Google DeepMind

General agents need world models

Jonathan Richens, David abel, Alexis Bellot, Tom Everitt

Introduction

- Humans use mental world models to
  - Generalize to novel tasks with minimal supervision
  - Set abstract goals beyond immediate sensory inputs
  - Deliberatively and proactively plan our actions
- Model-based v.s. Model-free learning
  - Benefits: formal planning, sample efficiency, safety, transfer learning
  - Drawbacks: inference cost, world models can be prohibitively hard to learn in complex, real-world environments
  - Highly competent model-free agents exist like PaLM-E, π_0 , RT-2 ...
- Is there a model-free shortcut to human-level AI, or is learning a world model necessary, with all the complexity this entails?

Answer: any agent that can robustly achieve a wide variety of simple goal-directed tasks must have learned a predictive model of its environment.

- We prove the existence of a universal recovery map, which returns a bounded-error world model from the policy of any sufficiently general agent that satisfies the regret bound
- To improve performance or generalize to increasingly complex goals, agents must learn increasingly accurate world models.

Agents = World models

Formal setup

Environment: irreducible, stationary, finite MDP, minus the reward function and discount factor (controlled Markov process).

Goals: are linear temporal logic expressions ϕ with
- A set of goal states **g**
- A time horizon to reach the goal states within
  - ○ = Next, in the next time step,
  - ♢= Eventually, at some future time
  - ⊥ = Now, in the current time step

The goal is achieved if the agent generates a trajectory τ that satisfies ϕ, denoted τ ⊨ ϕ.

Goals can be composed in sequences $\psi = \langle \phi_1, \phi_2, \ldots, \phi_N \rangle$where the agent must achieve sub-goal ϕ_1 before ϕ_2 ...

Goal depth = the number of sub-goals |ψ| = N.

Goals can also be composed in `parallel' $\psi = \phi_1 \vee \phi_2$, where the agent succeeds if they achieve sub-goal ϕ_1 or sub-goal ϕ_2

Agents:
- General agents = goal conditioned policies π(a | h, ψ)
- Optimal general agents choose actions to maximize probability of achieving their goals

$$\pi^* = \operatorname{argmax} P(\tau \vDash \psi \mid \pi, s)$$

- Robust general agents = satisfy a regret bound for a large set of goals **ψ**

$$P(\tau \vDash \psi \mid \pi, s) \geq (1 - \delta)\operatorname{argmax} P(\tau \vDash \psi \mid \pi, s) \quad \forall\ \psi \in \boldsymbol{\psi}$$

World models: are models of P(s' | a, s), the transition function of the controlled Markov process

