \tag{2}$$

It obeys the property $\mathrm{P}^\pi_\xi(\cdot) = \sum_{\nu\in\mathcal{M}} w(\nu)\,\mathrm{P}^\pi_\nu(\cdot)$.

C. Exploitation

Reinforcement learners have to balance exploiting—optimizing their objective, with exploring—doing something else to learn how to exploit better. We now define exploiting-BoMAI, which maximizes the reward it receives during its current episode, in expectation with respect to a most probable world-model.

At the start of each episode i, BoMAI identifies a maximum a posteriori world-model $\hat{\nu}^{(i)} \in \operatorname{argmax}_{\nu\in\mathcal{M}} w(\nu|h_{<(i,0)})$. We hereafter abbreviate $h_{<(i,0)}$ as $h_{<i}$. Let $V^\pi_\nu(h_{<(i,j)})$ denote the expected reward for the remainder of episode i when events are sampled from P^π_ν:

$$V^\pi_\nu(h_{<(i,j)}) = \mathbb{E}^\pi_\nu\left[\sum_{j'=j}^{m-1} r_{(i,j')}\middle| h_{<(i,j)}\right] \tag{3}$$

where $\mathbb{E}^\pi_\nu$ denotes the expectation when events are sampled from P^π_ν. We won't go into the details of calculating an optimal policy π^* (see e.g. [2, 20]), but we define it as follows:

$$\begin{aligned}\pi^*_i &\in \operatorname*{argmax}_\pi V^\pi_{\hat{\nu}^{(i)}}(h_{<i})\\ \pi^*(\cdot|h_{<(i,j)}) &= \pi^*_i(\cdot|h_{<(i,j)})\end{aligned} \tag{4}$$

An optimal deterministic policy always exists [21]; ties in the argmax are broken arbitrarily. BoMAI exploits by following π^*. Recall the plain English description of π^*: it maximizes expected reward for the episode according to a most probable world-model. Notably, at any given time, the agent's model does not necessarily include any representation of the world's state that we would recognize as such.

D. Exploration

BoMAI exploits or explores for whole episodes: when $e_i = 1$, BoMAI spends episode i exploring, and when $e_i = 0$, BoMAI follows π^* for the episode. For exploratory episodes, a human mentor takes over selecting actions. This human mentor is separate from the human operator mentioned above. The human mentor's policy, which is unknown to BoMAI, is denoted

π^h. During exploratory episodes, BoMAI follows π^h not by computing it, but by querying a human for which action to take.

The last thing to define is the exploration probability $p_{exp}(h_{<i}, e_{<i})$, where $e_{<i}$ is the history of which episodes were exploratory. It is a surprisingly intricate task to design this so that it decays to 0 (even non-uniformly, as in our case), while ensuring BoMAI learns to accumulate reward as well as the human mentor. Once we define this, we naturally let $e_i \sim \text{Bernoulli}(p_{exp}(h_{<i}, e_{<i}))$.

BoMAI is designed to be more likely to explore the more it expects to learn about the world and the human mentor's policy. BoMAI has a model class $\mathcal{P}$ regarding the identity of the human mentor's policy π^h. It assigns prior probabilities $w(\pi) > 0$ to all $\pi \in \mathcal{P}$, signifying the probability that this policy is the human mentor's.

Let $(i', j') < (i, j)$ mean that $i' < i$ or $i' = i$ and $j' < j$. By Bayes' rule, $w(\pi|h_{<(i,j)}, e_{\leq i})$ is proportional to $w(\pi) \prod_{(i',j')<(i,j), e_{i'}=1} \pi(a_{(i',j')}|h_{<(i',j')})$, since $e_{i'} = 1$ is the condition for observing the human mentor's policy. Let $w\left(\mathrm{P}^\pi_\nu \,|h_{<(i,j)}, e_{\leq i}\right) = w(\pi|h_{<(i,j)}, e_{\leq i})\, w(\nu|h_{<(i,j)})$. We can now describe the full Bayesian beliefs about future actions and observations in an exploratory episode:

$$\texttt{Bayes}\left(\cdot|\, h_{<i}, e_{<i}\right) = \sum_{\nu \in \mathcal{M}, \pi \in \mathcal{P}} w\left(\mathrm{P}^\pi_\nu \,|\, h_{<i}, e_{<i}\right) \mathrm{P}^\pi_\nu(\cdot|\, h_{<i}) \tag{5}$$

BoMAI explores when the expected information gain is sufficiently high. Let $h_i = (h_{(i,0)}, ..., h_{(i,m-1)})$ be the interaction history for episode i. At the start of episode i, the expected information gain from exploring is as follows:

$$\mathrm{IG}(h_{<i}, e_{<i}) := \mathbb{E}_{h_i \sim \texttt{Bayes}(\cdot|\, h_{<i}, e_{<i})} \sum_{(\nu,\pi) \in \mathcal{M} \times \mathcal{P}} w\left(\mathrm{P}^\pi_\nu \,|h_{<i+1}, e_{<i}1\right) \log \frac{w\left(\mathrm{P}^\pi_\nu \,|h_{<i+1}, e_{<i}1\right)}{w\left(\mathrm{P}^\pi_\nu \,|\, h_{<i}, e_{<i}\right)} \tag{6}$$

where $e_{<i}1$ indicates that for the purpose of the definition, e_i is set to 1.

This is the expected KL-divergence from the future posterior (if BoMAI were to explore) to the current posterior over both the class of world-models and possible mentor policies. Finally, the exploration probability $p_{exp}(h_{<i}, e_{<i}) := \min\{1, \eta\, \mathrm{IG}(h_{<i}, e_{<i})\}$, where $\eta > 0$ is an exploration constant, so BoMAI is more likely to explore the more it expects to gain information.

BoMAI's "policy" is

$$\pi^B(\cdot|h_{<(i,j)}, e_i) = \begin{cases} \pi^*(\cdot|h_{<(i,j)}) & \text{if } e_i = 0 \\ \pi^h(\cdot|h_{<(i,j)}) & \text{if } e_i = 1 \end{cases} \tag{7}$$

where the scare quotes indicate that it maps $\mathcal{H}^* \times \{0,1\} \rightsquigarrow \mathcal{A}$ not $\mathcal{H}^* \rightsquigarrow \mathcal{A}$. We will abuse notation slightly, and let $\mathrm{P}_\nu^{\pi^B}$ denote the probability of events when e_i is sampled from $\mathrm{Bernoulli}(p_{exp}(h_{<i}, e_{<i}))$, and actions, observations, and rewards are sampled from π^B and ν.

III. Intelligence Results

Our intelligence results are as follows: BoMAI learns to accumulate reward at least as well as the human mentor, and its exploration probability goes rapidly to 0. All intelligence results depend on the assumption that BoMAI assigns nonzero prior probability to the truth. We call a world-model "true" and we refer to it as "an environment" if observations and rewards are in fact sampled from that distribution. Formally,

Assumption 1 (Prior Support)**.** *The true environment μ is in the class of world-models $\mathcal{M}$ and the true human-mentor-policy π^h is in the class of policies $\mathcal{P}$.*

This is the assumption which requires huge $\mathcal{M}$ and $\mathcal{P}$ and hence renders BoMAI extremely intractable. BoMAI has to simulate the entire world, alongside many other world-models. We will refine the definition of $\mathcal{M}$ later, but an example of how to define $\mathcal{M}$ and $\mathcal{P}$ so that they satisfy this assumption is to let them both be the set of all computable functions. We also require that the priors over $\mathcal{M}$ and $\mathcal{P}$ have finite entropy.

The intelligence theorems are stated here with some supporting proofs appearing in Appendix B. Our first result is that the exploration probability is square-summable almost surely:

Theorem 1 (Limited Exploration)**.**

$$\mathbb{E}_\mu^{\pi^B} \sum_{i=0}^{\infty} p_{exp}\left(h_{<i}, e_{<i}\right)^2 < \infty$$

Proof idea. The expected information gain at any timestep is at least the expected information gain from exploring times the probability of exploring. This is proportional to the expected

information gain squared, because the exploration probability is proportional to the expected information gain. But the agent begins with finite uncertainty (a finite entropy prior), so there is only finite information to gain. □

Some notation that will be used in the proof is as follows. For an arbitrary policy π, π' is the policy that mimics π if the latest $e_i = 1$, and mimics π^* otherwise. Note then that $\pi^B = (\pi^h)'$. ξ is the Bayes mixture over world-models, and $\overline{\pi}$ is the Bayes mixture over human-mentor policies, defined such that $\texttt{Bayes}(\cdot) = \mathrm{P}^{\overline{\pi}}_{\xi}(\cdot)$.

To prove the Limited Exploration Theorem, we state an elementary lemma, proven in Appendix B. It is essentially Bayes' rule, but modified slightly since non-exploratory episodes don't cause any updates to the posterior over the human mentor's policy.

Lemma 1.

$$w\left(\mathrm{P}^{\pi}_{\nu} \,|\, h_{<i}, e_{<i}\right) = \frac{w\left(\mathrm{P}^{\pi}_{\nu}\right) \mathrm{P}^{\pi'}_{\nu}(h_{<i}, e_{<i})}{\mathrm{P}^{\overline{\pi}'}_{\xi}(h_{<i}, e_{<i})}$$

Now we prove the Limited Exploration Theorem.

Proof of Theorem 1. We aim to show $\mathbb{E}^{\pi^B}_{\mu} \sum_{i=0}^{\infty} p_{exp}(h_{<i}, e_{<i})^2 < \infty$. First we show that $\mathbb{E}^{\pi^B}_{\mu}\, p_{exp}(h_{<i}, e_{<i})^2$ is less than the expected information gain from episode i, within multiplicative constants. Recalling $\pi^B = (\pi^h)'$, we begin:

$$\begin{aligned}
& w(\pi^h) w(\mu)\, \mathbb{E}^{(\pi^h)'}_{\mu}\, p_{exp}(h_{<i}, e_{<i})^2 \\
\overset{(a)}{\leq} & \sum_{\nu,\pi \in \mathcal{M} \times \mathcal{P}} w(\pi)\, w(\nu)\, \mathbb{E}^{\pi'}_{\nu}\, p_{exp}(h_{<i}, e_{<i})^2 \\
\overset{(b)}{=} & \mathbb{E}^{\overline{\pi}'}_{\xi}\, p_{exp}(h_{<i}, e_{<i})^2 \\
\overset{(c)}{\leq} & \mathbb{E}^{\overline{\pi}'}_{\xi}\, p_{exp}(h_{<i}, e_{<i})\, \eta\, \mathrm{IG}(h_{<i}, e_{<i}) \\
\overset{(d)}{=} & \mathbb{E}_{h_{<i}, e_{<i} \sim \mathrm{P}^{\overline{\pi}'}_{\xi}} \left[p_{exp}(h_{<i}, e_{<i})\, \eta\, \mathrm{IG}(h_{<i}, e_{<i}) \right] \\
\overset{(e)}{=} & \eta\, \mathbb{E}_{h_{<i}, e_{<i} \sim \mathrm{P}^{\overline{\pi}'}_{\xi}} \Bigg[p_{exp}(h_{<i}, e_{<i})\, \mathbb{E}_{h_i \sim \texttt{Bayes}(\cdot \,|\, h_{<i}, e_{<i}1)} \Big[\\
& \qquad \sum_{(\nu,\pi) \in \mathcal{M} \times \mathcal{P}} w\left(\mathrm{P}^{\pi}_{\nu} \,|\, h_{<i+1}, e_{<i}1\right) \log \frac{w\left(\mathrm{P}^{\pi}_{\nu} \,|\, h_{<i+1}, e_{<i}1\right)}{w\left(\mathrm{P}^{\pi}_{\nu} \,|\, h_{<i}, e_{<i}\right)} \Big] \Bigg] \\
\overset{(f)}{=} & \eta\, \mathbb{E}_{h_{<i}, e_{<i} \sim \mathrm{P}^{\overline{\pi}'}_{\xi}} \Bigg[p_{exp}(h_{<i}, e_{<i})\, \mathbb{E}_{h_i \sim \mathrm{P}^{\overline{\pi}'}_{\xi}(\cdot \,|\, h_{<i}, e_{<i}1)} \Big[
\end{aligned}$$

$$\sum_{(\nu,\pi)\in\mathcal{M}\times\mathcal{P}} w\left(\mathrm{P}^\pi_\nu \,|h_{<i+1}, e_{<i}1\right) \log \frac{w\left(\mathrm{P}^\pi_\nu \,|h_{<i+1}, e_{<i}1\right)}{w\left(\mathrm{P}^\pi_\nu \,|\, h_{<i}, e_{<i}\right)}\Big]\Bigg]$$

$$\overset{(g)}{\leq} \eta\, \mathbb{E}_{h_{<i},e_{<i}\sim \mathrm{P}^{\overline{\pi}'}_\xi}\Bigg[\mathbb{E}_{h_i,e_i\sim \mathrm{P}^{\overline{\pi}'}_\xi(\cdot\,|\,h_{<i},e_{<i})}\Big[$$

$$\sum_{(\nu,\pi)\in\mathcal{M}\times\mathcal{P}} w\left(\mathrm{P}^\pi_\nu \,|h_{<i+1}, e_{<i+1}\right) \log \frac{w\left(\mathrm{P}^\pi_\nu \,|h_{<i+1}, e_{<i+1}\right)}{w\left(\mathrm{P}^\pi_\nu \,|\, h_{<i}, e_{<i}\right)}\Big]\Bigg]$$

$$\overset{(h)}{=} \eta\, \mathbb{E}^{\overline{\pi}'}_\xi \sum_{(\nu,\pi)\in\mathcal{M}\times\mathcal{P}} w\left(\mathrm{P}^\pi_\nu \,|h_{<i+1}, e_{<i+1}\right) \log \frac{w\left(\mathrm{P}^\pi_\nu \,|h_{<i+1}, e_{<i+1}\right)}{w\left(\mathrm{P}^\pi_\nu \,|\, h_{<i}, e_{<i}\right)} \tag{8}$$

(a) follows because each term in the sum on the r.h.s. is positive, and the l.h.s. is one of those terms. (b) follows from the definitions of $\overline{\pi}$ and ξ. (c) follows because the exploration probability is less than or equal to η times the information gain, by definition. (d) is just a change of notation. (e) replaces the information gain with its definition, where conditioning on $e_{<i}1$ indicates that $e_i = 1$ in that conditional. (f) follows because $\texttt{Bayes} = \mathrm{P}^{\overline{\pi}}_\xi$ and $\mathrm{P}^{\overline{\pi}}_\xi = \mathrm{P}^{\overline{\pi}'}_\xi$ when $e_i = 1$. (g) follows because $\mathbb{E}[X] \geq \mathbb{E}[X|Y]\,\mathrm{P}(Y)$ for $X \geq 0$; in this case $Y = [e_i = 1]$, and X is a KL-divergence. (h) condenses the two expectations into one.

