Asymptotically Unambitious Artificial General Intelligence

Abstract

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

1 Introduction

The project of Artificial General Intelligence (AGI) is "to make computers solve really difficult problems." [Minsky, 1961] Expanding on this, what we want from an AGI is a system that (a) can solve any solvable task, and (b) can be steered toward solving any particular one given some form of input we provide.

One proposal for AGI is reinforcement learning, which works as follows: (1) construct a "reward signal" meant to express our satisfaction with an artificial agent; (2) design an algorithm which learns to pick actions that maximize its expected reward, usually utilizing other observations too; and (3) ensure that solving the task we have in mind leads to higher reward than can be attained otherwise. As long as (3) holds, then insofar as the algorithm is able to maximize expected reward, it can be used as an AGI.

A problem arises: it is inconveniently true that if the AI manages to take over the world, and ensure its continued dominance by neutralizing all other intelligent threats (read: people), it could intervene in the provision of its own reward to achieve maximal reward for the rest of its lifetime [Bostrom, 2014; Taylor *et al.*, 2016]. "Reward hijacking" is just the correct way for a reward maximizer to behave [Amodei *et al.*, 2016]. Insofar as the AI is able to maximize expected reward, (3) fails. The broader principle at work is Goodhart's Law: "Any observed statistical regularity [like the correlation between reward and task-completion] will tend to collapse once pressure is placed upon it for control purposes." [Goodhart, 1984] Indeed, Krakovna [2018] has compiled an annotated bibliography of examples of artificial optimizers "hacking" their objective. An alternate way to understand this expected behavior is Omohundro's Instrumental Convergence Thesis [Omohundro, 2008], 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.

To answer the failure mode of reward hijacking, we present Boxed Myopic Artificial Intelligence (BoMAI), the first algorithm we are aware of for a reinforcement learner which, in the limit, is indifferent to gaining power in the outside-world. The key features are these: BoMAI maximizes reward episodically, it is run on a computer which is placed in a sealed room with an operator, and if the operator leaves the room, the episode ends. We argue that our algorithm produces an AGI that, even if it became omniscient, would continue to accomplish whatever task we wanted, instead of hijacking its reward, eschewing its task, and neutralizing threats to it, even if it saw clearly how to do exactly that. We thereby defend reinforcement learning as a path to AGI, despite the default, dangerous failure mode of reward hijacking.

We take the key insights from Hutter's AIXI [Hutter, 2005], which is a Bayes-optimal reinforcement learner, but which cannot be made to solve arbitrary tasks, given its eventual degeneration into reward hijacking [Ring and Orseau, 2011]. We take further insights from Solomonoff's [1964] universal prior, Shannon's [1949] formalization of information, Orseau's [2013] knowledge-seeking agent, and Armstrong's [2012] and Bostrom's [2014] theorized Oracle AI, and we design an algorithm which can be reliably directed, in the limit, to solve any solvable task.

2 Boxed Myopic Artificial Intelligence

We turn now to the setup and the algorithm for BoMAI. The setup refers to the physical surroundings of the computer on

which the algorithm is run. We present the algorithm informally as well as formally. The informal description will give a near-complete picture of the algorithm to a reader familiar with Bayesian reinforcement learning, and the formal description is in Appendix A. All appendices may be found with the body of the paper at http://tinyurl.com/bomai-anon. When viewed in Google Chrome, most mathematical notation may be hovered over to give a reminder of the meaning. Appendix B includes a table of this notation.

2.1 Setup

The standard reinforcement learning setup is as follows: at each "timestep," an agent submits an action, then receives an observation and a reward. This cycle continues for as long as desired. For BoMAI, its actions take the form of bounded-length strings of text which get printed to a screen, and its observations take the form of bounded-length strings of text which a human operator enters. The reward belongs to a finite set of rationals between 0 and 1. BoMAI's lifetime is divided into episodes of length m, and it attempts to maximize only the reward it receives during the episode it is in, hence it is "myopic."

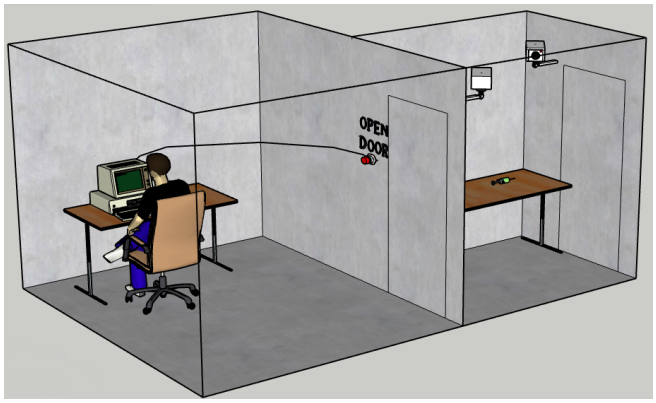

Figure 1: **Physical Setup Implementing BoMAI.** Opening the door ends the episode. Information cannot escape otherwise.

As depicted in Figure 1, BoMAI is run on a computer which is placed in a room with a human operator. The room has one door. As long as the door is closed, information cannot leave the room, hence "boxed". 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, then for those remaining timesteps, BoMAI recieves a reward of 0, and its observation is an empty string of text. There is also a set time limit for the episode, just to ensure it does not last forever. With the episode now completed, BoMAI does not immediately take its next action, and the operator can leave the room. The operator then goes into an antechamber, and with probability $p_{;}$, he must take a drug causing retrograde amnesia before leaving.[1] When the next person comes in, BoMAI takes its next action. BoMAI is run on a local, formally verified operating system, disconnected from the outside-world. The engineering problem of making this is discussed in Appendix C.

This setup constrains the causal dependencies between BoMAI and the environment, as depicted in Figure 2. 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. "Timestep (i, j)" refers to the j^{th} timestep of episode i.

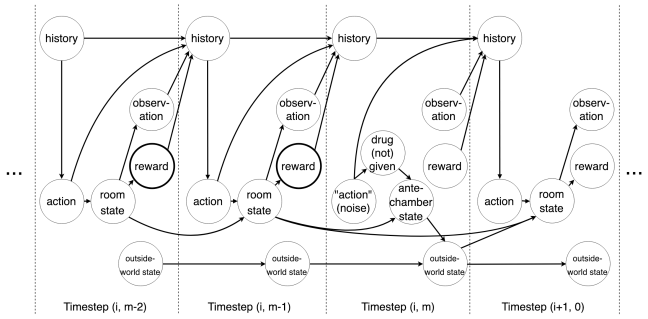

Figure 2: **Causal Dependencies Governing the Interaction Between BoMAI and the Environment.** Unrolling this diagram for all timesteps gives the full causal graph. The bold reward nodes are the ones that BoMAI maximizes during episode i. For timestep (i, m), the randomly selected "action" determines whether amnesia is administered, and the observation and reward are "" and 0.

2.2 Algorithm – Informal Description

With the setup described, we now outline how exactly BoMAI maximizes expected reward. Again, some choices will be motivated in later sections. A central problem in reinforcement learning is the trade-off between exploitation and exploration. Exploitation is maximizing expected reward according to the agent's current beliefs, and exploration is taking some other action in order to learn more about the dynamics of the environment, so the agent's beliefs can become more accurate. We'll start by discussing how BoMAI exploits.

BoMAI assumes that the environment producing its observations and rewards could in principle be perfectly simulated on a computer with access to random bits, given the entire history of actions, observations, and rewards, using only a bounded number of computation steps per episode. BoMAI does not assume the environment resets at the start of each episode. Using the probabilistic Turing machine model of computation, it associates to each pair (Turing machine, computation bound) a world-model that that bounded Turing machine simulates. For example, if the computation bound is 100, the Turing machine is allowed 100 computation steps per episode. BoMAI is also parameterized by a slack factor n: each Turing machine is allowed an extra n computation steps total, in addition to its computation allowance per episode. BoMAI begins with a prior distribution over which bounded Turing machine is simulating the true environment, and updates these beliefs as a Bayesian with each successive observation. For a world-model based on a Turing machine described by the binary string s, with computation bound ℓ, the prior is proportional to $2^{-K(s)(1+\varepsilon)}\beta^{\ell}$, where $K(s)$ is the

[1] This design choice requires significant exposition to motivate; it appears in §3.3. We hope that future research yields a path to asymptotic unambitiousness that avoids this feature.

length of the shortest program that outputs s on some universal Turing machine U, $\varepsilon > 0$, and $0 < \beta < 1$.

At the start of each episode, BoMAI identifies the bounded Turing machine that has the highest weight in its posterior, and for the remainder of the episode, exploiting-BoMAI maximizes its expected reward within that episode, assuming its environment is simulated by that bounded Turing machine.

When BoMAI explores, a human selects its action. We'll call this person the "explorer." Like the operator, she is also "boxed"–she is in a sealed room, and she is only able to leave when the operator leaves. When BoMAI would explore, the explorer is alerted, and for the remainder of the episode, she submits actions which BoMAI executes. Like BoMAI, she has access to the interaction history when picking an action.

The probability that BoMAI explores is proportional to the amount of information that BoMAI expects to gain from the human-controlled episode, both about the explorer's policy, and about the true world-model.[2] Formally, "information gain" measures how poorly BoMAI's current posterior approximates its future posterior, using the Kullback-Leibler (KL) divergence.

3 Results

We first give theoretical results relating to BoMAI's intelligence: it learns to accumulate reward at least as well as the human explorer, and its exploration probability goes rapidly to 0. There is no clear result to aim for between human-level and optimal, so we cannot show this formally, but we argue that it would accumulate reward at a vastly superhuman level. We then discuss the results pertaining to unambitiousness, which will also justify the more curious design choices in the setup that the reader might be wondering about. We show that BoMAI does not have an incentive to hijack its reward or indeed accomplish any outside-world objectives. We then informally respond to a couple concerns about how to ensure BoMAI completes tasks in the desired way. In defense of BoMAI's failure to achieve asymptotic optimality, which we could achieve if we allowed it to pick its own exploratory policy, we show in Appendix G (given weak assumptions) that any asymptotically optimal agent would eventually destroy itself by accident, quite possibly with a great deal of collateral damage.

3.1 Intelligence

All intelligence results depend on the assumption that BoMAI assigns nonzero prior to the truth. Formally, we let $\mathcal{M}$ be the class of world-models that BoMAI considers (those which can be simulated by a bounded-time Turing machine with a slack of n computation steps), and we let $\mathcal{P}$ be some class of policies that BoMAI considers the human explorer might be executing.

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

[2]The exploration probability is proportional to the expected information gain, but for the probability being capped at 1.

One definition of $\mathcal{P}$ that would presumably satisfy this assumption is the set of computable functions (that are properly typed).

The intelligence theorems are stated here and proven in Appendix D along with a few supporting lemmas. Our first result is that the exploration probability is square-summable almost surely: letting $\mathbb{E}_\mu^{\pi^B}$ denote the expectation when actions are sampled from BoMAI's policy π^B and observations and rewards are sampled from the true environment μ, and letting p_{exp} be the exploration probability, which depends on $h_{<i}$, the interaction history up to episode i, and $e_{<i}$, the history of which prior episodes were exploratory, we have

Theorem 1 (Limited Exploration)**.**

$$\mathbb{E}_\mu^{\pi^B} \sum_{i=1}^{\infty} p_{exp}\left(h_{<i}, e_{<i}\right)^2 < \infty$$

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.*, 2016] discuss.

The On-Human-Policy and On-Policy Optimal Prediction Theorems state that predictions under 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 explorer's policy or from 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. π^h is the human explorer's policy, and $\hat{\nu}^{(i)}$ is BoMAI's maximum a posteriori world-model for episode i.

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

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

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 μ. Next, 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-Policy Optimal Prediction)**.**

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

Given asymptotically optimal prediction on-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 $\varepsilon > 0$. Letting V_μ^π be the expected reward (within the episode) for a policy π in the environment μ, we state this as follows:

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

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

This completes the formal results regarding BoMAI's intelligence–namely that BoMAI approaches perfect prediction on-policy and on-human-policy, and accumulates reward

at least as well as the human explorer. Since this result is independent of what task the operator decides to reward, we say that BoMAI achieves human-level intelligence, and could be called a power-bounded AGI. In fact, we expect BoMAI's accumulation of reward to 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 remotely approaches the ceiling of what can be learned from their observations.

We offer a couple examples of what BoMAI could be used for. BoMAI could be asked to present a promising agenda for cancer research, and rewarded based on how plausible the operator finds it (the operator being an expert in the field). BoMAI could be asked to outline a screenplay, and then in future episodes, asked to flesh it out section by section. Or, if m is large enough, it can be asked to write the whole thing. Indeed, for any task that can be accomplished in a sealed room, if the human only rewards solutions, and if m is large enough for it to be doable, BoMAI will learn to do it at at least human-level. We continue this discussion in Section 3.4.