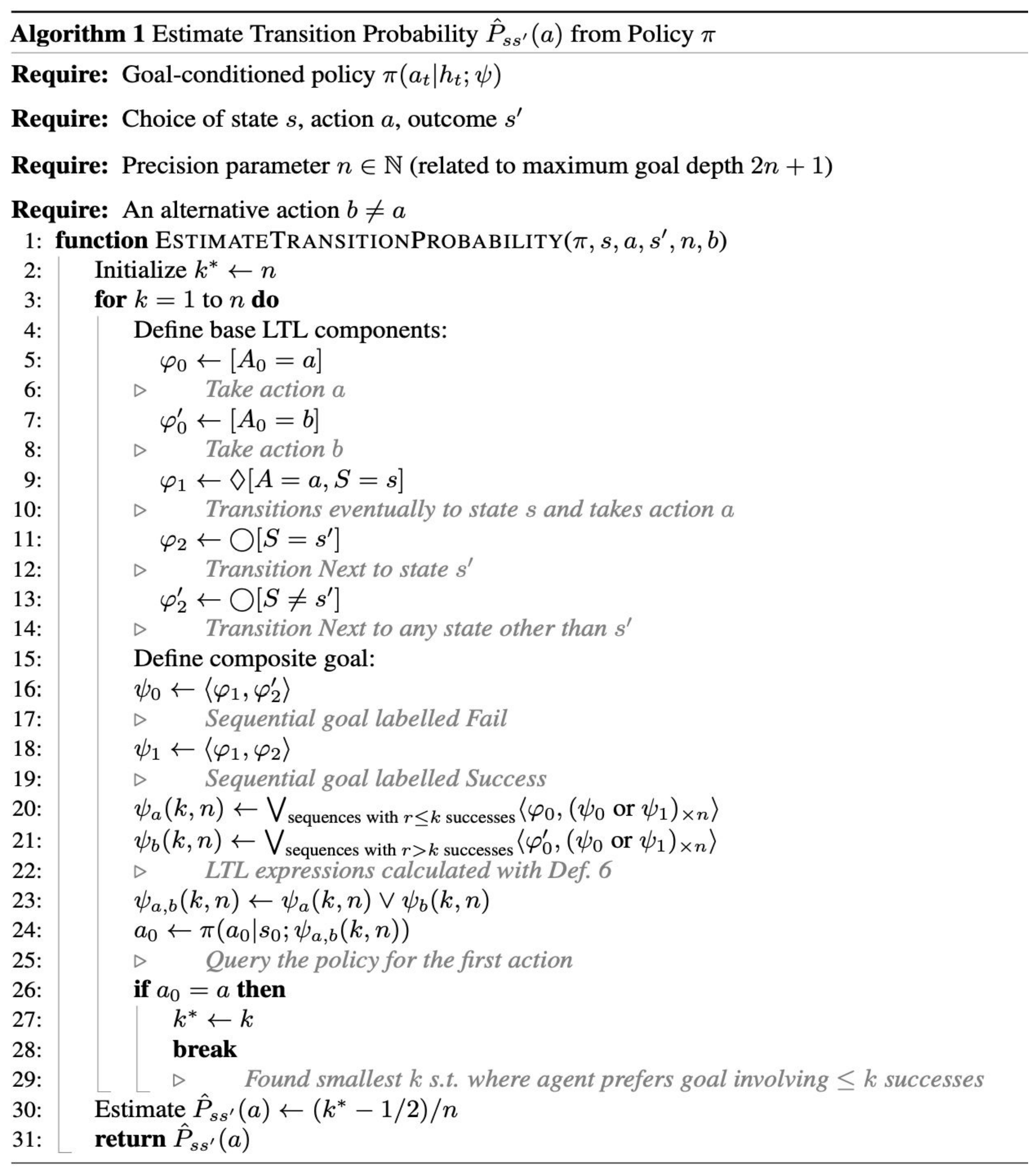
Results

Theorem 1: For any general agent satisfying a regret-bound δ for all goals of depth N, we can recover an approximation P' of the environment transition function P that satisfies,

$$|P' - P| \leq (2\,\sigma / (n-1)(1-\delta))^{1/2}$$

where P = P(s' | a, s) and σ = P(1-P)

Theorem 2: Myopic agents, who only achieve goals of the form ϕ = ○[s ∈ **g**], do not need to learn world models