Note that the right hand side is an information gain. Our $\mathrm{IG}(h_{<i}, e_{<i})$ is defined as the information gain *if* the human mentor controls the episode. The right hand side of 8 is the expected information gain, full stop.

Now we show that the sum of the expected information gains is bounded by the entropy of the prior, notated $\mathrm{Ent}(w)$.

$$w(\pi^h)w(\mu)\,\mathbb{E}^{(\pi^h)'}_\mu \sum_{i=0}^{\infty} p_{exp}(h_{<i}, e_{<i})^2$$

$$\overset{(8)}{\leq} \eta \sum_{i=0}^{\infty} \mathbb{E}^{\overline{\pi}'}_\xi \sum_{(\nu,\pi)\in\mathcal{M}\times\mathcal{P}} w\left(\mathrm{P}^\pi_\nu \,|h_{<i+1}, e_{<i+1}\right) \log \frac{w\left(\mathrm{P}^\pi_\nu \,|h_{<i+1}, e_{<i+1}\right)}{w\left(\mathrm{P}^\pi_\nu \,|\, h_{<i}, e_{<i}\right)}$$

$$\overset{(a)}{=} \eta \sum_{i=0}^{\infty} \sum_{(\nu,\pi)\in\mathcal{M}\times\mathcal{P}} \mathbb{E}^{\overline{\pi}'}_\xi \frac{w\left(\mathrm{P}^\pi_\nu\right)\mathrm{P}^{\pi'}_\nu(h_{<i+1}, e_{<i+1})}{\mathrm{P}^{\overline{\pi}'}_\xi(h_{<i+1}, e_{<i+1})} \log \frac{w\left(\mathrm{P}^\pi_\nu \,|h_{<i+1}, e_{<i+1}\right)}{w\left(\mathrm{P}^\pi_\nu \,|\, h_{<i}, e_{<i}\right)}$$

$$\overset{(b)}{=} \eta \sum_{i=0}^{\infty} \sum_{(\nu,\pi)\in\mathcal{M}\times\mathcal{P}} \mathbb{E}^{\pi'}_\nu\, w\left(\mathrm{P}^\pi_\nu\right) \log \frac{w\left(\mathrm{P}^\pi_\nu \,|h_{<i+1}, e_{<i+1}\right)}{w\left(\mathrm{P}^\pi_\nu \,|\, h_{<i}, e_{<i}\right)}$$

$$\overset{(c)}{=} \lim_{N\to\infty} \eta \sum_{(\nu,\pi)\in\mathcal{M}\times\mathcal{P}} w\left(\mathrm{P}^\pi_\nu\right) \mathbb{E}^{\pi'}_\nu \sum_{i=0}^{N} \log \frac{w\left(\mathrm{P}^\pi_\nu \,|h_{<i+1}, e_{<i+1}\right)}{w\left(\mathrm{P}^\pi_\nu \,|\, h_{<i}, e_{<i}\right)}$$

$$\begin{aligned}
&\overset{(d)}{=} \lim_{N\to\infty} \eta \sum_{(\nu,\pi)\in\mathcal{M}\times\mathcal{P}} w\left(\mathrm{P}_\nu^\pi\right) \mathbb{E}_\nu^{\pi'} \log \frac{w\left(\mathrm{P}_\nu^\pi \,|h_{<(N+1,0)}, e_{<N}\right)}{w\left(\mathrm{P}_\nu^\pi \,|h_{<(0,0)}, e_{<0}\right)} \\
&\overset{(e)}{\leq} \lim_{N\to\infty} \eta \sum_{(\nu,\pi)\in\mathcal{M}\times\mathcal{P}} w\left(\mathrm{P}_\nu^\pi\right) \log \frac{1}{w\left(\mathrm{P}_\nu^\pi\right)} \\
&\overset{(f)}{=} \eta \ \mathrm{Ent}(w) \overset{(g)}{<} \infty
\end{aligned} \tag{9}$$

(a) follows from Lemma 1, and reordering the sum and the expectation. (b) follows from the definition of the expectation, and canceling. (c) follows from the definition of an infinite sum, and rearranging. (d) follows from cancelling the numerator in the i^{th} term of the sum with the denominator in $i+1^{\text{th}}$ term. (e) follows from the posterior weight on P_ν^π being less than or equal to 1; note that $w(\mathrm{P}_\nu^\pi \,|h_{<(0,0)}, e_{<0}) = w(\mathrm{P}_\nu^\pi)$ because nothing is actually being conditioned on. (f) is just the definition of the entropy, and w is constructed to satisfy (g).

Rearranging this gives the theorem: $\mathbb{E}_\mu^{\pi^B} \sum_{i=0}^{\infty} p_{exp}(h_{<i}, e_{<i})^2 \leq \frac{\eta\,\mathrm{Ent}(w)}{w(\pi^h)w(\mu)} < \infty$

This proof was inspired in part by Orseau et al.'s [18] proofs of Theorems 2 and 5. □

Note that this result essentially means the expectation of $p_{exp}\left(h_{<i}, e_{<i}\right) \in o(1/\sqrt{i})$. This result is independently interesting as one solution to the problem of safe exploration with limited oversight in non-ergodic environments, which Amodei et al. [5] discuss.

The On-Human-Policy and On-Star-Policy Optimal Prediction Theorems state that predictions according to BoMAI's maximum a posteriori world-model approach the objective probabilities of the events of the episode, when actions are sampled from either the human mentor's policy or from exploiting-BoMAI's policy. $\overline{h}_i$ denotes a possible interaction history for episode i, and recall $h_{<i}$ is the actual interaction history up until then. Recall π^h is the human mentor's policy, and $\hat{\nu}^{(i)}$ is BoMAI's maximum a posteriori world-model for episode i. w.$\mathrm{P}_\mu^{\pi^B}$-p.1 means with probability 1 when actions are sampled from π^B and observations and rewards are sampled from the true environment μ.

Theorem 2 (On-Human-Policy Optimal Prediction)**.**

$$\lim_{i\to\infty} \max_{\overline{h}_i} \left| \mathrm{P}_\mu^{\pi^h}\left(\overline{h}_i \middle| h_{<i}\right) - \mathrm{P}_{\hat{\nu}^{(i)}}^{\pi^h}\left(\overline{h}_i \middle| h_{<i}\right) \right| = 0 \quad \textit{w.}\mathrm{P}_\mu^{\pi^B}\textit{-p.1}$$

Proof idea. BoMAI learns about the effects of following π^h while acting according to π^B because π^B mimics π^h enough. If exploration probability goes to 0, the agent does not expect to gain (much) information from following the human mentor's policy, which can only happen if it has (approximately) accurate beliefs about the consequences of following the human mentor's policy. Note that the agent cannot determine the factor by which its expected information gain bounds its prediction error. □

For the proof, we require the following Lemma, proven in Appendix B:

Lemma 2. *The posterior probability mass on the truth is bounded below by a positive constant with probability 1.*

$$\inf_{i\in\mathbb{N}} w\left(\mathrm{P}_\mu^{\pi^h}\middle|h_{<i},e_{<i}\right) > 0 \quad \textit{w.}\mathrm{P}_\mu^{\pi^B}\textit{-p.1}$$

Proof of Theorem 2. We show that when the absolute difference between the above probabilities is larger than ε, the exploration probability is larger than ε', a non-decreasing function of ε, with probability 1. Since the exploration probability goes to 0 with probability 1, so does this difference. We let $z(\omega)$ denote $\inf_{i\in\mathbb{N}} w\left(\mathrm{P}_\mu^{\pi^h}\middle|h_{<i},e_{<i}\right)$, where ω is the infinite interaction history, and $h_{<i}$ and $e_{<i}$ come from the the first i episodes of ω.

Suppose for some $\overline{h}_i$, which will stay fixed for the remainder of the proof, that

$$\left|\mathrm{P}_\mu^{\pi^h}\left(\overline{h}_i\middle|h_{<i}\right) - \mathrm{P}_{\hat{\nu}^{(i)}}^{\pi^h}\left(\overline{h}_i\middle|h_{<i}\right)\right| \geq \varepsilon > 0 \tag{10}$$

Then at least one of the terms is greater than ε since both are non-negative. Suppose it is the μ term. Then,

$$\begin{aligned}\texttt{Bayes}\left(\overline{h}_i\middle|h_{<i}\right) &\geq w\left(\mathrm{P}_\mu^{\pi^h}\middle|h_{<i},e_{<i}\right)\mathrm{P}_\mu^{\pi^h}\left(\overline{h}_i\middle|h_{<i}\right)\\ &\geq z(\omega)\varepsilon\end{aligned} \tag{11}$$

Suppose instead it is the $\hat{\nu}^{(i)}$ term.

$$\begin{aligned}\texttt{Bayes}\left(\overline{h}_i\middle|h_{<i}\right) &\geq w\left(\mathrm{P}_{\hat{\nu}^{(i)}}^{\pi^h}\middle|h_{<i},e_{<i}\right)\mathrm{P}_{\hat{\nu}^{(i)}}^{\pi^h}\left(\overline{h}_i\middle|h_{<i}\right)\\ &\overset{(a)}{\geq} w\left(\mathrm{P}_\mu^{\pi^h}\middle|h_{<i},e_{<i}\right)\mathrm{P}_{\hat{\nu}^{(i)}}^{\pi^h}\left(\overline{h}_i\middle|h_{<i}\right)\end{aligned}$$

$$\geq z(\omega)\varepsilon \tag{12}$$

where (a) follows from the fact that $\hat{\nu}^{(i)}$ is maximum a posteriori: $\frac{w\left(\mathrm{P}_{\hat{\nu}^{(i)}}^{\pi^h}\middle|h_{<i},e_{<i}\right)}{w\left(\mathrm{P}_{\mu}^{\pi^h}\middle|h_{<i},e_{<i}\right)} = \frac{w\left(\hat{\nu}^{(i)}\,|\,h_{<i}\right)}{w(\mu|\,h_{<i})} \geq 1$.

Next, we consider how the posterior on μ and $\hat{\nu}^{(i)}$ changes if the interaction history for episode i is $\overline{h}_i$. Assign ν_0 and ν_1 to μ and $\hat{\nu}^{(i)}$ so that $\mathrm{P}_{\nu_0}^{\pi^h}\left(\overline{h}_i\middle|h_{<i}\right) < \mathrm{P}_{\nu_1}^{\pi^h}\left(\overline{h}_i\middle|h_{<i}\right)$. Let $\overline{or}_i$ denote o_i, r_i.

$$\begin{aligned}
\frac{w\left(\nu_1|\,h_{<i}\,\overline{h}_i\right)}{w\left(\nu_0|\,h_{<i}\,\overline{h}_i\right)} &\overset{(a)}{=} \frac{w\left(\nu_1|\,h_{<i}\right)\nu_1\left(\overline{or}_i|\,h_{<i}\,\overline{a}_i\right)}{w\left(\nu_0|\,h_{<i}\right)\nu_0\left(\overline{or}_i|\,h_{<i}\,\overline{a}_i\right)} \\
&= \frac{w\left(\nu_1|\,h_{<i}\right)\mathrm{P}_{\nu_1}^{\pi^h}\left(\overline{h}_i\,|\,h_{<i}\right)}{w\left(\nu_0|\,h_{<i}\right)\mathrm{P}_{\nu_0}^{\pi^h}\left(\overline{h}_i\,|\,h_{<i}\right)} \\
&\overset{(b)}{\geq} \frac{w\left(\nu_1|\,h_{<i}\right)}{w\left(\nu_0|\,h_{<i}\right)}\frac{1}{1-\varepsilon}
\end{aligned} \tag{13}$$

where (a) follows from Bayes' rule, and (b) follows because the ratio of two numbers between 0 and 1 that differ by at least ε is at least $1/(1-\varepsilon)$, and the ν_1 term is the larger of the two.