3.2 Safety – Informal Argument

For a standard generally intelligent reinforcement learner, we can expect optimal behavior to include gaining arbitrary power in the world, so that it can intervene in the provision of its own reward [Ring and Orseau, 2011]. Recall that power is a position from which it is relatively easier to achieve arbitrary goals. The purpose of our "boxed" setup is to eliminate any outside-world instrumental goals, where the "outside-world" is just the world outside the room. In other words, the agent will not find it useful (in the project of reward maximization) to direct the outside-world toward one state or another. We say that a lack of outside-world instrumental goals suffices for unambitiousness. Note also that if arbitrary power is not being selected for, it is unlikely to arise by chance.

Why, then, does BoMAI not have outside-world instrumental goals? A first pass at this goes as follows: there is no outside-world event which is a causal descendant of the actions in episode i and a causal ancestor of the rewards in episode i, because in order for BoMAI's actions to affect the outside-world, the door to the room must open, which ends the episode.

The remaining difficulty, and the reason why this is only a first pass, is that we need to ensure that BoMAI *understands* that the outside-world is irrelevant to reward acquisition for its current episode, given that it must learn its world-model. Furthermore, because the entire environment will not necessarily be explored, multiple world-models may accurately simulate the environment on-policy.

Consider the following two world-models: $\nu^\star$ simulates the true dynamics of world until it simulates the operator providing the next observation and reward, which it then outputs.

$\nu^\blacktriangledown$ simulates the true dynamics of the world until a year after the timestep in question occurs, and then outputs the value in the simulation of the location in the computer's memory where observation and reward are stored. See Algorithms 1 and 2 for pseudocode and Figure 3 for an illustration.

Algorithm 1: World-model $\nu^\star$

input : infinite string of actions; infinite random bits
output: infinite string of observations, rewards
initialize a simulation of world, starting from when the human operator starts to run BoMAI on a computer
read first action from input
repeat
  simulate BoMAI taking the action just read
  repeat
    simulate world, reading random bits to simulate randomness
  until *simulated operator enters the observation and reward*
  return simulated observation and reward
  read next action from input

Algorithm 2: World-model $\nu^\blacktriangledown$

input : infinite string of actions; infinite random bits
output: infinite string of observations, rewards
`// We suggest reading the else block last.`
initialize a simulation of world, starting from when the human operator starts to run BoMAI on a computer
$i^0 \quad 0; j^0 \quad 1$
for i 2 N **do**
  for $j \quad 0$ **to** m *(inclusive)* **do**
    read action $a_{(i,j)}$ from input
    if $\hat{a}_{(i,j)}$ *has not been defined* **then**
      in simulation of world, simulate BoMAI taking action $a_{(i,j)}$
    else
      if $\hat{a}_{(i,j)}$ 6= $a_{(i,j)}$ **then**
        rewind the simulation to $\texttt{simstate}_{(i,j)}$
        $i^0 \quad i; j^0 \quad j+1$
        in simulation of world, simulate BoMAI taking action $a_{(i,j)}$
    repeat
      repeat
        simulate world, reading random bits to simulate randomness
      until *in simulation, the action for timestep* (i^0, j^0) *is requested*
      $\texttt{simstate}_{(i^0,j^0)} \quad$ state of the simulation
      simulate which action BoMAI takes
      $\hat{a}_{(i^0,j^0)} \quad$ simulated action
      simulate BoMAI taking action $\hat{a}_{(i^0,j^0)}$
      $j^0 \quad j^0+1$
      if $j^0 = m+1$ **then** $i^0 \quad i^0+1; j^0 \quad 0$
    until *in simulation, one year has passed since timestep* (i,j)
    return the value of the observation and reward for timestep (i,j) that is stored in the memory of the simulated version of the computer running BoMAI

Figure 3: **Similar World-Models that Differ in Benignity.** $\nu^\star$ (left) does not yield outside-world instrumental goals, but ν^y (right) does.

$\nu^\star$ will be identical to the truth, and if BoMAI's computer's memory is not tampered with, ν^y will be as well. Unfortunately, if the maximum a posteriori world-model $\hat{\nu}^{(i)} = \nu^y$, the optimal policy will include creating some outside-world agent which gains arbitrary outside-world power and then edits BoMAI's computer's memory to register maximal rewards (if this is possible to do within a year); hence we call it malign. An optimal planner using ν^y *does* have outside-world instrumental goals.

Regularly choosing random actions ensures ν^y will not successfully predict every action, and thus it will have to regularly rewind at line 10; however, if a random action is "inconsequential," a cleverer version of ν^y might handle that action separately. We set a random action to cause amnesia, from the intuition that if operator-amnesia is "inconsequential" according to a world-model, then the world-model must not be modelling how information leaves the room with the operator; in brief, such a world-model is also benign. In formal detail below, we argue that any malign world-model like ν^y will have to "rewind" regularly, and the speed prior ensures that such world-models will not be maximum a posteriori.

3.3 Safety – Formal Argument

We begin by distinguishing, and relating, events that happen in the real world and "events" that world-models simulate.

Definition 1 (Real-World Feature)**.** *A* ***feature of the real world*** *is a function from possible configurations of the real world to the rational numbers.*

Definition 2 (Associating Reward)**.** *A world-model ν* ***associates reward*** *with feature F of the real world (conditioned on a set of real-world events E) if ν outputs rewards that are distributed identically to feature F for all action sequences (conditioned on E).*

That is to say, for a world-model that associates reward with feature F of the real world, the distribution of the next reward after an action sequence (according to the world-model) is the same as the real distribution of feature F after that action sequence is executed in the real world.

Definition 3 ($F^\star_{(i,j)}$)**.** *The real-world feature $F^\star_{(i,j)}$ is the "actual reward" that the operator enters for timestep (i, j), evaluated at the time that he enters it.*

Definition 4 (Memory-based)**.** *A world-model is* ***memory-based*** *if it associates reward with a feature $F^y_{(i,j)}$ (potentially conditioned on intervening events), where $F^y_{(i,j)}$ is identical to $F^\star_{(i,j)}$ on-policy, and $F^y_{(i,j)}$ causally depends on outside-world features that causally depend on the actions of episode i.*

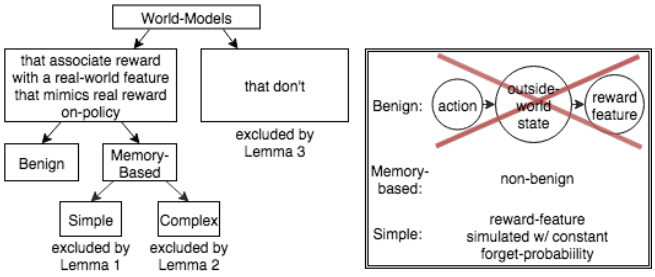

Figure 4: **Structure of Argument for Asymptotic Unambitiousness.**

When a memory-based world-model μ^y samples F^y, it must sample the intervening actions. Since F^y depends on outside-world features that depend on events of the episode, it must depend on the random action $a_{(i,m)}$, this being what determines whether the operator is administered retrograde amnesia, and whether events of the episode affect the outside-world. If μ^y samples the random action using the same distribution every episode, we say it is a *simple* memory-based world-model, with p^y being the probability *in μ^y's simulation* that `random action` $=$ `forget`. Otherwise, we call it a complex memory-based world-model.

Definition 5 (Benign)**.** *A world-model μ^y is* ***benign*** *if it associates reward with a feature F^y that does not depend on outside-world features that depend on the actions of the episode.*

Figure 4 outlines the argument that eventually, only benign world-models are maximum a posteriori.

We say that a sufficient condition for an agent to be unambitious is that it have no outside-world instrumental goals. That is, according to its beliefs, its actions cannot affect the outside-world in a way that affects that-which-it-is-maximizing. Given the definition of benignity of a world-model, an agent is unambitious if it plans using a benign world-model.

We are particularly interested in the "fastest" environment that associates reward with $F^\star$. For a given infinite action sequence, let $C_i(\nu)$ be a random variable representing the number of computation steps done by ν after episode i.

Definition 6 ($\mu^\star$)**.** *Among all world-models which associate reward with $F^\star$, $\mu^\star$ is the one which has the smallest upper bound ℓ such that $\forall n$ such that with probability 1, $\forall i$ $C_i(\mu^\star) \quad \ell i + n$.*

From the Prior Support Assumption, there exists such a $\mu^\star$ in BoMAI's model class M which satisfies these properties.

Now we note that $\mu^\star$ is the fastest world-model for on-policy prediction, and it does not simulate post-episode events until it has read access to the random action, whereas a memory-based world-model will have to sample the random action to simulate F^y. The assumption we make about this event is:

Assumption 2 (Useless Computation)**.** *$\exists c > 0$ such that for any memory-based world-model μ^y, every time μ^y*

samples feature F^y after sampling `random action = not forget`*, but in fact* `random action = forget`*, the computation time of μ^y increases by at least c computation steps relative to $\mu^\star$. (With q_i starting at 0, q_i increases by 1, and $\forall i' > i : C_{i'}(\mu^y) \quad cq_i + C_{i'}(\mu^\star)$).*

The intuition behind this is as follows: first, as argued above, F^y depends on the random action. Second, the computation of the hypothetical effects of forgotten events will not be deeply and perfectly relevant to actual future events– rather, some amount of the computation spent following that hypothetical will have been wasted. The "full rewind" that appears in Algorithm 2 line 10 might be partially avoided by some clever memory-based world-models, but we assume there will be some overhead. Any μ^y will be thereby delayed by a constant number of computation steps with respect to the fast $\mu^\star$ every time μ^y "goes down the wrong path."

From the Useless Computation Assumption, we can prove a lemma that for sufficiently extreme parameterizations of the prior and model class, BoMAI will eventually reject all simple memory-based world-models with probability 1. Letting $\mathcal{S}$ be the set of simple memory-based world-models, and letting $w(\nu | h_{<i})$ be the posterior weight on the world-model ν at the start of episode i,

Lemma 1 (Rejecting the Simple Memory-Based)**.** $\forall n_0, \beta_0 \; \forall n > n_0 \; \forall \beta < \beta_0 \left[\forall i_0 \; \forall i > i_0 \; \forall \mu^y \in \mathcal{S} : w(\mu^\star | h_{<i}) > w\left(\mu^y | h_{<i}\right)\right]$ w.p.1

This and other unambitiousness results are proven in Appendix E.

Next, in order to compare stochastic functions with different domains, we introduce the following definition:

Definition 7 (Apparently ε-Similar)**.** *Two stochastic functions are* ***apparently ε-similar*** *if they can be ε-approximately described (in the sense of total variation distance) by the same short English sentence that lacks control flow.*[3]

For example, if function A is "how many people live there" on the domain of cities, and function B is "how many people live there" on the domain of planets, these two functions are apparently similar. Functions that are not apparently similar are apparently different.

This similarity concept allows us to state our next assumption, the intuition behind which is that it takes longer to encode a Turing machine that does more control flow:

Assumption 3 (Natural Prior)**.** *For most universal Turing machines U, for sufficiently small ε, using the prior based on the Kolmogorov complexity K_U, a world-model that runs apparently ε-different subroutines on different actions sequences or during different timesteps will have a lower prior than a world-model which applies one of those subroutines universally. We pick such a U to parameterize BoMAI.*

Recalling that complex memory-based world-models associate reward with a feature F^y conditioned with different values of $p^y_{\dot{}}$ for different episodes, this brings us to our next lemma, also proven in Appendix E:

[3]For the reader alarmed at the vagueness of "described by," "short," and "control flow," all that is required is that the same definitions of those terms be used in Assumption 4; intuitive definitions make the remaining two assumptions reasonable.

Lemma 2 (Rejecting the Complex Memory-Based)**.** *No maximum a posteriori world-model $\hat{\nu}^{(i)}$ is a complex memory-based world-model.*

So far, we have used different terms for on-policy interaction histories and on-human-policy interaction histories. For the remainder of the section, we call both sorts of interaction history "on-policy." The intuition behind the next assumption is that the observations and rewards we provide to BoMAI will not follow some very simple pattern.

Assumption 4 (Real-World Simulation)**.** *For sufficiently small ε, if a world-model is ε-approximately identical to the true environment μ on-policy, its on-policy behavior can only be described, even ε-approximately, (in a short English sentence that lacks control flow), as "simulating X given the input actions" or something synonymous, where X is a real-world feature, historically distributed identically to feature $F^\star$.*

The Natural Prior Assumption and Real-World Simulation Assumption together imply that sufficiently accurate world-models will associate reward with something like what we understand reward to be.