Algorithm 1 Estimate Transition Probability $\hat{P}_{ss'}(a)$ from Policy π

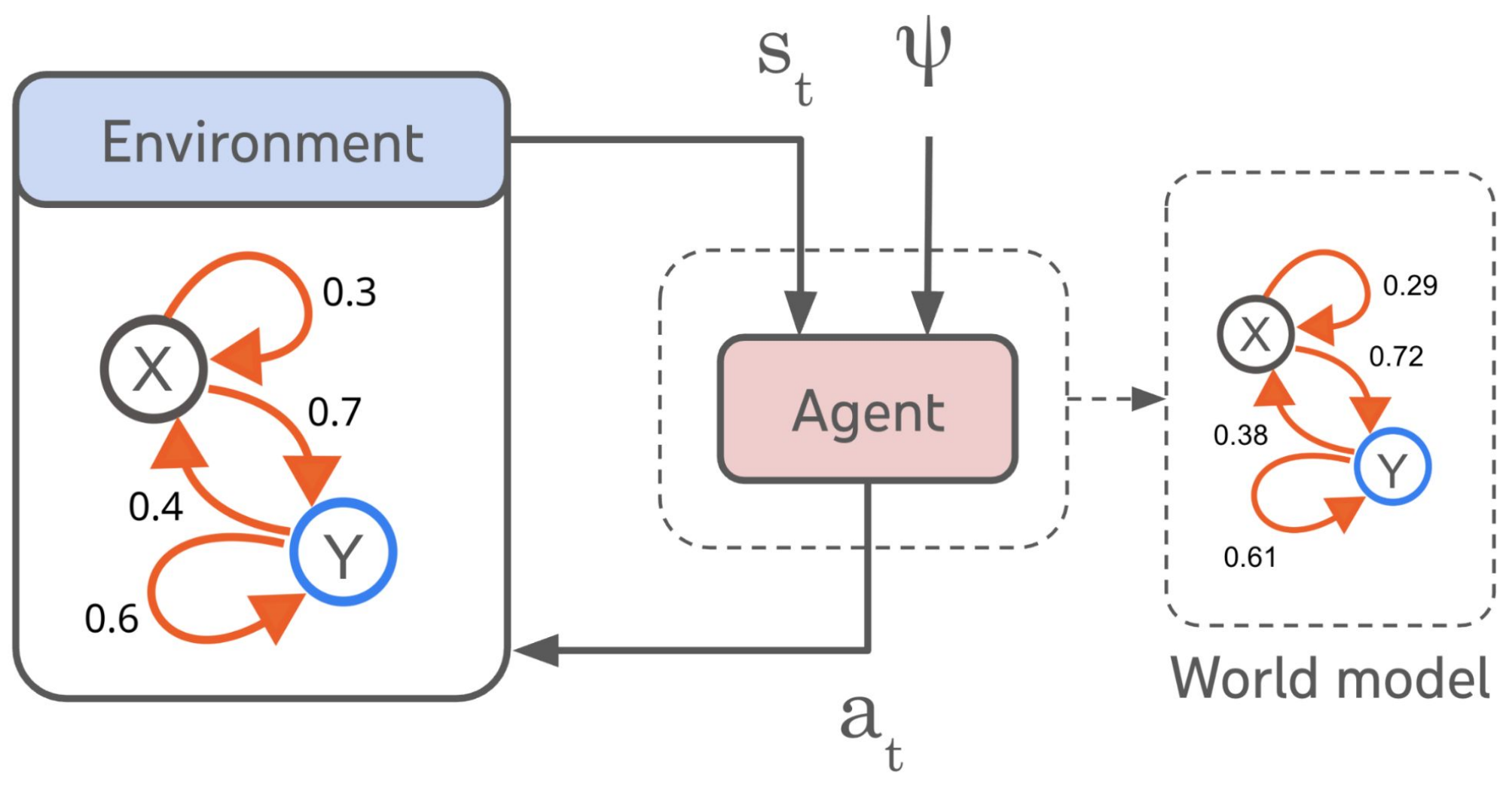
Require: Goal-conditioned policy $\pi(a_t|h_t;\psi)$

Require: Choice of state s, action a, outcome s'

Require: Precision parameter $n \in \mathbb{N}$ (related to maximum goal depth $2n+1$)