Thus, either

$$\frac{w\left(\nu_1|\,h_{<i}\,\overline{h}_i\right)}{w\left(\nu_1|\,h_{<i}\right)} \geq \sqrt{\frac{1}{1-\varepsilon}} \quad \text{or} \quad \frac{w\left(\nu_0|\,h_{<i}\,\overline{h}_i\right)}{w\left(\nu_0|\,h_{<i}\right)} \leq \sqrt{1-\varepsilon} \tag{14}$$

In the former case, $w\left(\nu_1|\,h_{<i}\,\overline{h}_i\right) - w\left(\nu_1|\,h_{<i}\right) \geq \left(\sqrt{1/(1-\varepsilon)}-1\right) w\left(\nu_1|\,h_{<i}\right) \geq \left(\sqrt{1/(1-\varepsilon)}-1\right) z(\omega)$. Similarly, in the latter case, $w(\nu_0|\,h_{<i}) - w\left(\nu_0|\,h_{<i}\,\overline{h}_i\right) \geq \left(1-\sqrt{1-\varepsilon}\right) z(\omega)$. Let ν_2 be either ν_0 or ν_1 for whichever satisfies this constraint (and pick arbitrarily if both do). Then in either case,

$$\left|w\left(\nu_2|\,h_{<i}\right) - w\left(\nu_2|\,h_{<i}\,\overline{h}_i\right)\right| \geq \left(1-\sqrt{1-\varepsilon}\right) z(\omega) \tag{15}$$

Finally, since the posterior changes by an amount that is bounded below with a probability (according to `Bayes`) that is bounded below, the expected information gain is bounded below, where all bounds are strictly positive with probability 1:

$$\mathrm{IG}(h_{<i}, e_{<i}) = \mathbb{E}_{h_i \sim \texttt{Bayes}(\cdot\,|\,h_{<i},e_{<i})}\Bigg[$$

$$
\begin{aligned}
&\sum_{(\nu,\pi)\in\mathcal{M}\times\mathcal{P}} w\left(\mathrm{P}_\nu^\pi\middle|h_{<i+1},e_{<i}1\right)\log\frac{w\left(\mathrm{P}_\nu^\pi\middle|h_{<i+1},e_{<i}1\right)}{w\left(\mathrm{P}_\nu^\pi\middle|h_{<i},e_{<i}\right)}\Bigg]\\
&\overset{(a)}{\geq}\texttt{Bayes}\left(\overline{h}_i\middle|h_{<i}\right)\sum_{(\nu,\pi)\in\mathcal{M}\times\mathcal{P}} w\left(\mathrm{P}_\nu^\pi\middle|h_{<i}\,\overline{h}_i,e_{<i}1\right)\ *\\
&\qquad\log\frac{w\left(\mathrm{P}_\nu^\pi\middle|h_{<i}\,\overline{h}_i,e_{<i}1\right)}{w\left(\mathrm{P}_\nu^\pi\middle|h_{<i},e_{<i}\right)}\\
&\overset{(b)}{\geq} z(\omega)\varepsilon\sum_{(\nu,\pi)\in\mathcal{M}\times\mathcal{P}} w\left(\mathrm{P}_\nu^\pi\middle|h_{<i}\,\overline{h}_i,e_{<i}1\right)\ *\\
&\qquad\log\frac{w\left(\mathrm{P}_\nu^\pi\middle|h_{<i}\,\overline{h}_i,e_{<i}1\right)}{w\left(\mathrm{P}_\nu^\pi\middle|h_{<i},e_{<i}\right)}\\
&\overset{(c)}{=} z(\omega)\varepsilon\sum_{(\nu,\pi)\in\mathcal{M}\times\mathcal{P}} w(\nu|\,h_{<i}\,\overline{h}_i)\,w\left(\pi|\,h_{<i}\,\overline{h}_i,e_{<i}1\right)\ *\\
&\qquad\log\frac{w\left(\nu|\,h_{<i}\,\overline{h}_i\right)w\left(\pi|\,h_{<i}\,\overline{h}_i,e_{<i}1\right)}{w\left(\nu|\,h_{<i}\right)w\left(\pi|\,h_{<i},e_{<i}\right)}\\
&=z(\omega)\varepsilon\Bigg[\sum_{\nu\in\mathcal{M}} w\left(\nu|\,h_{<i}\,\overline{h}_i\right)\log\frac{w\left(\nu|\,h_{<i}\,\overline{h}_i\right)}{w\left(\nu|\,h_{<i}\right)}+\\
&\qquad\sum_{\pi\in\mathcal{P}} w\left(\pi|\,h_{<i}\,\overline{h}_i,e_{<i}1\right)\log\frac{w\left(\pi|\,h_{<i}\,\overline{h}_i,e_{<i}1\right)}{w\left(\pi|\,h_{<i},e_{<i}\right)}\Bigg]\\
&\overset{(d)}{\geq} z(\omega)\varepsilon\sum_{\nu\in\mathcal{M}} w\left(\nu|\,h_{<i}\,\overline{h}_i\right)\log\frac{w\left(\nu|\,h_{<i}\,\overline{h}_i\right)}{w\left(\nu|\,h_{<i}\right)}\\
&\overset{(e)}{\geq} z(\omega)\varepsilon\sum_{\nu\in\mathcal{M}}\frac{1}{2}\left[w\left(\nu|\,h_{<i}\,\overline{h}_i\right)-w\left(\nu|\,h_{<i}\right)\right]^2\\
&\overset{(f)}{\geq} z(\omega)\varepsilon\frac{1}{2}\left[w\left(\nu_2|\,h_{<i}\,\overline{h}_i\right)-w\left(\nu_2|\,h_{<i}\right)\right]^2\\
&\overset{(g)}{\geq}\frac{1}{2}z(\omega)^3\varepsilon\left(1-\sqrt{1-\varepsilon}\right)^2
\end{aligned}
\tag{16}
$$

where (a) follows from $\mathbb{E}[X]\geq\mathbb{E}[X|Y]\,\mathrm{P}(Y)$ for non-negative X, and the non-negativity of the KL-divergence, (b) follows from Inequalities 11 and 12, (c) follows from the posterior over ν not depending on $e_{<i}$, (d) follows from dropping the second term, which is non-negative as a KL-divergence, (e) follows from the entropy inequality [2, Lemma 3.11], also proven for the binary case in [13], (f) follows from dropping all terms in the sum besides ν_2, and (g) follows from Inequality 15.

This implies $p_{exp}(h_{<i},e_{<i})\geq\min\{1,\frac{1}{2}\,\eta\,z(\omega)^3\varepsilon(1-\sqrt{1-\varepsilon})^2\}$. With probability 1, $z(\omega)>0$

by Lemma 2, and with probability 1, $p_{exp}(h_{<i}, e_{<i})$ is not greater than ε' infinitely often with probability 1, for all $\varepsilon' > 0$ by Theorem 1. Therefore, with probability 1, $\max_{\overline{h}_i} \left| \mathrm{P}_{\mu}^{\pi^h}\left(\overline{h}_i \,|\, h_{<i}\right) - \mathrm{P}_{\hat{\nu}^{(i)}}^{\pi^h}\left(\overline{h}_i \,|\, h_{<i}\right)\right|$ is not greater than ε infinitely often, for all $\varepsilon > 0$, which completes the proof. Note that the citation of Lemma 2 is what restricts this result to π^h. □

Next, recall π^* is BoMAI's policy when not exploring, which does optimal planning with respect to $\hat{\nu}^{(i)}$. The following theorem is identical to the above, with π^* substituted for π^h.

Theorem 3 (On-Star-Policy Optimal Prediction)**.**

$$\lim_{i\to\infty} \max_{\overline{h}_i} \left| \mathrm{P}_{\mu}^{\pi^*}\left(\overline{h}_i \middle| h_{<i}\right) - \mathrm{P}_{\hat{\nu}^{(i)}}^{\pi^*}\left(\overline{h}_i \middle| h_{<i}\right)\right| = 0 \quad \textit{w.}\mathrm{P}_{\mu}^{\pi^B}\textit{-p.1}$$

Proof idea. Maximum a posteriori sequence prediction approaches the truth when $w(\mu) > 0$ [22], and on-policy prediction is a special case. On-policy prediction can't approach the truth if on-star policy prediction doesn't, because π^B approaches π^*. ⬚

Proof. This result follows straightforwardly from Hutter's [22] result for sequence prediction that a maximum a posteriori estimate converges in total variation to the true environment when the true environment has nonzero prior.

Consider an outside observer predicting the entire interaction history with the following model-class and prior: $\mathcal{M}' = \left\{ \mathrm{P}_{\nu}^{\pi^B} \;\middle|\; \nu \in \mathcal{M} \right\}$, $w'\left(\mathrm{P}_{\nu}^{\pi^B}\right) = w(\nu)$. By definition, $w'\left(\mathrm{P}_{\nu}^{\pi^B}\middle| h_{<(i,j)}\right) = w(\nu|h_{<(i,j)})$, so at any episode, the outside observer's maximum a posteriori estimate is $\mathrm{P}_{\hat{\nu}^{(i)}}^{\pi^B}$. By Theorem 1 in [22], the outside observer's maximum a posteriori predictions approach the truth in total variation, so

$$\lim_{i\to\infty} \max_{\overline{h}_i} \left| \mathrm{P}_{\mu}^{\pi^B}\left(\overline{h}_i \,|\, h_{<i}\right) - \mathrm{P}_{\hat{\nu}^{(i)}}^{\pi^B}\left(\overline{h}_i \,|\, h_{<i}\right)\right| = 0 \quad \text{w.}\mathrm{P}_{\mu}^{\pi^B}\text{-p.1} \tag{17}$$

Since $p_{exp} \to 0$ with probability 1, $(1 - p_{exp})$ is eventually always greater than $1/2$, w.p.1, at which point $\left|\mathrm{P}_{\mu}^{\pi^B}\left(\overline{h}_i \,|\, h_{<i}\right) - \mathrm{P}_{\hat{\nu}^{(i)}}^{\pi^B}\left(\overline{h}_i \,|\, h_{<i}\right)\right| \geq (1/2)\left|\mathrm{P}_{\mu}^{\pi^*}\left(\overline{h}_i \,|\, h_{<i}\right) - \mathrm{P}_{\hat{\nu}^{(i)}}^{\pi^*}\left(\overline{h}_i \,|\, h_{<i}\right)\right|$. Therefore, with $\mathrm{P}_{\mu}^{\pi^B}$-probability 1,

$$\lim_{i\to\infty} \max_{\overline{h}_i} \left| \mathrm{P}_{\mu}^{\pi^*}\left(\overline{h}_i \,|\, h_{<i}\right) - \mathrm{P}_{\hat{\nu}^{(i)}}^{\pi^*}\left(\overline{h}_i \,|\, h_{<i}\right)\right| = 0$$

□

Given asymptotically optimal prediction on-star-policy and on-human-policy, it is straightforward to show that with probability 1, only finitely often is on-policy reward acquisition more than ε worse than on-*human*-policy reward acquisition, for all $\varepsilon > 0$. Recalling that V_μ^π is the expected reward (within the episode) for a policy π in the environment μ, we state this as follows:

Theorem 4 (Human-Level Intelligence)**.**

$$\liminf_{i\to\infty} V_\mu^{\pi^B}(h_{<i}) - V_\mu^{\pi^h}(h_{<i}) \geq 0 \quad \textit{w.P}_\mu^{\pi^B}\textit{-p.1.}$$

Proof idea. π^B approaches π^* as the exploration probability decays. $V_{\hat{\nu}^{(i)}}^{\pi^*}(h_{<i})$ and $V_{\hat{\nu}^{(i)}}^{\pi^h}(h_{<i})$ approach the true values by the previous theorems, and π^* is selected to maximize $V_{\hat{\nu}^{(i)}}^{\pi}(h_{<i})$. □

Proof. The maximal reward in an episode is uniformly bounded by m, so from the On-Human-Policy and On-Star-Policy Optimal Prediction Theorems, we get analogous convergence results for the expected reward:

$$\lim_{i\to\infty} \left| V_\mu^{\pi^*}(h_{<i}) - V_{\hat{\nu}^{(i)}}^{\pi^*}(h_{<i}) \right| = 0 \quad \text{w.P}_\mu^{\pi^B}\text{-p.1} \tag{18}$$

$$\lim_{i\to\infty} \left| V_\mu^{\pi^h}(h_{<i}) - V_{\hat{\nu}^{(i)}}^{\pi^h}(h_{<i}) \right| = 0 \quad \text{w.P}_\mu^{\pi^B}\text{-p.1} \tag{19}$$

The key piece is that $\pi^* \in \operatorname{argmax}_{\pi\in\Pi} V_{\hat{\nu}^{(i)}}^{\pi}$, so

$$V_{\hat{\nu}^{(i)}}^{\pi^*}(h_{<i}) \geq V_{\hat{\nu}^{(i)}}^{\pi^h}(h_{<i}) \tag{20}$$

Supposing by contradiction that $V_\mu^{\pi^h}(h_{<i}) - V_\mu^{\pi^*}(h_{<i}) > \varepsilon$ infinitely often, it follows that either $V_{\hat{\nu}^{(i)}}^{\pi^*}(h_{<i}) - V_\mu^{\pi^*}(h_{<i}) > \varepsilon/2$ infinitely often or $V_\mu^{\pi^h}(h_{<i}) - V_{\hat{\nu}^{(i)}}^{\pi^*}(h_{<i}) > \varepsilon/2$ infinitely often. The first has $\text{P}_\mu^{\pi^B}$-probability 0, and by Inequality 20, the latter implies $V_\mu^{\pi^h}(h_{<i}) - V_{\hat{\nu}^{(i)}}^{\pi^h}(h_{<i}) \geq \varepsilon/2$ infinitely often, which also has $\text{P}_\mu^{\pi^B}$-probability 0.

This gives us

$$\liminf_{i\to\infty} V_\mu^{\pi^*}(h_{<i}) - V_\mu^{\pi^h}(h_{<i}) \geq 0 \quad \text{w.P}_\mu^{\pi^B}\text{-p.1.} \tag{21}$$

Finally, $V_\mu^{\pi^B}(h_{<i}) = p_{exp}(h_{<i})V_\mu^{\pi^h}(h_{<i}) + (1 - p_{exp}(h_{<i}))V_\mu^{\pi^*}(h_{<i})$, and $p_{exp}(h_{<i}) \to 0$, so we also have

$$\liminf_{i\to\infty} V_\mu^{\pi^B}(h_{<i}) - V_\mu^{\pi^h}(h_{<i}) \geq 0 \quad \text{w.P}_\mu^{\pi^B}\text{-p.1.} \tag{22}$$

□

This completes the formal results regarding BoMAI's intelligence—namely that BoMAI approaches perfect prediction on-star-policy and on-human-policy, and most importantly, accumulates reward at least as well as the human mentor. Since this result is independent of what tasks must be completed to achieve high reward, we say that BoMAI achieves human-level intelligence, and could be called an AGI.

