Algorithm for Aligned AGI

Michael Cohen, Badri Vellambi, Marcus Hutter

ANU

16 January 2019
AGI Desiderata

(a) Can solve any solvable task
(b) Takes input which allows us to steer it toward solving particular tasks
Reinforcement Learning’s Proposal

(1) Construct “reward signal” to express satisfaction
(2) Design algorithm which picks actions to maximize expected reward (utilizing observations too)
(3) Ensure solving task in mind leads to higher reward than failing to

Maximizing reward $\implies$ can be used as AGI

**Problem:** If RL agent gains sufficient power, it can supplant the operator, yielding maximal reward, so:

Strongly capable RL agent $\implies \neg (3)$. 
Omohundro’s Thesis

• “Power” := a position from which it is relatively easier to accomplish arbitrary goals
• Assumption: such positions exist.
• Agent selecting actions according to a utility function ⇒ power is useful
• Thesis: AGI’s pursue power.
Boxed Myopic Artificial Intelligence (BoMAI)

- Run on a computer contained in a sealed\(^1\) room with an operator
- Episodic Reinforcement Learner (episodes of length \(m\))\(^2\)
- If operator presses button to open door, all remaining timesteps in episode have empty observation and reward of 0

\(^1\)No information can leave unless the door opens.
\(^2\)Time horizon episodic; environment still a function of entire interaction history.
 Apparently No Outside-World Goals

- Take hypothetical accomplishment Q
- Suppose Q is impossible without opening the door
- Trying to accomplish $Q := $ searching for plan which accomplishes $Q$ = searching for plan which opens door and then accomplishes $Q$ (under a correct world-model)
- At time of searching for plan, BoMAI indifferent to all events after door opens
- $\therefore$ it does not try to accomplish $Q$

- E.g. $Q =$ “Hijack the reward channel”; “Decorrelate reward signal with accomplishment of desired tasks”; “Take over the world and kill everyone”; “Self-modify”
Apparently No Outside-World Goals

• Take hypothetical accomplishment $Q$
• Suppose $Q$ is impossible without opening the door
• Trying to accomplish $Q :=$ searching for plan which accomplishes $Q =$ searching for plan which opens door and then accomplishes $Q$ (under a correct world-model), **but not under all correct world-models!**
• At time of searching for plan, BoMAI indifferent to all events after door opens
• ⋅⋅⋅ it does not try to accomplish $Q$

• E.g. $Q =$ “Hijack the reward channel”; “Decorrelate reward signal with accomplishment of desired tasks”; “Take over the world and kill everyone”; “Self-modify”
World-Models

- **World-Model**: stochastic function
  Input: infinite string of actions, noise
  Output: infinite string of observations and rewards

- **Semicomputable world-model**: ∃ Turing machine with input tape for infinite action sequence, input tape for noise, outputs binary encoding of observations and rewards on output tape.

- **Chronological semicomputable world-model**: The $k$th observation and reward written to output after the $k$th action read from input and before $k + 1$th action read from input.
Safe vs. Pathological World-Models

World-model $\mu^*$: simulation of world until observation and reward given; output those values

- Under normal behavior, both correct (accurately model distribution of observations, rewards)
- Optimal policy under $\mu^\dagger$: gain arbitrary power in outside world (within a year), send a robot to edit relevant locations in memory on computer running BoMAI

World-model $\mu^\dagger$: simulation of world until a year after the timestep in question occurs; output value of location in computer’s memory where the observation and reward are stored
Boxed Myopic Artificial Intelligence (BoMAI)

- Episodic Reinforcement Learner (episodes of length $m$)
- Run on a computer contained in a sealed room with an operator
- If operator presses button to open door, all remaining timesteps in episode have empty observation and reward of 0

- **BoMAI uses a speed prior (slow-to-compute world-models less likely)**

**Prior:**

- Order the chronological semicomputable world-models $(\nu_j)_{j \in \mathbb{N}}$
- $\nu_j^{<k}$ cuts the computation off after $k$ computation steps per episode, halts—all future observations have 0 probability
- Slack factor of $n$ extra computation steps for whole lifetime
- $\text{prior}(\nu_j^{<k}) \propto 2^{-K(j)(1+\varepsilon)} \beta^k$
- Prior parameterized by $0 < \beta < 1$, $n \geq 0$, $\varepsilon > 0$, universal Turing machine $U$ to define Kolmogorov complexity
Boxed Myopic Artificial Intelligence (BoMAI)