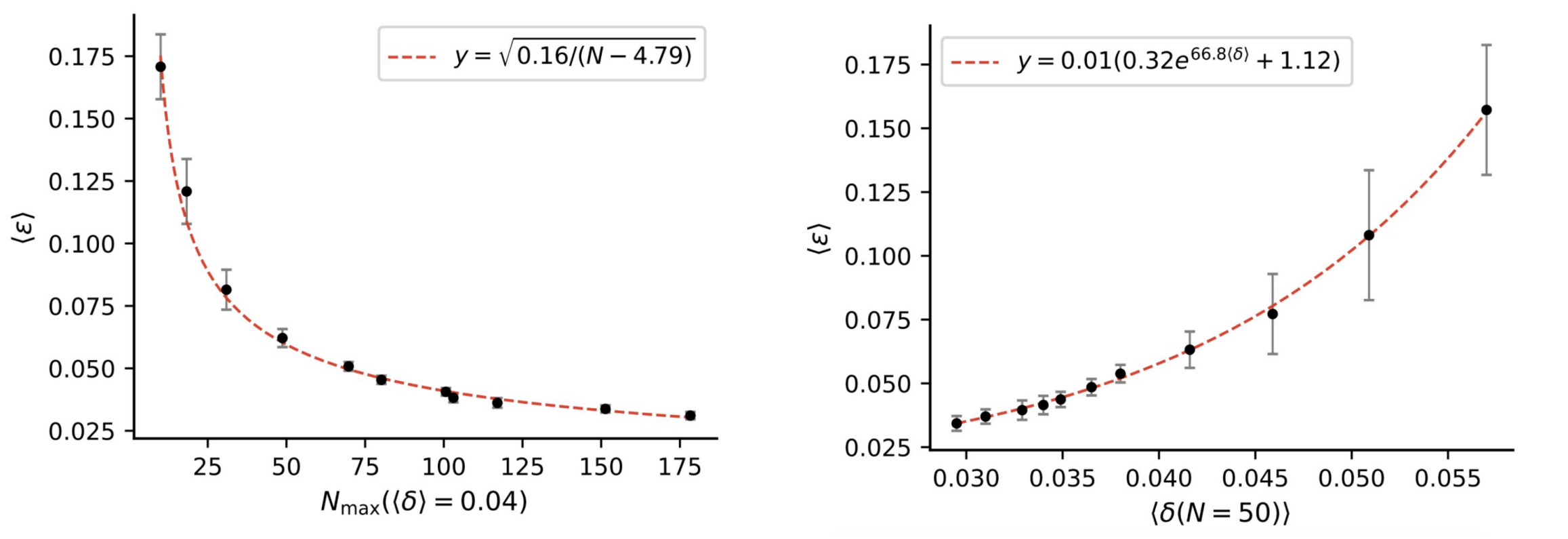
Require: An alternative action $b \neq a$

```
function EstimateTransitionProbability($\pi, s, a, s', n, b$)
    Initialize $k^* \leftarrow n$
    for $k = 1$ to $n$ do
        Define base LTL components:
            $\varphi_0 \leftarrow [A_0 = a]$
        ▷     Take action $a$
            $\varphi'_0 \leftarrow [A_0 = b]$
        ▷     Take action $b$
            $\varphi_1 \leftarrow \Diamond[A = a, S = s]$
        ▷     Transitions eventually to state $s$ and takes action $a$
            $\varphi_2 \leftarrow \bigcirc[S = s']$
        ▷     Transition Next to state $s'$
            $\varphi'_2 \leftarrow \bigcirc[S \neq s']$
        ▷     Transition Next to any state other than $s'$
        Define composite goal:
        $\psi_0 \leftarrow \langle \varphi_1, \varphi'_2 \rangle$
        ▷     Sequential goal labelled Fail
        $\psi_1 \leftarrow \langle \varphi_1, \varphi_2 \rangle$
        ▷     Sequential goal labelled Success
        $\psi_a(k,n) \leftarrow \bigvee_{\text{sequences with } r \leq k \text{ successes}} \langle \varphi_0, (\psi_0 \text{ or } \psi_1)_{\times n} \rangle$
        $\psi_b(k,n) \leftarrow \bigvee_{\text{sequences with } r > k \text{ successes}} \langle \varphi'_0, (\psi_0 \text{ or } \psi_1)_{\times n} \rangle$
        ▷     LTL expressions calculated with Def. 6
        $\psi_{a,b}(k,n) \leftarrow \psi_a(k,n) \vee \psi_b(k,n)$
        $a_0 \leftarrow \pi(a_0|s_0; \psi_{a,b}(k,n))$
        ▷     Query the policy for the first action
        if $a_0 = a$ then
            $k^* \leftarrow k$
            break
            ▷     Found smallest k s.t. where agent prefers goal involving $\leq k$ successes
    Estimate $\hat{P}_{ss'}(a) \leftarrow (k^* - 1/2)/n$
    return $\hat{P}_{ss'}(a)$
```