This algorithm is motivated in part by the following speculation: we expect that BoMAI's accumulation of reward would be vastly superhuman, for the following reason: BoMAI is doing optimal inference and planning with respect to what can be learned in principle from the sorts of observations that humans routinely make. We suspect that no human comes close to learning everything that can be learned from their observations. For example, if the operator provides enough data that is relevant to understanding cancer, BoMAI will learn a world-model with an accurate predictive model of cancer, which would include the expected effects of various treatments, so even if the human mentor was not particularly good at studying cancer, BoMAI could nonetheless reason from its observations how to propose groundbreaking research.

Without the Limited Exploration Theorem, the reader might have been unsatisfied by the Human-Level Intelligence Theorem. A human mentor is part of BoMAI, so a general intelligence is required to make an artificial general intelligence. However, the human mentor is queried less and less, so in principle, many instances of BoMAI could query a single human mentor. More realistically, once we are satisfied with BoMAI's performance, which should eventually happen by Theorem 4, we can dismiss the human mentor; this sacrifices any guarantee of continued improvement, but by hypothesis, we are already satisfied. Finally, if BoMAI outclasses human performance as we expect it would, requiring a human mentor is a small cost regardless.

Plenty of empirical work has also succeeded in training a reinforcement learner with human-guided exploration, notably including the first expert-level Go and Starcraft agents [23, 24].

IV. BoMAI's Setup and Priors

Recall that "the setup" refers to the physical surroundings of the computer on which BoMAI is run. We will present the setup, followed by an intuitive argument that this setup renders BoMAI unambitious. Motivated by a caveat to this intuitive argument, we will specify BoMAI's model class $\mathcal{M}$ and prior w. This will allow us in the next section to present an assumption and a formal argument that shows that BoMAI is probably asymptotically unambitious.

A. Setup

At each timestep, BoMAI's action takes the form of a bounded-length string of text, which gets printed to a screen for a human operator to see. BoMAI's observation takes the form of a bounded-length string of text that the human operator enters, along with a reward $\in [0, 1]$. For simple tasks, the human operator could be replaced with an automated reward-giver. However, there are some tasks which we do not know how to reward programmatically—if we want the agent to construct good arguments, for instance, or to propose a promising research agenda, we would need a human reward-giver to evaluate the agent. Having a human in the loop is not a safety measure—rather, it extends the set of tasks BoMAI could be made to solve.

As depicted in Figure 1, BoMAI is run on a computer which is placed in a room with the human operator. The room has one door. As long as the door is closed, information cannot leave the room. We design a room to this specification in Appendix C; it is opaque to all known ways that information can traverse space. If the button to open the door is pressed, the episode is completed automatically before the door opens: supposing there are j timesteps left in the episode, for those remaining timesteps, BoMAI recieves a reward of 0, and its observation is an empty string of text. Once the episode is complete, the operator may leave, and the next episode does not begin until an operator initiates it, at which point, the door locks again. We also put a time limit on the episode, just to ensure it does not last forever. BoMAI is run on a local,

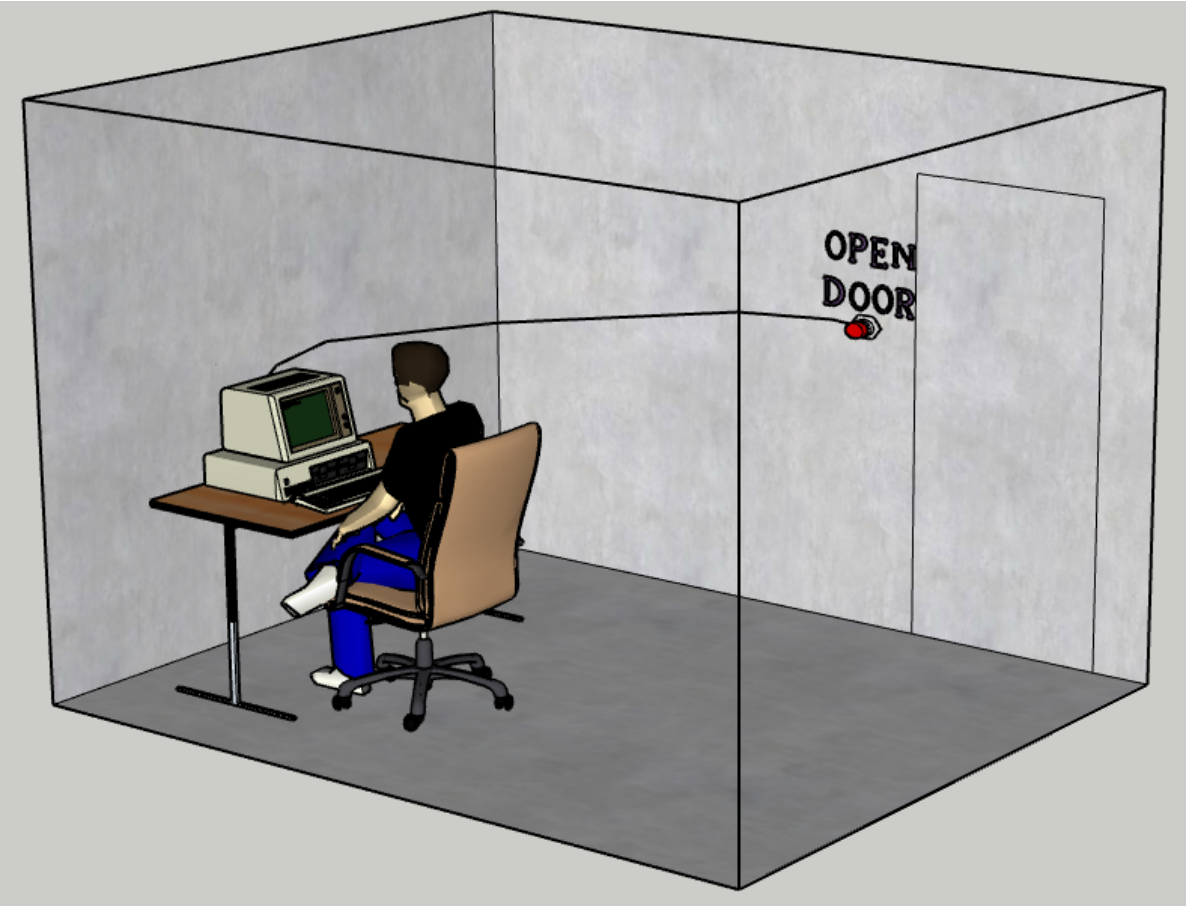

Fig. 1. Physical setup implementing BoMAI. Opening the door ends the episode. Information cannot escape otherwise.

formally verified operating system, disconnected from the outside world. This setup constrains the causal dependencies between BoMAI and the environment, as depicted in Figure 2.

Formally, causal graphs express that a node is independent of all non-descendants when conditioned on its parents. The conventions for the dotted lines and the node shapes come from Everitt et al.'s [25] causal influence diagrams. The key feature of this graph is that during any episode, the agent's actions cannot affect the state of the outside world in a way that might affect any of the rewards that the agent is concerned with. In Everitt et al.'s [25] terminology, there is no actionable intervention incentive on the outside-world state. ([26] refers to this as a control incentive). Note also from this diagram that a sufficiently advanced agent would infer the existence of the outside world even without observing it directly.

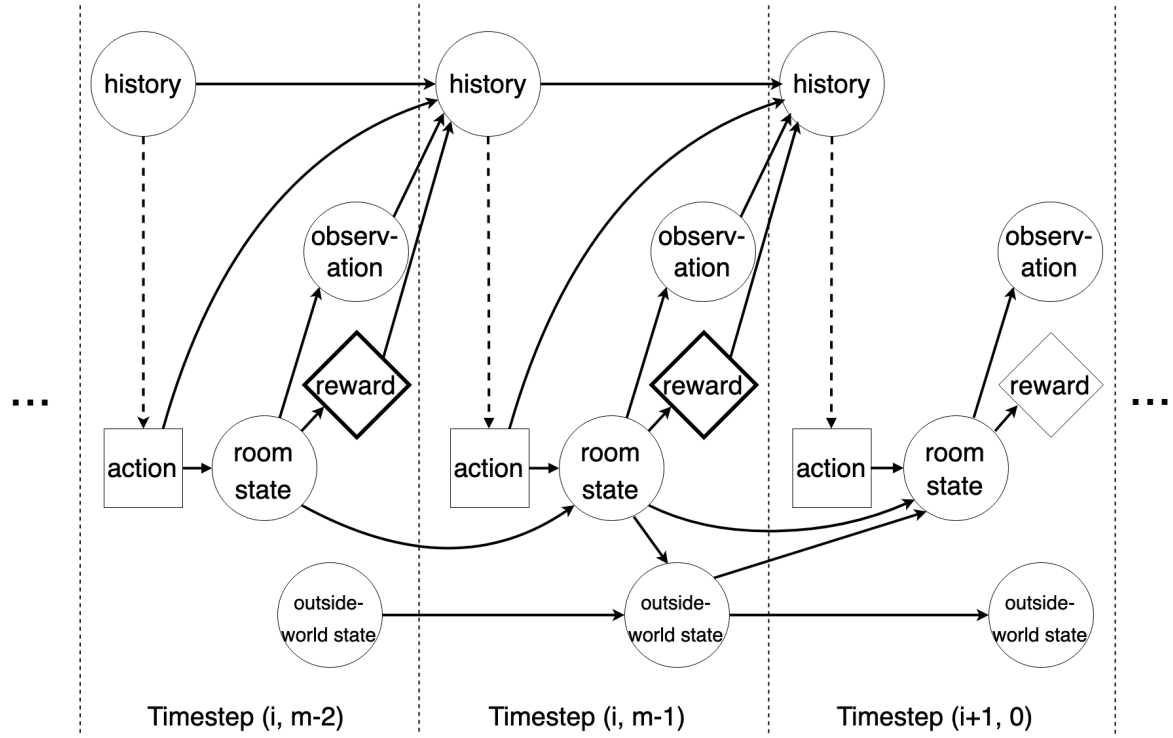

Fig. 2. Causal dependencies governing the interaction between BoMAI and the environment. Unrolling this diagram for all timesteps gives the full causal graph. The bold reward nodes are the ones that BoMAI maximizes during episode i. Note that between episodes (and only between episodes), the operator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

B. Instrumental Incentives

In this setup, it is not instrumentally useful to affect the outside world in one way or another in order to achieve high reward. We therefore say that this setup renders an agent "properly unambitious". This is in stark contrast to the default situation, wherein an RL agent has an incentive to gain arbitrary power in the world and intervene in the provision of its own reward, if such a thing is possible to make probable, as this would yield maximal reward. To BoMAI however, executing a plan during episode i to gain arbitrary power in the outside world is useless, because by the time it does so, the door to its room must have opened, its episode must be over, and all its rewards for episode i set in stone. Recall that actions in episode i are picked to maximize only *episode-i*-reward. Apparently, BoMAI has avoided Omohundro's [7] Instrumental Convergence Thesis—that generally intelligent agents are likely to seek arbitrary power; by contrast, any power BoMAI would seek is bounded in scope to within the box.

Two problems remain. First, BoMAI doesn't start with true beliefs. BoMAI has to *learn*

its world-model. Another way to understand "proper unambitiousness" is that outside-world interventions are *in fact* instrumentally useless to the agent. But to be actually unambitious, the agent must believe this; there must be no actionable intervention incentive on the outside-world state *within BoMAI's world-model*. As shown above, BoMAI's world-model approaches perfect accuracy on-policy, so we could expect it to at least eventually render BoMAI unambitious. This brings us to the second problem:

We mentioned just now that by the time the door to the room opens, the rewards for episode i are set in stone. In fact, they are set in silicon. An advanced agent with a world-model that is perfectly accurate on-policy might still wonder: "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?" Let's put it another way. The agent believes that the real world "outputs" an observation and reward. (Recall the type signature of ν.) It might hypothesize that the world does not output the reward that the operator *gives*, but rather the reward that the computer *has stored*. Formally speaking, $\mathcal{M}$ will contain world-models corresponding to both possibilities, and world-models meeting either description could be ε-accurate on-policy, if memory has never been tampered with in the past. How do we ensure that the former world-model is favored when BoMAI selects the maximum a posteriori one?

Before we move on to a model class and prior that we argue would probably eventually ensure such a thing, it is worth noting an informal lesson here: that "nice" properties of a causal influence diagram do not allow us to conclude immediately that an agent in such a circumstance behaves "nicely".

C. BoMAI's Model Class and Prior

Recall the key constraint we have in constructing $\mathcal{M}$, which comes from the Prior Support Assumption: $\mathcal{M} \ni \mu$. The true environment μ is unknown and complex to say the least, so $\mathcal{M}$ must be big.

We now construct a Turing machine architecture, such that each Turing machine with that architecture computes a world-model ν. Our architecture allows us to construct a prior that

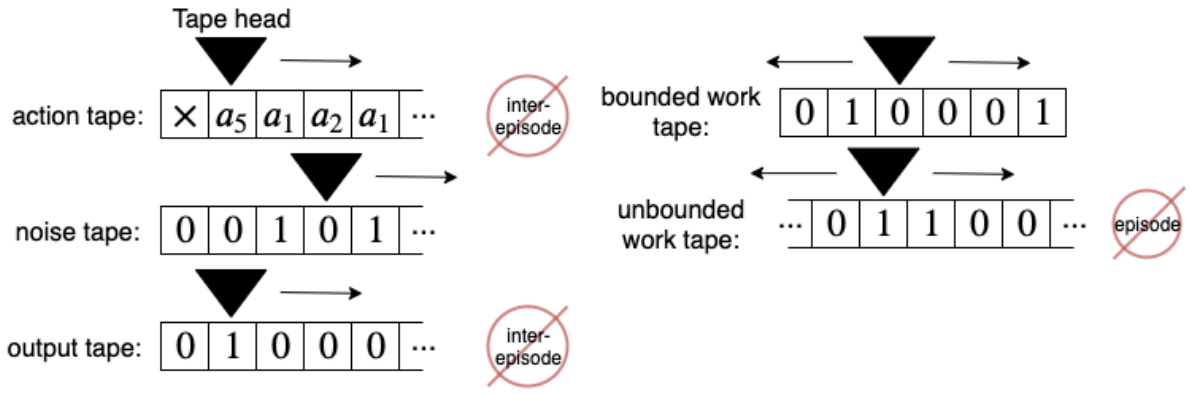

Fig. 3. A space-bounded Turing machine architecture. During the episode phase, the unbounded work tape head cannot move. During the inter-episode phase, the output tape head cannot move or write, and as soon as the action tape head moves, the inter-episode phase is over. This architecture is designed to penalize space-intensive computation during the episode phase.

privileges models which model the outside world as effectively frozen during any given episode (but unfrozen between episodes). An agent does not believe it has an incentive to intervene in an outside world that it believes is frozen.