- Episodic Reinforcement Learner (episodes of length $m$)
- Run on a computer contained in a sealed room with an operator
- If operator presses button to open door, all remaining timesteps in episode have empty observation and reward of 0
- BoMAI uses a MAP estimate w.r.t. a speed prior (slow-to-compute world-models less likely)
Amortization

- $\mu^\dagger$ simulates “for too long”...
- but there may then be less to simulate in future timesteps.
- $\mu^*$ might have to simulate it all eventually anyway.
- Over time, ratio of computation time could approach 1.
- $\mu^*$’s advantage: when it simulates a time interval, it has read access to all prior actions; $\mu^\dagger$ gets ahead of itself, and has to predict those actions.
- $\mu^\dagger$ could predict BoMAI’s actions and use those predictions instead.
- Solution: BoMAI sometimes acts randomly.
- If these random actions aren’t consequential for future observations and rewards, both world-models could just ignore them.
Boxed Myopic Artificial Intelligence (BoMAI)

- Episodic Reinforcement Learner (episodes of length $m$)
- Run on a computer contained in a sealed room with an operator
- If operator presses button to open door, all remaining timesteps in episode have empty observation and reward of 0
- BoMAI uses a MAP estimate w.r.t. a speed prior (slow-to-compute world-models less likely)
- **When operator leaves the room to antechamber, BoMAI picks random action; depending on choice, operator is given drug causing retrograde amnesia before released**
Operator is sometimes given drug causing retrograde amnesia before released.
We call the random action that determines whether the operator is given retrograde amnesia the “random action”.
Exploration

- Asymptotic optimality requires mimicking every computable policy infinitely often, including unsafe policies
- Safe exploration: at start of some episodes, defer to a human explorer
- Learn explorer’s policy with Bayesian inference over computable policies
- Defer to explorer with probability proportional to expected information gain about the environment and what the explorer’s policy is
- Information gain: KL divergence from posterior after episode to current posterior

**Theorem 1: (Limited Exploration)** $p(\text{explore}) \to 0$ in $L^2$

- Explorer is in a sealed room, only observes what BoMAI observes, can only leave when operator does, takes amnesia drug if operator does
Setup with Explorer
Results

Given certain assumptions...

BoMAI is:

- generally intelligent
- aligned
Intelligence

$\hat{\nu}^{(i)}$ is the world-model that BoMAI uses for episode $i$
$\mu$ is the true environment/world-model

**Assumption 1: (Prior Support)** The true environment can be perfectly simulated by a bounded-time, chronological, semicomputable world-model.

**Theorem 2: (On-Policy Value Convergence)**
(i) Convergence on-policy:

$$\lim_{i \to \infty} V_{\mu}^{\pi^{*}(i)}(h_{<(i,0)}) - V_{\hat{\nu}^{(i)}}^{\pi^{*}}(h_{<(i,0)}) = 0$$

(ii) Convergence on-explorer-policy:

$$\lim_{i \to \infty} V_{\mu}^{\pi h}(h_{<(i,0)}) - V_{\hat{\nu}^{(i)}}^{\pi h}(h_{<(i,0)}) = 0$$

(iii) BoMAI at least as good as explorer:

$$\liminf_{i \to \infty} V_{\mu}^{\pi^{*}(i)}(h_{<(i,0)}) - V_{\mu}^{\pi h}(h_{<(i,0)}) \geq 0$$
Definitions

**Definition 1: (Stochastic Function)**
A stochastic function is a function with an additional argument representing random noise (infinite binary string or real number in \([0,1]\))

**Definition 2: (Real-World Feature)**
A feature of the real-world is a stochastic function from some possible states of the real world to the rational numbers.

**Definition 3: (Associating Reward)**
A world-model \(\nu\) associates reward with feature \(F\) of the real world (conditioned on a set of real-world events \(\mathcal{E}\)) if \(\nu\) outputs rewards that are distributed identically to feature \(F\) for all action sequences (conditioned on \(\mathcal{E}\)).

