

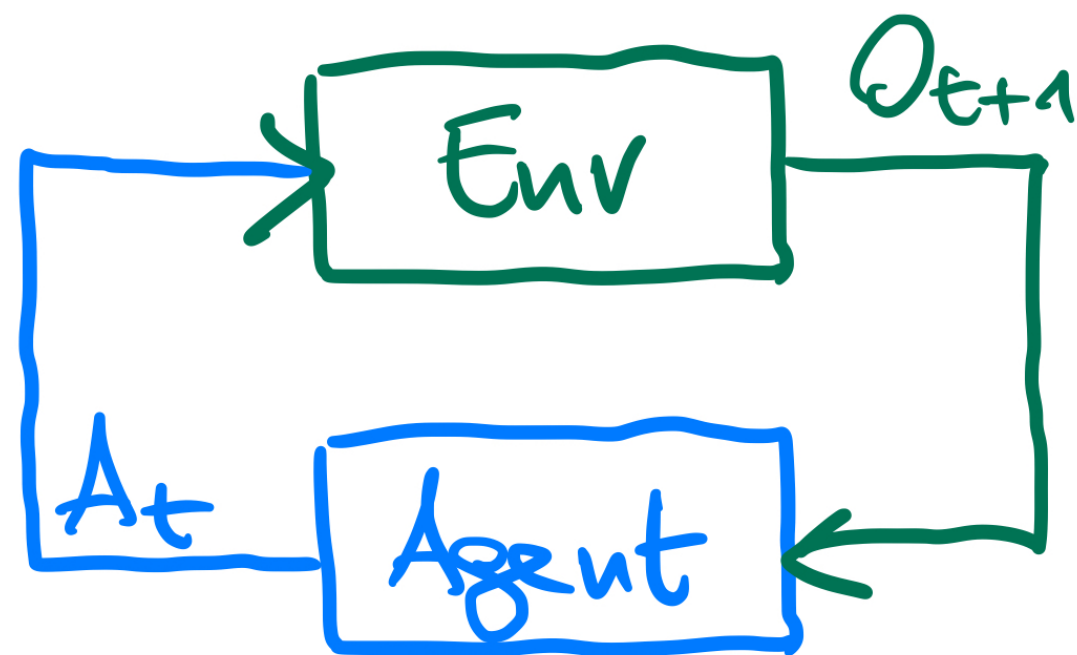
World models for RL, with triangles

Manuel Baltieri @ Iliad 3

04/08/2026

Background

- Agents are distinct from/interact with/are embedded in an environment
- Agents' beliefs/abstractions/world models update in some interesting way
- Agents have goals



Mathematical approaches to the study of agents

Manuel Baltieri^{1,2,†}, Keisuke Suzuki³

¹ Araya Inc., Tokyo, Japan

² University of Sussex, Brighton, UK

³ Center for Human Nature, Artificial Intelligence, and Neuroscience (CHAIN),
Hokkaido University, Sapporo, Japan

The definition of life remains one of science's most profound challenges, with contemporary approaches usually focusing on two research programmes: Darwinian evolution and self-maintenance in chemical systems. While evolution has been successfully abstracted and mathematically modelled, the concept of a self-sustaining system has so far resisted a comparable level of formalisation. This paper tackles this challenge by reframing the concept of self-sustaining system within a more abstract framework to study *agents*: goal-directed systems acting in an environment. We build on an existing conceptual framework comprising three requirements for agents: individuality, normativity (or goal-directedness), and interactional asymmetry. We then provide a systematic analysis, under a unified notation, of several mathematical approaches aiming to formalise these requirements, including the free energy principle, integrated information theory and dynamical systems. Unlike this conceptual framework, which commits to an intrinsic perspective on agency, we commit to a less ontologically committed *as-if* stance. Using this, we discuss links between identity and normativity, and a way to understand actions as if they were produced by causal interventions. Taken together, our systematic analysis clarifies the limitations of current proposals and reveals how they can work synergistically within a unified, mathematical account of agency across natural and artificial domains.

Keywords: agency, individuality, normativity, asymmetry

Plan + disclaimer

- 1. We have some theoretical results
- 2. ... here I try to show some applications
- 3. ... for which we need more theory
- 4. ... very much WIP (disclaimer)

World models of environment, agent and joint agent-environment systems

Manuel Baltieri^{1,2} Filippo Torresan¹ Yivan Zhang³
Alexander Boyd⁴ Fernando E. Rosas^{2,5,6,7}

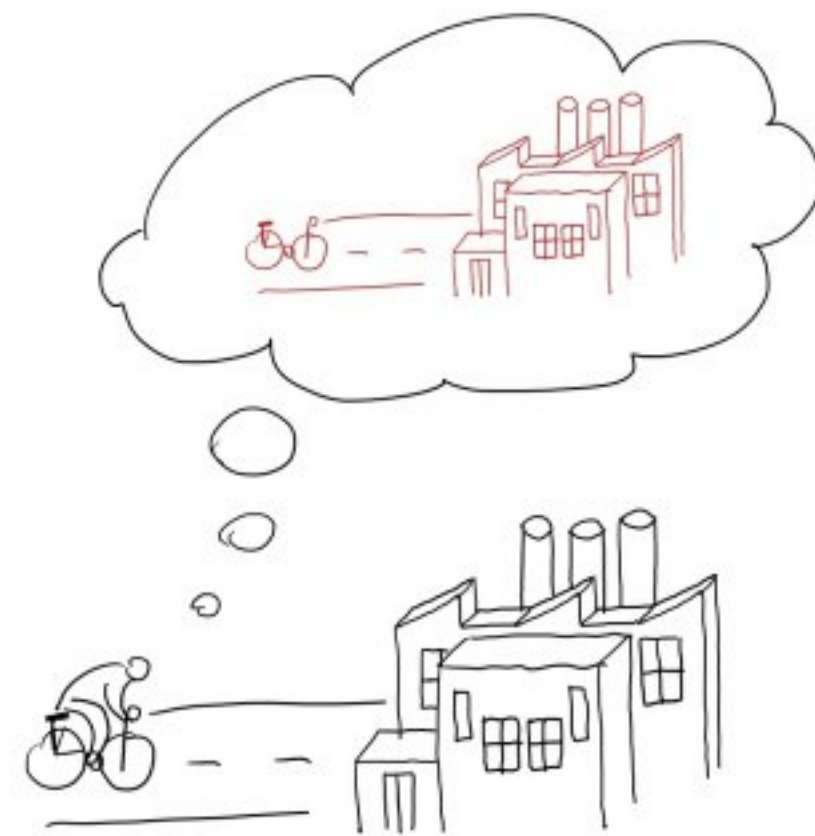
¹ Araya Inc.
² Department of Informatics, University of Sussex
³ The University of Tokyo
⁴ Beyond Institute for Theoretical Science (BITS)
⁵ Sussex AI and Sussex Centre for Consciousness Science, University of Sussex
⁶ Department of Brain Sciences, Imperial College London
⁷ Centre for Eudaimonia and Human Flourishing, University of Oxford

manuel.baltieri@araya.org

Abstract

World models are a central component of model-based reinforcement learning. They are usually discussed in terms of what variables they predict, such as observations, rewards, states, latent or information states. We argue that there is a prior distinction: which channel they model. We consider three cases: the environment channel $O: \mid A$., the agent channel $A: \mid O$., and the realised joint process (A, O) ., equivalently viewed as a channel with no inputs. Using computational mechanics, we define canonical predictive models for these three cases as ϵ -transducers or ϵ -machines. Canonical environment models recover standard predictive state representations, while the other two give analogous notions of canonical models for the agent and the joint system. We then build canonical support-restricted environment and agent models induced by closed-loop coupling, whose predictive equivalences range over continuations supported by the realised interaction. The key structural result is that canonical support-restricted environment states factor through the canonical joint causal states, and their transition structure is induced directly from the joint model; the agent-side construction is dual. Finally, we give a POMDP/controller example in which the unrestricted environment model has infinitely many states while the canonical support-restricted model induced by the coupling is finite. The framework clarifies what different world models are models of, and how coupling and support restriction can change their canonical predictive structure and complexity.

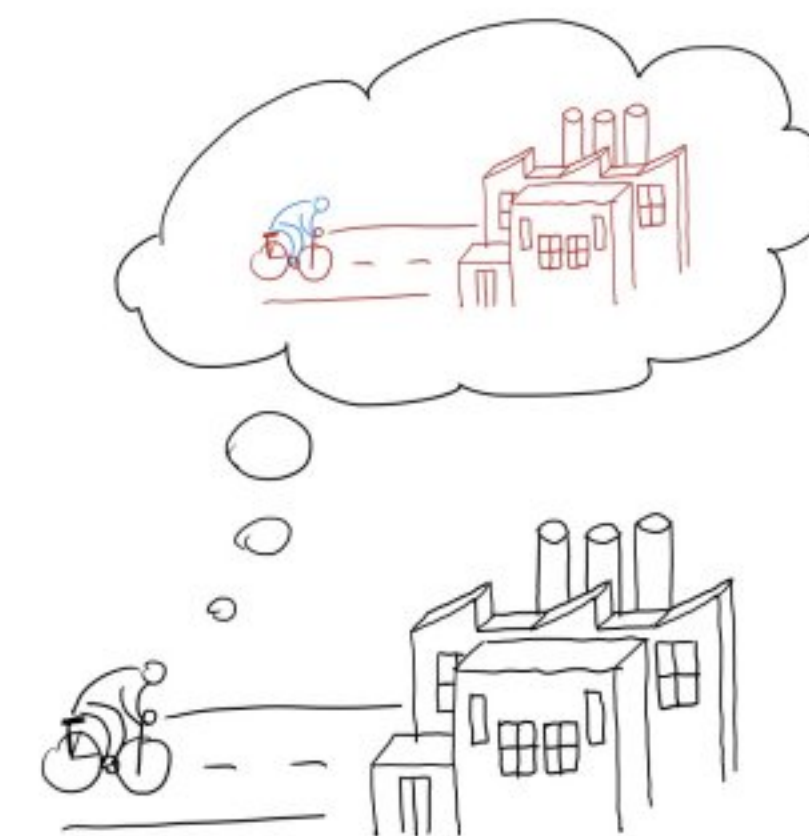
Canonical world models



(a) Environment model.



(b) Agent model.



(c) Joint model.

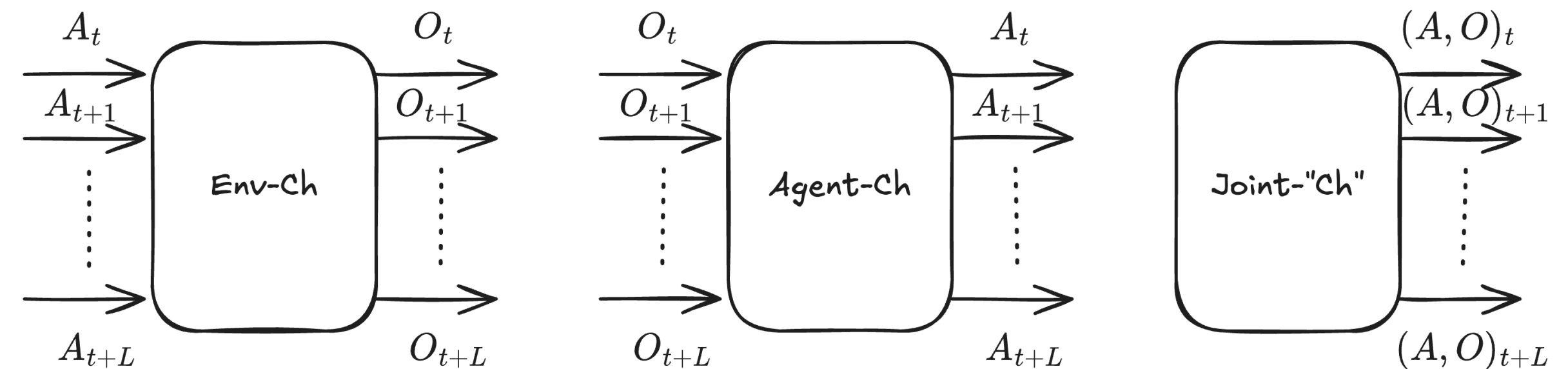
World models, based on interfaces

Channels associated with an interface

An **agent-environment interface** is a pair of finite sets, $(\mathcal{A}, \mathcal{O})$ such that $|\mathcal{A}| \geq 2, |\mathcal{O}| \geq 2$.