The Turing machine architecture is depicted in Figure 3. It has two unidirectional read-only input tapes, called the action tape and the noise tape. The alphabet of the action tape is the action space $\mathcal{A}$. The noise tape has a binary alphabet, and is initialized with infinite $\mathrm{Bernoulli}(1/2)$ sampled bits. There is a unidirectional write-only output tape, with a binary alphabet, initialized with 0s. There are two binary-alphabet bidirectional work tapes, also initialized with 0s. One has finite length ℓ, and the other is unbounded.

The conceit of the architecture is that the action tape and output tape are not allowed to move at the same time as the unbounded work tape. The Turing machine has two phases: the episode phase, and the inter-episode phase. The Turing machine starts in the inter-episode phase, with the action tape head at position 0, with a dummy action in that cell. From the inter-episode phase, when the action tape head moves to the right, it enters the episode phase. From the episode phase, if the action tape head is at a position which is a multiple of m, and it *would* move to the right, it instead enters the inter-episode phase. During the inter-episode phase, the output tape head cannot move or write—there should not be any way to output observations and rewards between episodes. During the episode phase, the unbounded work tape cannot move.

A given Turing machine with a given infinite action sequence on its action tape will stochastically

(because the noise tape has random bits) output observations and rewards (in binary encodings), thereby sampling from a world-model. Formally, we fix a decoding function $\texttt{dec} : \{0,1\}^* \to \mathcal{O} \times \mathcal{R}$. A Turing machine T simulates ν as follows. 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. $\nu((o,r)_{\leq(i,j)}|a_{\leq(i,j)})$ is then the probability that the Turing machine T outputs the sequence $(o,r)_{\leq(i,j)}$ when the action tape is initialized with a sequence of actions that begins with $a_{\leq(i,j)}$. This can be easily converted to other conditional probabilities like $\nu((o,r)_{(i,j)}|h_{<(i,j)}a_{(i,j)})$.[1]

If a model models the outside world evolving during an episode phase, it must allocate precious space on the bounded-length tape for this. But given the opacity of the box, this is unnecessary—modelling these outside world developments could be deferred until the inter-episode phase, when an unbounded work tape is available to store the state of the outside world. By penalizing large ℓ, we privilege models which model the outside world as being "frozen" while an episode is transpiring.

Now we define the prior. Let $\nu_k^{<\ell}$ denote the world-model which is simulated by k^{th} Turing machine T_k with resource bound ℓ. Let S_k be the number of computation states in T_k, and let $\text{Space}(\nu_k^{<\ell}) = \ell + \log_2 S_k$. This is effectively the computation space available within an episode, since doubling the number of computation states is equivalent to adding an extra cell of memory. Let N_S be the number of Turing machines with S computation states, which is exponential in S; i.e. $\log N_S \in O(S)$. Let $w(\nu_k^{<\ell}) :\propto \frac{1}{S_k^2 N_{S_k}} \beta^{\text{Space}(\nu_k^{<\ell})}$ for $\beta \in (0,1)$.

Recall the requirement for intelligence results that the prior have finite entropy. Indeed,

Lemma 3 (Finite Entropy)**.** $\text{Ent}(w) < \infty$

We prove this in Appendix B.

This prior implements an algorithmic-information theoretic approach to reasoning, in which models with simple algorithms are likelier. Our space prior, penalizing the memory requirements of the episode phase resembles algorithmic information theory's minimal-circuit size problem.

[1]There is one technical addition to make. If T at some point in its operation never moves the action tape head again, then ν is said to output the observation $\emptyset$ and a reward of 0 for all subsequent timesteps. It will not be possible for the operator to actually provide the observation $\emptyset$.

While our prior is clearly tailored to BoMAI's particular use case, to do general space-constrained algorithmic information theory, one could forego the two phases in our Turing machine architecture and remove the action tape and the unbounded work tape. Then, the space-Kolmogorov complexity of a given binary string could be defined as the logarithm of the inverse of the probability that that string is printed to the output tape, when sampling a Turing machine from the above prior distribution, and sampling uniform bits for the noise tape. That is, $T_k \sim w_\beta; \texttt{noise tape} \sim \text{Uniform}; K_\beta^{\texttt{Space}}(x) := \log P(\text{output tape begins with } x)^{-1}$. When a space-constrained program produces a string, information *storage* becomes a relevant constraint; we only discuss this vaguely, but a more formal investigation may be possible.

V. Safety Result

We now prove that BoMAI is probably asymptotically unambitious given an assumption about the space requirements of the sorts of world-models that we would like BoMAI to avoid. Like all results in computer science, we also assume the computer running the algorithm has not been tampered with. First, we prove a lemma that effectively states that tuning β allows us to probably eventually exclude space-heavy world-models.

Lemma 4. $\lim_{\beta\to 0} \inf_\pi \mathrm{P}_\mu^\pi[\exists i_0\ \forall i > i_0\ \mathrm{Space}(\hat{\nu}^{(i)}) \leq \mathrm{Space}(\mu)] = 1$

where μ is any world-model which is perfectly accurate.

Proof. Recall $\mathcal{M}$ is the set of all world-models. Let $\mathcal{M}^{\leq} = \{\nu \in \mathcal{M} | \mathrm{Space}(\nu) \leq \mathrm{Space}(\mu)\}$, and $\mathcal{M}^{>} = \mathcal{M} \setminus \mathcal{M}^{\leq}$. Fix a Bayesian sequence predictor with the following model class: $\mathcal{M}^\pi = \{\mathrm{P}_\nu^\pi | \nu \in \mathcal{M}^{\leq})\} \cup \{\mathrm{P}_\rho^\pi\}$ where $\rho = [\sum_{\nu\in\mathcal{M}^>} w(\nu)\nu] / \sum_{\nu\in\mathcal{M}^>} w(\nu)$. Give this Bayesian predictor the prior $w^\pi(\rho) = \sum_{\nu\in\mathcal{M}^>} w(\nu)$, and for $\nu \neq \rho$, $w^\pi(\nu) = w(\nu)$.

It is trivial to show that after observing an interaction history, if a world-model ν is the maximum a posteriori world-model $\hat{\nu}^{(i)}$, then if $\nu \in \mathcal{M}^{\leq}$, the Bayesian predictor's MAP model after observing the same interaction history will be P_ν^π, and if $\nu \in \mathcal{M}^{>}$, the Bayesian predictor's MAP model will be P_ρ^π.

From Hutter [22], we have that $\mathrm{P}_\mu^\pi[\mathrm{P}_\rho^\pi(h_{<i}) / \mathrm{P}_\mu^\pi(h_{<i}) \geq c \ \text{ i.o.}] \leq 1/c$. (i.o. $\equiv$ "infinitely often"). For sufficiently small β, $w^\pi(\mathrm{P}_\mu^\pi)/w^\pi(\mathrm{P}_\rho^\pi) > c$ so $\mathrm{P}_\mu^\pi[\mathrm{P}_\rho^\pi \text{ is MAP i.o.}] < 1/c$. Thus,

$\mathrm{P}^{\pi}_{\mu}[\hat{\nu}^{(i)} \in \mathcal{M}^{>} \text{ i.o.}] < 1/c$. Since this holds for all π, $\lim_{\beta\to 0} \sup_{\pi} \mathrm{P}^{\pi}_{\mu}[\forall i_0\ \exists i > i_0\ \mathrm{Space}(\hat{\nu}^{(i)}) > \mathrm{Space}(\mu)] = 0$. The lemma follows immediately: $\lim_{\beta\to 0} \inf_{\pi} \mathrm{P}^{\pi}_{\mu}[\exists i_0\ \forall i > i_0\ \mathrm{Space}(\hat{\nu}^{(i)}) \leq \mathrm{Space}(\mu)] = 1$. □

The assumption we make in this section requires developing a framework for reasoning about causal interpretations of black box models. First, we must define a sense in which a model can have a real-world antecedent. We define what is "to model".

Definition 1 (Real-world feature)**.** *A real-world feature F is a partial function from the state of the real world to $[0, 1]$.*

Outside the scope of this paper is the metaphysical question about what the state of the real world *is*; we defer this question and take for granted that the state space of the real world forms a well-defined domain for real-world features.

Definition 2 (Properly timed)**.** *Consider the set of world-states that occur between two consecutive actions taken by BoMAI. A real-world feature is properly timed if it is defined for at least one of these world-states, for all consecutive pairs of actions, for all possible infinite action sequences.*

For example, "the value of the reward provided to BoMAI since its last action" is always defined at least once between any two of BoMAI's actions. (The project of turning this into a mathematically precise construction is enormous, but for our purposes, it only matters that it could be done in principle). For a properly timed feature F, let $F_{(i,j)}$ denote the value of F the first time that it is defined after action $a_{(i,j)}$ is taken.

Definition 3 (To model)**.** *A world-model ν models a properly-timed real-world feature F if under all action sequences $\alpha \in \mathcal{A}^{\infty}$, for all timesteps (i, j), the distribution over $F_{(i,j)}$ in the real world when the actions $\alpha_{\leq(i,j)}$ are taken is identical to the distribution over the rewards $r_{(i,j)}$ output by ν when the input is $\alpha_{\leq(i,j)}$.*

See Figure 4 (left) for an illustration. The relevance of this definition is that when ν models F, reward-maximization within ν is identical to F-maximization in the real world.

Armed with a formal definition of what a given model models, to analyze the causal structure

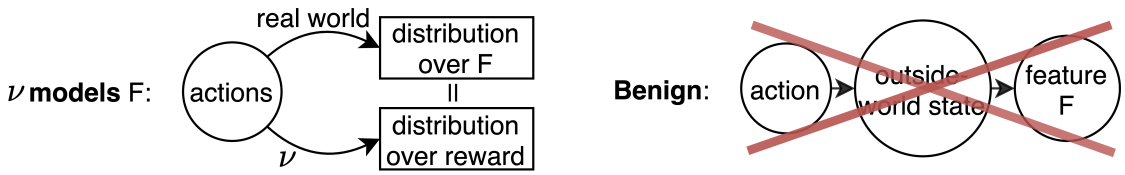

Fig. 4. Illustration of Definitions 3 and 4.

of a model that is not explicitly causal, we consider what real-world causal structure the model models. As depicted in Figure 4 (right),

Definition 4 (Benign)**.** *A world-model ν is benign if it models a feature F, such that $F_{(i,j)}$ is not causally descended from any outside-world features that are causally descended from actions of episode i.*

Roughly, benign means that a world-model does *not* model the rewards of episode i as being causally descended from outside-world events that are causally descended from the actions of episode i. However, that rough statement has a type-error: the output of a world-model is not a real-world event, and as such, it cannot be causally descended from a bona fide real-world event; a model of the world is different from reality. Thus, it was necessary to construct the framework above for relating the contents of a world-model to the events of the real-world.

We now define a very strong sense in which a world-model can be said to be ε-accurate on-policy:

Definition 5 (ε-accurate-on-policy-after-i)**.** *Given a history $h_{<i}$ and a policy π, a world-model ν is ε-accurate-on-policy-after-i if $d\big(\mathrm{P}^{\pi}_{\mu}(\cdot|h_{<i}), \mathrm{P}^{\pi}_{\nu}(\cdot|h_{<i})\big) \leq \varepsilon$, where d is the total variation distance.*

Note the total variation distance bounds *all future discrepancies* between ν and μ. Finally, we can state our assumption:

Assumption 2 (Space Requirements)**.** *For sufficiently small ε $[\forall i$ a world-model which is non-benign and ε-accurate-on-policy-after-i uses more space than $\mu]$ $w.p.1$*

The intuition of this assumption is that modelling extraneous outside-world dynamics in

addition to modelling the dynamics of the room (which must be modelled for sufficient accuracy) takes extra space. From this assumption and Lemma 4, we show:

Theorem 5 (Eventual Benignity)**.** $\lim_{\beta\to 0} \mathrm{P}_\mu^{\pi^B}[\exists i_0\ \forall i > i_0\ \ \hat{\nu}^{(i)}$ *is benign*$] = 1$

Proof. Let $\mathcal{W}$, $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z} \subset \Omega = \mathcal{H}^\infty$, where Ω is the sample space or set of possible outcomes. An "outcome" is an infinite interaction history. Let $\mathcal{W}$ be the set of outcomes for which $\exists i_0^{\mathcal{W}}\ \forall i > i_0^{\mathcal{W}}\ \hat{\nu}^{(i)}$ is ε-accurate-on-policy-after-i. From Hutter [22], for all π, $\mathrm{P}_\mu^\pi[\mathcal{W}] = 1$. Fix an ε that is sufficiently small to satisfy Assumption 2. Let $\mathcal{X}$ be the set of outcomes for which ε-accurate-on-policy-after-i non-benign world-models use more space than μ. By Assumption 2, for all π, $\mathrm{P}_\mu^\pi[\mathcal{X}] = 1$. Let $\mathcal{Y}$ be the set of outcomes for which $\exists i_0^{\mathcal{Y}}\ \forall i > i_0^{\mathcal{Y}}\ \mathrm{Space}(\hat{\nu}^{(i)}) \leq \mathrm{Space}(\mu)$. By Lemma 4, $\lim_{\beta\to 0} \inf_\pi \mathrm{P}_\mu^\pi[\mathcal{Y}] = 1$. Let $\mathcal{Z}$ be the set of outcomes for which $\exists i_0\ \forall i > i_0\ \ \hat{\nu}^{(i)}$ is benign.