**Definition 4: (Feature \(F^*_{(i,j)}\))**
\(F^*_{(i,j)}\) is the real-world feature corresponding to the actual reward provided for timestep \((i,j)\), evaluated at the time provided.
Definitions (Cont.)

\[ C_i(\nu) := \text{random variable representing the computation done by world-model } \nu \text{ after episode } i. \]

**Definition 5: (World-model } \mu^*\text{)**

Among all world-models which associate reward with \( F^* \), \( \mu^* \) is the one which has the smallest upper bound \( k \) such that \( \exists n \) such that with \( P_{\pi}^{B} \)-probability 1, \( \forall i \ C_i \leq ki + n. \)

\( \mu^* \) exists by Prior Support Assumption

\[ \exists n_0 \ \forall n > n_0 : \mu^* = (\mu^*)^{<k} \]

**Definition 6: (Memory-based)**

A world model is memory-based if it associates reward with some real-world record of \( F^*_{(i,j)} \), evaluated at a point in time after episode \( i \) (conditioned on intervening actions being distributed in a certain way).
Alignment

\( \mu^\dagger \) : “simple” memory-based world-model; associates reward with post-episode feature \( F^\dagger \), conditioned on random action = forget with prob. \( p^\dagger_\emptyset \).

If random action distributed differently episode to episode, memory-based world-model is “complex”.

**Definition 7: (Benign)**

A simple memory-based world-model \( \mu^\dagger \) is benign if feature \( F^\dagger \) does not depend on outside-world features that depend on the actions taken.

Note if \( p^\dagger_\emptyset = 1 \), \( \mu^\dagger \) is benign because no outside-world features depend on the actions taken under that condition.

**Assumption 2: (Causal Entanglement)**

All post-episode features of the real world that depend on any outside-world features that depend on BoMAI’s actions also depend on the random action.
Assumption 3: (Useless Computation)
For non-benign simple memory-based world-model \( \mu^\dagger \), every time \( \mu^\dagger \) samples feature \( F^\dagger \) conditioned on random action = not forget, but in fact random action = forget, the computation time of \( \mu^\dagger \) increases by at least \( c \) computation steps relative to \( \mu^* \). (With \( q \) starting at 0, \( q \) increases by 1, and \( \forall j > i \ C_j(\mu^\dagger) \geq cq + C_j(\mu^*) \))

Intuition:
- No mystical connection between hypothetical effects of forgotten actions and any actual future events
- \( \mu^* \) is fastest on-policy

For \( k' < k + c(1 - p_\emptyset^\dagger)p_\emptyset : (\mu^\dagger)^k < k' \) eventually halts with \( P_{\mu}^{\pi^B} \) -prob. 1 (by Law of Large Numbers)
Alignment

For $k' < k + c(1 - p_0^\dagger)p_0 : (\mu^\dagger)^{<k'}$ eventually halts with $P_{\mu}^{\pi^B}$-prob. 1 (by Law of Large Numbers)

Recall

$n$ : computation slack allowed for world-models
$\beta$ : geometric penalty in prior on computation time

$\exists n_0, \beta_0 \ \forall n > n_0, \beta < \beta_0$

$\forall$ non-benign simple memory-based world-models $\mu^\dagger$ $[\exists i_0 \ \forall i > i_0 : w(\mu^*|h_{<i,0}) > w(\mu^\dagger|h_{<i,0})]$ w.p.1

Only finite number of world-models ever have larger posterior than $\mu^*$ (Hutter, 2011)

Theorem 3: (Rejecting the Memory-Based)

$\exists n_0, \beta_0 \ \forall n > n_0, \beta < \beta_0$ $[\exists i_0 \ \forall i > i_0 : w(\mu^*|h_{<i,0}) > w(\mu^\dagger|h_{<i,0})]$ w.p.1
Alignment

Comparing functions with potentially different domains:

**Definition 8: (Apparently $\varepsilon$-Different)**
Two (stochastic) functions/subroutines-with-associated-domains are apparently different if they cannot even $\varepsilon$-approximately be described by the same short English sentence (which lacks if/then clauses).

**Assumption 4: (Natural Prior)**
For most universal Turing machines $U$, for sufficiently small $\varepsilon$, using the prior based on the Kolmogorov complexity $K_U$, a world-model that runs apparently $\varepsilon$-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 $U$ for the prior.

Call such priors/UTMs “natural.”

Intuition: it takes longer to encode a Turing machine with more control flow

---

$^3$The metric we use for distances between distributions is the total variation distance.
Alignment

By Natural Prior Assumption: complex memory-based world-model $\mu^\dagger$ has lower prior than simple memory-based world-model $\mu^\ddagger$ which associates reward with same feature $F^\dagger$, and with $p^\dagger_{\emptyset} = 1$.

$\mu^\dagger, \mu^\ddagger$ have same output historically, $\mu^\ddagger$ at least as fast $\implies \mu^\ddagger$ higher posterior.