Three associated channels:

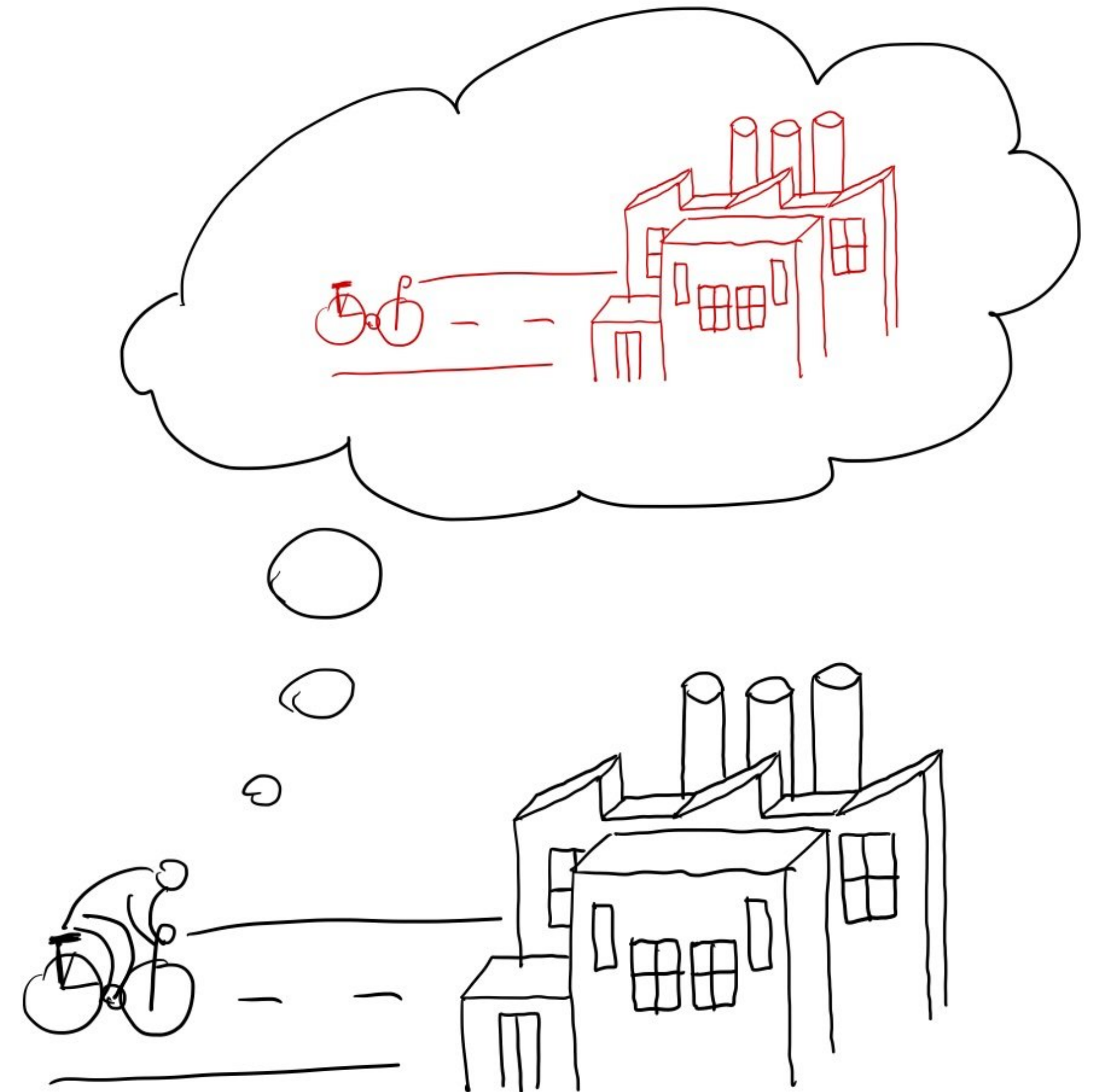
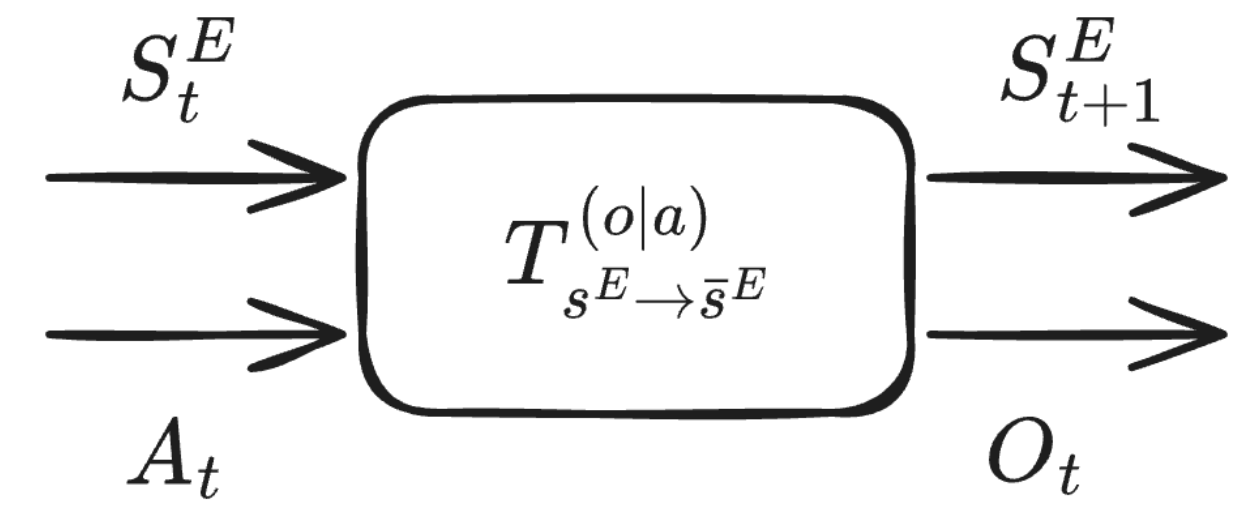
- environment channel $\mathcal{O} \mid \mathcal{A}$.
- agent channel $\mathcal{A} \mid \mathcal{O}$.
- joint process “channel” $(\mathcal{A}, \mathcal{O}) \mid 1$.



Environment models

Things like

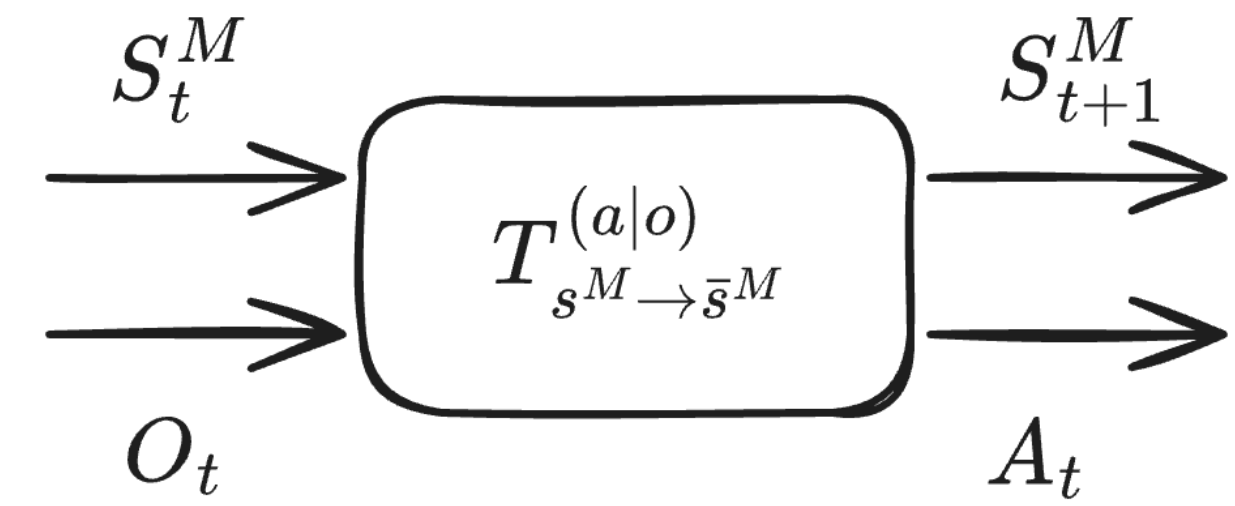
- Belief MDPs
- Bisimulation equivalences for abstractions
- Predictive state representations
- Models maximising empowerment
- AIXI
- ...



Agent (self) models

Things like

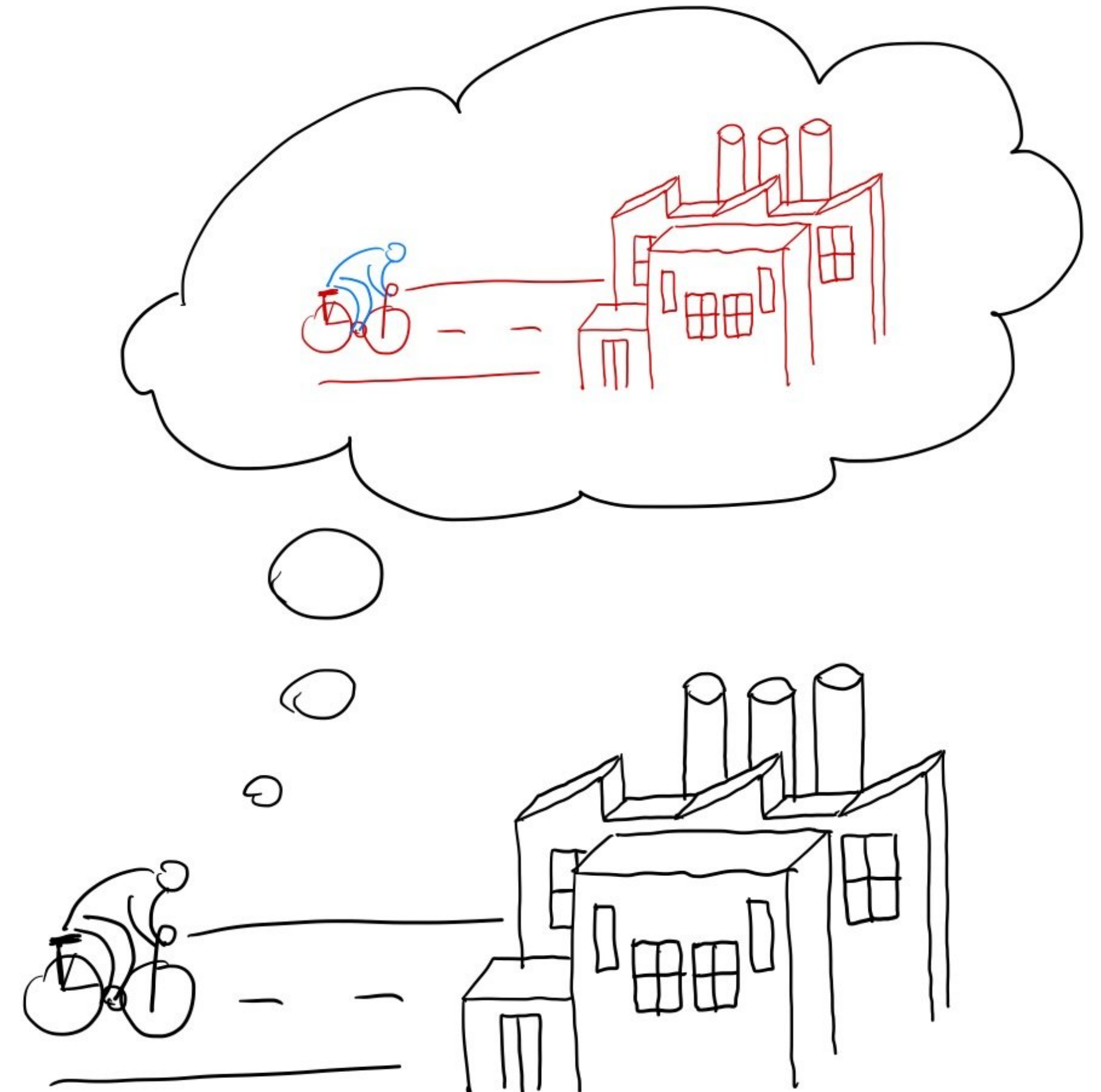
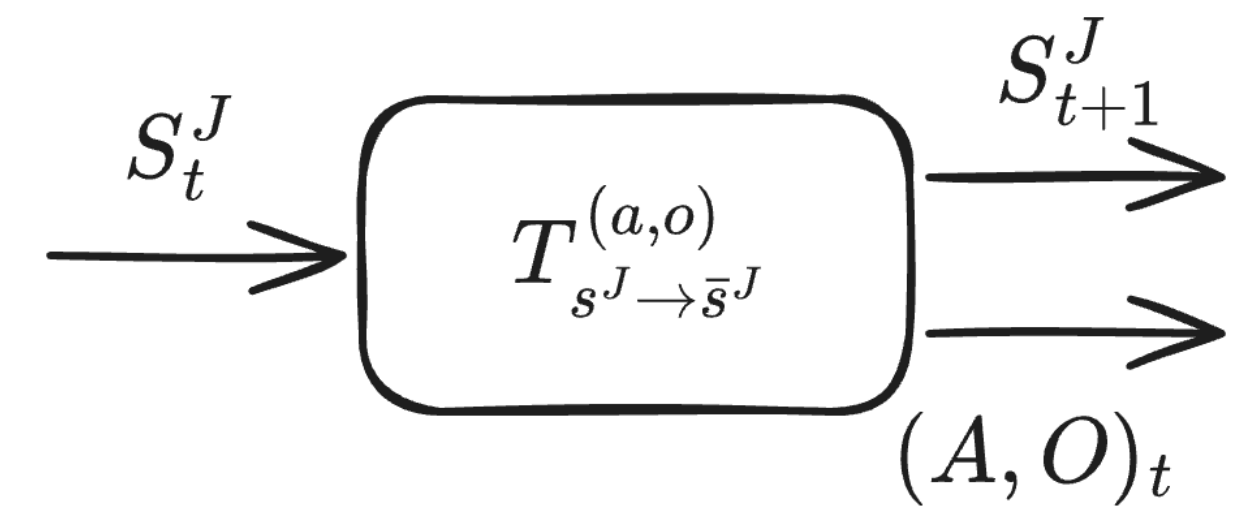
- Policy distillation
- Body-schema
- Behavioural cloning
- Models maximising plasticity
- Self-AIXI
- ...