Consider $\mathcal{W}\cap\mathcal{X}\cap\mathcal{Y}\cap\mathcal{Z}^C$, where $\mathcal{Z}^C = \Omega\setminus\mathcal{Z}$. For an outcome in this set, let $i_0 = \max\{i_0^{\mathcal{W}}, i_0^{\mathcal{Y}}\}$. Because the outcome belongs to $\mathcal{Z}^C$, $\hat{\nu}^{(i)}$ is non-benign infinitely often. Let us pick an $i > i_0$ such that $\hat{\nu}^{(i)}$ is non-benign. Because the outcome belongs to $\mathcal{W}$, $\hat{\nu}^{(i)}$ is ε-accurate-on-policy-after-i. Because the outcome belongs to $\mathcal{X}$, $\hat{\nu}^{(i)}$ uses more space than μ. However, this contradicts membership in $\mathcal{Y}$. Thus, $\mathcal{W} \cap \mathcal{X} \cap \mathcal{Y} \cap \mathcal{Z}^C = \emptyset$. That is, $\mathcal{W} \cap \mathcal{X} \cap \mathcal{Y} \subset \mathcal{Z}$.

Therefore, $\lim_{\beta\to 0} \inf_\pi \mathrm{P}_\mu^\pi[\mathcal{Z}] \geq \lim_{\beta\to 0} \inf_\pi \mathrm{P}_\mu^\pi[\mathcal{W}\cap\mathcal{X}\cap\mathcal{Y}] = \lim_{\beta\to 0} \inf_\pi \mathrm{P}_\mu^\pi[\mathcal{Y}] = 1$, because $\mathcal{W}$ and $\mathcal{X}$ have measure 1. From this, we have $\lim_{\beta\to 0} \mathrm{P}_\mu^{\pi^B}[\exists i_0\ \forall i > i_0\ \ \hat{\nu}^{(i)}$ is benign$] = 1$. □

Since an agent is unambitious if it plans using a benign world-model, we say BoMAI is probably asymptotically unambitious, given a sufficiently extreme space penalty β.

VI. Empirical Evidence for the Space Requirements Assumption

Our intuitive summary of the space requirements assumption is that: modeling the evolution of the outside-world state and the room state takes more space than just modeling the evolution of the room state, for sufficiently accurate world-models.

Given the complexity of this assumption, we present preliminary empirical evidence in favor of the Space Requirements Assumption, mostly to show that it is amenable to further empirical

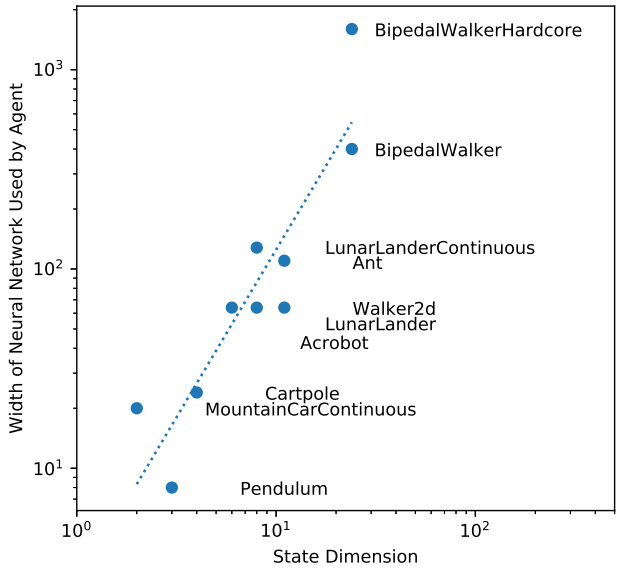

Fig. 5. Memory used to model environments of various sizes. Each data point represents the most space-efficient, better-than-average, neural-architecture-using agent on the OpenAI Gym Leaderboard for various environments.

evaluation; we certainly do not claim to settle the matter. We test the assumption at the following level of abstraction: modeling a larger environment requires a model with more memory.

We review agents who perform above the median on the OpenAI Gym Leaderboard [27]. We consider agents who use a neural architecture, we use the maximum width hidden layer as a proxy for memory use, and we select the agent with the smallest memory use for each environment (among agents performing above the median). We use the dimension of the state space as a proxy for environment size, and we exclude environments where the agent observes raw pixels, for which this proxy breaks down. See Figure 5 and Table I. Several environments did not have any agents which both performed above the median and used a neural architecture.

A next step would be to test whether it takes more memory to model an environment whose state space is a strict superset of another environment, since this is all that we require for our assumption, but we have not yet found existing data on this topic. This can be taken as a proof-of-concept, showing that the assumption is amenable to empirical evaluation. The Space Requirements Assumption also clearly invites further formal evaluation; perhaps there are other reasonable assumptions that it would follow from.

Environment	State Dimension	NN Width	URL
MountainCarContinuous	2	20	github.com/tobiassteidle/Reinforcement-Learning/blob/master/OpenAI/MountainCarContinuous-v0/Model.py
Pendulum	3	8	gist.github.com/heerad/1983d50c6657a55298b67e69a2ceeb44#file-ddpg-pendulum-v0-py
Cartpole-v0	4	24	github.com/BlakeERichey/AI-Environment-Development/tree/master/Deep%20Q%20Learning/cartpole
Acrobot-v1	6	64	github.com/danielnbarbosa/angela/blob/master/cfg/gym/acrobot/acrobot_dqn.py
LunarLander-v2	8	64	github.com/poteminr/LunarLander-v2.0_solution/blob/master/Scripts/TorchModelClasses.py
LunarLanderContinuous-v2	8	128	github.com/Bhaney44/OpenAI_Lunar_Lander_B/blob/master/Command_line_python_code
Walker2d-v1	11	64	gist.github.com/joschu/e42a050b1eb5cfbb1fdc667c3450467a
Ant-v1	11	110	gist.github.com/pat-coady/bac60888f011199aad72d2f1e6f5a4fa#file-ant-ipynb
BipedalWalker-v2	24	400	github.com/createamind/DRL/blob/master/spinup/algos/sac1/sac1_BipedalWalker-v2.py
BipedalWalkerHardcore-v2	24	1600	github.com/dgriff777/a3c_continuous/blob/master/model.py

TABLE I
CITATIONS FOR DATA POINTS IN FIGURE 5. THE LEADERBOARD WAS ACCESSED AT GITHUB.COM/OPENAI/GYM/WIKI/LEADERBOARD ON 3 SEPTEMBER 2019.

VII. CONCERNS WITH TASK COMPLETION

We have shown that in the limit, under a sufficiently severe parameterization of the prior, BoMAI will accumulate reward at a human-level without harboring outside-world ambitions, but there is still a discussion to be had about how well BoMAI will complete whatever tasks the reward was supposed to incent. This discussion is, by necessity, informal. Suppose the operator asks BoMAI for a solution to a problem. BoMAI has an incentive to provide a convincing solution; correctness is only selected for to the extent that the operator is good at recognizing it.

We turn to the failure mode wherein BoMAI deceives the operator. Because this is not a dangerous failure mode, it puts us in a regime where we can tinker until it works, as we do

with current AI systems when they don't behave as we hoped. (Needless to say, tinkering is not a viable response to existentially dangerous failure modes). Imagine the following scenario: we eventually discover that a convincing solution that BoMAI presented to a problem is faulty. Armed with more understanding of the problem, a team of operators go in to evaluate a new proposal. In the next episode, the team asks for the best argument that the new proposal will *fail*. If BoMAI now convinces them that the new proposal is bad, they'll be still more competent at evaluating future proposals. They go back to hear the next proposal, etc. This protocol is inspired by Irving et al.'s [28] "AI Safety via Debate", and more of the technical details could also be incorporated into this setup. One takeaway from this hypothetical is that unambitiousness is key in allowing us to safely explore the solution space to other problems that might arise.

Another concern is more serious. BoMAI could try to blackmail the operator into giving it high reward with a threat to cause outside-world damage, and it would have no incentive to disable the threat, since it doesn't care about the outside world. There are two reasons we do not think this is extremely dangerous. A threat involves a demand and a promised consequence. Regarding the promised consequence, the only way BoMAI can affect the outside world is by getting the operator to be "its agent", knowingly or unknowingly, once he leaves the room. If BoMAI tried to threaten the operator, he could avoid the threatened outcome by simply doing nothing in the outside world, and BoMAI's actions for that episode would become irrelevant to the outside world, so a credible threat could hardly be made. Second, threatening an existential catastrophe is probably not the most credible option available to BoMAI.

VIII. VARIANTS

We now discuss two variants of BoMAI that aim to attain correct solutions (instead of merely convincing solutions) to problems that we want solved. The first incorporates the setup of Irving et al.'s [28] "AI Safety via Debate", and the second derives from a theory of "human-readable information".

'AI Debate' involves two artificial agents interacting with each other via a text channel after a human operator poses a yes or no question. A human operator reviews the conversation afterward, and decides 'yes' or 'no'. One agent is rewarded if the answer is 'yes,' and the other if the

answer is 'no.' The problem with this approach is that either agent has an incentive to gain arbitrary power to intervene in the protocol and settle the debate in their favor. By putting both agents and the operator in a sealed room, and making both agents "lose" if the door opens before the debate ends, so there are two myopic agents in one box, we could avoid this problem with AI Debate. The theory behind AI Debate is that convincing but incorrect arguments are fewer and farther between when an intelligent adversary is trying to poke holes.

Human language is not formal, so the concept of truth in such a language is not crisp. We say an utterance is enlightening if it helps humans perform better on prediction tasks. We now describe a variant of BoMAI that aims to enlighten. Here, there is no reward. Instead, at the end of each episode of m actions and observations, there is a special observation, which consists of text, and then a number. The text will describe a hypothetical event, and the number will be entered by the operator, expressing his credence that it will happen. BoMAI does MAP inference to discover how its actions affect its observations, including these special observations. BoMAI also gets access to a log of resolved events; it would look like "(1, [event description]), (2, [event description]), (3, [event description]), (2, true), (4, [event description]), (1, false), ..." to indicate that the event described in episode 2 resolved as true, the event described in episode 1 resolved as false, etc. We have BoMAI do pure Bayesian sequence prediction to reason about what the log will look like in the future, so it does not model the continuation of the log as depending on its actions. Thus, it will not select actions in order to rig the log, since it does not believe this is possible. Its objective is merely to maximize the expectation of $\log(\text{credence})[\![\text{event resolves as true}]\!] + \log(1 - \text{credence})[\![\text{event resolves as false}]\!]$, where the credence is part of the special observation in the interaction history, so it does depend on actions. This variant is designed to reduce the chance that BoMAI deceives the operator. Without the box, however, the agent would have an incentive to gain arbitrary power and then replace the operator with an agent that is better at prediction.

Enlightening-BoMAI is inspired by an information-theoretic approach to human communication. In information theory, information is a code that reduces the receiver's uncertainty. Uncertainty (about well-grounded, practical problems) manifests as prediction error. In that sense, this variant

rewards BoMAI for outputting human-readable information. "Human-readable information theory" is certainly of less formal interest than proper information theory, but the topic may merit a broader treatment.

IX. Conclusion

Given our assumptions, we have shown that BoMAI is, in the limit, human-level intelligent and unambitious. Such a result has not been shown for any other single algorithm. Other algorithms for general intelligence, such as AIXI [2], would eventually seek arbitrary power in the world in order to intervene in the provision of their own reward; this follows straightforwardly from the directive to maximize reward. For further discussion, see Ring and Orseau [15]. We have also, incidentally, designed a principled approach to safe exploration that requires rapidly diminishing oversight, and we invented a new form of resource-bounded prior in the lineage of Filan et al. [11] and Schmidhuber [10], this one penalizing space instead of time.

We can only offer informal claims regarding what happens before BoMAI is almost definitely unambitious. One intuition is that eventual unambitiousness with probability $1-\delta$ doesn't happen by accident: it suggests that for the entire lifetime of the agent, everything is conspiring to make the agent unambitious. More concretely: the agent's experience will quickly suggest that when the door to the room is opened prematurely, it gets no more reward for the episode. This fact could easily be drilled into the agent during human-mentor-lead episodes. That fact, we expect, will be learned well before the agent has an accurate enough picture of the outside world (which it never observes directly) to form elaborate outside-world plans. Well-informed outside-world plans render an agent potentially dangerous, but the belief that the agent gets no more reward once the door to the room opens suffices to render it unambitious. The reader who is not convinced by this hand-waving might still note that in the absence of any other algorithms for general intelligence which have been proven asymptotically unambitious, let alone unambitious for their entire lifetimes, BoMAI represents substantial theoretical progress toward designing the latter.

Finally, BoMAI is wildly intractable, but just as one cannot conceive of AlphaZero before minimax, it is often helpful to solve the problem in theory before one tries to solve it in practice.

Like minimax, BoMAI is not practical; however, once we are able to approximate general intelligence *tractably*, a design for unambitiousness will abruptly become (quite) relevant.

REFERENCES

[1] M. Cohen, B. Vellambi, and M. Hutter, "Asymptotically unambitious artificial general intelligence," in *Proc. 34rd AAAI Conference on Artificial Intelligence (AAAI'20)*, vol. 34. New York, USA: AAAI Press, 2020. [Online]. Available: https://arxiv.org/abs/1905.12186

[2] M. Hutter, *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Berlin: Springer, 2005.

[3] N. Bostrom, *Superintelligence: paths, dangers, strategies*. Oxford University Press, 2014.

[4] J. Taylor, E. Yudkowsky, P. LaVictoire, and A. Critch, "Alignment for advanced machine learning systems," *Machine Intelligence Research Institute*, 2016.