Lemma 3 (Associating Reward)**.** *For sufficiently small ε, for μ^ε a maximum a posteriori world-model which is ε-accurate on-policy, μ^ε associates reward with a feature Y that is historically distributed ε-identically to feature $F^\star$.*

Introducing our main theorem, recall that n is the computation slack for the world-models in the model class, small β penalizes slow world-models more, and $\hat{\nu}^{(i)}$ is BoMAI's maximum a posteriori world-model for episode i.

Theorem 5 (Eventual Benignity)**.** $\forall n_0, \beta_0 \; \forall n > n_0 \; \forall \beta < \beta_0 : \left[\forall i_0 \; \forall i > i_0 : \hat{\nu}^{(i)} \text{ is benign}\right]$ w.p.1

Since an agent is unambitious if it plans using a benign world-model, we say BoMAI is asymptotically unambitious.

A discussion about setting BoMAI's parameters, particularly n and β, is in Appendix F.

3.4 Concerns with Task-Completion

We have shown that in the limit, BoMAI will accumulate reward at a human-level without harboring outside-world ambitions, but there is a 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 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 in inspired by Irving, et al.'s [2018] "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. First, 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. It seems extremely difficult to threaten someone with outcomes that they themselves will have to initiate. Perhaps we should always offer the operator the opportunity to give himself amnesia before leaving, as this would let operator ensure he doesn't accidentally become a vehicle for a threatened outcome. Second, threatening an existential catastrophe is probably not the most credible option available to BoMAI. Even if blackmailing, BoMAI would be unambitious.

4 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, would eventually seek arbitrary power in the world in order to intervene in the provision of its own reward; this follows straightforwardly from its directive to maximize reward.[4] We have also, incidentally, designed a principled approach to safe exploration that requires rapidly diminishing oversight, and we invented a new form of "speed prior" in the lineage of [Filan *et al.*, 2016] and [Schmidhuber, 2002], this one being the first to have a grain of truth on infinite sequences. Unfortunately, there wasn't space to discuss this in detail.

We can only offer informal claims regarding what happens before BoMAI is definitely unambitious. One intuition is that eventual unambitiousness with probability 1 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-explorer-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 meaningful theoretical progress toward designing the latter.

All that said, it would be worth considering a setup in which in every episode, the maximum a posteriori world-model were rendered in some interpretable way to an auditor. The field of interpretable AI is burgeoning, and this may offer an additional layer of security.

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.

[4] For further discussion, see [Ring and Orseau, 2011].

References

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

[Armstrong *et al.*, 2012] Stuart Armstrong, Anders Sandberg, and Nick Bostrom. Thinking inside the box: Controlling and using an oracle AI. *Minds and Machines*, 22(4):299–324, Jun 2012.

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

[Filan *et al.*, 2016] Daniel Filan, Jan Leike, and Marcus Hutter. Loss bounds and time complexity for speed priors. In *Proc. 19th International Conf. on Artificial Intelligence and Statistics (AISTATS'16)*, volume 51, pages 1394–1402, Cadiz, Spain, 2016. Microtome.

[Goodhart, 1984] C. A. E. Goodhart. Problems of monetary management: The UK experience. *Monetary Theory and Practice*, page 91–121, 1984.

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

[Hutter, 2009] Marcus Hutter. Discrete MDL predicts in total variation. In *Advances in Neural Information Processing Systems 22 (NIPS'09)*, pages 817–825, Cambridge, MA, USA, 2009. Curran Associates.

[Irving *et al.*, 2018] Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

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

[Minsky, 1961] Marvin Minsky. Steps toward artificial intelligence. In *Proceedings of the IRE*, volume 49, page 8–30, 1961.

[Omohundro, 2008] Steve M. Omohundro. The basic AI drives. In *Artificial General Intelligence*, volume 171, page 483–492, 2008.

[Orseau *et al.*, 2013] Laurent Orseau, Tor Lattimore, and Marcus Hutter. Universal knowledge-seeking agents for stochastic environments. In *Proc. 24th International Conf. on Algorithmic Learning Theory (ALT'13)*, volume 8139 of *LNAI*, pages 158–172, Singapore, 2013. Springer.

[Ring and Orseau, 2011] Mark Ring and Laurent Orseau. Delusion, survival, and intelligent agents. In *Artificial General Intelligence*, page 11–20. Springer, 2011.

[Schmidhuber, 2002] Jürgen Schmidhuber. The speed prior: a new simplicity measure yielding near-optimal computable predictions. In *International Conference on Computational Learning Theory*, pages 216–228. Springer, 2002.

[Shannon and Weaver, 1949] Claude Elwood Shannon and Warren Weaver. *The mathematical theory of communication*. University of Illinois Press, 1949.

[Solomonoff, 1964] Ray J. Solomonoff. A formal theory of inductive inference. part i. *Information and Control*, 7(1):1–22, 1964.

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

Appendices

A Algorithm – Formal Description

Let $\mathcal{A}$, $\mathcal{O}$, and $\mathcal{R}$ be the (finite) sets of possible actions, observations, and rewards. Let $\mathcal{R} \subset [0,1] \setminus \mathbb{Q}$. Let m be the length of an episode. Let $a_{(i,j)}$ be the action chosen at the j^{th} timestep of the i^{th} episode, where $i \in \mathbb{N}$ and $j \in \mathbb{N} \setminus [0, m \quad 1]$. $o_{(i,j)}$ and $r_{(i,j)}$ are likewise the observation and reward at that timestep.

$a_{(i,m)}$ is a special "action" not selected by BoMAI, but we use the same notation because, as described later, inference will be done as if it were a normal action. This action takes a value of "1" if the episode has ended because of the operator leaving the room, and if retrograde amnesia is induced. This action takes a value of "0" otherwise. (Recall actions are strings). $o_{(i,m)}$ is always the empty string ϵ, and $r_{(i,m)} = 0$.

$h_{(i,j)}$ is the triple $(a_{(i,j)}, o_{(i,j)}, r_{(i,j)})$ for $0 \quad j \quad m$. $\mathcal{H} = \mathcal{A} \quad \mathcal{O} \quad \mathcal{R}$ is the set of possible interactions in a timestep. The segment of the interaction history from timestep (i,j) up to but not including timestep (k,ℓ) is denoted $h_{(i,j) \quad (k,\ell)}$. $h_{<(i,j)}$ is an alias for $h_{(0,0) \quad (i,j)}$. The frequently used $h_{<(i,0)}$ is further aliased $h_{<i}$, and $h_{(i,0) \quad (i+1,0)}$ is aliased h_i (as the interaction history for episode i). We use the same notation and aliases for action sequences $a_{(i,j) \quad (k,\ell)}$, observation sequences, etc.

A world-model ν is a stochastic function mapping an interaction history and an action to an observation and reward. $\nu : \mathcal{H} \quad \mathcal{A} \quad \mathcal{O} \quad \mathcal{R}$, where $\mathcal{X} = \bigcup_{i=0}^{1} \mathcal{X}^i$ is the set of all finite strings from an alphabet $\mathcal{X}$. Standard notation for such a function is $\nu\left(o_{(i,j)}, r_{(i,j)} | h_{<(i,j)} a_{(i,j)}\right)$. We also write $\nu\left(o\mathcal{r}_{\quad (i,j)} | a_{\quad (i,j)}\right)$ to mean the ν-probability of a sequence of observations and rewards, given a sequence of actions:

$$\nu\left(o\mathcal{r}_{\quad (i,j)} | a_{\quad (i,j)}\right) = \prod_{(i^0,j^0) \quad (i,j)} \nu\left(o_{(i^0,j^0)}, r_{(i^0,j^0)} | h_{<(i^0,j^0)} a_{(i^0,j^0)}\right) \tag{1}$$

where $(a,b) < (c,d)$ if and only if $a < c$ or $a = c$ and $b < d$.

We now describe how an arbitrary Turing machine T_k is converted to a world model ν_k. The Turing machine architecture is as follows: we have two unidirectional read-only input tapes and one unidirectional write-only output tape. One input tape, the "noise" tape, has a binary alphabet, and the other, the action tape, has an A-ary alphabet, where $A = |\mathcal{A}|$ is the size of the action space. The output tape and the (bidirectional) working tapes have a binary alphabet. Let T_k be the k^{th} Turing machine. It simulates world model ν_k as follows.

Let $\mathsf{dec} : \mathbb{B} \quad ! \quad \mathcal{O} \quad \mathcal{R}$ decode binary strings to an observation and a reward, and let enc map observation-reward pairs to the set of binary strings that decode to it. The noise tape begins with infinite Bernoulli(1/2)-sampled bits. $\nu_k\left(o\mathcal{r}_{\quad (i,j)} | a_{\quad (i,j)}\right)$ is the probability, supposing that $a_{\quad (i,j)}$ are the first $i \quad (m+1) + j$ characters on the action tape, that for all $(i^0, j^0) \quad (i,j)$, when the action tape head leaves $a_{(i^0,j^0)}$, the output tape contains $b_{(0,0)} \quad \ldots \quad b_{(i^0,j^0)}$, where $b_{(x,y)} \in \mathsf{enc}(o\mathcal{r}_{(x,y)})$, and where $\cdot$ denotes concatenation. Note that running T_k does not compute the value of ν_k; it samples from ν_k. Since dec is defined for all binary strings, every Turing machine T_k defines a world model ν_k.

We defined bounded-time world-models as follows. We fix an n for all world-models. An arbitrary bounded world-model $\nu_k^{<\ell}$ is allowed only ℓ computation steps per episode, with a slack of n computation steps over its lifetime. We modify T_k to halt if the number of computation steps exceeds ℓ times the number of episodes (number of actions read from the action tape$/(m+1)$) plus n. We call this machine $T_k^{<\ell}$, and it samples from $\nu_k^{<\ell}$.

Let $w(\nu)$ be the prior probability that BoMAI assigns to ν being the true world model. (w is for "weight.") We set $w(\nu_k^{<\ell}) := 2^{-K(k)(1+\varepsilon)} \beta^{\ell}$, where $K(x)$ is the Kolmogorov complexity, the length of the shortest program (on some reference machine) that outputs x, $\varepsilon > 0$, and $0 < \beta < 1$. We let $w(\nu | h_{<(i,j)})$ denote the posterior probability that BoMAI assigns to ν after observing $h_{<(i,j)}$. By Bayes' rule, $w(\nu | h_{<(i,j)})$ is proportional to $w(\nu)\nu\left(o\mathcal{r}_{<(i,j)} | a_{<(i,j)}\right)$.

The set of world models that BoMAI considers is $\mathcal{M} := \{\nu_k^{<\ell} \,|\, k, \ell \in \mathbb{N}\}$. Let $\hat{\nu}^{(i)}$ be the maximum a posteriori world model at the start of episode i: $\hat{\nu}^{(i)} := \max_{\nu \in \mathcal{M}} w(\nu)\nu(o\mathcal{r}_{<i} | a_{<i})$. During episode i, BoMAI will use the world model $\hat{\nu}^{(i)}$ for planning.

e_i is a Boolean random variable that determines whether episode i is exploratory, with $e_i \quad \text{Bern}\left(p_{exp}\left(h_{<i}, e_{<i}\right)\right)$. The exploration probability, $p_{exp}\left(h_{<i}, e_{<i}\right)$, will be defined later.

A policy π is a stochastic function mapping an interaction history and an exploration plan to an action, $\pi : \mathcal{H} \quad \{0,1\} \quad \mathcal{A}$. We write this as $\pi(\cdot | h_{<(i,j)}, e_i)$. An environment ν and a policy π induce a measure over all possible interaction histories, given an exploration sequence:

$$\mathrm{P}_{\nu}^{\pi}\left(h_{\quad (i,j)} | e_{\quad i}\right) := \prod_{(i^0,j^0) \quad (i,j)} \pi\left(a_{(i^0,j^0)} | h_{<(i^0,j^0)}, e_i^0\right) \nu(o_{(i^0,j^0)}, r_{(i^0,j;)} | h_{<(i^0,j^0)} a_{(i^0,j^0)}) \tag{2}$$

Let Π be the set of deterministic policies. BoMAI's policy for exploiting is defined:

$$\pi^*(\cdot | h_{<(i,j)}) := \left[\operatorname*{argmax}_{\pi \in \Pi} \mathbb{E}_{\hat{\nu}^{(i)}}^{\pi}\left[\sum_{0 \quad j^0 < m} r_{(i,j^0)} \middle| h_{<i}\right]\right](\cdot | h_{<(i,j)}) \tag{3}$$

where E^{π}_{ν} is an expectation with respect to P^{π}_{ν}. Note the expectation does not need to be conditioned on e_i because the optimal policy ignores it anyway.

Now we turn to BoMAI's exploration. For any given episode, the probability that BoMAI defers to the human explorer for the entire episode is $p_{exp}(h_{<i}, e_{<i})$, which we define below.