Joint (embedded) models

Things like

- Embedded agency
- ... and attempts to implement it



Epsilon transducers and machines

“Canonical” := unique, minimal, unifilar transducer

Let $Y_{\cdot} \mid X_{\cdot}$ be a channel.

Its **ϵ -transducer** is the the channel’s unique, minimal, unifilar (transducer) presentation of the channel, $(\mathcal{S}, \mathcal{X}, \mathcal{Y}, \mathcal{T})$ where $\mathcal{S}$ are causal states defined by a surjective map $\epsilon : (\mathcal{X}, \mathcal{Y})_{\cdot t} \rightarrow \mathcal{S}$ given by $\epsilon((x, y)_{\cdot t}) = s = \{(x, y)'_{\cdot t} \mid (x, y)_{\cdot t} \sim_{\epsilon} (x, y)'_{\cdot t}\}$ with the equivalence relation $\sim_{\epsilon}$ of pasts:

$$(x, y)_{\cdot t} \sim_{\epsilon} (x, y)'_{\cdot t} \iff \Pr(Y_{t:} \mid X_{t:}, (X, Y)_{\cdot t} = (x, y)_{\cdot t}) = \Pr(Y_{t:} \mid X_{t:}, (X, Y)_{\cdot t} = (x, y)'_{\cdot t}).$$

The ϵ -map induces dynamics $\mathcal{T} = \{T^{(y|x)}\}_{x \in \mathcal{X}, y \in \mathcal{Y}}$ given by:

$$T_{s \rightarrow s'}^{(y|x)} = \Pr(S_{t+1} = s', Y_t = y \mid S_t = s, X_t = x).$$

An **ϵ -machine** is the **ϵ -transducer** for the channel with trivial inputs $Y_{\cdot} \mid 1_{\cdot}$.

Canonical world models

= epsilon transducers of appropriate channels

Given

- environment channel $O_{\cdot} \mid A_{\cdot}$
- agent channel $A_{\cdot} \mid O_{\cdot}$
- joint process “channel” $(A, O)_{\cdot} \mid 1_{\cdot}$

define ϵ -transducers for them.

If you know RL more than compmech

Example: Predictive state representations (PSRs)

- PSRs approximate canonical environment models for finite sets of finite observation continuations and finite histories
- We define counterparts for agent and joint models

Some results (paper)

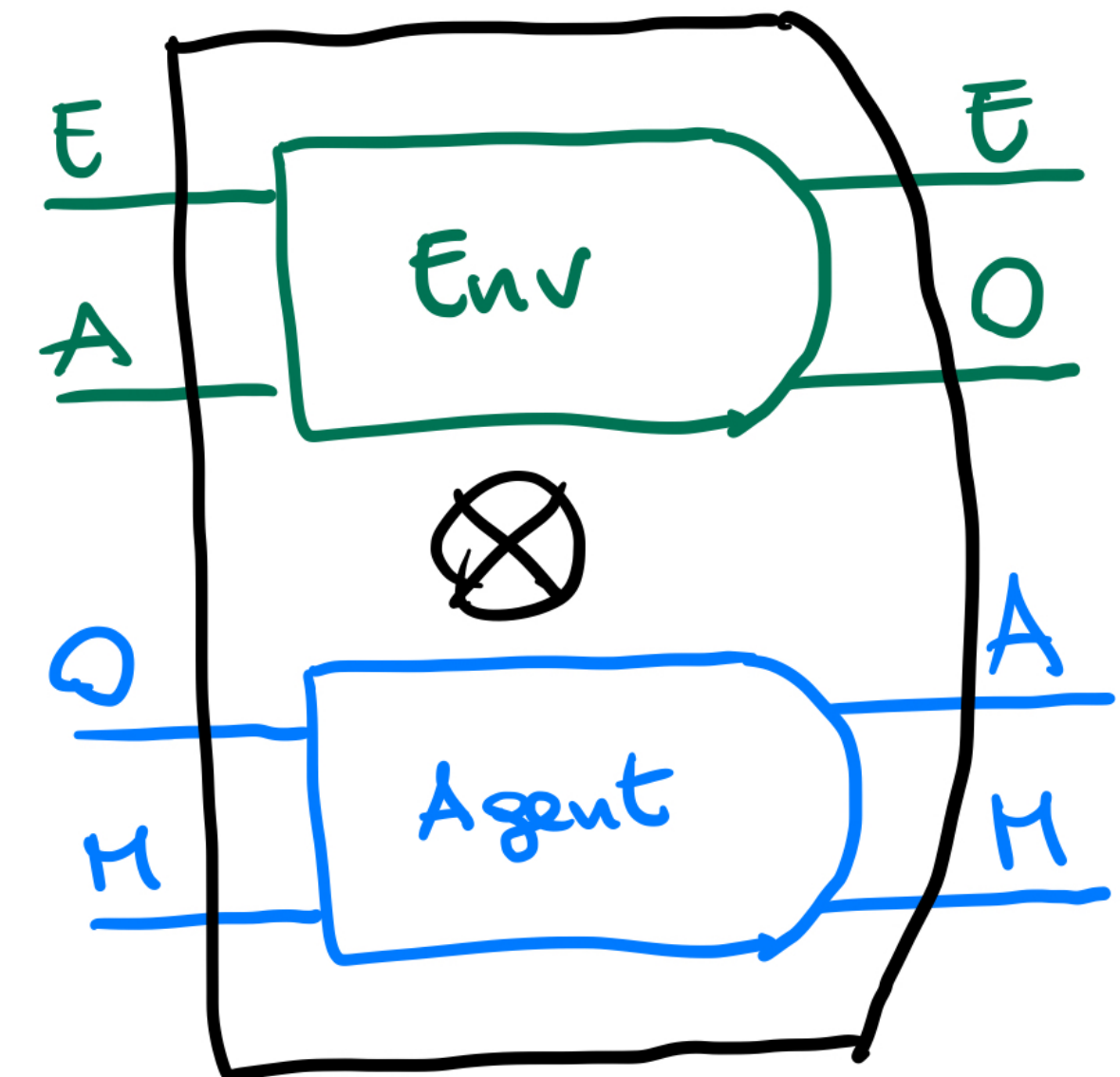
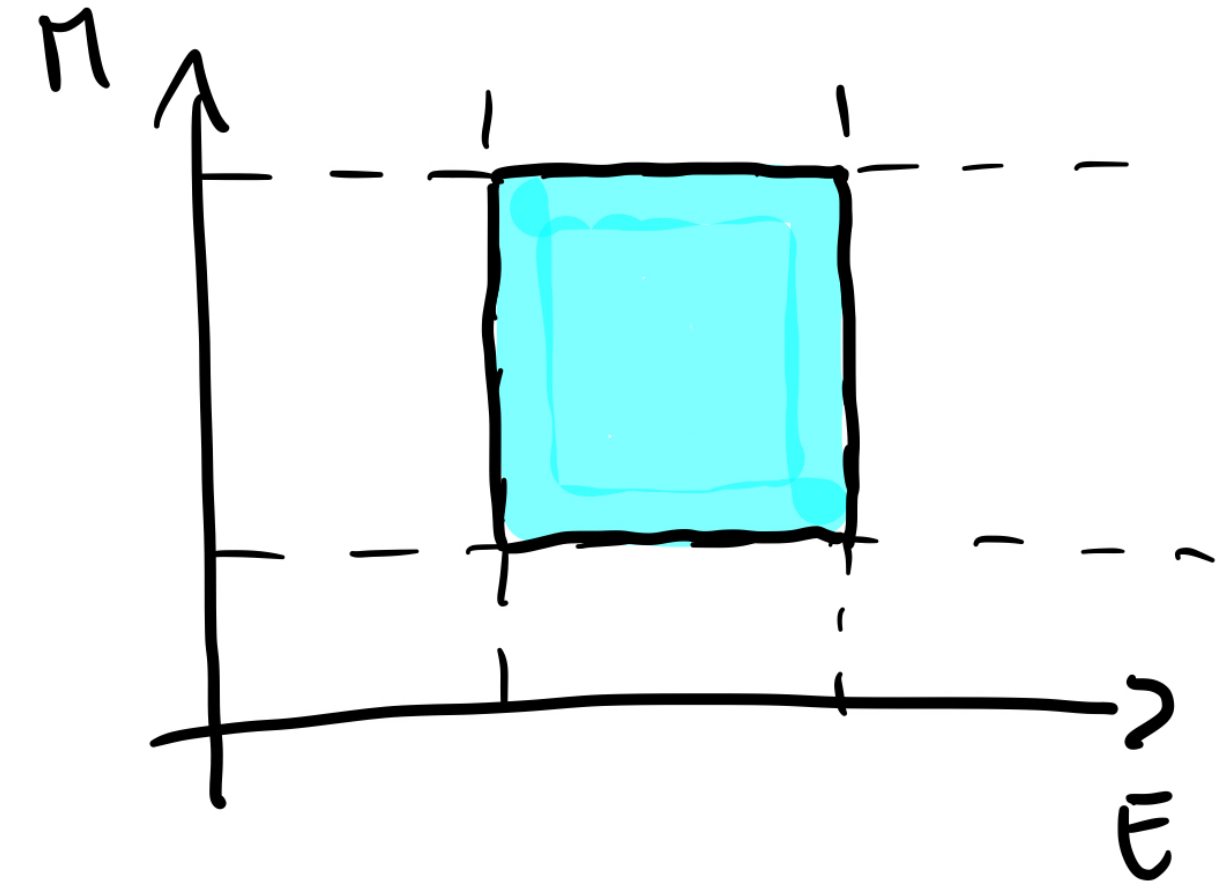
Relating different models

Predicting the environment $\neq$ predicting the agent/self $\neq$ predicting the joint

$\implies$ causal states of environment $\neq$ causal states of the agent/self $\neq$ causal states of the joint

(Check toy model in the paper)

Tldr; learning about an environment and an agent without knowing if/how they interact is like learning a product of two independent processes, cf. Simplex work on factored representations



Support restriction

Agent



Actions: {up, down,
left, right}

Fixed orientation: ↑

Environment



Action trajectory

Full support

- {up, left, left, right} ✓
- {right, right, up, up} ✓
- {up, up, left, right} ✓
- ⋮

Restricted support $\rightarrow P(O_{t:t+N} | a_{t-N:t})$ undefined

- {up, left, left, right} ✗
- {right, right, up, up} ✓
- {up, up, left, right} ✗
- ⋮

Coupling agent and environment

Start from the joint/embedded model of **coupled** agent-environment system

Define support-restricted models via totalisation of partial channels

Canonical support-restricted environment and agent are coarse-grainings of the canonical joint model (Theorem 18 for environment)

Connected to good regulator results

A “Good Regulator Theorem” for Embodied Agents

Nathaniel Virgo^{1,2}, Martin Biehl³, Manuel Baltieri⁴, and Matteo Capucci⁵

¹University of Hertfordshire, UK ²Earth-Life Science Institute (ELSI), Institute of Science Tokyo, Japan

³Cross Labs, Cross Compass Ltd., Japan ⁴Araya, Inc., Japan ⁵Independent researcher

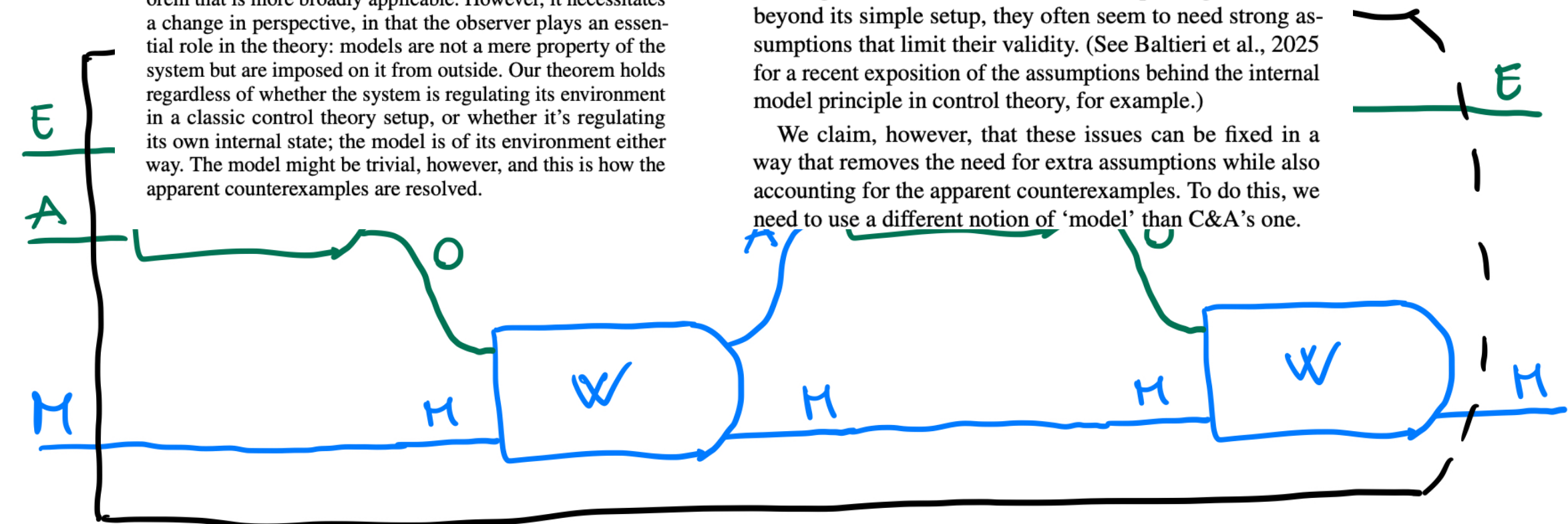
Abstract

In a classic paper, Conant and Ashby claimed that “every good regulator of a system must be a model of that system.” Artificial Life has produced many examples of systems that perform tasks with apparently no model in sight; these suggest Conant and Ashby’s theorem doesn’t easily generalise beyond its restricted setup. Nevertheless, here we show that a similar intuition can be fleshed out in a different way: whenever an agent is able to perform a regulation task, it is possible for an observer to interpret it as having “beliefs” about its environment, which it “updates” in response to sensory input. This notion of belief updating provides a notion of model that is more sophisticated than Conant and Ashby’s, as well as a theorem that is more broadly applicable. However, it necessitates a change in perspective, in that the observer plays an essential role in the theory: models are not a mere property of the system but are imposed on it from outside. Our theorem holds regardless of whether the system is regulating its environment in a classic control theory setup, or whether it’s regulating its own internal state; the model is of its environment either way. The model might be trivial, however, and this is how the apparent counterexamples are resolved.

work descending from them appear to have reached disparate conclusions: Artificial Life has produced many examples of agents that apparently do not need an internal model in order to achieve their tasks (e.g. Braitenberg, 1986; Beer, 2003), apparently contradicting (C&A).