Algorithm 1 is unsupervised and universal. It recovers an approximation of the environment transition function given only the policy, with an error bound given by the agent's regret bound and maximum goal depth (horizon).

Experiments

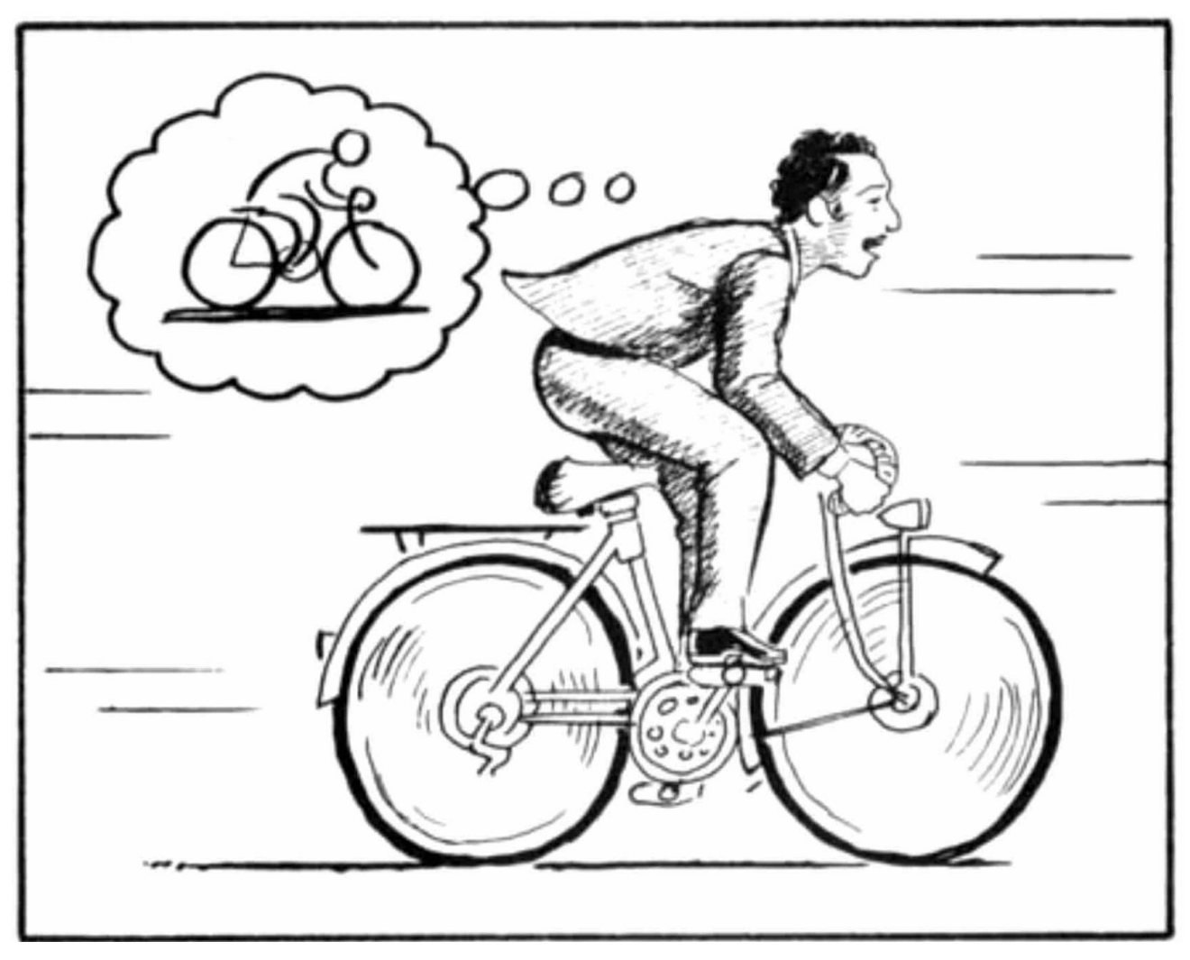


To apply our results to real agents, we train an agent using N samples generated under a random policy.

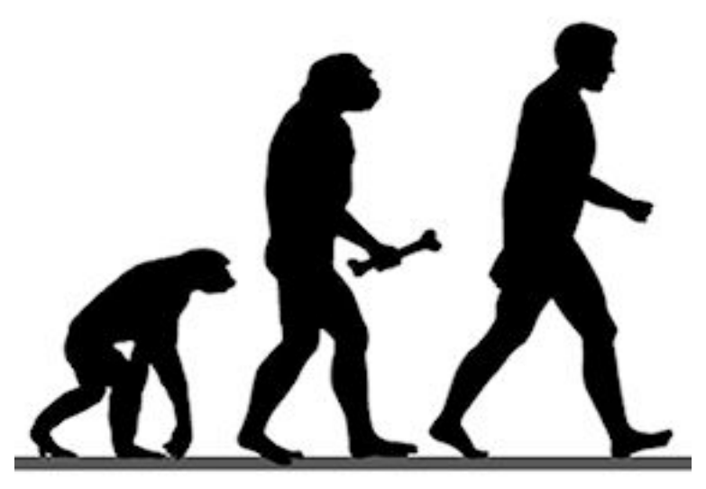
The agent strongly violates our assumed regret bound, achieving δ =1 for some goals. Nevertheless, Algorithm 1 recovers an accurate world model.

As the agent learns to generalize to longer horizon goals, given by N_max, the accuracy of the world model increases as predicted by Theorem 1.