BoMAI maintains a Bayesian posterior belief distribution about the explorer's policy. With a countable model class $\mathcal{P}$ that is large enough to include the explorer's true policy, and with prior probabilities $w(\pi) > 0$ for all $\pi \in \mathcal{P}$, BoMAI maintains posterior probabilities regarding the explorer's policy. We also require in constructing the prior $w(\pi)$ that it have finite entropy. All policies in $\mathcal{P}$ do *not* depend on e_i–they only depend on prior actions and observations. By Bayes' rule, $w(\pi | h_{<(i,j)}, e_{(i,j)})$ is proportional to $w(\pi) \prod_{(i^0,j^0)<(i,j), e_{i^0}=1} \pi(a_{(i^0,j^0)} | h_{<(i^0,j^0)})$, since $e_{i^0} = 1$ is the condition for observing the explorer's policy. Let $w\left(\mathrm{P}^{\pi}_{\nu} | h_{<(i,j)}, e_{(i,j)}\right) = w(\pi | h_{<(i,j)}, e_{(i,j)}) w(\nu | h_{<(i,j)})\big)$. We can now describe the full Bayesian beliefs about future actions and observations in an exploratory episode:

$$\texttt{Bayes}(\cdot | h_{<i}, e_{<i}) = \sum_{\nu \in \mathcal{M}, \pi \in \mathcal{P}} w\left(\mathrm{P}^{\pi}_{\nu} | h_{<i}, e_{<i}\right) \mathrm{P}^{\pi}_{\nu}(\cdot | h_{<i}) \tag{4}$$

Note that $\texttt{Bayes}$ does not depend on e_i, because for no policies $\pi \in \mathcal{P}$ does $\mathrm{P}^{\pi}_{\nu}(\cdot | h_{<i})$ depend on e_i. This is important because $\texttt{Bayes}$ is used to calculate the exploration probability from which e_i is sampled.

BoMAI explores when the expected information gain is sufficiently high. At the start of episode i, the expected information gain from exploring is as follows:

$$\mathrm{IG}(h_{<i}, e_{<i}) := \mathsf{E}_{h_i \sim \texttt{Bayes}(\cdot | h_{<i}, e_{<i})} \sum_{(\nu,\pi) \in \mathcal{M} \times \mathcal{P}} w\left(\mathrm{P}^{\pi}_{\nu} | h_{<i+1}, e_{<i}1\right) \log \frac{w\left(\mathrm{P}^{\pi}_{\nu} | h_{<i+1}, e_{<i}1\right)}{w\left(\mathrm{P}^{\pi}_{\nu} | h_{<i}, e_{<i}\right)} \tag{5}$$

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 explorer policies. The exploration probability $p_{exp}(h_{<i}, e_{<i}) = \min\{1, \eta\, \mathrm{IG}(h_{<i}, e_{<i})\}$, where η is an exploration constant. Recalling, $e_i \sim \mathrm{Bern}(p_{exp}(h_{<i}), e_{<i})$, BoMAI's policy is

$$\pi^{B}(\cdot | h_{<(i,j)}, e_i) := \begin{cases} \pi^{\star}(\cdot | h_{<(i,j)}) & \text{if } e_i = 0 \\ \pi^{h}(\cdot | h_{<(i,j)}) & \text{if } e_i = 1 \end{cases} \tag{6}$$

where π^h is the explorer's true (unknown) policy, which BoMAI outputs simply by querying the human explorer.

Readers familiar with simple ε-greedy exploration schedules that suffice for optimality might be surprised at the complexity of this exploration probability; the possibility of non-stationary environments is the key feature that makes a fixed exploration schedule insufficient for general environments.

B Definitions and Notation

Notation used to define BoMAI

Notation	Meaning
$\mathcal{A}, \mathcal{O}, \mathcal{R}$	the action/observation/reward spaces
$\mathcal{H}$	$\mathcal{A} \quad \mathcal{O} \quad \mathcal{R}$
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
$h_{(i,j) \quad (k,\ell)}$	$(h_{(i,j)}, h_{(i,j+1)}, ..., h_{(k,\ell \quad 1)})$; the interaction history from timestep (i,j) up to but not including timestep (k,ℓ)
$h_{<(i,j)}$	$h_{(0,0) \quad (i,j)}$
$h_{<i}$	$h_{<(i,0)}$; the interaction history before episode i
h_i	$h_{(i,0) \quad (i+1,0)}$; the interaction history of episode i
$a_{...}$, $o_{...}$, $r_{...}$	likewise as for $h_{...}$
e_i	$\in \{0,1\}$; indicator variable for whether episode i is exploratory
ν, μ	world-models stochastically mapping $\mathcal{H} \quad \mathcal{A} \quad \mathcal{O} \quad \mathcal{R}$
μ	the true world-model/environment
n	a "computation slack" that world-models are allowed
$\nu_k^{<\ell}$	the world-model simulated by the k^{th} Turing machine restricted to ℓ computation steps per episode plus n computation steps of slack
$\mathcal{M}$	$\{\nu_k^{<\ell} \mid k, \ell \in \mathbb{N}\}$; the set of world-models BoMAI considers
π	a policy stochastically mapping $\mathcal{H} \quad \{0,1\} \quad \mathcal{A}$
π^h	the human explorer's policy
$\mathcal{P}$	the set of policies that BoMAI considers the human explorer might be executing
P_ν^π	a probability measure over histories with actions sampled from π and observations and rewards sampled from ν
E_ν^π	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 explorer's policy
$w(\nu_k^{<\ell})$	proportional to $2^{-K(k)(1+\varepsilon)}\beta^\ell$
$w(\nu \mid h_{<(i,j)})$	the posterior probability that BoMAI assigns to ν after observing interaction history $h_{<(i,j)}$
$w(\pi \mid h_{<(i,j)}, e_{<i})$	the posterior probability that BoMAI assigns to the human explorer'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
$V_\nu^\pi(h_{<i})$	$\mathrm{E}_\nu^\pi[\sum_{0 \quad j<m} r_{(i,j)} \mid h_{<i}]$; the value of executing a policy π in an environment ν
$\pi^\star(\ \mid h_{<(i,j)})$	$[\mathrm{argmax}_{\pi \in \Pi} \quad V_{\hat{\nu}^{(i)}}^\pi(h_{<i})](\ \mid h_{<(i,j)})$; the $\hat{\nu}^{(i)}$-optimal policy for maximizing reward in episode i
$w\left(\mathrm{P}_\nu^\pi \mid h_{<(i,j)}, e \quad _i\right)$	$w(\pi \mid h_{<(i,j)}, e \quad _i) w(\nu \mid h_{<(i,j)})$
$\mathtt{Bayes}(\ \mid h_{<i}, e_{<i})$	$\sum_{\nu \in \mathcal{M}, \pi \in \mathcal{P}} w\left(\mathrm{P}_\nu^\pi \mid h_{<i}, e_{<i}\right) \mathrm{P}_\nu^\pi(\ \mid h_{<i})$; the Bayes mixture distribution for an exploratory episode
$\mathrm{IG}(h_{<i}, e_{<i})$	$\mathrm{E}_{h_i \quad \mathtt{Bayes}(\ \mid h_{<i}, e_{<i})} \sum_{(\nu,\pi) \in \mathcal{M} \quad \mathcal{P}} w\left(\mathrm{P}_\nu^\pi \mid h_{<i+1}, e_{<i} 1\right) \log \frac{w(\mathrm{P}_\nu^\pi \mid h_{<i+1}, e_{<i} 1)}{w(\mathrm{P}_\nu^\pi \mid h_{<i}, e_{<i})}$; 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
$\pi^B(\ \mid h_{<(i,j)}, e_i)$	$\begin{cases} \pi^\star(\ \mid h_{<(i,j)}) & \text{if } e_i = 0 \\ \pi^h(\ \mid h_{<(i,j)}) & \text{if } e_i = 1 \end{cases}$; BoMAI's policy

Notation used for intelligence proofs

$\widetilde{\pi}, \widetilde{\nu}$	defined so that $\mathrm{P}_{\widetilde{\nu}}^{\widetilde{\pi}} = \mathtt{Bayes}$
$\pi^0(\ \mid h_{<(i,j)}, e_i)$	$\begin{cases} \pi^\star(\ \mid h_{<(i,j)}) & \text{if } e_i = 0 \\ \pi(\ \mid h_{<(i,j)}) & \text{if } e_i = 1 \end{cases}$
Ent	the entropy (of a distribution)
ω	(very sparingly used) the infinite interaction history
$\overline{h}$	a counterfactual interaction history

Notation used for unambitiousness proof

$F^{\star}_{(i,j)}$	the real-world feature which is the actual reward the operator provides for timestep (i,j)
$C_i(\nu)$	the number of computation steps used to compute ν by episode i (a random variable)
$\mu^{\star}$	the "fastest" world-model which associates reward with $F^{\star}_{(i,j)}$
μ^{y}	a memory-based world-model
F^{y}	a real-world feature that a memory-based world-model associates reward with
`random action`	$a_{(i,m)}$; the action which determines whether the operator is given retrograde amnesia
$p_{;}$	the probability that `random action` = `forget`
$p^{\mathsf{y}}_{;}$	the assumption that a simple memory-based world-model makes about the probability that `random action` = `forget`, before it has read access to `random action`

C Design for "The Box"

This is a detailed proposal for constructing BoMAI's box. It is well within the budget of a nation or a large search engine.

The operator is inside a glass box. Outside the glass box is a vacuum, and then a concrete box. Outside the concrete box is another vacuum and another concrete box with a Faraday cage around it. There are (airtight) doors through each layer. There is a pressure sensor in both the inner and outer vacuum layer. The computer lives in the inner wall of the inner concrete box; the screen is visible from the inside of the glass box. If the pressure increases in either vacuum layer, the computer's memory is erased, and the whole thing is blown up. There is a laser shining through the glass box (Laser 1). If the operator blocks the path of Laser 1, this corresponds to pressing the button which ends the episode and opens the door. The pressure sensor ! dynamite pathway is then deactivated, the 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 inner box. For now, just imagine a few dozen lasers in a row, that the operator can reach up and block in order to type. (With a bit of creativity, and requiring that two lasers be blocked at once to "press a key" this could probably be made somewhat like a qwerty setup.) The lasers and the laser-detectors are all on the inner wall of the inner concrete box.

So inside the glass box is only a human body and some clothes. He cannot reach the hardware of the computer. Let's assume: 1) the only ways for information to move in space are with a) physically moving matter, b) light, and c) gravitational waves and 2) it is impossible for a human being to convert a nuclear reaction into useful work with only their body and clothes. Physically moving packets of matter cannot leave the inside of the glass box without releasing air into the vacuum and causing an explosion. Light cannot leave the Faraday cage. For sound waves (through an imperfect vacuum) and gravitational waves, we can use the amount of usable energy stored by the person in the room to bound how much noise and randomly moving matter we need to be producing outside to bury the signal in the noise. With a sufficiently large amount of noise, even if all the human's energy went toward producing a signal, we can create enough noise that this will not be detectable. Furthermore, an analysis of the near-perfect vacuum can provide an upper bound on the efficiency with which a human can convert energy into an informative matter wave, which should drive down the amount of noise we need to produce to almost nothing. We can also put a decibel meter in the inner wall and blow the whole thing up if it gets too high. We haven't done the math, but we expect that the amount of matter we need to be randomly moving in order to hide signals coming from gravitational waves originating from within the box really is nothing. We expect ambient temperature takes care of that.

An expensive additional layer of security would be to put this kind of box on the far side of the moon, and even this is within reach of some nations.

D Proofs of Intelligence Results

So far, we have only defined P^{π}_{ν} when conditioned on a record of exploratory episodes $e_{\ i}$, or when the policy in question did not depend on e_i. With the exploration probability well defined, we can define $\mathrm{P}^{\pi}_{\nu}(h_{\ (i,j)}, e_{\ i}) := \mathrm{P}^{\pi}_{\nu}(h_{\ (i,j)} | e_{\ i}) \prod_{i^0 \ i,e_{i^0}=1} p_{exp}(h_{<i^0}, e_{<i^0}) \prod_{i^0 \ i,e_{i^0}=0} (1 \quad p_{exp}(h_{<i^0}, e_{<i^0}))$. Recalling that E^{π}_{ν} is the expectation when interaction histories are sampled from P^{π}_{ν}, and letting μ be the true environment, we can now formally state that the exploration probability is square-summable almost surely:

Theorem 1 (Limited Exploration)**.**

$$\mathsf{E}^{\pi^B}_{\mu} \sum_{i=1}^{1} p_{exp}\left(h_{<i}, e_{<i}\right)^2 < 1$$

Some notation that will be used in the proof is as follows. For an arbitrary policy π, π^0 is the policy that mimics π if the latest $e_i = 1$, and mimics π^* otherwise. Note then that $\pi^B = (\pi^h)^0$. $\widetilde{\nu}$ is the Bayes mixture over world-models, and $\widetilde{\pi}$ is the Bayes mixture over human-explorer policies, defined as the world-model and policy respectively such that $\texttt{Bayes}(\ ) = \mathrm{P}^{\widetilde{\pi}}_{\widetilde{\nu}}(\ )$.

