Summary of *General Agents Contain World Models*

J. Richens, D. Abel, A. Bellot and T. Everitt, ICML 2025

This is the follow-up to *Robust Agents Learn Causal World Models*. That paper asked what an agent must know to stay near-optimal when the *environment* shifts; this one asks what it must know to pursue many different *goals* in a fixed environment. The answer is again that competence forces a model, but the model is now a predictive (transition) model rather than a causal one, and its accuracy is forced up by the *depth* of the goals the agent can achieve.

Basic setup

The environment is a *controlled Markov process* (cMP): an MDP with no reward and no discount, given by a tuple $(\boldsymbol{S}, \boldsymbol{A}, P)$ with finite state space $\boldsymbol{S}$, action space $\boldsymbol{A}$, and transition function

$$P_{ss'}(a) = P(S_{t+1} = s' \mid S_t = s,\ A_t = a).$$

Assumption 1. The cMP is finite, stationary (transition probabilities do not change over time), *communicating* (every state is reachable from every other under some policy), has at least two actions, and is fully observed.

A *world model* is an approximation $\widehat{P}$ of the transition function with bounded error, $|\widehat{P}_{ss'}(a) - P_{ss'}(a)| \leq \varepsilon$ for all s, s', a. A one-step predictor of this kind is enough to predict the evolution of the environment under any policy, by rolling it forward.

An *agent* is a deterministic goal-conditioned policy $\pi(a_t \mid h_t; \psi)$: given the history h_t and a goal ψ, it outputs an action. The paper's question is whether an agent that can pursue a rich family of goals must contain a model of P.

Goals

Goals are built in three layers, from linear temporal logic.

Base goals. A base goal is $\varphi = \mathcal{O}\big([(s,a) \in \boldsymbol{g}]\big)$, where $\boldsymbol{g}$ is a set of goal state–action pairs and $\mathcal{O} \in \{\bigcirc, \Diamond, \top\}$ is a temporal operator: *next*, *eventually*, or *now*. For instance "eventually visit a state in $\boldsymbol{g}$".

Sequential goals. $\psi = \langle \varphi_1, \ldots, \varphi_n \rangle$ requires the sub-goals to be achieved in order, φ_i before φ_{i+1}. Its *depth* is $\mathrm{depth}(\psi) = n$: the number of steps the agent must chain together.

Composite goals. $\psi = \bigvee_{i=1}^{m} \psi_i$ is a disjunction of sequential goals; its depth is the largest depth among them. $\boldsymbol{\Psi}_n$ denotes all composite goals of depth at most n.

Write $\tau \models \psi$ when trajectory τ satisfies ψ; the success probability of a policy from s_0 is $P(\tau \models \psi \mid \pi, s_0)$. Two features of this goal language do all the work later: depth n measures how many steps ahead a task reaches, and a disjunction offers the agent *alternatives it must choose between*, which is what makes its preferences observable.

Agents and regret

An *optimal* goal-conditioned agent maximises success on every goal from every start state, $\pi^* = \arg\max_\pi P(\tau \models \psi \mid \pi, s_0)$ for all $\psi \in \boldsymbol{\Psi}$ and all s_0 with $P(s_0) > 0$.

The object of study is weaker. A *bounded goal-conditioned agent* with maximum goal depth n and failure rate $\delta \in [0,1]$ is any policy satisfying

$$P(\tau \models \psi \mid \pi, s_0) \;\geq\; (1-\delta)\max_{\pi'} P(\tau \models \psi \mid \pi', s_0) \qquad \text{for all } \psi \in \mathbf{\Psi}_n.$$

This is the counterpart of δ-optimality across domain shifts in the previous paper, with goals now playing the role that interventions played. Note the regret here is *multiplicative* (a fraction of the optimum) rather than additive, and that the bound must hold uniformly over *all* goals up to depth n. As before, the agent need not know the optimum; regret is an external measure.

Main theorems

Theorem 1 (world models from bounded regret). Let the environment satisfy Assumption 1 and let π be a bounded goal-conditioned agent with failure rate δ on all goals in $\mathbf{\Psi}_n$, with $n > 1$. Then π fully determines a model $\widehat{P}$ with

$$\left|\widehat{P}_{ss'}(a) - P_{ss'}(a)\right| \;\leq\; \sqrt{\frac{2\,P_{ss'}(a)\big(1-P_{ss'}(a)\big)}{(n-1)(1-\delta)}} \;\sim\; O\Big(\frac{\delta}{\sqrt{n}}\Big) + O\Big(\frac{1}{n}\Big) \quad (\delta \ll 1,\; n \gg 1).$$

$$\boxed{\text{bounded regret on goals of depth } n \implies \text{a world model accurate to } O(\delta/\sqrt{n}) + O(1/n).}$$

Three things to read off. The error vanishes as the agent improves ($\delta \to 0$) *or* as the goals it can achieve get deeper ($n \to \infty$); in particular any fixed $\delta < 1$ is overcome by large enough n, so even a mediocre agent that can chain long tasks must carry an accurate model. The bound is on absolute error, so rare transitions ($P_{ss'}(a) \ll 1$) may be poorly resolved in relative terms.