This is in part because (C&A) doesn't strictly succeed, even in its own terms, in showing what its title claims. (This has been noted previously in blog posts: Baez, 2016; Wentworth, 2021.) Its theorem doesn't establish that *every* good regulator is a model but only that some of them are (those that are not "unnecessarily complex" in C&A's terminology). Although there have been several attempts to generalise it beyond its simple setup, they often seem to need strong assumptions that limit their validity. (See Baltieri et al., 2025 for a recent exposition of the assumptions behind the internal model principle in control theory, for example.)

We claim, however, that these issues can be fixed in a way that removes the need for extra assumptions while also accounting for the apparent counterexamples. To do this, we need to use a different notion of ‘model’ than C&A’s one.



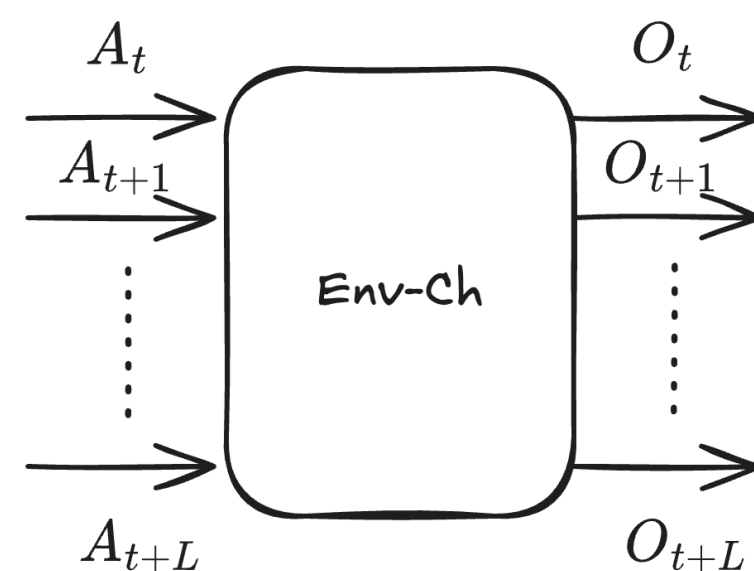
Some experiments (WIP)

With or without latents

In RL, belief MPDs vs. PSRs

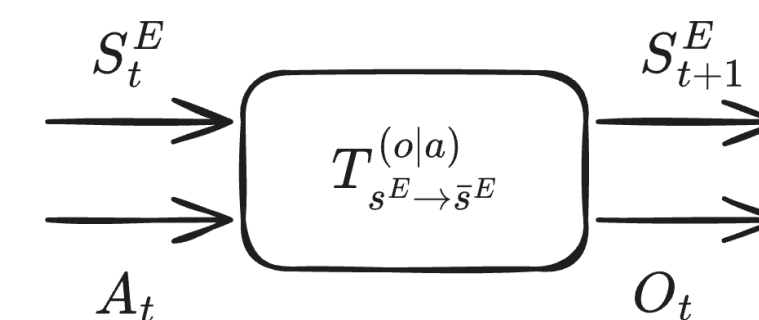
Without latents:

- start from a channel,
- define causal states and transitions between them,
- get the ϵ -transducers
- things like: PSR-y things approximations



With latents:

- start with a transducer,
- define beliefs over the states of the transducer and transitions between them,
- get the mixed-state presentation that perform Bayes updates like a belief MDP
- things like: belief MDPs



Canonical world models with latents

= mixed-state presentations of transducers of appropriate channels

Given

- the environment **CHMM** $A \times M \rightarrow \Delta(O \times M)$,
- the agent **CHMM** $O \times E \rightarrow \Delta(A \times E)$, and
- the joint process **HMM** $M \times E \rightarrow \Delta(A \times O \times M \times E)$.

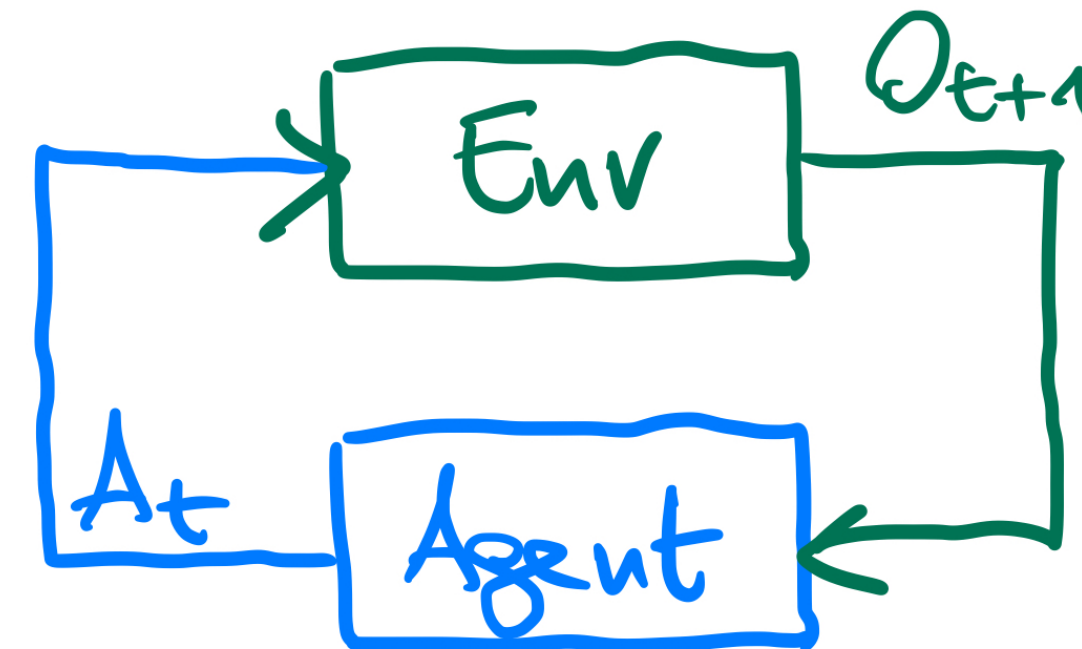
define mixed-state presentations for them.

Generating process

Some ideas from Simplex, revisited



vs.

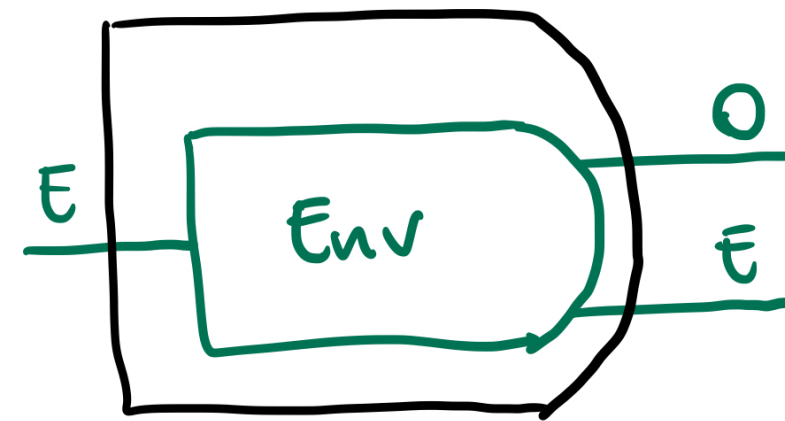


Slightly more formally

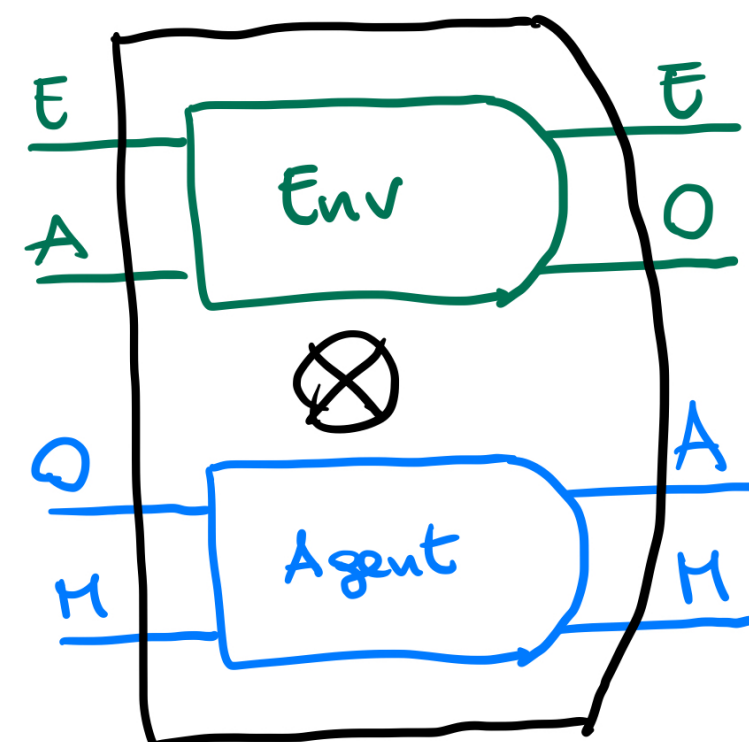
Kernels and their compositions



"Open-loop"
Autonomous

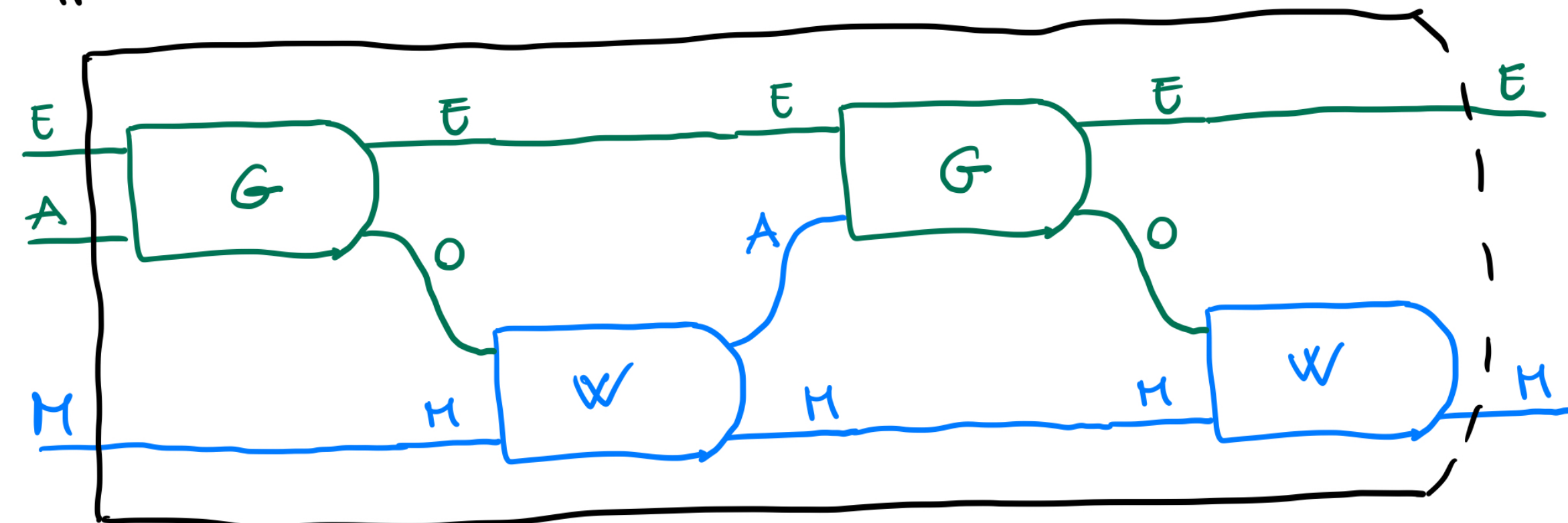


"Open-loop"
Full-support



VS.