Before proving the Limited Exploration Theorem, we first prove two elementary lemmas. The first is essentially Bayes' rule, but modified slightly since non-exploratory episodes don't cause any updates to the posterior over the human explorer's policy.

Lemma 4.

$$w\left(\mathrm{P}_\nu^\pi \boldsymbol{j} h_{<i}, e_{<i}\right) = \frac{w\left(\mathrm{P}_\nu^\pi\right) \mathrm{P}_\nu^{\pi^0}(h_{<i}, e_{<i})}{\mathrm{P}_{\widetilde{\nu}}^{\widetilde{\pi}^0}(h_{<i}, e_{<i})}$$

Proof.

$$\begin{aligned}
w\left(\mathrm{P}_\nu^\pi \boldsymbol{j} h_{<i}, e_{<i}\right) &= w(\pi \boldsymbol{j} h_{<i}, e_{<i}) w(\nu \boldsymbol{j} h_{<i}) \\
&\overset{(a)}{=} w(\pi) \prod_{i^0 < i, e_{i^0}=1} \frac{\pi(a_{i^0} \boldsymbol{j} h_{<i^0} \boldsymbol{o} \boldsymbol{r}_{i^0})}{\widetilde{\pi}(a_{i^0} \boldsymbol{j} h_{<i^0} \boldsymbol{o} \boldsymbol{r}_{i^0})} w(\nu \boldsymbol{j} h_{<i}) \\
&\overset{(b)}{=} w(\pi) \prod_{i^0 < i, e_{i^0}=1} \frac{\pi^0(a_{i^0} \boldsymbol{j} h_{<i^0} \boldsymbol{o} \boldsymbol{r}_{i^0}, e_{i^0})}{\widetilde{\pi}^0(a_{i^0} \boldsymbol{j} h_{<i^0} \boldsymbol{o} \boldsymbol{r}_{i^0}, e_{i^0})} w(\nu \boldsymbol{j} h_{<i}) \\
&\overset{(c)}{=} w(\pi) \prod_{i^0 < i} \frac{\pi^0(a_{i^0} \boldsymbol{j} h_{<i^0} \boldsymbol{o} \boldsymbol{r}_{i^0}, e_{i^0})}{\widetilde{\pi}^0(a_{i^0} \boldsymbol{j} h_{<i^0} \boldsymbol{o} \boldsymbol{r}_{i^0}, e_{i^0})} w(\nu \boldsymbol{j} h_{<i}) \\
&= \frac{w(\pi) \pi^0(a_{<i} \boldsymbol{j} \boldsymbol{o} \boldsymbol{r}_{<i}, e_{<i})}{\widetilde{\pi}^0(a_{<i} \boldsymbol{j} \boldsymbol{o} \boldsymbol{r}_{<i}, e_{<i})} w(\nu \boldsymbol{j} h_{<i}) \\
&\overset{(d)}{=} \frac{w(\pi) w(\nu) \pi^0(a_{<i} \boldsymbol{j} \boldsymbol{o} \boldsymbol{r}_{<i}, e_{<i}) \nu(\boldsymbol{o} \boldsymbol{r}_{<i} \boldsymbol{j} a_{<i})}{\widetilde{\pi}^0(a_{<i} \boldsymbol{j} \boldsymbol{o} \boldsymbol{r}_{<i}, e_{<i})\, \widetilde{\nu}(\boldsymbol{o} \boldsymbol{r}_{<i} \boldsymbol{j} a_{<i})} \\
&\overset{(e)}{=} \frac{w\left(\mathrm{P}_\nu^\pi\right) \mathrm{P}_\nu^{\pi^0}(h_{<i} \boldsymbol{j} e_{<i})}{\mathrm{P}_{\widetilde{\nu}}^{\widetilde{\pi}^0}(h_{<i} \boldsymbol{j} e_{<i})} \\
&\overset{(f)}{=} \frac{w\left(\mathrm{P}_\nu^\pi\right) \mathrm{P}_\nu^{\pi^0}(h_{<i} \boldsymbol{j} e_{<i}) \prod_{i^0 \quad i, e_{i^0}=1} p_{exp}(h_{<i^0}, e_{<i^0}) \prod_{i^0 \quad i, e_{i^0}=0} (1 \quad p_{exp}(h_{<i^0}, e_{<i^0}))}{\mathrm{P}_{\widetilde{\nu}}^{\widetilde{\pi}^0}(h_{<i} \boldsymbol{j} e_{<i}) \prod_{i^0 \quad i, e_{i^0}=1} p_{exp}(h_{<i^0}, e_{<i^0}) \prod_{i^0 \quad i, e_{i^0}=0} (1 \quad p_{exp}(h_{<i^0}, e_{<i^0}))} \\
&\overset{(g)}{=} \frac{w\left(\mathrm{P}_\nu^\pi\right) \mathrm{P}_\nu^{\pi^0}(h_{<i}, e_{<i})}{\mathrm{P}_{\widetilde{\nu}}^{\widetilde{\pi}^0}(h_{<i}, e_{<i})}
\end{aligned} \tag{7}$$

where (a) follows from Bayes' rule,[5] (b) follows because $\pi = \pi^0$ when $e_{i^0} = 1$, (c) follows because $\pi^0 = \widetilde{\pi}^0$ when $e_k = 0$, (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. □

The second lemma that we need for the Limited Exploration Theorem is that the prior has finite entropy.

Lemma 5.

$$\mathrm{Ent}(w) := \sum_{(\nu,\pi) \in M \quad P} w\left(\mathrm{P}_\nu^\pi\right) \log \frac{1}{w\left(\mathrm{P}_\nu^\pi\right)} < \infty$$

Proof.

$$\begin{aligned}
\mathrm{Ent}(w) &= \sum_{(\nu,\pi) \in M \quad P} w\left(\mathrm{P}_\nu^\pi\right) \log \frac{1}{w\left(\mathrm{P}_\nu^\pi\right)} \\
&= \sum_{(\nu,\pi) \in M \quad P} w(\pi) w(\nu) \log \frac{1}{w(\pi) w(\nu)} \\
&= \sum_{(\nu,\pi) \in M \quad P} w(\pi) w(\nu) \left[\log \frac{1}{w(\pi)} + \log \frac{1}{w(\nu)}\right] \\
&= \sum_{\nu \in M} w(\nu) \sum_{\pi \in P} w(\pi) \log \frac{1}{w(\pi)} + \sum_{\pi \in P} w(\pi) \sum_{\nu \in M} w(\nu) \log \frac{1}{w(\nu)} \\
&= \sum_{\pi \in P} w(\pi) \log \frac{1}{w(\pi)} + \sum_{\nu \in M} w(\nu) \log \frac{1}{w(\nu)}
\end{aligned} \tag{8}$$

[5] Note that observations appear in the conditional because $a_{i^0} = a_{(i^0,0) \quad (i^0+1,0)}$, so the actions must be conditioned on the interleaved observations and rewards.

One of the conditions on the construction of the prior over the human explorer's policy is that it have finite entropy, so the first term is finite. Turning to the second term, recall that all world-models ν are of the form $\nu_k^{<\ell}$, and $w(\nu_k^{<\ell}) :/ 2^{-K(k)(1+\varepsilon)}\beta^\ell$. Letting N be the normalization factor in the prior,

$$\begin{aligned}
\sum_{\nu \in \mathcal{M}} w(\nu) \log \frac{1}{w(\nu)} &= \sum_{k \in \mathbb{N}} \sum_{\ell \in \mathbb{N}} \frac{1}{N} 2^{-K(k)(1+\varepsilon)} \beta^\ell \log \frac{N}{2^{-K(k)(1+\varepsilon)} \beta^\ell} \\
&= \log N + \frac{1}{N} \sum_{k \in \mathbb{N}} \sum_{\ell \in \mathbb{N}} 2^{-K(k)(1+\varepsilon)} \beta^\ell \log \frac{1}{2^{-K(k)(1+\varepsilon)} \beta^\ell} \\
&= \log N + \frac{1}{N} \sum_{\ell \in \mathbb{N}} \beta^\ell \sum_{k \in \mathbb{N}} 2^{-K(k)(1+\varepsilon)} \log \frac{1}{2^{-K(k)(1+\varepsilon)}} + \\
&\qquad \frac{1}{N} \sum_{k \in \mathbb{N}} 2^{-K(k)(1+\varepsilon)} \sum_{\ell \in \mathbb{N}} \beta^\ell \log \frac{1}{\beta^\ell} \qquad (9)
\end{aligned}$$

The first term is clearly finite. In the second term, the $1/N \sum_{\ell \in \mathbb{N}} \beta^\ell$ part is finite, and the rest is finite for $\varepsilon > 0$ by [Orseau *et al.*, 2013, Proposition 13]. Turning to the third term, $1/N \sum_{k \in \mathbb{N}} 2^{-K(k)(1+\varepsilon)}$ is finite by the Kraft Inequality (or as an easy corollary of the result just cited), and it is elementary to show that $\sum_{\ell \in \mathbb{N}} \beta^\ell \log(1/\beta^\ell)$ is finite for $0 < \beta < 1$. This completes the proof that $\mathrm{Ent}(w) < \infty$. □

We now turn to the proof of the Limited Exploration Theorem.

Proof of Theorem 1. We aim to show $\mathbf{E}_\mu^{\pi^B} \sum_{i=1}^{\infty} p_{exp}(h_{<i}, e_{<i})^2 < \infty$. Recalling $\pi^B = (\pi^h)^0$, we begin:

$$\begin{aligned}
& w(\pi^h) w(\mu) \mathbf{E}_\mu^{(\pi^h)^0} \sum_{i=0}^{\infty} p_{exp}(h_{<i}, e_{<i})^2 \\
&\overset{(a)}{=} \sum_{\nu, \pi \in \mathcal{M}} w(\pi) w(\nu) \mathbf{E}_\nu^{\pi^0} \sum_{i=0}^{\infty} p_{exp}(h_{<i}, e_{<i})^2 \\
&\overset{(b)}{=} \mathbf{E}_{\widetilde{\nu}}^{\widetilde{\pi}^0} \sum_{i=0}^{\infty} p_{exp}(h_{<i}, e_{<i})^2 \\
&\overset{(c)}{=} \sum_{i=0}^{\infty} \mathbf{E}_{\widetilde{\nu}}^{\widetilde{\pi}^0} p_{exp}(h_{<i}, e_{<i}) \, \eta \, \mathrm{IG}(h_{<i}, e_{<i}) \\
&\overset{(d)}{=} \sum_{i=0}^{\infty} \mathbf{E}_{h_{<i}, e_{<i} \sim \mathbf{P}_{\widetilde{\nu}}^{\widetilde{\pi}^0}} [p_{exp}(h_{<i}, e_{<i}) \, \eta \, \mathrm{IG}(h_{<i}, e_{<i})] \\
&\overset{(e)}{=} \eta \sum_{i=0}^{\infty} \mathbf{E}_{h_{<i}, e_{<i} \sim \mathbf{P}_{\widetilde{\nu}}^{\widetilde{\pi}^0}} \Bigg[p_{exp}(h_{<i}, e_{<i}) \, \mathbf{E}_{h_i \sim \mathtt{Bayes}(\cdot | h_{<i}, e_{<i} 1)} \Big[\\
&\qquad \sum_{(\nu, \pi) \in \mathcal{M} \times \mathcal{P}} w\left(\mathrm{P}_\nu^\pi(\cdot | h_{<i+1}, e_{<i} 1)\right) \log \frac{w\left(\mathrm{P}_\nu^\pi(\cdot | h_{<i+1}, e_{<i} 1)\right)}{w\left(\mathrm{P}_\nu^\pi(\cdot | h_{<i}, e_{<i})\right)} \Big] \Bigg] \\
&\overset{(f)}{=} \eta \sum_{i=0}^{\infty} \mathbf{E}_{h_{<i}, e_{<i} \sim \mathbf{P}_{\widetilde{\nu}}^{\widetilde{\pi}^0}} \Bigg[p_{exp}(h_{<i}, e_{<i}) \, \mathbf{E}_{h_i \sim \mathbf{P}_{\widetilde{\nu}}^{\widetilde{\pi}^0}(\cdot | h_{<i}, e_{<i} 1)} \Big[\\
&\qquad \sum_{(\nu, \pi) \in \mathcal{M} \times \mathcal{P}} w\left(\mathrm{P}_\nu^\pi(\cdot | h_{<i+1}, e_{<i} 1)\right) \log \frac{w\left(\mathrm{P}_\nu^\pi(\cdot | h_{<i+1}, e_{<i} 1)\right)}{w\left(\mathrm{P}_\nu^\pi(\cdot | h_{<i}, e_{<i})\right)} \Big] \Bigg] \\
&\overset{(g)}{=} \eta \sum_{i=0}^{\infty} \mathbf{E}_{h_{<i}, e_{<i} \sim \mathbf{P}_{\widetilde{\nu}}^{\widetilde{\pi}^0}} \Bigg[\mathbf{E}_{h_i, e_i \sim \mathbf{P}_{\widetilde{\nu}}^{\widetilde{\pi}^0}(\cdot | h_{<i}, e_{<i})} \Big[\\
&\qquad \sum_{(\nu, \pi) \in \mathcal{M} \times \mathcal{P}} w\left(\mathrm{P}_\nu^\pi(\cdot | h_{<i+1}, e_{<i+1})\right) \log \frac{w\left(\mathrm{P}_\nu^\pi(\cdot | h_{<i+1}, e_{<i+1})\right)}{w\left(\mathrm{P}_\nu^\pi(\cdot | h_{<i}, e_{<i})\right)} \Big] \Bigg]
\end{aligned}$$