Theorem 2 (myopic agents are exempt). For an optimal agent restricted to depth $n = 1$ (goals demanding the goal state immediately), the only bound on the transition probabilities derivable from π^* is the trivial one, $\varepsilon = 1$, and this is tight. A myopic agent need not learn any world model.

So goal-directedness alone forces nothing; it is *depth*, the need to chain steps, that forces a model. The converse direction is the familiar one: given P, or a good enough $\widehat{P}$, goal-optimal policies can be computed by planning, so the paper's summary is that a world model cannot be avoided by any agent that generalises across deep goals.

Extraction from the policy

As in the previous paper, the mechanism turns the agent's *choices* into numbers, and the parallel is exact once seen. Offer the agent a composite goal $\psi_a \vee \psi_b$ whose two branches cannot both be pursued. An optimal agent commits to the branch with the larger maximal success probability; a δ-bounded agent commits to one within a factor $(1-\delta)$ of the best. Each such query is therefore a (slightly noisy) comparison of two success probabilities, read off from a single action.

Measuring one transition. To measure $p := P_{ss'}(a)$, pose a goal that makes the agent run n independent trials of the same sub-task: go to s, take a, observe whether the outcome is s', and return to s (possible because the cMP is communicating). The number of trials that land on s' is

$$K \sim \text{Binomial}(n, p).$$

Now compare two incompatible sub-goals parameterised by a threshold r: $\psi_{\leq r}$, "at most r of the n trials succeed", and $\psi_{\geq r+1}$, "at least $r+1$ succeed". Offered the choice, the agent prefers whichever is likelier. As r sweeps from 0 to n the preferred branch flips exactly once, at the threshold r_* where the two are balanced,

$$P(K \leq r_*) \approx P(K \geq r_* + 1) \quad \iff \quad r_* \text{ is the median of } \text{Binomial}(n, p),$$

and the binomial median lies within one of its mean np. Hence the observed switching threshold gives

$$p \approx \frac{r_*}{n} \qquad \text{up to } O(1/n).$$

Regret blurs the switch: a δ-bounded agent may take a branch that is merely within a factor $(1-\delta)$ of the best, so the flip is smeared over the window of r on which the two probabilities are that close. Binomial concentration, with standard deviation $\sqrt{np(1-p)}$, makes that window $O(\delta\sqrt{n})$ wide in r, i.e. $O(\delta/\sqrt{n})$ in p. The two effects add to exactly the bound of Theorem 1. This is why depth helps: n is both the depth of the goal and the number of trials, and more trials give a sharper binomial and a finer switch. It is the same role the *continuum* of mixing weights q played before; there resolution came from sweeping q continuously, here it comes from repeating a sub-goal n times, and a myopic agent ($n = 1$) has a single coin flip and no resolution at all, which is Theorem 2.

choice between branches $\to$ switching threshold $r_* \to$ binomial median $\to P_{ss'}(a) \to \widehat{P}$.

The resulting procedure (the paper's Algorithm 1) is unsupervised and universal: its only input is the policy π, and it works for every agent satisfying the regret bound and every environment satisfying Assumption 1.

Relation to the previous paper

The two results are the two halves of one picture. *Robust Agents* concerns *domain* generalisation: adapting to a large family of interventions forces a *causal* model (G, P). This paper concerns *task* generalisation: achieving a large family of goals in a fixed environment forces a *predictive* model $P_{ss'}(a)$. The authors draw a striking corollary from the pair: domain generalisation requires *strictly more* knowledge than task generalisation. A transition function is an observational object, and the causal structure of the environment, how it responds to being intervened on, is not identifiable from it. A task-general agent therefore need not know the causal graph; a domain-general one must.

Interpretation and consequences

The theorem does not say the agent stores an explicit transition table, only that a sufficiently capable agent's policy contains enough information for an external observer to *extract* one, with guaranteed accuracy. The paper draws four consequences. For *safety*, a checkable world model can be elicited from any sufficiently capable model-free agent, which many verification and safe-exploration proposals presuppose. For *capability bounds*, building a general agent is at least as hard as learning an accurate world model, so regret-bounded generality is confined to domains whose dynamics are learnable. For *emergence*, it offers a mechanism for implicit world models in foundation models: minimising regret across diverse training tasks forces one. And it removes a motivation for model-free methods, since the model is learned either way.

Two limits are worth keeping in view for discussion. The result assumes a fully observed, stationary Markov environment, a deterministic policy, and a regret bound holding uniformly over *every* goal up to depth n; what an agent in a partially observed environment must learn about latent state is left open. And Theorem 2 marks the boundary sharply: the model is forced by multi-step structure, so an agent that only ever needs one-step goals is genuinely exempt.