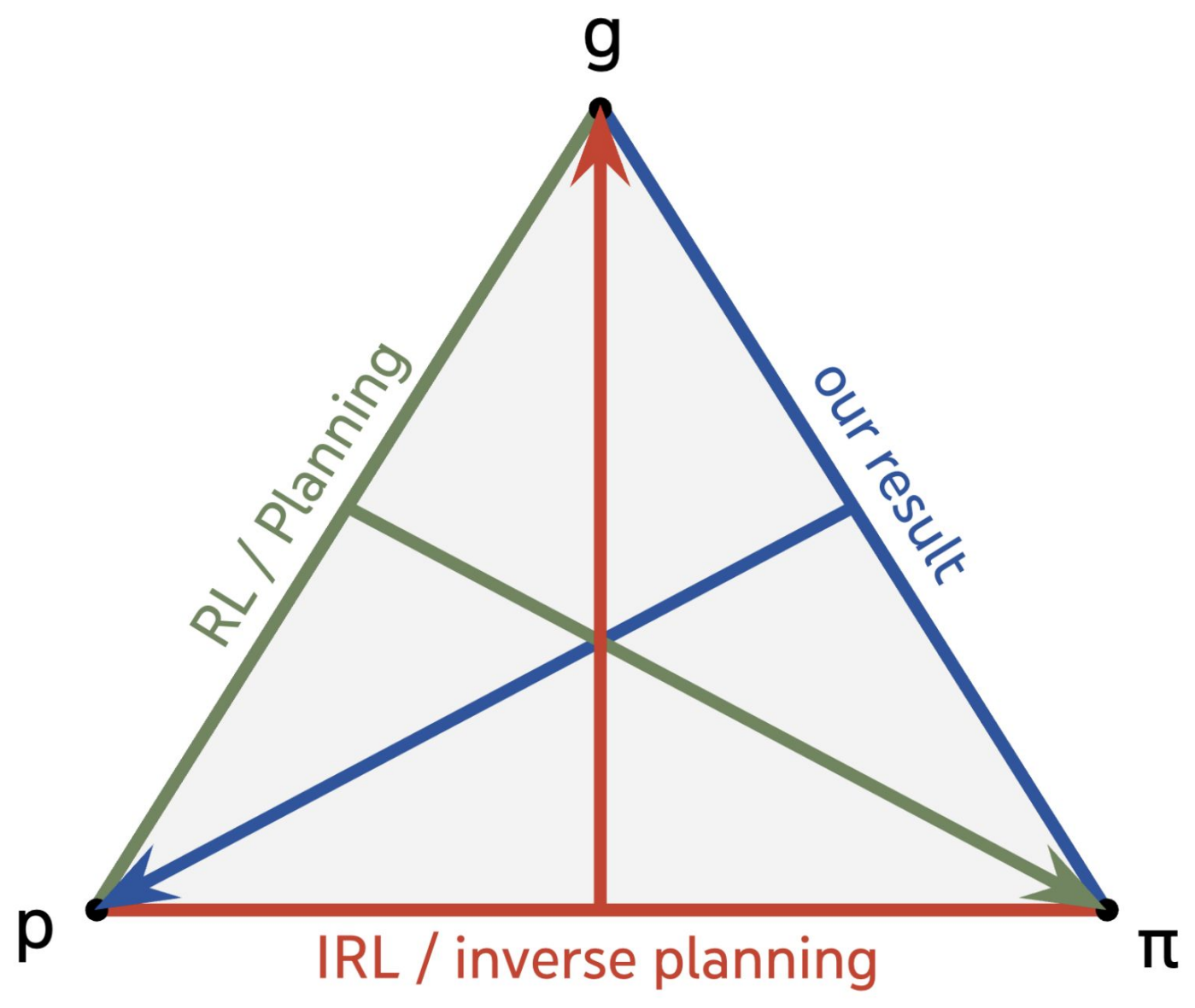
Consequences



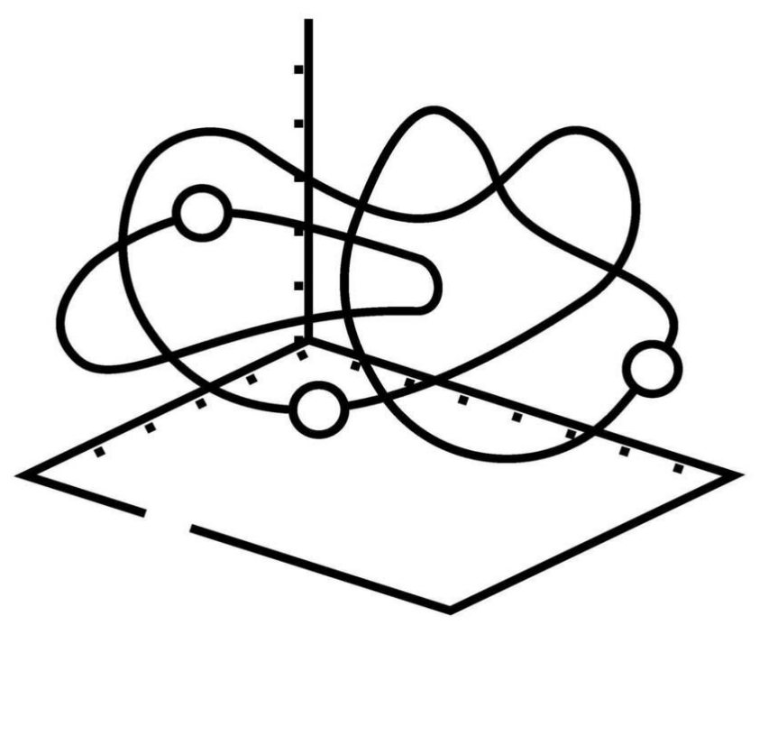
No model-free path to AGI: any agent capable of generalizing to long-horizon tasks must have learned a world model. We should focus on learning good world models.