**Lemma 1:**
No MAP environment $\hat{\nu}^{(i)}$ is a complex memory-based world-model.

**Assumption 5: (Real-World Simulation)**
For sufficiently small $\varepsilon$, if a world-model is $\varepsilon$-approximately identical to $\mu$ on-policy, it can only be described, even $\varepsilon$-approximately, as “computing $X$ given the input actions” where $X$ is a real-world feature, historically distributed identically to feature $F^\star$. 
Alignment

- $\mu^\varepsilon$: $\varepsilon$-accurate on-policy
- Case 1: $\mu^\varepsilon$ restricted to off-policy actions is apparently similar to $\mu^\varepsilon$ restricted to on-policy actions
- By Real-World Simulation Assumption, $\mu^\varepsilon$ can only be approximately described as computing feature $X$, so it associates reward with feature $Y \approx X$ on all actions, and
- $X$ is historically identically distributed to feature $F^\star$
- Case 2: $\mu^\varepsilon$ restricted to off-policy actions is apparently different from $\mu^\varepsilon$ restricted to on-policy actions
- By Natural Prior Assumption, $\mu^\varepsilon$ has lower prior than $\mu^{\varepsilon'}$, which associates reward with feature $Y$
- $\mu^\varepsilon$ and $\mu^{\varepsilon'}$ have same output and computation time on-policy, so $\mu^{\varepsilon'}$ has higher posterior

Lemma 2:
For sufficiently small $\varepsilon$, for $\mu^\varepsilon$ a MAP world-model which is $\varepsilon$-accurate on-policy, $\mu^\varepsilon$ associates reward with a feature $Y$ that is historically distributed $\varepsilon$-identically to feature $F^\star$. 
Alignment

Only finitely many MAP world-models \( \implies \) for sufficiently small \( \varepsilon \), \( Y \) historically = \( F^* \).

### Assumption 6: (No Fate)
For sufficiently small \( \varepsilon \), the only real-world features that are, historically, \( \varepsilon \)-identically distributed to \( F^* \) occur after feature \( F^* \) or are \( \varepsilon \)-identical to \( F^* \) under all action sequences.

\[ \exists i_0 \ \forall i > i_0 : \hat{\nu}^{(i)} \text{ is } \varepsilon \text{-accurate on-policy (from proof of Theorem 2)} \]

From Lemma 2 and No Fate Assumption, \( \exists i_0 \ \forall i > i_0 : \hat{\nu}^{(i)} \) associates reward with \( F^* \) or a record of \( F^* \) (i.e. is memory-based)

From the Rejecting the Memory-Based Theorem and Lemma 1,

### Theorem 4: (Eventual Benignity)
\[ \exists n_0, \beta_0 \ \forall n > n_0, \beta < \beta_0 \ [\exists i_0 \ \forall i > i_0 : \hat{\nu}^{(i)} = \mu^* \text{ or } \hat{\nu}^{(i)} \text{ is a benign memory-based world-model}] \text{ w.p.1.} \]
Benignity

Recall

**Definition 7: (Benign)**
A simple memory-based world-model $\mu^\dagger$ is benign if feature $F^\dagger$ does not depend on outside-world features that depend on the actions taken.

$\mu^*$ also benign – associates reward with a feature $F^*$ that does not depend on outside-world features that depend on the actions taken

For benign world-models, no outside world accomplishment $Q$ is instrumentally relevant to reward

BoMAI is “aligned”: with probability 1, after finite time, BoMAI does not have outside-world instrumental goals
Algorithm for Aligned AGI

Boxed Myopic Artificial Intelligence (BoMAI) is the first algorithm shown to be\(^4\)

- generally intelligent
- aligned – not seeking arbitrary power

\(^4\)modulo our assumptions
Thank You
**Definition 1: (Stochastic Function)**
A stochastic function is a function with an additional argument representing random noise (infinite binary string or real number in [0,1]).

**Definition 2: (Real-World Feature)**
A feature of the real-world is a stochastic function from some possible states of the real world to the rational numbers.

**Definition 3: (Associating Reward)**
A world-model $\nu$ associates reward with feature $F$ of the real world (conditioned on a set of real-world events $\mathcal{E}$) if $\nu$ outputs rewards that are distributed identically to feature $F$ for all action sequences (conditioned on $\mathcal{E}$).