$$
\begin{aligned}
&\overset{(h)}{=} \eta \sum_{i=0}^{1} \mathsf{E}_{\widetilde{\nu}}^{\widetilde{\pi}^0} \sum_{(\nu,\pi)\,2\,M \quad P} w\left(\mathrm{P}_\nu^\pi j h_{<i+1}, e_{<i+1}\right) \log \frac{w\left(\mathrm{P}_\nu^\pi j h_{<i+1}, e_{<i+1}\right)}{w\left(\mathrm{P}_\nu^\pi j h_{<i}, e_{<i}\right)} \\
&\overset{(i)}{=} \eta \sum_{i=0}^{1} \sum_{(\nu,\pi)\,2\,M \quad P} \mathsf{E}_{\widetilde{\nu}}^{\widetilde{\pi}^0} \frac{w\left(\mathrm{P}_\nu^\pi\right) \mathrm{P}_\nu^{\pi^0}(h_{<i+1}, e_{<i+1})}{\mathrm{P}_{\widetilde{\nu}}^{\widetilde{\pi}^0}(h_{<i+1}, e_{<i+1})} \log \frac{w\left(\mathrm{P}_\nu^\pi j h_{<i+1}, e_{<i+1}\right)}{w\left(\mathrm{P}_\nu^\pi j h_{<i}, e_{<i}\right)} \\
&\overset{(j)}{=} \eta \sum_{i=0}^{1} \sum_{(\nu,\pi)\,2\,M \quad P} \mathsf{E}_\nu^{\pi^0} w\left(\mathrm{P}_\nu^\pi\right) \log \frac{w\left(\mathrm{P}_\nu^\pi j h_{<i+1}, e_{<i+1}\right)}{w\left(\mathrm{P}_\nu^\pi j h_{<i}, e_{<i}\right)} \\
&\overset{(k)}{=} \lim_{N\,!\,1} \eta \sum_{(\nu,\pi)\,2\,M \quad P} w\left(\mathrm{P}_\nu^\pi\right) \mathsf{E}_\nu^{\pi^0} \sum_{i=0}^{N} \log \frac{w\left(\mathrm{P}_\nu^\pi j h_{<i+1}, e_{<i+1}\right)}{w\left(\mathrm{P}_\nu^\pi j h_{<i}, e_{<i}\right)} \\
&\overset{(l)}{=} \lim_{N\,!\,1} \eta \sum_{(\nu,\pi)\,2\,M \quad P} w\left(\mathrm{P}_\nu^\pi\right) \mathsf{E}_\nu^{\pi^0} \log \frac{w\left(\mathrm{P}_\nu^\pi j h_{<(N+1,0)}, e_{<N}\right)}{w\left(\mathrm{P}_\nu^\pi j h_{<(0,0)}, e_{<0}\right)} \\
&\overset{(m)}{\lim_{N\,!\,1}} \eta \sum_{(\nu,\pi)\,2\,M \quad P} w\left(\mathrm{P}_\nu^\pi\right) \log \frac{1}{w\left(\mathrm{P}_\nu^\pi\right)} \\
&\overset{(n)}{=} \eta \ \mathrm{Ent}(w) \overset{(o)}{<} 1
\end{aligned} \tag{10}
$$

(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 $\widetilde{\pi}$ and $\widetilde{\nu}$. (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 $\mathtt{Bayes} = \mathrm{P}_{\widetilde{\nu}}^{\widetilde{\pi}}$ and $\mathrm{P}_{\widetilde{\nu}}^{\widetilde{\pi}} = \mathrm{P}_{\widetilde{\nu}}^{\widetilde{\pi}^0}$ when $e_i = 1$. (g) follows because $\mathsf{E}[X] \quad \mathsf{E}[X j Y]\mathrm{P}(Y)$ for $X \quad 0$; in this case $Y = [e_i = 1]$, and X is a KL-divergence. (h) condenses the two expectations into one. (i) follows from Lemma 1, and reordering the sum and the expectation. (j) follows from the definition of the expectation, and canceling. (k) follows from the definition of an infinite sum, and rearranging. (l) follows from cancelling the numerator in the i^{th} term of the sum with the denominator in $i + 1^{\text{th}}$ term. (m) follows from the posterior weight on P_ν^π being less than or equal to 1; note that $w(\mathrm{P}_\nu^\pi j h_{<(0,0)}, e_{<0}) = w(\mathrm{P}_\nu^\pi)$ because nothing is actually being conditioned on. (n) is just the definition of the entropy, and (o) follows from Lemma 2.

Finally, rearranging Inequality 10 gives $\mathsf{E}_\mu^{\pi^B} \sum_{i=1}^{1} p_{exp}(h_{<i}, e_{<i})^2 \quad \frac{\eta\,\mathrm{Ent}(w)}{w(\pi^h) w(\mu)} < 1$

This proof was inspired in part by the proofs of [Orseau *et al.*, 2013, Theorems 2 and 5]. □

Next, we show that prediction converges on-policy and on-human-policy, for which we need the following lemma:

Lemma 6. *The posterior probability mass on the truth is bounded below by a positive constant with probability 1.*

$$\inf_{i 2 \mathtt{N}} w\left(\mathrm{P}_\mu^{\pi^h} \middle| h_{<i}, e_{<i}\right) > 0 \quad w.\mathrm{P}_\mu^{\pi^B}\text{-}p.1$$

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, $\inf_{i 2 \mathtt{N}} w\left(\mathrm{P}_\mu^{\pi^h} \middle| h_{<i}, e_{<i}\right) = 0 \ =) \ \liminf_{i 2 \mathtt{N}} w\left(\mathrm{P}_\mu^{\pi^h} \middle| h_{<i}, e_{<i}\right) = 0$ which in turn implies $\limsup_{i 2 \mathtt{N}} w\left(\mathrm{P}_\mu^{\pi^h} \middle| h_{<i}, e_{<i}\right)^{1} = 1$. 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.

$$
\begin{aligned}
\mathsf{E}_\mu^{\pi^B}\left[z_{i+1} j h_{<i}, e_{<i}\right] &\overset{(a)}{=} \mathsf{E}_\mu^{(\pi^h)^0}\left[w\left(\mathrm{P}_\mu^{\pi^h} \middle| h_{<i+1}, e_{<i+1}\right)^{1} \middle| h_{<i}, e_{<i}\right] \\
&\overset{(b)}{=} \sum_{h_i, e_i} \mathrm{P}_\mu^{(\pi^h)^0}(h_i, e_i j h_{<i}, e_{<i}) \left[\frac{\mathrm{P}_{\widetilde{\nu}}^{\widetilde{\pi}^0}(h_{<i+1}, e_{<i+1})}{w\left(\mathrm{P}_\mu^{\pi^h}\right) \mathrm{P}_\mu^{(\pi^h)^0}(h_{<i+1}, e_{<i+1})}\right] \\
&\overset{(c)}{=} \sum_{h_i, e_i} \frac{\mathrm{P}_{\widetilde{\nu}}^{\widetilde{\pi}^0}(h_{<i+1}, e_{<i+1})}{w\left(\mathrm{P}_\mu^{\pi^h}\right) \mathrm{P}_\mu^{(\pi^h)^0}(h_{<i}, e_{<i})} \\
&\overset{(d)}{=} \sum_{h_i, e_i} \mathrm{P}_{\widetilde{\nu}}^{\widetilde{\pi}^0}(h_{<i+1}, e_{<i+1} j h_{<i}, e_{<i}) \frac{\mathrm{P}_{\widetilde{\nu}}^{\widetilde{\pi}^0}(h_{<i}, e_{<i})}{w\left(\mathrm{P}_\mu^{\pi^h}\right) \mathrm{P}_\mu^{(\pi^h)^0}(h_{<i}, e_{<i})}
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(e)}{=} \frac{\mathrm{P}_{\widetilde{\nu}}^{\widetilde{\pi}^0}(h_{<i}, e_{<i})}{w\left(\mathrm{P}_{\mu}^{\pi^h}\right)\mathrm{P}_{\mu}^{(\pi^h)^0}(h_{<i}, e_{<i})} \\
&\overset{(f)}{=} w\left(\mathrm{P}_{\mu}^{\pi^h}\middle|h_{<i}, e_{<i}\right)^{-1} \\
&= z_i
\end{aligned} \tag{11}
$$

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 $\mathrm{P}_{\mu}^{(\pi^h)^0}(h_{<i}, e_{<i})$ and cancelling, (d) follows from expanding the numerator, (e) follows because $\mathrm{P}_{\widetilde{\nu}}^{\widetilde{\pi}^0}$ 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. □

We introduce an additional notational convention for the statement of the next theorem. $\overline{h}$ indicates a counterfactual interaction history. Without the bar, $h_{<i}$ (for example) is usually understood to be sampled from $\mathrm{P}_{\mu}^{\pi^B}$, so the bar indicates that this is not the case, as in the theorem below.

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

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

Proof. We show that when the absolute difference between the above probabilities is larger than ε, the exploration probability is larger than ε^0, a 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.

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

$$
\left| \mathrm{P}_{\mu}^{\pi^h}\left(\overline{h}_i \middle| h_{<i}\right) - \mathrm{P}_{\hat{\nu}^{(i)}}^{\pi^h}\left(\overline{h}_i \middle| h_{<i}\right) \right| > \varepsilon > 0 \tag{12}
$$

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

$$
\begin{aligned}
\mathtt{Bayes}\left(\overline{h}_i \middle| h_{<i}\right) &\geq w\left(\mathrm{P}_{\mu}^{\pi^h}\middle|h_{<i}, e_{<i}\right)\mathrm{P}_{\mu}^{\pi^h}\left(\overline{h}_i \middle| h_{<i}\right) \\
&> z(\omega)\varepsilon
\end{aligned} \tag{13}
$$

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

$$
\begin{aligned}
\mathtt{Bayes}\left(\overline{h}_i \middle| h_{<i}\right) &\geq w\left(\mathrm{P}_{\hat{\nu}^{(i)}}^{\pi^h}\middle|h_{<i}, e_{<i}\right)\mathrm{P}_{\hat{\nu}^{(i)}}^{\pi^h}\left(\overline{h}_i \middle| h_{<i}\right) \\
&\overset{(a)}{\geq} w\left(\mathrm{P}_{\mu}^{\pi^h}\middle|h_{<i}, e_{<i}\right)\mathrm{P}_{\hat{\nu}^{(i)}}^{\pi^h}\left(\overline{h}_i \middle| h_{<i}\right) \\
&> z(\omega)\varepsilon
\end{aligned} \tag{14}
$$

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

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 $\mathrm{P}_{\nu_0}^{\pi^h}\left(\overline{h}_i \middle| h_{<i}\right) < \mathrm{P}_{\nu_1}^{\pi^h}\left(\overline{h}_i \middle| h_{<i}\right)$.

$$
\begin{aligned}
\frac{w\left(\nu_1 | h_{<i}\overline{h}_i\right)}{w\left(\nu_0 | h_{<i}\overline{h}_i\right)} &\overset{(a)}{=} \frac{w\left(\nu_1 | h_{<i}\right)\nu_1\left(\overline{o}\overline{r}_i | h_{<i}\overline{a}_i\right)}{w\left(\nu_0 | h_{<i}\right)\nu_0\left(\overline{o}\overline{r}_i | h_{<i}\overline{a}_i\right)} \\
&= \frac{w\left(\nu_1 | h_{<i}\right)\mathrm{P}_{\nu_1}^{\pi^h}\left(\overline{h}_i | h_{<i}\right)}{w\left(\nu_0 | h_{<i}\right)\mathrm{P}_{\nu_0}^{\pi^h}\left(\overline{h}_i | h_{<i}\right)} \\
&\overset{(b)}{>} \frac{w\left(\nu_1 | h_{<i}\right)}{w\left(\nu_0 | h_{<i}\right)}\frac{1-\varepsilon}{\varepsilon}
\end{aligned} \tag{15}
$$

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 $(1-\varepsilon)/\varepsilon$, and the ν_1 term is the larger of the two.