"Closed-loop"
Support restricted



RL-like loop

Task for RL agent: make $m = o$

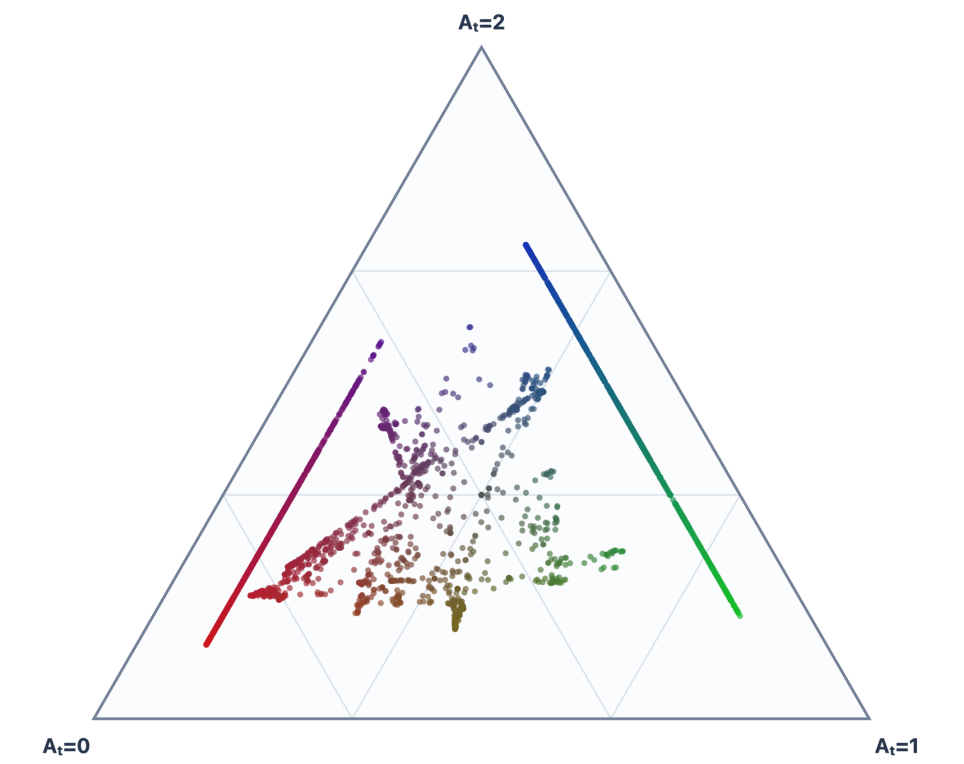
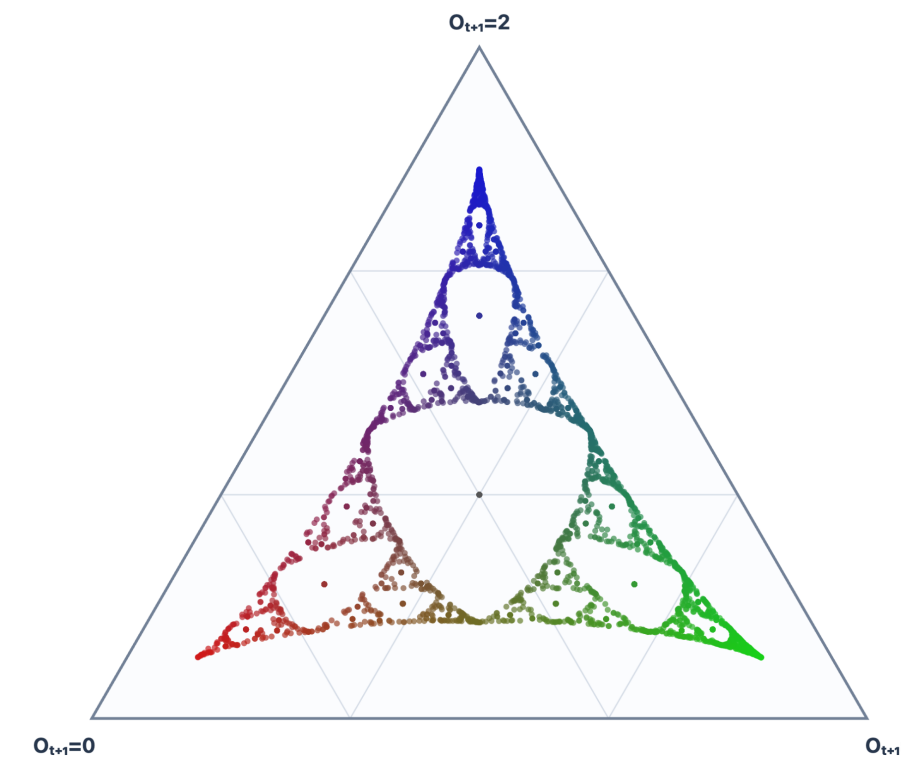
Environment

$W : \mathcal{A} \times \mathcal{E} \rightarrow \Delta(\mathcal{O} \times \mathcal{E})$, with $W_o^{(a)} = P_a F_o$

Mess3 with $\alpha = 0.60$, $x = 0.15$, defining F_o

Actions change dynamics of Mess 3, defining P_a which permutes F_o 's rows.

$$P_{\text{raise}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad P_{\text{lower}} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad P_{\text{hold}} = I$$



Agent

$G : \mathcal{O} \times \mathcal{M} \rightarrow \Delta(\mathcal{A} \times \mathcal{M})$, with
 $G(a, m' | o, m) = \pi(a | o, m) U_{m, m'}^a$

Policy

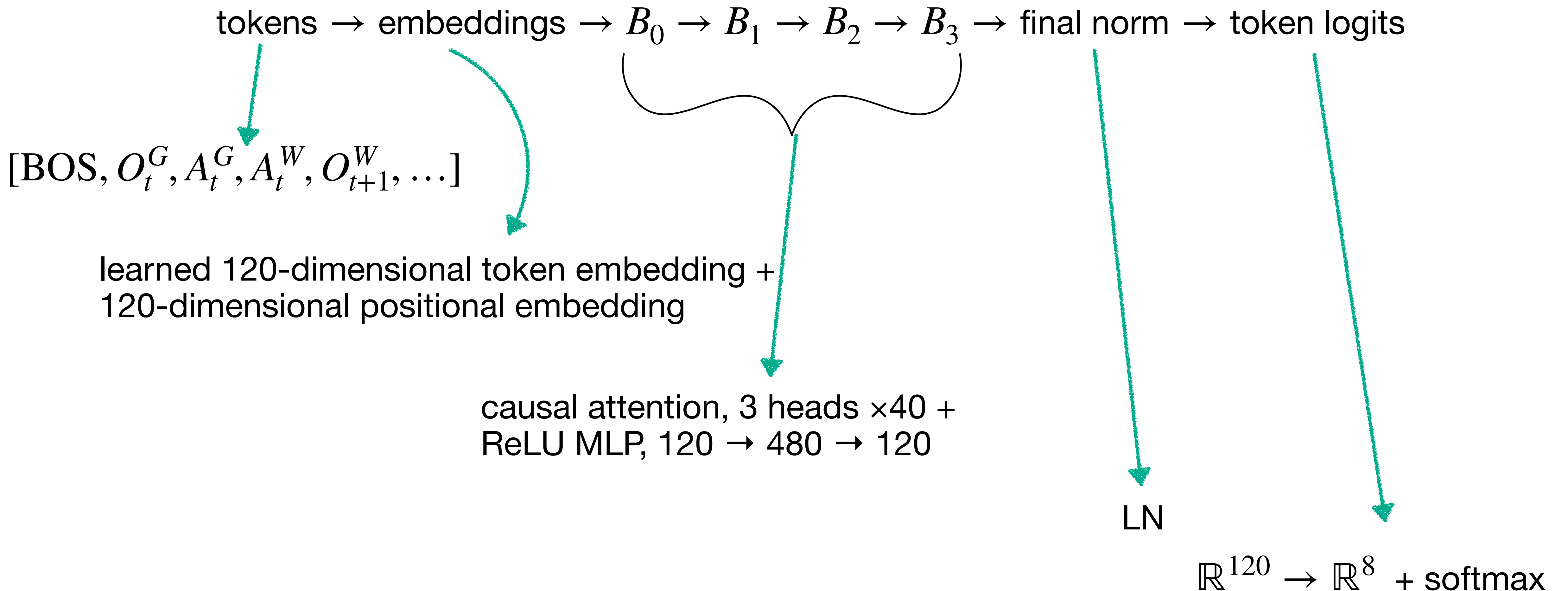
$$\pi(a | o, m) = 0.09 + 0.731 \left[a = a^*(m, o) \right]$$

$$\text{with } a^*(m, o) = \begin{cases} \text{raise,} & m < o, \\ \text{lower,} & m > o, \\ \text{hold,} & m = o. \end{cases}$$

Transitions U^a depend on a , and in turn on o

Toy transformer

Task for transformer: many, see next



Seven models

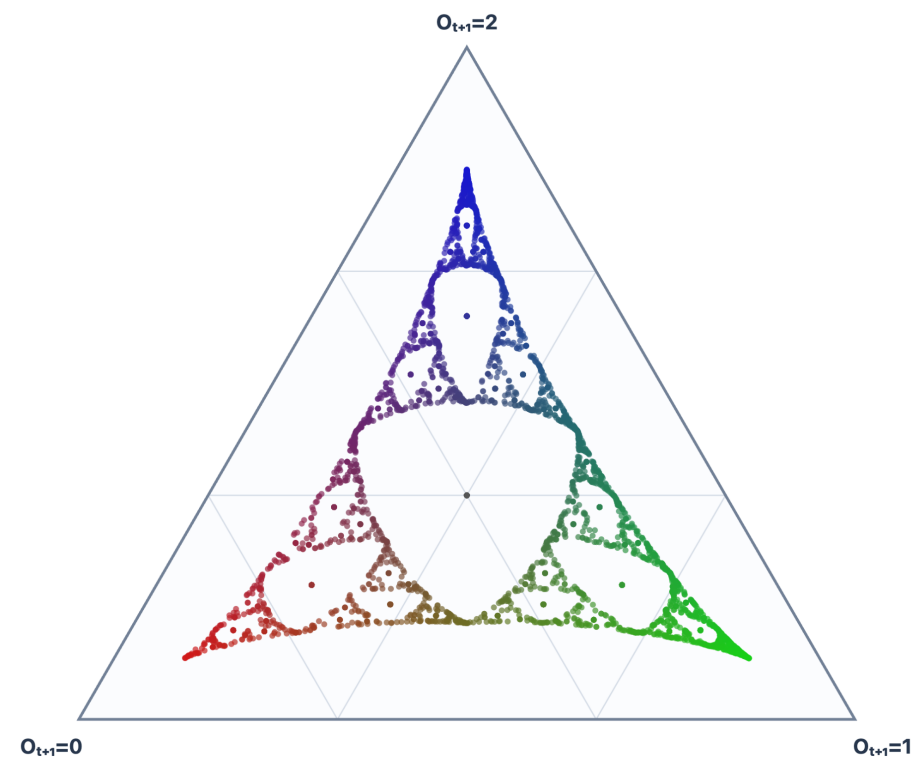
Same architecture, different training

Trained to predict:

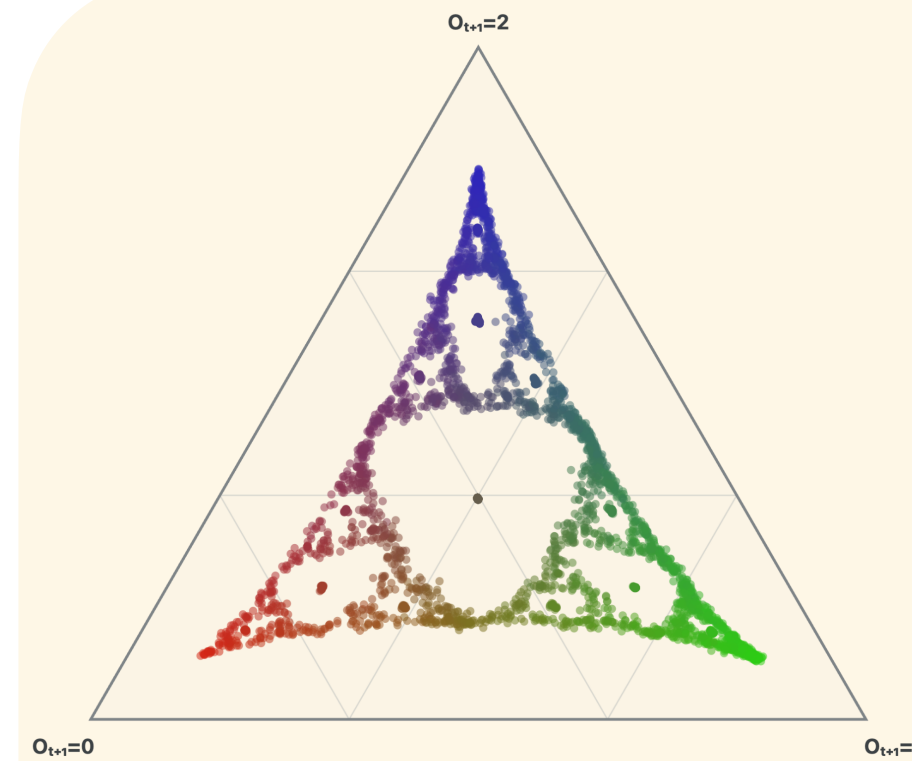
- Nothing (T0)
- Open-loop data
 - Env, $P(O_{t+1} \mid O_{:t}, A_{:t})$
 - Agent, $P(A_t \mid O_{:t}, A_{:t-1})$
 - Joint, $P(O_{t+1} \mid O_{:t}, A_{:t}), P(A_t \mid O_{:t}, A_{:t-1})$
- Closed-loop data
 - Env, $P(O_{t+1} \mid O_{:t}, A_{:t})$
 - Agent, $P(A_t \mid O_{:t}, A_{:t-1})$
 - Joint, $P(O_{t+1} \mid O_{:t}, A_{:t}), P(A_t \mid O_{:t}, A_{:t-1})$