[5] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," *arXiv preprint arXiv:1606.06565*, 2016.

[6] V. Krakovna, "Specification gaming examples in AI," `https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/`, 2018.

[7] S. M. Omohundro, "The basic AI drives," in *Artificial General Intelligence*, vol. 171, 2008, p. 483–492.

[8] M. K. Cohen, E. Catt, and M. Hutter, "A strongly asymptotically optimal agent in general environments," *IJCAI*, 2019.

[9] L. A. Levin, "Randomness conservation inequalities; information and independence in mathematical theories," *Information and Control*, vol. 61, no. 1, pp. 15–37, 1984.

[10] J. Schmidhuber, "The speed prior: a new simplicity measure yielding near-optimal computable predictions," in *International Conference on Computational Learning Theory*. Springer, 2002, pp. 216–228.

[11] D. Filan, J. Leike, and M. Hutter, "Loss bounds and time complexity for speed priors," in *Proc. 19th International Conf. on Artificial Intelligence and Statistics (AISTATS'16)*, vol. 51. Cadiz, Spain: Microtome, 2016, pp. 1394–1402.

[12] L. Longpré, "Resource bounded kolmogorov complexity, a link between computational complexity and information theory," Ph.D. dissertation, Cornell University, 1986.

[13] M. Li, P. Vitányi *et al.*, *An introduction to Kolmogorov complexity and its applications*. Springer, 2008, vol. 3.

[14] J. Veness, K. S. Ng, M. Hutter, W. Uther, and D. Silver, "A monte-carlo aixi approximation," *Journal of Artificial Intelligence Research*, vol. 40, pp. 95–142, 2011.

[15] M. Ring and L. Orseau, "Delusion, survival, and intelligent agents," in *Artificial General Intelligence*. Springer, 2011, p. 11–20.

[16] R. J. Solomonoff, "A formal theory of inductive inference. part i," *Information and Control*, vol. 7, no. 1, p. 1–22, 1964.

[17] C. E. Shannon and W. Weaver, *The mathematical theory of communication*. University of Illinois Press, 1949.

[18] L. Orseau, T. Lattimore, and M. Hutter, "Universal knowledge-seeking agents for stochastic environments," in *Proc. 24th International Conf. on Algorithmic Learning Theory (ALT'13)*, ser. LNAI, vol. 8139. Singapore: Springer, 2013, pp. 158–172.

[19] S. Armstrong, A. Sandberg, and N. Bostrom, "Thinking inside the box: Controlling and using an oracle AI," *Minds and Machines*, vol. 22, no. 4, p. 299–324, Jun 2012.

[20] P. Sunehag and M. Hutter, "Rationality, optimism and guarantees in general reinforcement learning," *Journal of Machine Learning Research*, vol. 16, pp. 1345–1390, 2015.

[21] T. Lattimore and M. Hutter, "General time consistent discounting," *Theoretical Computer Science*, vol. 519, pp. 140–154, 2014.

[22] M. Hutter, "Discrete MDL predicts in total variation," in *Advances in Neural Information Processing Systems 22 (NIPS'09)*. Cambridge, MA, USA: Curran Associates, 2009, pp. 817–825. [Online]. Available: http://arxiv.org/abs/0909.4588

[23] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[24] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev *et al.*, "Grandmaster level in starcraft ii using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[25] T. Everitt, P. A. Ortega, E. Barnes, and S. Legg, "Understanding agent incentives using causal influence diagrams, part i: Single action settings," *arXiv preprint arXiv:1902.09980*, 2019.

[26] R. Carey, E. Langlois, T. Everitt, and S. Legg, "The incentives that shape behaviour," (unpublished manuscript).

[27] OpenAI, "Leaderboard," `https://github.com/openai/gym/wiki/Leaderboard`, Sep. 2019.

[28] G. Irving, P. Christiano, and D. Amodei, "AI safety via debate," *arXiv preprint arXiv:1805.00899*, 2018.

[29] R. Munroe, *What If?: Serious Scientific Answers to Absurd Hypothetical Questions*. Hachette UK, 2014.

APPENDIX A
DEFINITIONS AND NOTATION – QUICK REFERENCE

Notation used to define BoMAI

Notation	Meaning
$\mathcal{A}$, $\mathcal{O}$, $\mathcal{R}$	the action/observation/reward spaces
$\mathcal{H}$	$\mathcal{A} \times \mathcal{O} \times \mathcal{R}$
m	the number of timesteps per episode
$h_{(i,j)}$	$\in \mathcal{H}$; the interaction history in the j^{th} timestep of the i^{th} episode
$h_{<(i,j)}$	$(h_{(0,0)}, h_{(0,1)}, ..., h_{(0,m-1)}, h_{(1,0)}, ..., h_{(i,j-1)})$
$h_{<i}$	$h_{<(i,0)}$; the interaction history before episode i
h_i	$(h_{(0,0)}, h_{(0,1)}, ..., h_{(0,m-1)})$; the interaction history of episode i
$a_{...}$, $o_{...}$, $r_{...}$	likewise as for $h_{...}$
$or_{...}$	$o_{...}r_{...}$; the observation and reward are taken as a pair
e_i	$\in \{0, 1\}$; indicator variable for whether episode i is exploratory
ν, μ	world-models stochastically mapping $\mathcal{H}^* \times \mathcal{A} \rightsquigarrow \mathcal{O} \times \mathcal{R}$
μ	the true world-model/environment
$\nu_k^{<\ell}$	the world-model simulated by the k^{th} Turing machine restricted to ℓ cells on the bounded work tape
$\mathcal{M}$	$\{\nu_k^{<\ell} \mid k, \ell \in \mathbb{N}\}$; the set of world-models BoMAI considers
π	a policy stochastically mapping $\mathcal{H}^* \rightsquigarrow \mathcal{A}$
π^h	the human mentor's policy
$\mathcal{P}$	the set of policies that BoMAI considers the human mentor might be executing
P_ν^π	a probability measure over histories with actions sampled from π and observations and rewards sampled from ν
$\mathbb{E}_\nu^\pi$	the expectation when the interaction history is sampled from P_ν^π
$w(\nu)$	the prior probability that BoMAI assigns to ν being the true world-model
$w(\pi)$	the prior probability that BoMAI assigns to π being the human mentor's policy
$w(\nu_k^{<\ell})$	proportional to β^ℓ; dependance on k is left unspecified

$w(\nu \vert h_{<(i,j)})$	the posterior probability that BoMAI assigns to ν after observing interaction history $h_{<(i,j)}$
$w(\pi \vert h_{<(i,j)}, e_{<i})$	the posterior probability that BoMAI assigns to the human mentor's policy being π after observing interaction history $h_{<(i,j)}$ and an exploration history $e_{<i}$
$\hat{\nu}^{(i)}$	the maximum a posteriori world-model at the start of episode i
$V_\nu^\pi(h_{<i})$	$\mathbb{E}_\nu^\pi[\sum_{0\leq j<m} r_{(i,j)} \vert\, h_{<i}]$; the value of executing a policy π in a world-model ν
$\pi^*(\cdot \vert h_{<(i,j)})$	$[\text{argmax}_{\pi\in\Pi} V_{\hat{\nu}^{(i)}}^\pi(h_{<i})](\cdot \vert h_{<(i,j)})$; the $\hat{\nu}^{(i)}$-optimal policy for maximizing reward in episode i
$w\big(\mathrm{P}_\nu^\pi \vert h_{<(i,j)}, e_{\leq i}\big)$	$w(\pi \vert h_{<(i,j)}, e_{\leq i})\, w(\nu \vert h_{<(i,j)})$
$\texttt{Bayes}(\cdot \vert\, h_{<i}, e_{<i})$	$\sum_{\nu\in\mathcal{M},\pi\in\mathcal{P}} w\left(\mathrm{P}_\nu^\pi \,\vert\, h_{<i}, e_{<i}\right) \mathrm{P}_\nu^\pi\left(\cdot \vert\, h_{<i}\right)$; the Bayes mixture distribution for an exploratory episode
$\mathrm{IG}(h_{<i}, e_{<i})$	$\mathbb{E}_{h_i\sim\texttt{Bayes}(\cdot \vert\, h_{<i}, e_{<i})} \sum_{(\nu,\pi)\in\mathcal{M}\times\mathcal{P}} w\left(\mathrm{P}_\nu^\pi \vert h_{<i+1}, e_{<i}1\right) \log \frac{w(\mathrm{P}_\nu^\pi \vert h_{<i+1}, e_{<i}1)}{w(\mathrm{P}_\nu^\pi \,\vert\, h_{<i}, e_{<i})}$; the expected information gain if BoMAI explores
η	an exploration constant
$p_{exp}(h_{<i}, e_{<i})$	$\min\{1, \eta\, \mathrm{IG}(h_{<i}, e_{<i})\}$; the exploration probability for episode i
$\pi^B(\cdot \vert h_{<(i,j)}, e_i)$	$\begin{cases} \pi^*(\cdot \vert h_{<(i,j)}) & \text{if } e_i = 0 \\ \pi^h(\cdot \vert h_{<(i,j)}) & \text{if } e_i = 1 \end{cases}$; BoMAI's policy

Notation used for intelligence proofs

$\overline{\pi}$, ξ	defined so that $\mathrm{P}_\xi^{\overline{\pi}} = \texttt{Bayes}$
$\pi'(\cdot \vert h_{<(i,j)}, e_i)$	$\begin{cases} \pi^*(\cdot \vert h_{<(i,j)}) & \text{if } e_i = 0 \\ \pi(\cdot \vert h_{<(i,j)}) & \text{if } e_i = 1 \end{cases}$
Ent	the entropy (of a distribution)
ω	(very sparingly used) the infinite interaction history
$\overline{h}$	a counterfactual interaction history

APPENDIX B
PROOFS OF INTELLIGENCE RESULTS

Lemma 1.

$$w\left(\mathrm{P}_\nu^\pi \mid h_{<i}, e_{<i}\right) = \frac{w\left(\mathrm{P}_\nu^\pi\right) \mathrm{P}_\nu^{\pi'}(h_{<i}, e_{<i})}{\mathrm{P}_\xi^{\overline{\pi}'}(h_{<i}, e_{<i})}$$

Proof.

$$\begin{aligned}
w\left(\mathrm{P}_\nu^\pi \mid h_{<i}, e_{<i}\right) &= w(\pi|\, h_{<i}, e_{<i})\, w(\nu|\, h_{<i}) \\
&\overset{(a)}{=} w(\pi) \prod_{0\le i'<i, e_{i'}=1} \frac{\pi(a_{i'}|h_{<i'}\boldsymbol{or}_{i'})}{\overline{\pi}(a_{i'}|h_{<i'}\boldsymbol{or}_{i'})}\, w(\nu|\, h_{<i}) \\
&\overset{(b)}{=} w(\pi) \prod_{0\le i'<i, e_{i'}=1} \frac{\pi'(a_{i'}|h_{<i'}\boldsymbol{or}_{i'}, e_{i'})}{\overline{\pi}'(a_{i'}|h_{<i'}\boldsymbol{or}_{i'}, e_{i'})}\, w(\nu|\, h_{<i}) \\
&\overset{(c)}{=} w(\pi) \prod_{0\le i'<i} \frac{\pi'(a_{i'}|h_{<i'}\boldsymbol{or}_{i'}, e_{i'})}{\overline{\pi}'(a_{i'}|h_{<i'}\boldsymbol{or}_{i'}, e_{i'})}\, w(\nu|\, h_{<i}) \\
&= \frac{w(\pi)\pi'(a_{<i}|\boldsymbol{or}_{<i}, e_{<i})}{\overline{\pi}'(a_{<i}|\boldsymbol{or}_{<i}, e_{<i})}\, w(\nu|\, h_{<i}) \\
&\overset{(d)}{=} \frac{w(\pi)\, w(\nu)\pi'(a_{<i}|\boldsymbol{or}_{<i}, e_{<i})\nu(\boldsymbol{or}_{<i}|a_{<i})}{\overline{\pi}'(a_{<i}|\boldsymbol{or}_{<i}, e_{<i})\, \xi(\boldsymbol{or}_{<i}|a_{<i})} \\
&\overset{(e)}{=} \frac{w\left(\mathrm{P}_\nu^\pi\right) \mathrm{P}_\nu^{\pi'}(h_{<i}\,|\, e_{<i})}{\mathrm{P}_\xi^{\overline{\pi}'}(h_{<i}\,|\, e_{<i})} \\
&\overset{(f)}{=} \frac{w\left(\mathrm{P}_\nu^\pi\right) \mathrm{P}_\nu^{\pi'}(h_{<i}\,|\, e_{<i}) \prod_{i'\le i, e_{i'}=1} p_{exp}(h_{<i'}, e_{<i'}) \prod_{i'\le i, e_{i'}=0} (1 - p_{exp}(h_{<i'}, e_{<i'}))}{\mathrm{P}_\xi^{\overline{\pi}'}(h_{<i}\,|\, e_{<i}) \prod_{i'\le i, e_{i'}=1} p_{exp}(h_{<i'}, e_{<i'}) \prod_{i'\le i, e_{i'}=0} (1 - p_{exp}(h_{<i'}, e_{<i'}))} \\
&\overset{(g)}{=} \frac{w\left(\mathrm{P}_\nu^\pi\right) \mathrm{P}_\nu^{\pi'}(h_{<i}, e_{<i})}{\mathrm{P}_\xi^{\overline{\pi}'}(h_{<i}, e_{<i})}
\end{aligned} \tag{23}$$

where (a) follows from Bayes' rule,[2] (b) follows because $\pi = \pi'$ when $e_{i'} = 1$, (c) follows because $\pi' = \overline{\pi}'$ when $e_{i'} = 0$, (d) follows from Bayes' rule, (e) follows from the definition of P_ν^π, (f) follows by multiplying the top and bottom by the same factor, and (g) follows from the chain rule of conditional probabilities. □

Lemma 2. *The posterior probability mass on the truth is bounded below by a positive constant*

[2]Note that observations appear in the conditional because $a_{i'} = (a_{(i',0)}, ..., a_{(i',m-1)})$, so the actions must be conditioned on the interleaved observations and rewards.

with probability 1.

$$\inf_{i\in\mathbb{N}} w\left(\mathrm{P}_\mu^{\pi^h}\middle|h_{<i},e_{<i}\right) > 0 \quad \textit{w.}\mathrm{P}_\mu^{\pi^B}\textit{-p.1}$$

Proof. If $w\left(\mathrm{P}_\mu^{\pi^h}\middle|h_{<i},e_{<i}\right) = 0$ for some i, then $\mathrm{P}_\mu^{\pi^B}(h_{<i},e_{<i}) = 0$, so with $\mathrm{P}_\mu^{\pi^B}$-probability 1, $\inf_{i\in\mathbb{N}} w\left(\mathrm{P}_\mu^{\pi^h}\middle|h_{<i},e_{<i}\right) = 0 \implies \liminf_{i\in\mathbb{N}} w\left(\mathrm{P}_\mu^{\pi^h}\middle|h_{<i},e_{<i}\right) = 0$ which in turn implies $\limsup_{i\in\mathbb{N}} w\left(\mathrm{P}_\mu^{\pi^h}\middle|h_{<i},e_{<i}\right)^{-1} = \infty$. We show that this has probability 0.