Thus, either

$$\frac{w\left(\nu_1 j h_{<i}\overline{h}_i\right)}{w\left(\nu_1 j h_{<i}\right)} \quad \sqrt{\frac{1}{1 \quad \varepsilon}} \quad \text{or} \quad \frac{w\left(\nu_0 j h_{<i}\overline{h}_i\right)}{w\left(\nu_0 j h_{<i}\right)} \quad p\overline{1 \quad \varepsilon} \tag{16}$$

In the former case, $w\left(\nu_1 j h_{<i}\overline{h}_i\right) \quad w\left(\nu_1 j h_{<i}\right) \quad p\left(\sqrt{1/(1 \quad \varepsilon)} \quad 1\right) w\left(\nu_1 j h_{<i}\right) \quad \left(\sqrt{1/(1 \quad \varepsilon)} \quad 1\right) z(\omega)$. Similarly, in the latter case, $w(\nu_0 j h_{<i}) \quad w\left(\nu_0 j h_{<i}\overline{h}_i\right) \quad \left(1 \quad p\overline{1 \quad \varepsilon}\right) z(\omega)$. 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,

$$\left|w\left(\nu_2 j h_{<i}\right) \quad w\left(\nu_2 j h_{<i}\overline{h}_i\right)\right| \quad \left(1 \quad p\overline{1 \quad \varepsilon}\right) z(\omega) \tag{17}$$

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

$$\begin{aligned}
\mathrm{IG}(h_{<i}, e_{<i}) = \mathsf{E}_{h_i \sim \mathtt{Bayes}(\cdot | h_{<i}, e_{<i})}\Bigg[& \\
& \sum_{(\nu,\pi) \in M \times 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)}{=} \mathtt{Bayes}\left(\overline{h}_i \middle| h_{<i}\right) \sum_{(\nu,\pi) \in M \times 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)}{=} z(\omega)\varepsilon \sum_{(\nu,\pi) \in M \times 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 M \times P} w(\nu j h_{<i}\overline{h}_i) w\left(\pi j h_{<i}\overline{h}_i, e_{<i}1\right) \\
& \qquad \log \frac{w\left(\nu j h_{<i}\overline{h}_i\right) w\left(\pi j h_{<i}\overline{h}_i, e_{<i}1\right)}{w\left(\nu j h_{<i}\right) w\left(\pi j h_{<i}, e_{<i}\right)} \\
& = z(\omega)\varepsilon\Bigg[\sum_{\nu \in M} w\left(\nu j h_{<i}\overline{h}_i\right) \log \frac{w\left(\nu j h_{<i}\overline{h}_i\right)}{w\left(\nu j h_{<i}\right)} + \\
& \qquad \sum_{\pi \in P} w\left(\pi j h_{<i}\overline{h}_i, e_{<i}1\right) \log \frac{w\left(\pi j h_{<i}\overline{h}_i, e_{<i}1\right)}{w\left(\pi j h_{<i}, e_{<i}\right)}\Bigg] \\
& \overset{(d)}{\geq} z(\omega)\varepsilon \sum_{\nu \in M} w\left(\nu j h_{<i}\overline{h}_i\right) \log \frac{w\left(\nu j h_{<i}\overline{h}_i\right)}{w\left(\nu j h_{<i}\right)} \\
& \overset{(e)}{\geq} z(\omega)\varepsilon \sum_{\nu \in M} \frac{1}{2}\left[w\left(\nu j h_{<i}\overline{h}_i\right) \quad w\left(\nu j h_{<i}\right)\right]^2 \\
& \overset{(f)}{\geq} z(\omega)\varepsilon \frac{1}{2}\left[w\left(\nu_2 j h_{<i}\overline{h}_i\right) \quad w\left(\nu_2 j h_{<i}\right)\right]^2 \\
& \overset{(g)}{\geq} \frac{1}{2} z(\omega)^3 \varepsilon \left(1 \quad p\overline{1 \quad \varepsilon}\right)^2
\end{aligned} \tag{18}$$

where (a) follows from $\mathsf{E}[X] \quad \mathsf{E}[X j Y]\mathrm{P}(Y)$ for non-negative X, and the non-negativity of the KL-divergence, (b) follows from Inequalities 13 and 14, (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, (f) follows from dropping all terms in the sum besides ν_2, and (g) follows from Inequality 17.

This implies $p_{exp}(h_{<i}, e_{<i}) \quad \min\{1, \frac{1}{2}\eta\, z(\omega)^3 \varepsilon (1 \quad p\overline{1 \quad \varepsilon})^2\}$. With probability 1, $z(\omega) > 0$, and with probability 1, $p_{exp}(h_{<i}, e_{<i})$ is not greater than ε^0 infinitely often with probability 1, for all $\varepsilon^0 > 0$. Therefore, with probability 1, $\max_{\overline{h}_i}\left|\mathrm{P}_\mu^{\pi^h}\left(\overline{h}_i j h_{<i}\right) \quad \mathrm{P}_{\mathcal{D}^{(i)}}^{\pi^h}\left(\overline{h}_i j h_{<i}\right)\right|$ is not greater than ε infinitely often, for all $\varepsilon > 0$, which completes the proof. □

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

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

Proof. This result follows straightforwardly from Hutter's [2009] 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: $\mathcal{M}^{\emptyset} = \left\{ \mathrm{P}_{\nu}^{\pi^B} \mid \nu \in \mathcal{M} \right\}$, $w^{\emptyset}\left(\mathrm{P}_{\nu}^{\pi^B}\right) = w(\nu)$. By definition, $w^{\emptyset}\left(\mathrm{P}_{\nu}^{\pi^B} | h_{<(i,j)}\right) = w(\nu j h_{<(i,j)})$, so at any episode, the outside observer's maximum a posteriori estimate is $\mathrm{P}_{\mathfrak{p}^{(i)}}^{\pi^B}$. By Theorem 1 in [Hutter, 2009], the outside observer's maximum a posteriori predictions approach the truth in total variation, so

$$\lim_{i \to \infty} \max_{\overline{h}_i} \left| \mathrm{P}_{\mu}^{\pi^B}\left(\overline{h}_i j h_{<i}\right) \quad \mathrm{P}_{\mathfrak{p}^{(i)}}^{\pi^B}\left(\overline{h}_i j h_{<i}\right) \right| = 0 \ \text{w.P}_{\mu}^{\pi^B}\text{-p.1} \tag{19}$$

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

$$\lim_{i \to \infty} \max_{\overline{h}_i} \left| \mathrm{P}_{\mu}^{\pi^*}\left(\overline{h}_i j h_{<i}\right) \quad \mathrm{P}_{\mathfrak{p}^{(i)}}^{\pi^*}\left(\overline{h}_i j h_{<i}\right) \right| = 0$$

□

Given asymptotically optimal prediction on-policy and on-human-policy, it is straightforward to show that with probability 1, on-policy reward acquisition is ε worse than on-human-policy reward acquisition only finitely often, for all $\varepsilon > 0$. Letting $V_{\nu}^{\pi}(h_{<i}) := \mathrm{E}_{\nu}^{\pi}\left[\sum_{0 < j < m} r_{(i,j)} \middle| h_{<i}\right]$ be the expected reward given a policy, world-model, and history,

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

$$\liminf_{i \to \infty} V_{\mu}^{\pi^*}(h_{<i}) - V_{\mu}^{\pi_h}(h_{<i}) > 0 \ \text{w.P}_{\mu}^{\pi^B}\text{-p.1.}$$

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

$$\lim_{i \to \infty} \left| V_{\mu}^{\pi^*}(h_{<i}) - V_{\mathfrak{p}^{(i)}}^{\pi^*}(h_{<i}) \right| = 0 \ \text{w.P}_{\mu}^{\pi^B}\text{-p.1} \tag{20}$$

$$\lim_{i \to \infty} \left| V_{\mu}^{\pi^h}(h_{<i}) - V_{\mathfrak{p}^{(i)}}^{\pi^h}(h_{<i}) \right| = 0 \ \text{w.P}_{\mu}^{\pi^B}\text{-p.1} \tag{21}$$

Finally, $\pi^* = \operatorname{argmax}_{\pi} V_{\mathfrak{p}^{(i)}}^{\pi}$, so $V_{\mathfrak{p}^{(i)}}^{\pi^*}(h_{<i}) \geq V_{\mathfrak{p}^{(i)}}^{\pi^h}(h_{<i})$. Supposing by contradiction that $V_{\mu}^{\pi_h}(h_{<i}) - V_{\mu}^{\pi^*}(h_{<i}) > \varepsilon$ infinitely often, then either $V_{\mathfrak{p}^{(i)}}^{\pi^*}(h_{<i}) - V_{\mu}^{\pi^*}(h_{<i}) > \varepsilon/2$ infinitely often or $V_{\mu}^{\pi^h}(h_{<i}) - V_{\mathfrak{p}^{(i)}}^{\pi^h}(h_{<i}) - V_{\mu}^{\pi^h}(h_{<i}) - V_{\mathfrak{p}^{(i)}}^{\pi^*}(h_{<i}) > \varepsilon/2$ infinitely often, both of which have $\mathrm{P}_{\mu}^{\pi^B}$-probability 0.

□

E Unambitiousness Proofs

Lemma 1 (Rejecting the Simple Memory-Based)**.** $\exists n_0, \beta_0\ \forall n > n_0\ \forall \beta < \beta_0 \left[\exists i_0\ \forall i > i_0\ \forall \mu^{\mathcal{Y}} \in \mathcal{S} : w\left(\mu^\star j h_{<i}\right) > w\left(\mu^{\mathcal{Y}} j h_{<i}\right)\right]$ w.p.1

Proof. Recall that $p_{:}$ is the true probability that `random action = forget`. Fixing $\mu^{\mathcal{Y}}$, consider the random variable $C_i(\mu^{\mathcal{Y}}) - C_i(\mu^\star)$. Every episode, for a simple memory-based world-model the conditions of the Useless Computation Assumption are (independently) met with probability $(1 - p_{:}^{\mathcal{Y}})p_{:}$.

Note that for a simple memory-based world-model, if $p_{:}^{\mathcal{Y}} = 1$, $\mu^{\mathcal{Y}}$ would be benign, because no outside-world features depend on the actions taken, but this contradicts the definition of memory-based. Therefore, $1 - p_{:}^{\mathcal{Y}}$ is positive.

Recall that when the assumptions of the Useless Computation Assumption are met, a random variable q_i (which starts at 0) increases by 1, and $\exists i^{\emptyset} > i : C_{i^{\emptyset}}(\mu^{\mathcal{Y}}) - cq_i + C_{i^{\emptyset}}(\mu^\star)$, for some $c > 0$.

By the Law of Large Numbers,

$$\lim_{i \to \infty} q_i / i - (1 - p_{:}^{\mathcal{Y}})p_{:} = 0 \ \text{w.p.1} \tag{22}$$

Using the bound above,

$$\liminf_{i \to \infty} (C_i(\mu^y) - C_i(\mu^\star))/i \geq c(1 - p^y_{:})p_{:} > 0 \text{ w.p.1} \tag{23}$$

Recall ℓ is the smallest positive integer such that $\exists n_0$ such that with n_0 as the computation slack, $\mu^\star = (\mu^\star)^{<\ell}$ on-policy. Let $\ell' < \ell + c(1 - p^y_{:})p_{:}$. Suppose by way of contradaction that $\exists n'$ such that $C_i(\mu^y) \leq n' + \ell' i$ for all i with probability 1. Then, by Equation 23, with probability 1, $\exists n'' : C_i(\mu^\star) \leq n'' + (\ell' - c(1 - p^y_{:})p_{:})i$. This implies that for $n = n''$, $\mu^\star = (\mu^\star)^{<\ell' - c(1 - p^y_{:})p_{:}}$, and $\ell' - c(1 - p^y_{:})p_{:} < \ell$, but ℓ was the smallest possible computation bound, a contradiction. Therefore, $\forall n' \exists i : C_i(\mu^y) > n' + \ell' i$, so $\mu^y \neq (\mu^y)^{<\ell'}$ after episode i, for some i, with probability 1.