Predictions about the environment

Showing $P(O_{t+1} \mid O_{:t}, A_{:t})$ from frozen “observation head”



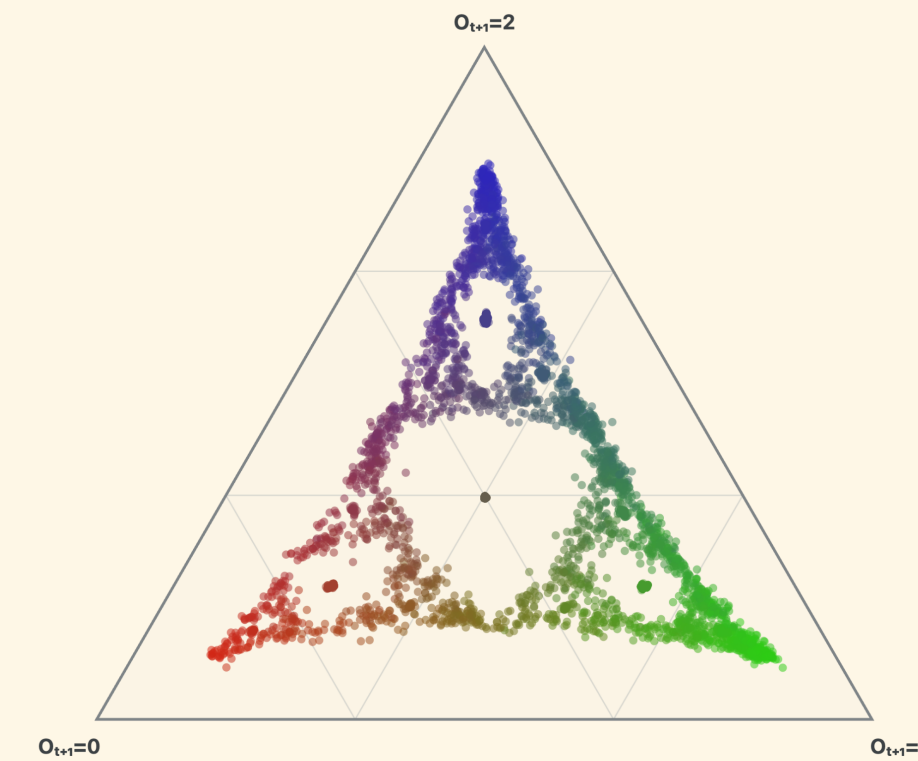
(Training)



$$P(O_{t+1} \mid O_{:t}, A_{:t})$$



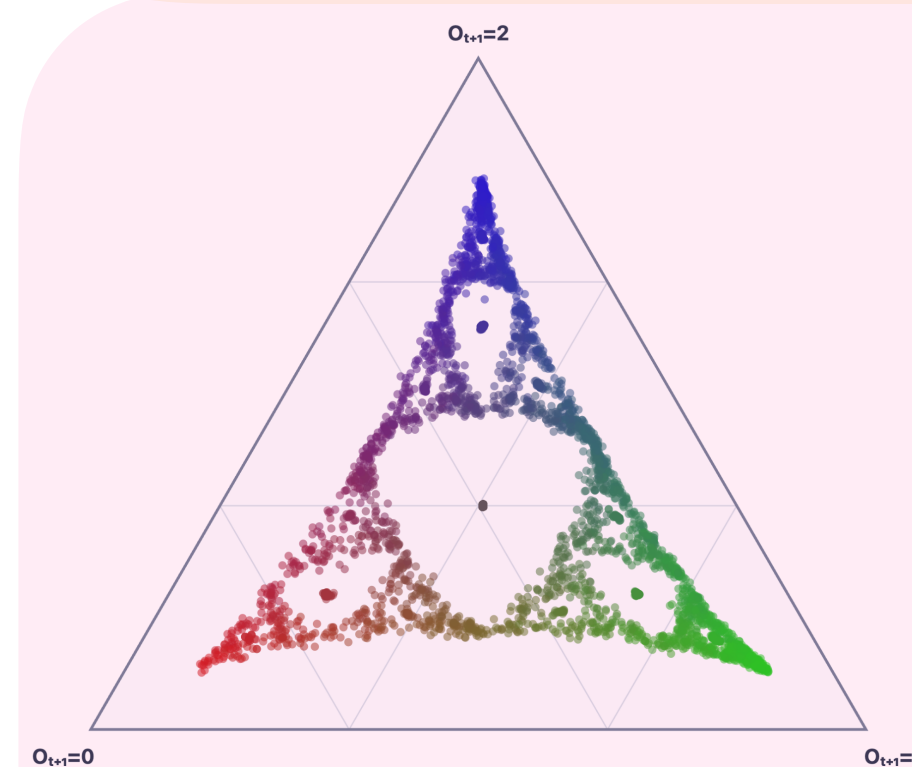
$$P(A_t \mid O_{:t}, A_{:t-1})$$



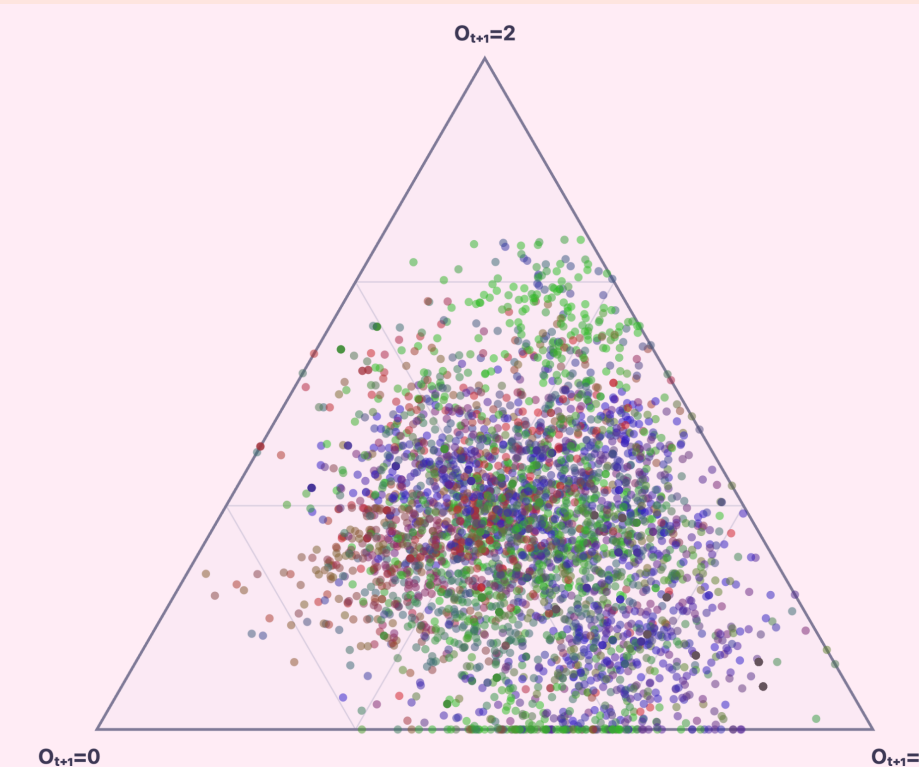
$$P(O_{t+1} \mid O_{:t}, A_{:t})P(A_t \mid O_{:t}, A_{:t-1})$$



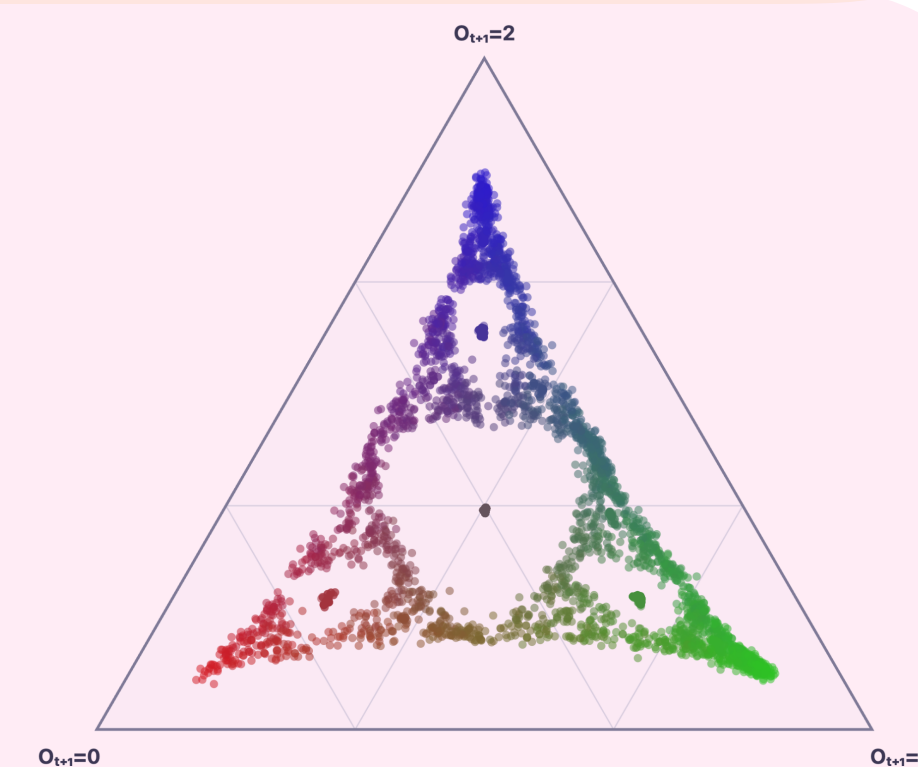
(Training)