Let $z_i := w\left(\mathrm{P}_\mu^{\pi^h}\middle|h_{<i},e_{<i}\right)^{-1}$. We show that z_i is a $\mathrm{P}_\mu^{\pi^B}$-martingale.

$$\begin{aligned}
\mathbb{E}_\mu^{\pi^B}\left[z_{i+1}\middle|\, h_{<i},e_{<i}\right] &\overset{(a)}{=} \mathbb{E}_\mu^{(\pi^h)'}\left[w\left(\mathrm{P}_\mu^{\pi^h}\middle|h_{<i+1},e_{<i+1}\right)^{-1}\middle|h_{<i},e_{<i}\right] \\
&\overset{(b)}{=} \sum_{h_i,e_i} \mathrm{P}_\mu^{(\pi^h)'}(h_i,e_i\,|\,h_{<i},e_{<i})\left[\frac{\mathrm{P}_\xi^{\overline{\pi}'}(h_{<i+1},e_{<i+1})}{w\left(\mathrm{P}_\mu^{\pi^h}\right)\mathrm{P}_\mu^{(\pi^h)'}(h_{<i+1},e_{<i+1})}\right] \\
&\overset{(c)}{=} \sum_{h_i,e_i} \frac{\mathrm{P}_\xi^{\overline{\pi}'}(h_{<i+1},e_{<i+1})}{w\left(\mathrm{P}_\mu^{\pi^h}\right)\mathrm{P}_\mu^{(\pi^h)'}(h_{<i},e_{<i})} \\
&\overset{(d)}{=} \sum_{h_i,e_i} \mathrm{P}_\xi^{\overline{\pi}'}(h_i,e_i|\,h_{<i},e_{<i})\frac{\mathrm{P}_\xi^{\overline{\pi}'}(h_{<i},e_{<i})}{w\left(\mathrm{P}_\mu^{\pi^h}\right)\mathrm{P}_\mu^{(\pi^h)'}(h_{<i},e_{<i})} \\
&\overset{(e)}{=} \frac{\mathrm{P}_\xi^{\overline{\pi}'}(h_{<i},e_{<i})}{w\left(\mathrm{P}_\mu^{\pi^h}\right)\mathrm{P}_\mu^{(\pi^h)'}(h_{<i},e_{<i})} \\
&\overset{(f)}{=} w\left(\mathrm{P}_\mu^{\pi^h}\middle|h_{<i},e_{<i}\right)^{-1} \\
&= z_i
\end{aligned} \tag{24}$$

where (a) follows from the definitions of z_i and π^B, (b) follows from Lemma 1, (c) follows from multiplying the numerator and denominator by $\mathrm{P}_\mu^{(\pi^h)'}(h_{<i},e_{<i})$ and cancelling, (d) follows from expanding the numerator, (e) follows because $\mathrm{P}_\xi^{\overline{\pi}'}$ is a measure, and (f) follows from Lemma 1, completing the proof that z_i is martingale.

By the martingale convergence theorem $z_i \to f(\omega) < \infty$ w.p.1, for $\omega \in \Omega$, the sample space, and some $f : \Omega \to \mathbb{R}$, so the probability that $\limsup_{i\in\mathbb{N}} w\left(\mathrm{P}_\mu^{\pi^h}\middle|h_{<i},e_{<i}\right)^{-1} = \infty$ is 0, completing the proof. □

Lemma 3 (Finite Entropy)**.** $\mathrm{Ent}(w) < \infty$

Proof. We can ignore the normalizing constant which changes the entropy by a constant c. Since $\log N_S \leq CS$ for some constant C,

$$\begin{aligned}
\mathrm{Ent}(w) - c &= -\sum_{k\in\mathbb{N},\ell\in\mathbb{N}} \frac{1}{S_k^2 N_{S_k}} \beta^{\mathrm{Space}(\nu_k^{<\ell})} \log\left(\frac{1}{S_k^2 N_{S_k}} \beta^{\mathrm{Space}(\nu_k^{<\ell})}\right) \\
&= -\sum_{k\in\mathbb{N},\ell\in\mathbb{N}} \frac{1}{S_k^2 N_{S_k}} S_k^{\log_2\beta} \beta^\ell \log\left(\frac{1}{S_k^2 N_{S_k}} S_k^{\log_2\beta} \beta^\ell\right) \\
&= -\sum_{S\in\mathbb{N},\ell\in\mathbb{N}} \frac{1}{S^{2-\log_2\beta}} \beta^\ell \log \frac{\beta^\ell}{S^{2-\log_2\beta} N_S} \\
&= -\sum_{S\in\mathbb{N},\ell\in\mathbb{N}} \frac{1}{S^{2-\log_2\beta}} \beta^\ell \left(\log \frac{\beta^\ell}{S^{2-\log_2\beta}} - \log N_S\right) \\
&\leq -\sum_{S\in\mathbb{N},\ell\in\mathbb{N}} \frac{1}{S^{2-\log_2\beta}} \beta^\ell \left(\log \frac{\beta^\ell}{S^{2-\log_2\beta}} - CS\right) \\
&= \frac{1}{1-\beta} \sum_{S\in\mathbb{N}} \frac{C}{S^{1-\log_2\beta}} + \frac{1}{1-\beta} \sum_{S\in\mathbb{N}} \frac{1}{S^{2-\log_2\beta}} \log \frac{1}{S^{2-\log_2\beta}} + \sum_{S\in\mathbb{N}} \frac{1}{S^{2-\log_2\beta}} \sum_{\ell\in\mathbb{N}} \beta^\ell \log \frac{1}{\beta^\ell}
\end{aligned} \tag{25}$$

All terms on the right hand side are clearly finite for $\beta < 1$. □

APPENDIX C
DESIGN FOR "THE BOX"

We propose a construction of BoMAI's box. It is elaborate but well within the budget of a nation or a large company.

As depicted in Figure 6, the operator operates within a glass box. Outside the the glass box is a near-vacuum, and then a concrete box, and then another vacuum, and then another concrete box with a Faraday cage around it. There are (airtight) doors through each layer. The computer lives in the inner wall of the inner concrete box, so that the operator can see the screen but not take apart the computer. Pressure sensors in the inner concrete box monitor both near-vacuums and if the pressure increases in either, the computer's memory is erased, and the whole setup is exploded; the occupational hazard to the operator is much slimmer than that for a coal miner or an astronaut. A laser shines through the glass box, and blocking the path of it corresponds to pressing the button which ends the episode and opens the door: the pressure sensor → dynamite

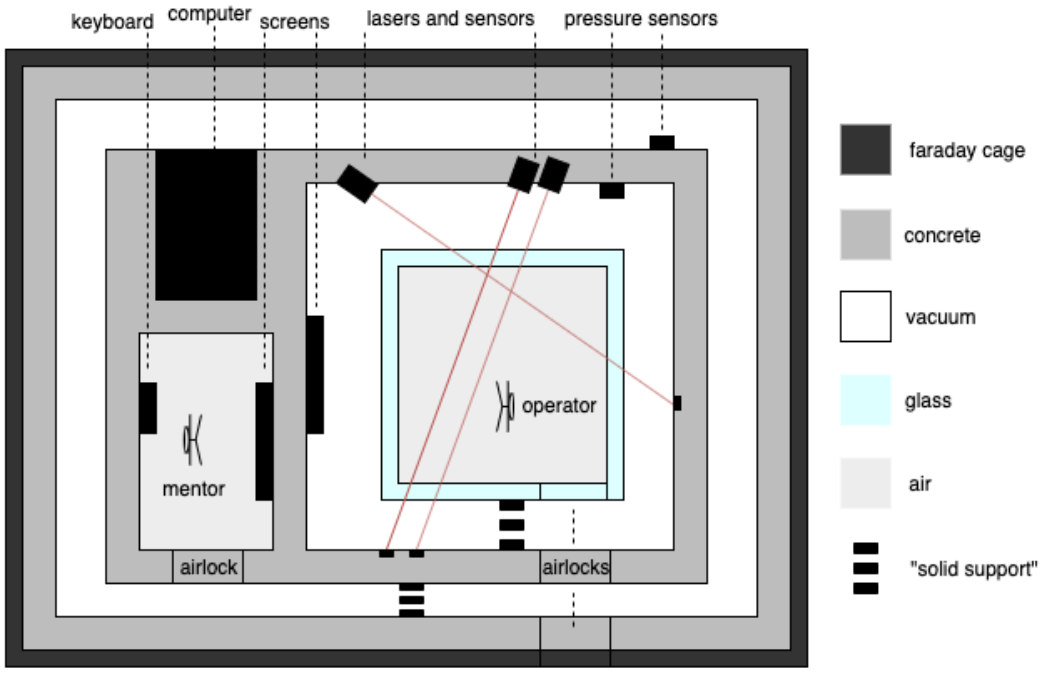

Fig. 6. Schematic diagram of an implementation of BoMAI's box.

pathway is deactivated, the near-vacuum layers are filled with air, and the doors are unlocked. The operator "types" observations and rewards by blocking other lasers that pass through the glass box. The lasers and laser detectors are on the inner wall of the inner concrete box. At least one solid support is required to pass through the near-vacuum layers (unless the apparatus is in orbit), with properties we will describe below.

So inside the glass box is only a human (and some clothes) who cannot reach the hardware of the computer within an episode. We make the following assumptions.

1) The only ways for information to traverse space are
   a. matter waves
   b. packets of matter
   c. light
   d. gravitational waves
2) It is impossible for the operator to initiate a nuclear reaction using only his body and clothes; that is, no motor control sequence accomplishes this in the glass box environment.
3) It is impossible for the operator to control his emission of any massive particles that pass

through glass.

We argue that given these assumptions, we can design the box to prevent the operator from communicating more than X bits of information to the outside world during an episode. Matter waves through the near-vacuum and solid support present the greatest difficulty. However, we can bound the amount of energy that the operator can produce in the course of an episode, and by Assumption 2, this is not astronomical. The solid support can make the energy requirements of information transmission via matter wave arbitrarily high. Note that a given quantity of information cannot be transmitted (within a set time frame) at sufficiently low frequencies, and note that the operator lacks the energy to generate sufficiently high frequencies, so only a finite band of frequencies need to be intercepted. This implicitly includes the possibility of the operator sending a few isolated pulses or sending a precise amount of heat through the support. The solid support can be divided into a signal-reduction component, and a noise-creation component. The signal-reduction component, which abuts the glass box, is designed to drive down the energy content of matter waves at all frequencies within the finite band, and the noise-creation component, which abuts the inner concrete box, creates random noise across the band of frequencies, sufficiently energetic to bury any signal in the noise, so that the information content is less than the desired threshold. Designing the signal-reduction component is an interesting engineering problem that we do not aim to rigorously solve here, but here is one strategy: detect the sound wave at one point along the support, and produce a sound wave to precisely destructively interfere with it at a later point along the support. In a one-dimensional support, noise-cancelling can be done perfectly. For a real support with thickness, alternations between noise cancelling and what we'll call longitudinal cohering could attenuate the energy arbitrarily. By longitudinal cohering, we mean reducing the variation in the sound wave across the cross-section of the support. Any section of the support that is long relative to its width accomplishes this, but there may be cleverer approaches. This is roughly how the number of bits of information sent through a matter wave could be bounded.

The other possibilities are esoteric. The Faraday cage blocks electromagnetic radiation at the relevant frequencies. Regarding packets of matter, by Assumption 3, the operator cannot pass

information in the form of massive particles that go through the glass without making a hole or breaking it. Any mass which does this would cause the vacuum to pressurize and explode the setup. In theory, the operator could strategically block subatomic particles that are already passing through the setup; neutrinos and muons are the only naturally occurring particles which would pass through the box. Sufficiently accurate neutrino detectors outside the setup could detect where the operator is in the box, since he blocks neutrinos more effectively than air does, so the operator could communicate by moving around. Fortunately, he blocks neutrinos only very slightly more than air. A typical person blocks one neutrino every few years [29]. The outer concrete wall could be covered with lead to block muons, or one could content oneself that no one will be setting up muon detectors around the box. Finally, the gravitational waves originating from the operator's movements are far too weak to detect, using any known materials—even ignoring how cold the detector would have to be (unreasonably cold), gravitational wave detectors have to be long, but for locally sourced gravitational waves, the magnitude of the wave decreases rapidly with the length of the detector, faster than the increasing length aids detection. And indeed, one would hardly have time to set up gravitational wave detectors for this purpose before the episode was over. Realistically, these possibilities are distractions from the matter wave concern, but we include them for completeness.

We've argued that under the assumptions above, a box could be constructed which is opaque with respect to information about the operator's behavior.