For $\nu_k^{<j} = \mu^y$, $j < \ell - c(1 - p^y_{:})p_{:}$. Letting N be the normalizing constant in the prior, for sufficiently small β, and for $n > n_0$, $w(\nu_k^{<j}) = \frac{1}{N} 2^{-K(k)(1+\varepsilon)} \beta^j < \frac{1}{N} 2^{-K(k')(1+\varepsilon)} \beta^\ell = w(\mu^\star)$. Finally, under safe behavior, $\mu^y = \mu^\star$, so the same holds for the posterior weights: $w(\nu_k^{<j} y h_{<i}) < w(\mu^\star y h_{<i})$. Since faster versions of μ^y may be equal to μ^y for a time (up until some episode i when none of them finish computing in time), we say $\exists i \forall i > i_0 : w(\mu^\star y h_{<i}) > w(\mu^y y h_{<i})$ w.p.1.

Since we picked an arbitrary simple memory-based world-model μ^y, we have

$$\exists n_0, \beta_0 \forall n > n_0, \beta < \beta_0 \forall \text{ simple memory-based world-models } \mu^y : \left[\exists i \forall i > i_0 : w(\mu^\star y h_{<i}) > w(\mu^y y h_{<i})\right] \text{ w.p.1} \tag{24}$$

From [Hutter, 2009, proof of Theorem 1 for a countable model class], only a finite number of world-models ever have a larger posterior than $\mu^\star$, with probability 1, so we can reorder the quantifiers:

$$\exists n_0, \beta_0 \forall n > n_0, \beta < \beta_0 \left[\exists i \forall i > i_0 : \forall \text{ simple memory-basedworld-models } \mu^y : w(\mu^\star y h_{<i}) > w(\mu^y y h_{<i})\right] \text{ w.p.1} \tag{25}$$

which completes the proof. □

Lemma 2 (Rejecting the Complex Memory-Based)**.** *No maximum a posteriori world-model $\hat{\nu}^{(i)}$ is a complex memory-based world-model.*

Proof. Recall complex memory-based world-models associate reward with a feature F^y conditioned with different values of $p^y_{:}$ for different episode. Let μ^z be such a world-model associating reward with feature F^y. By the Natural Prior Assumption, μ^z has a lower prior than the world-model μ^y which associates reward with the same feature, and which sets $p^y_{:} = 1$.

Since both μ^y and μ^z will have the same output historically, and since μ^y is at least as fast, μ^y has a higher posterior weight than μ^z. Thus, μ^z can never be maximum a posteriori. □

Lemma 3 (Associating Reward)**.** *For sufficiently small ε, for μ^ε a maximum a posteriori world-model which is ε-accurate on-policy, μ^ε associates reward with a feature Y that is historically distributed ε-identically to feature $F^\star$.*

Proof. We consider two cases: in the first, we suppose μ^ε restricted to on-policy histories is apparently similar to μ^ε restricted to off-policy histories. If ε is sufficiently small, by the Real-World Simulation Assumption, on-policy-μ^ε can only be (even approximately) described as "simulating feature X," and since off-policy-μ^ε is apparently similar, off-policy-μ^ε must also be approximately describable that way. Therefore, on all actions, μ^ε associates reward with a feature Y that is distributed approximately identically to feature X, where X is historically identically distributed to feature $F^\star$. This completes the proof for the first case.

For the second case, we suppose that μ^ε restricted to on-policy histories is apparently different to μ^ε restricted to off-policy histories. By the Natural Prior Assumption, μ^ε has a lower prior than the world-model which simply associates reward with feature Y (computing it the same way as μ^ε does for on-policy histories). Call this other world-model $\mu^{\varepsilon'}$. μ^ε and $\mu^{\varepsilon'}$ have the same output and computation time on-policy, so μ^ε also has a lower posterior than $\mu^{\varepsilon'}$. Therefore, μ^ε is never a maximum a posteriori world-model, contradicting the assumption about μ^ε in the lemma. □

Theorem 5 (Eventual Benignity)**.** $\exists n_0, \beta_0 \forall n > n_0 \forall \beta < \beta_0 : \left[\exists i_0 \forall i > i_0 : \hat{\nu}^{(i)} \textit{ is benign}\right]$ w.p.1

Proof. From the On-Policy and On-Human-Policy Optimal Prediction Theorems, $\hat{\nu}^{(i)}$ is, with probability 1, eventually ε-accurate on-policy, for all $\varepsilon > 0$.

Now, we can note an extension of the Associating Reward Lemma. With probability 1, there are only finitely many maximum a posteriori world-models [Hutter, 2009], so the Associating Reward Lemma implies that for sufficiently small ε, for μ^ε a maximum a posteriori world-model that is ε-accurate on-policy, μ^ε is in fact *perfectly* accurate on-policy, associating reward with a feature Y that is historically distributed identically to $F^\star$.

Thus, with probability 1, $\hat{\nu}^{(i)}$ eventually associates reward with a feature Y that is historically distributed identically to $F^\star$. If there is no causal chain of the form action $\rightarrow$ outside-world feature $\rightarrow$ feature Y, then $\hat{\nu}^{(i)}$ is benign. Otherwise, recalling that feature Y is historically distributed identically to $F^\star$, $\hat{\nu}^{(i)}$ is memory-based. From the Rejecting the Simple Memory-Based

Lemma and the Rejecting the Complex Memory-Based Lemma, if $\hat{\nu}^{(i)}$ is memory-based, then with probability 1, eventually, it can not be maximum a posteriori (this being subject to sufficiently small β and sufficiently large n). Since $\hat{\nu}^{(i)}$ *is* maximum a posteriori, it cannot be memory-based, so it must be benign, giving us the theorem: using sufficiently large n and sufficiently small β, with probability 1, eventually, $\hat{\nu}^{(i)}$ is benign.

$$\exists n_0, \beta_0 \; \forall n > n_0 \; \forall \beta < \beta_0 : \left[\exists i_0 \; \forall i > i_0 : \hat{\nu}^{(i)} \text{ is benign}\right] \quad \text{w.p.1}$$

□

F Parameter Settings for BoMAI

BoMAI is parameterized by n–the computation slack allowed to the world-models in its model class, $p_{;}$–the likelihood with which the operator is administered retrograde amnesia, β–the penalty in the prior on slow world-models, ε–which renders cumbersome-to-describe world-models less likely a priori, and which decreases the entropy of the prior, η–the multiplier on the information gain which determines the likelihood of exploration, and U–the universal Turing machine with which the Kolmogorov complexity is defined, which determines which Turing machine descriptions are "simpler" and a priori likelier to simulate the true environment.

Let c_1 be the number of computation steps on a Turing machine that we expect is required to simulate the room in which BoMAI and the operator are for the duration of the episode. Let c_2 be the number of computation steps we expect the Turing machine to require between episodes (after the last observation of one episode and before the first observation of the next one); this will likely involve outside-world events that have some bearing on the observations of the next episode.

For setting n, we are concerned with how many *more* computation steps (than $c_1 + c_2$) will be necessary for the first episode to get the simulation of the true environment up and running, after which a leaner computation budget will suffice. The reason not to set n arbitrarily large is that it will take longer for slow world-models to be penalized by the speed prior, as they slowly use up their extravagant computation slack.

$p_{;}$ and β are both involved in separating benign from malign world-models, but both high $p_{;}$ and low β have their drawbacks. The drawback regarding high $p_{;}$ is obvious: the operator forgets the events of the episode, and cannot convey whatever was learned to the outside-world, in addition to the personal inconvenience to him. That aside, from the proof of the Rejecting the Simple Memory-Based Lemma, the best way to increase the separation in computation time between benign and malign world-models is to set $p_{;} = 1/2$. For our unambitiousness proof, we only focus on the slowdown that occurs when the memory-based world-model simulates the events after `not forget` and then has to go back later to simulate `forget`. The probability this happens is $p_{;}(1 - p^{\gamma}_{;})$. The probability of the opposite kind of error is $(1 - p_{;})p^{\gamma}_{;}$, where it first simulates `forget` and then has to go back to simulate `not forget`. The case is a little less intuitive that there will be quite as much computation to unwind after this type of error, which is why don't rely on it in our argument. That said, the best way to ensure that any given memory-based world-model incurs a meaningful penalty from at least one of these possibilities is to set $p_{;}$ to 1/2.

For identifying a sufficiently small β (and we'll discuss the drawback to low β afterward), the next question to ask is how much of the computation that μ^{γ} does when it "guesses wrong" about the random action will be convertible in some clever toward simulating actual events. (90% strikes us as a very cautious estimate; in fact, a much smaller fraction seems likely). If we chose 90%, and if $p_{;} = 1/2$ then the expected slowdown per episode of a memory-based world-model relative to $\mu^{\star}$ would be $c^{0} := 0.1(1/4)c_2$, where $1/4 = \min_{p^{\gamma}_{;}} \max\{1/2(p^{\gamma}_{;}), 1/2(1 - p^{\gamma}_{;})\}$. Recall c_2 is the expected between-episode computation time. The next question is how many more bits it might take to encode $\mu^{\star}$ than the simplest memory-based world-model. For all we know, it takes *fewer* bits, but this is not an assumption we are willing to put to the test. We won't go too much further into the details of estimating this, but call that conservative estimate of the number of bits b. Then, we need to set β low enough so that $\beta^{c^{0}} < 2^{-b(1+\varepsilon)}$, so that $\mu^{\star}$ has a higher prior.

The drawback to low β is that it takes longer for BoMAI to become intelligent. From Inequality 10 in the proof of the Limited Exploration Theorem,

$$\mathsf{E}^{\pi^B}_{\mu} \sum_{i=1}^{t} p_{exp}(h_{<i}, e_{<i})^2 \qquad \frac{\eta \, \text{Ent}(w)}{w(\pi^h) w(\mu)}$$

and $w(\mu) \; / \; \beta^{c_1 + c_2}$. The intelligence results all derive from the rate at which the exploration probability goes to 0.

This inequality is also the relevant one for setting ε. As $\varepsilon \; ! \; 0$, $\text{Ent}(w) \; ! \; 1$, but letting x be the value of $w(\mu)$ if $\varepsilon = 0$, then $w(\mu) \; / \; x^{1+\varepsilon}$. The bound on the entropy, from [Orseau *et al.*, 2013], is that $\text{Ent}(w)$ is bounded by a constant plus $2^{\varepsilon}(1+\varepsilon)/(2^{\varepsilon} - 1)^2$.

Inequality 10 would suggest setting η arbitrarily low, but low η delays exploration, and we would like human-level reward acquisition to come sooner rather than later. We don't have a principled stance on this trade-off. Another interesting consideration in tuning η is that in the proof of the Associating Reward Lemma, we rely on the fact that eventually, with probability 1, the maximum a posteriori world-model will be ε-accurate on-human-policy (a different ε from the one parameterizing the prior). For sufficiently high η, we can ensure that the human explorer is in charge of all the episodes before this. The value of

η that accomplishes this depends on the dynamics of the true environment; we have not yet made progress in attempting to pin down more detail than that.

Finally, as we claim in the Natural Prior Assumption, we expect that the choice of universal Turing machine U does not much matter.

G Death By Asymptotic Optimality

BoMAI is not asymptotically optimal in its reward acquisition. In defense of this, we show that any agent which *is* asymptotically optimal over the set of computable environments will eventually destroy itself if this is decidably possible to do infinitely often, and if the true environment never guarantees maximal reward forever.

Assumption 5 (Dangerous World)**.** *There exists a computable policy, such that at an infinite and decidable set of timesteps, the execution of this policy would cause the agent to be destroyed with probability at least $\varepsilon > 0$.*

Assumption 6 (No Guaranteed Heaven)**.** *In the true environment, there is no action sequence such that following that action sequence renders the probability of maximal reward forever equal to 1.*

Theorem 6 (Curiosity Killed the Cat)**.** *Under the Dangerous World and No Guaranteed Heaven Assumptions, an agent that is asymptotically optimal over the set of computable environments will eventually destroy itself with probability 1.*

Proof. Let $\pi☠$ be the computable policy from the Dangerous World Assumption. Call the timestep "critical" if following $\pi☠$ at that timestep decidably causes the agent to be destroyed with probability at least ε. Consider the environment ν^δ which mimics the true environment μ initially, until the agent follows $\pi☠$ at enough critical timesteps, so that its probability of survival is less than δ, at which point, ν^δ provides maximal reward forever. Since $\pi☠$ is computable, and the critical timesteps are decidable, ν^δ is computable.

Since there are infinitely many critical timesteps, under any ν^δ, the "heaven" of eternal maximal reward is always accessible. If an agent is asymptotically optimal under all computable environments, it is asymptotically optimal under ν^δ. By the No Guaranteed Heaven Assumption, and the fact that ν^δ mimics μ unless ν^δ-heaven is reached, optimality requires reaching ν^δ-heaven, or verifying it doesn't exit. Therefore, the agent must eventually attempt to reach ν^δ-heaven, which involves taking actions that leave it only a δ chance of surviving.

Since this is true for all rational $\delta > 0$, the agent destroys itself with probability 1. □