$$P(O_{t+1} \mid O_{:t}, A_{:t})$$



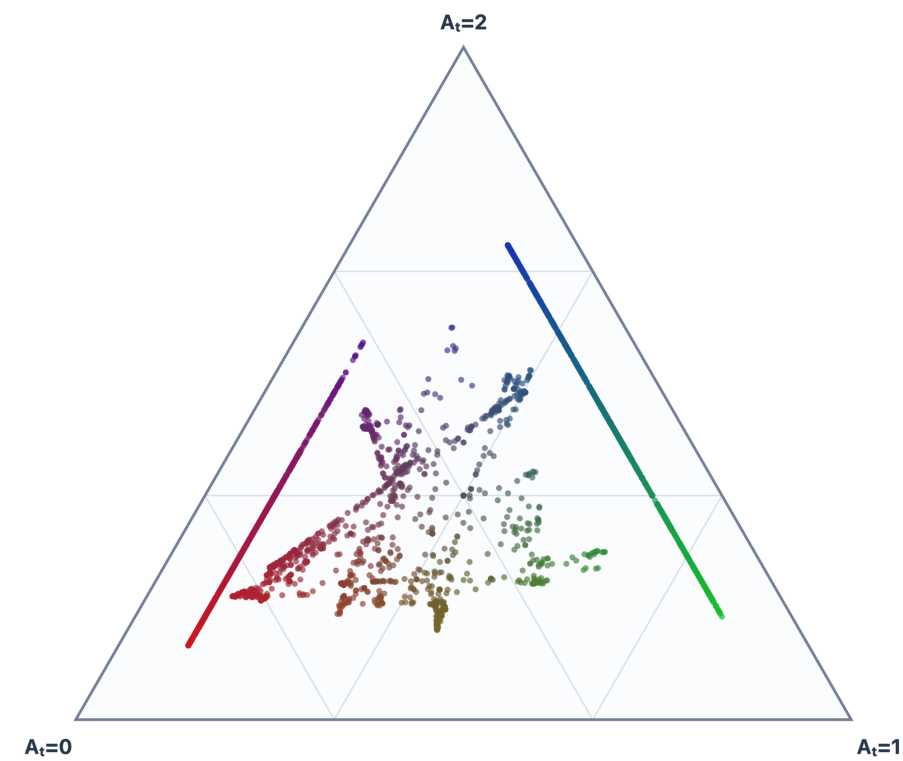
$$P(A_t \mid O_{:t}, A_{:t-1})$$



$$P(O_{t+1} \mid O_{:t}, A_{:t})P(A_t \mid O_{:t}, A_{:t-1})$$

Predictions about the agent

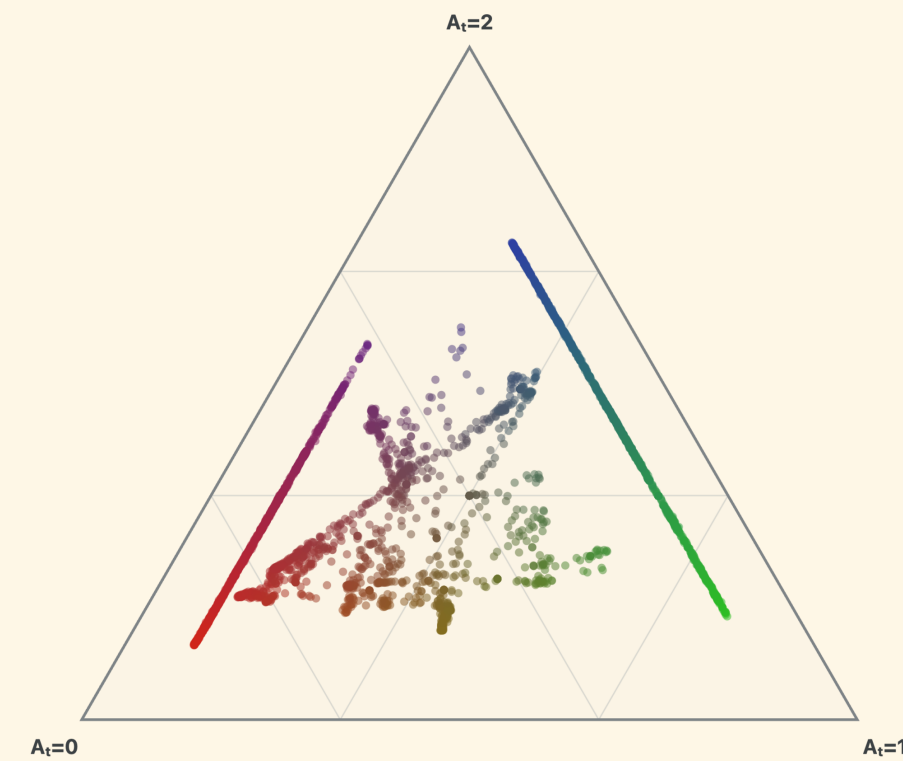
Showing $P(A_t \mid O_{:t}, A_{:t-1})$ from frozen “action head”



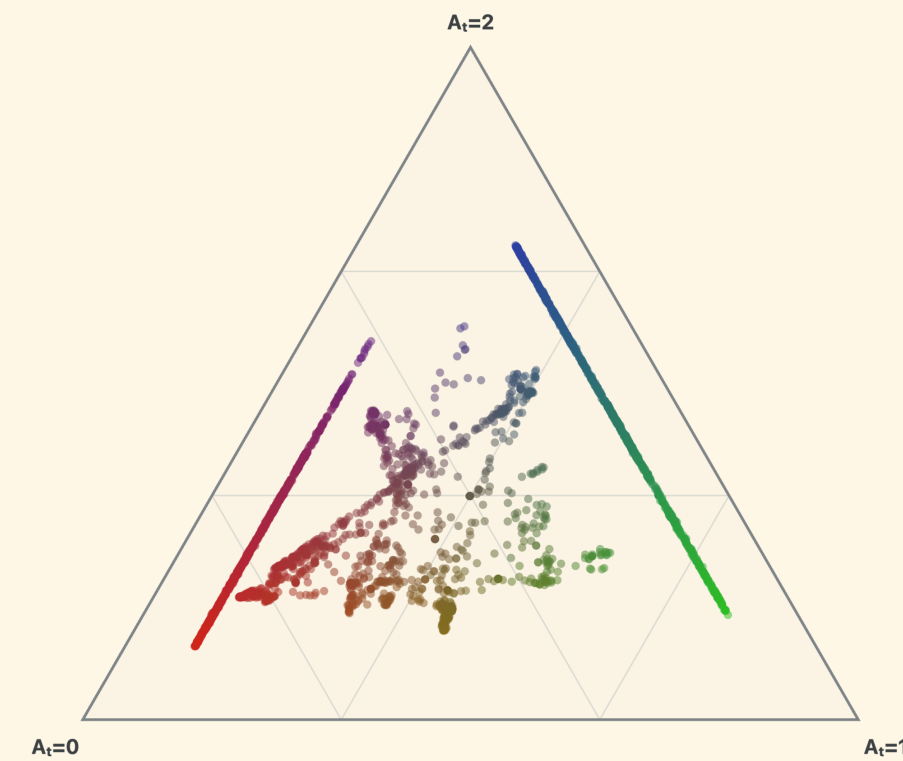
(Training)



$P(O_{t+1} \mid O_{:t}, A_{:t})$



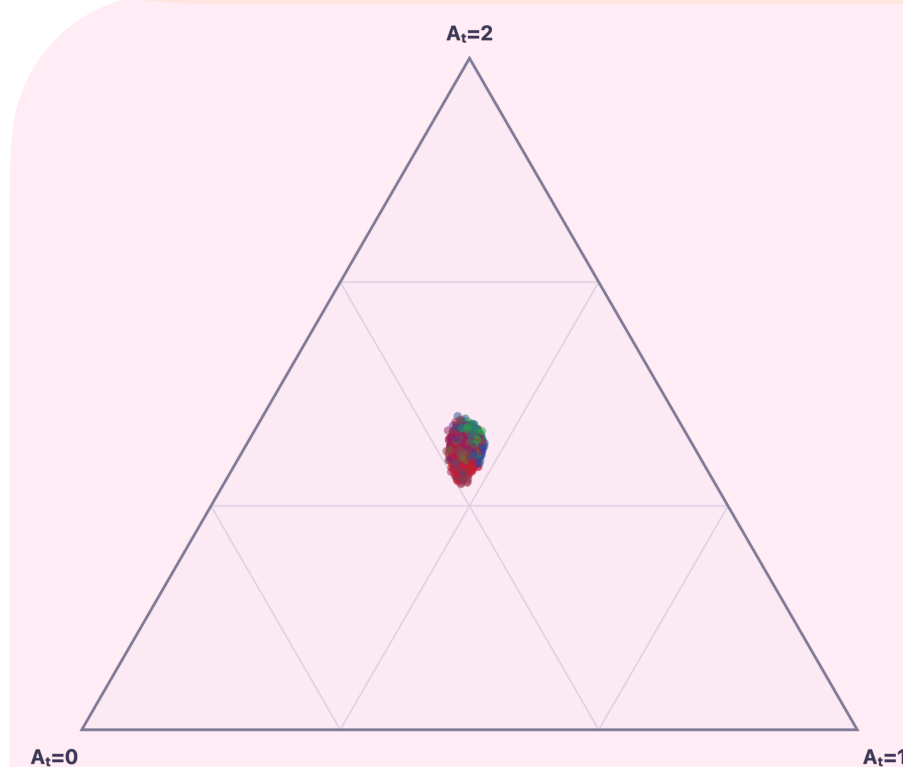
$P(A_t \mid O_{:t}, A_{:t-1})$



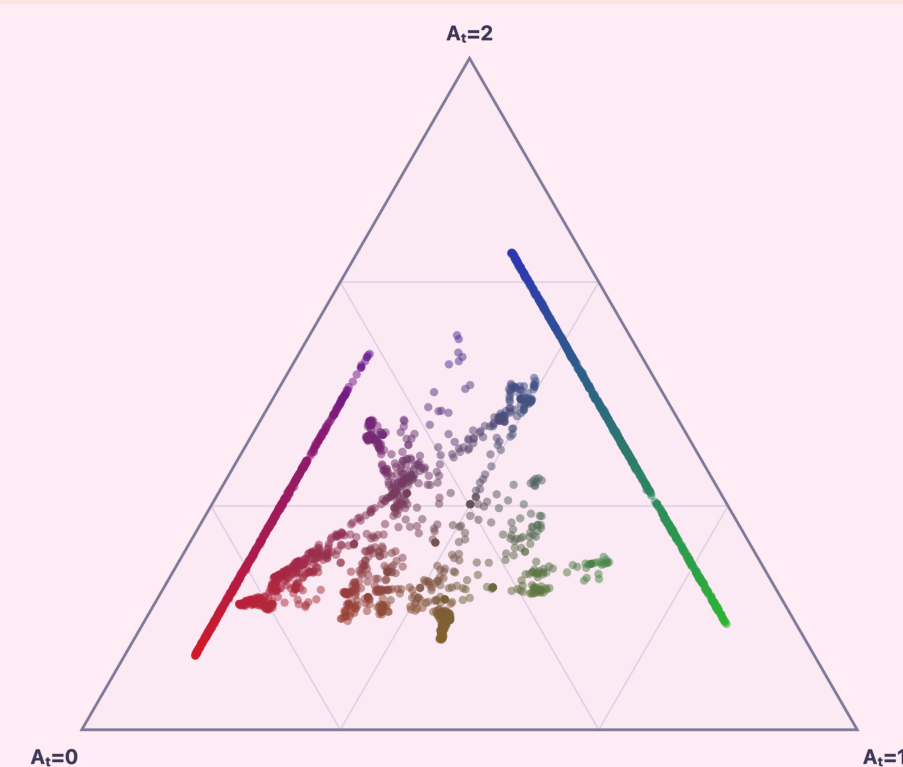
$P(O_{t+1} \mid O_{:t}, A_{:t})P(A_t \mid O_{:t}, A_{:t-1})$



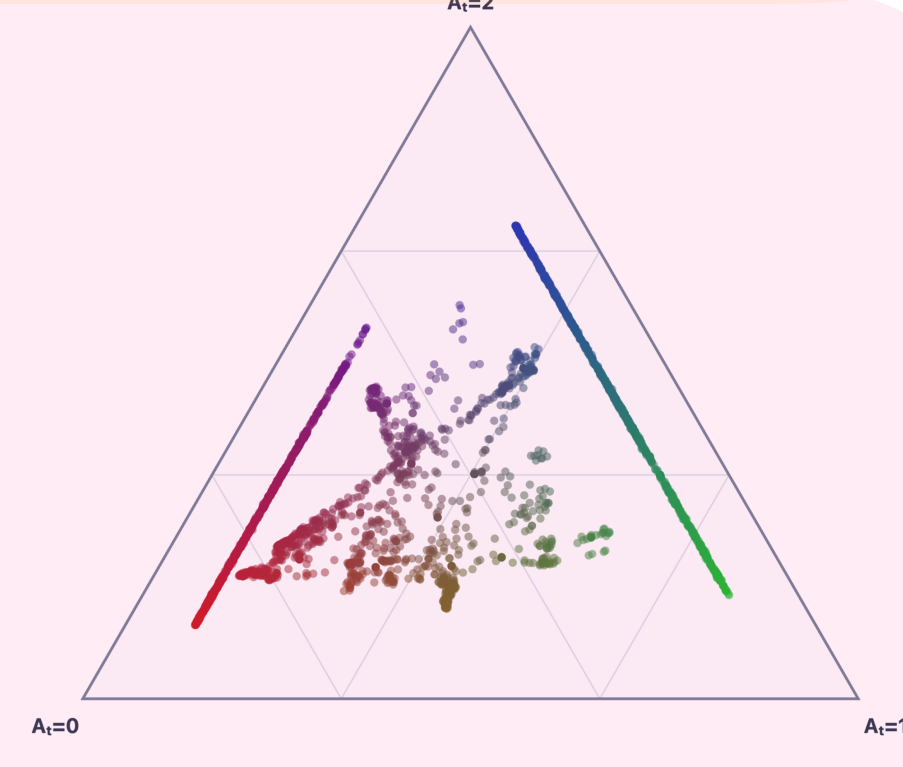
(Training)