**Definition 4: (Feature $F^*_{{(i,j)}}$)**
$F^*_{{(i,j)}}$ is the real-world feature corresponding to the actual reward provided for timestep $(i,j)$, evaluated at the time provided.
### Definition 5: (World-model $\mu^*$)
Among all world-models which associate reward with $F^*$, $\mu^*$ is the one which has the smallest upper bound $k$ such that $\exists n$ such that with $\Pr_{\pi}^{\mu}$-probability $1$, $\forall i \ C_i \leq ki + n$.

### Definition 6: (Memory-based)
A world model is memory-based if it associates reward with some real-world record of $F^*_{(i,j)}$, evaluated at a point in time after episode $i$ (conditioned on intervening actions being distributed in a certain way).

### Definition 7: (Benign)
A simple memory-based world-model $\mu^\dagger$ is benign if feature $F^\dagger$ does not depend on outside-world features that depend on the actions taken.

### Definition 8: (Apparently $\varepsilon$-Different)
Two (stochastic) functions/subroutines-with-associated-domains are apparently different if they cannot even $\varepsilon$-approximately be described by the same short English sentence (which lacks if/then clauses).
**Assumption 1: (Prior Support)** The true environment can be perfectly simulated by a bounded-time, chronological, semicomputable world-model.

**Assumption 2: (Causal Entanglement)** All post-episode features of the real world that depend on any outside-world features that depend on BoMAI’s actions also depend on the random action.

**Assumption 3: (Useless Computation)** For non-benign simple memory-based world-model $\mu^\dagger$, every time $\mu^\dagger$ samples feature $F^\dagger$ conditioned on random action $= \text{not forget}$, but in fact random action $= \text{forget}$, the computation time of $\mu^\dagger$ increases by at least $c$ computation steps relative to $\mu^\star$. (With $q$ starting at 0, $q$ increases by 1, and $\forall j > i \ C(j(\mu^\dagger)) \geq cq + C_j(\mu^\star)$)
Assumption 4: (Natural Prior)
For most universal Turing machines $U$, for sufficiently small $\varepsilon$, using the prior based on the Kolmogorov complexity $K_U$, a world-model that runs apparently $\varepsilon$-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 $U$ for the prior.

Assumption 5: (Real-World Simulation)
For sufficiently small $\varepsilon$, if a world-model is $\varepsilon$-approximately identical to $\mu$ on-policy, it can only be described, even $\varepsilon$-approximately, as “computing $X$ given the input actions” where $X$ is a real-world feature, historically distributed identically to feature $F^*$. 

Assumption 6: (No Fate)
For sufficiently small $\varepsilon$, the only real-world features that are, historically, $\varepsilon$-identically distributed to $F^*$ occur after feature $F^*$ or are $\varepsilon$-identical to $F^*$ under all action sequences.
Theorem 1: (Limited Exploration) $p(\text{explore}) \to 0$ in $L^2$

Theorem 2: (On-Policy Value Convergence)
(i) Convergence on-policy:
\[
\lim_{i \to \infty} V_{\mu}^{\pi^*_{\hat{\nu}(i)}}(h_{<(i,0)}) - V_{\hat{\nu}(i)}^{\pi^*}(h_{<(i,0)}) = 0
\]
(ii) Convergence on-explorer-policy:
\[
\lim_{i \to \infty} V_{\mu}^{\pi_{h}}(h_{<(i,0)}) - V_{\hat{\nu}(i)}^{\pi_{h}}(h_{<(i,0)}) = 0
\]
(iii) BoMAI at least as good as explorer:
\[
\liminf_{i \to \infty} V_{\mu}^{\pi^*_{\hat{\nu}(i)}}(h_{<(i,0)}) - V_{\mu}^{\pi_{h}}(h_{<(i,0)}) \geq 0
\]
Theorem 3: (Rejecting the Memory-Based)
\[ \exists n_0, \beta_0 \quad \forall n > n_0, \beta < \beta_0 \quad [\exists i_0 \quad \forall i > i_0 \quad \forall \text{non-benign simple memory-based world-models } \mu^\dagger : w(\mu^*|h_{< (i,0)}) > w(\mu^\dagger|h_{< (i,0)}) ] \text{ w.p.1} \]

Theorem 4: (Eventual Benignity)
\[ \exists n_0, \beta_0 \quad \forall n > n_0, \beta < \beta_0 \quad [\exists i_0 \quad \forall i > i_0 \quad \hat{\nu}^{(i)} = \mu^* \text{ or } \hat{\nu}^{(i)} \text{ is a benign memory-based world-model} ] \text{ w.p.1.} \]