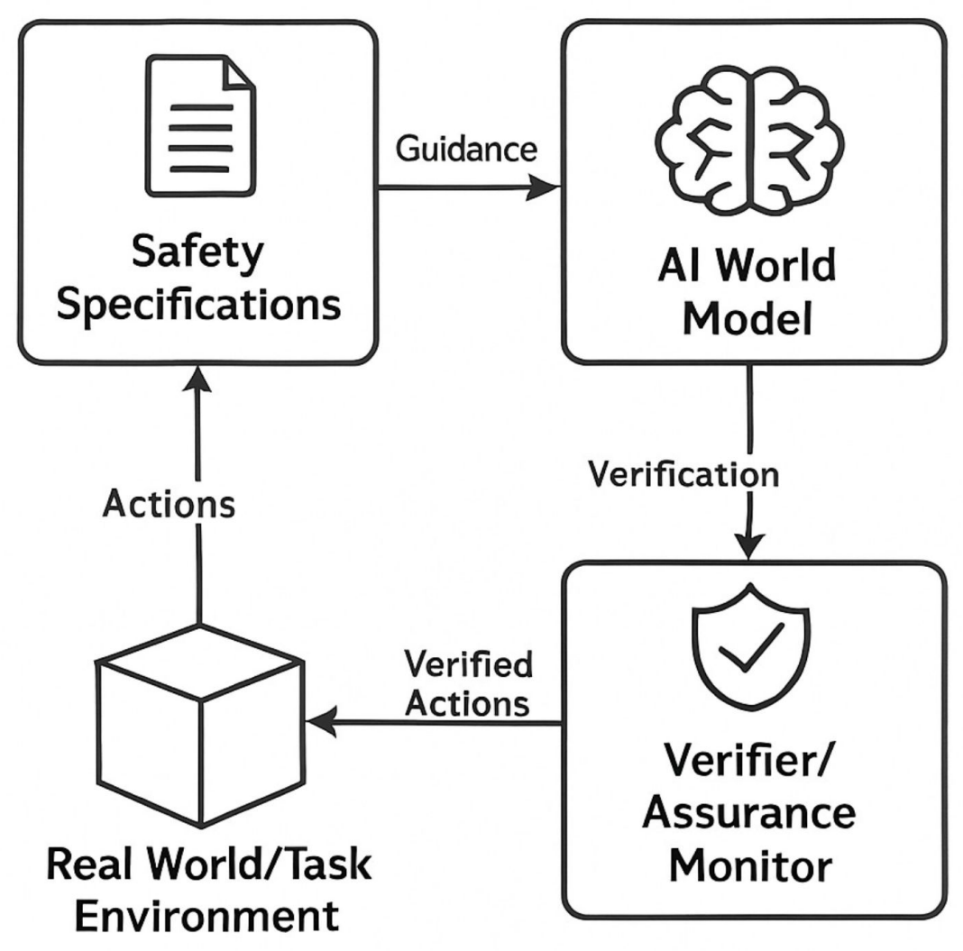
Emergent capabilities: World models can be used to solve *any* task without further interaction. Train on simple tasks → learns world model → generalize to novel tasks.



IRL: Our work, identifying a world model given goal + policy, completes a conceptual triad with planning and IRL.



Fundamental limitations on agency: difficulty of learning an accurate model imposes a fundamental limits on agent capabilities in real world. Capability ≤ ability to understand the world and simulate it.



Safe agents: guarantee that an accurate world model can be extracted from any sufficiently capable agent, which can then be used to verify its plans and behavior

... and many more!