$P(O_{t+1} \mid O_{:t}, A_{:t})$



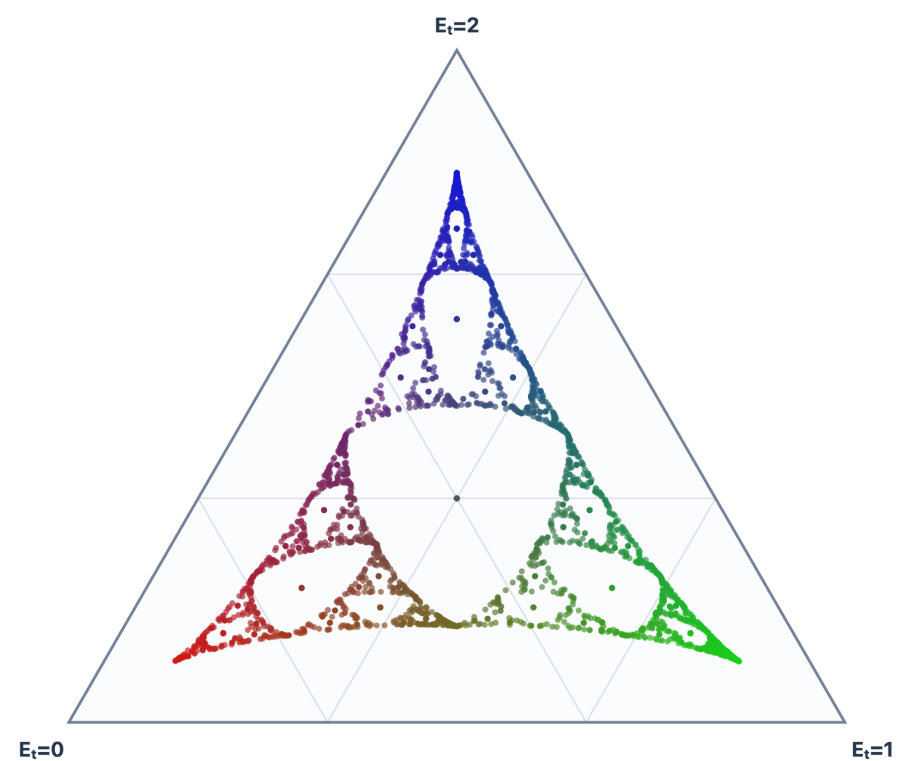
$P(A_t \mid O_{:t}, A_{:t-1})$



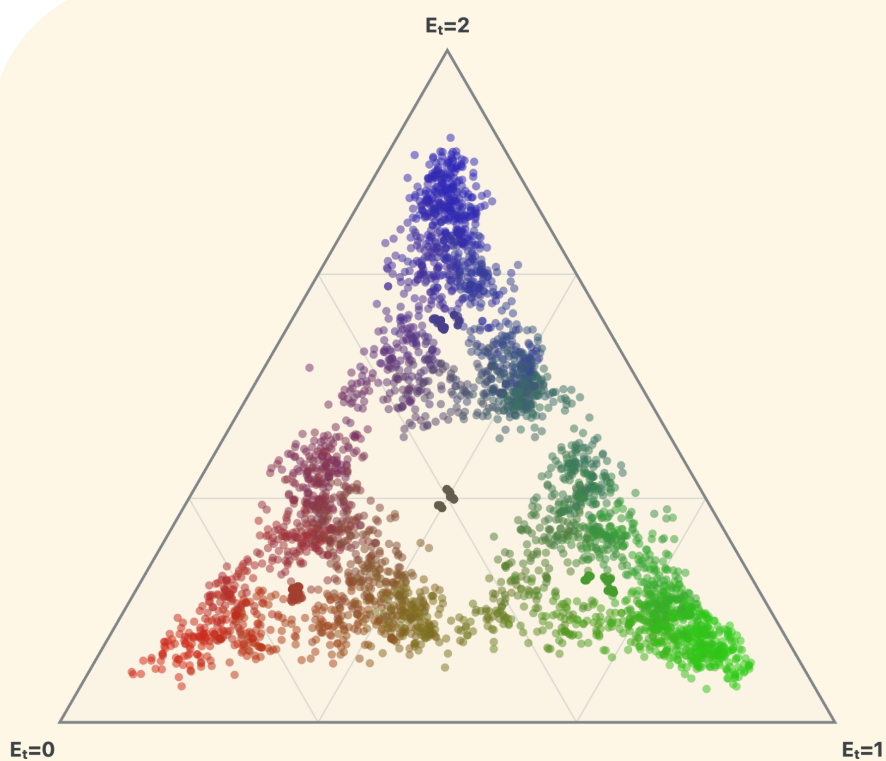
$P(O_{t+1} \mid O_{:t}, A_{:t})P(A_t \mid O_{:t}, A_{:t-1})$

Beliefs about the environment

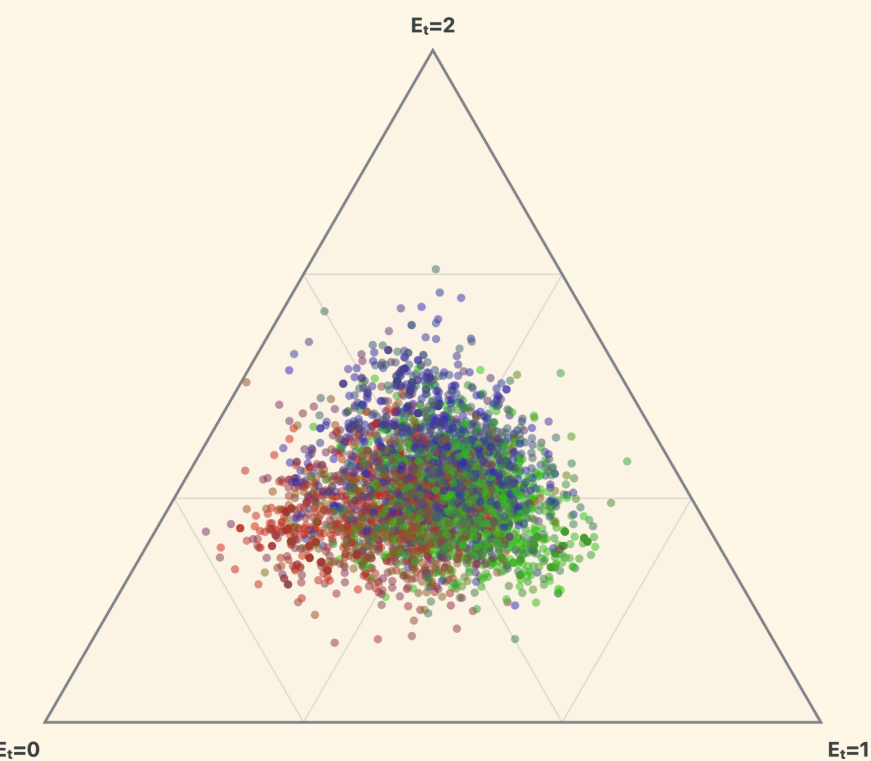
Showing $b_t^E = P(E_t \mid O_{:t}, A_{:t-1})$, ridge reconstruction after B_3



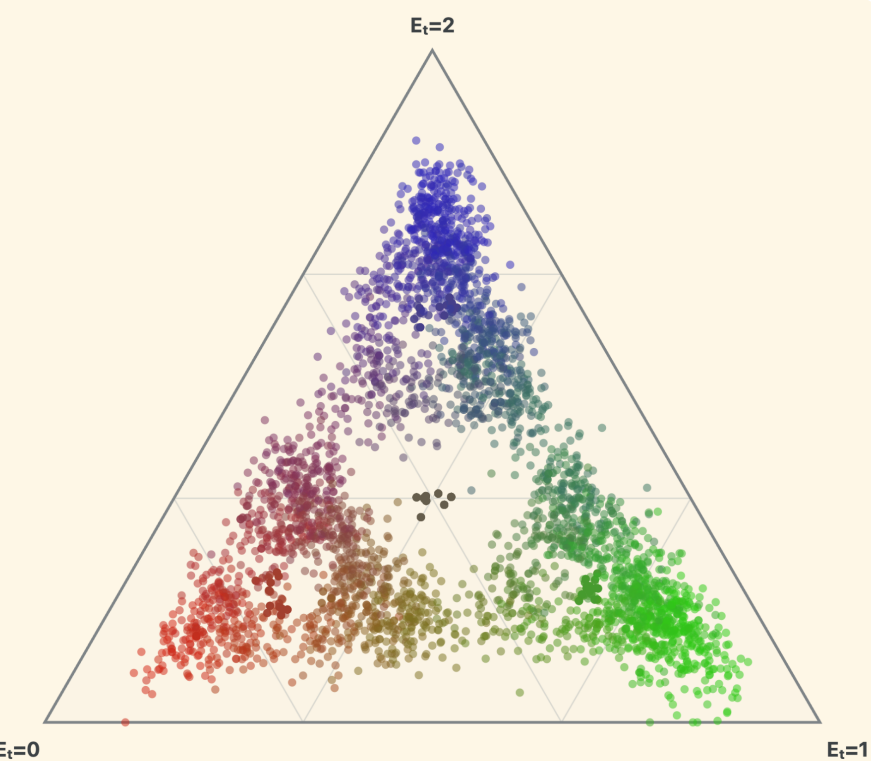
(Training)



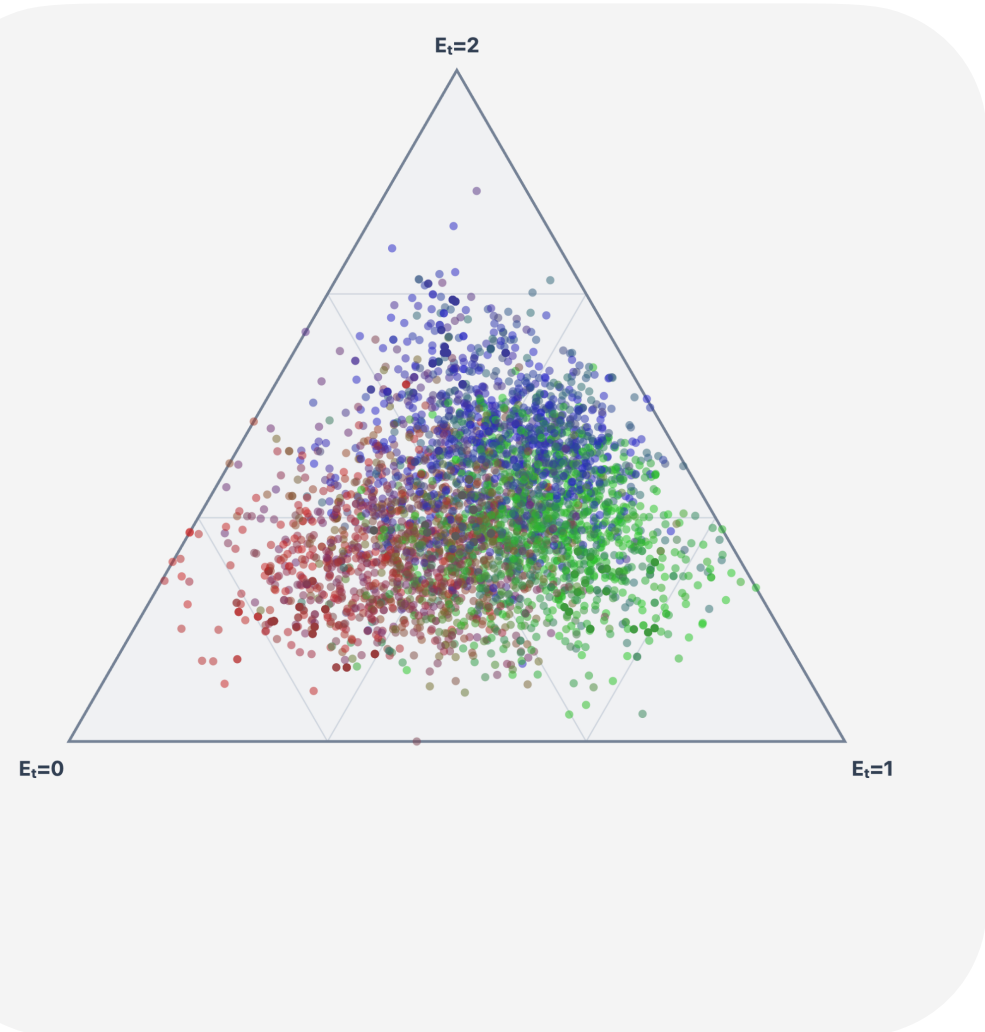
$P(O_{t+1} \mid O_{:t}, A_{:t})$



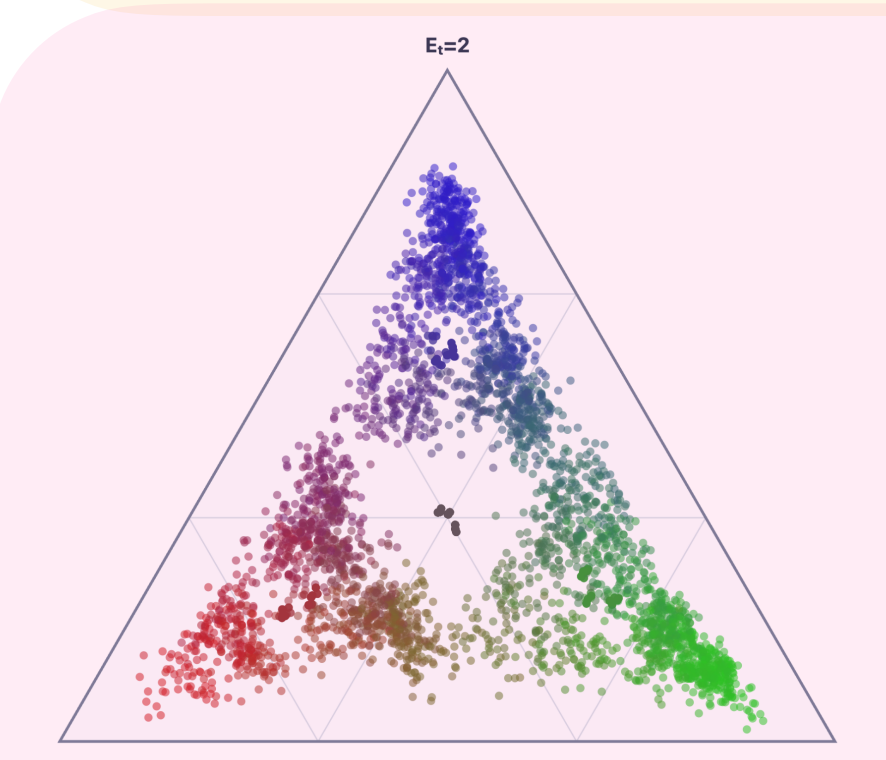
$P(A_t \mid O_{:t}, A_{:t-1})$



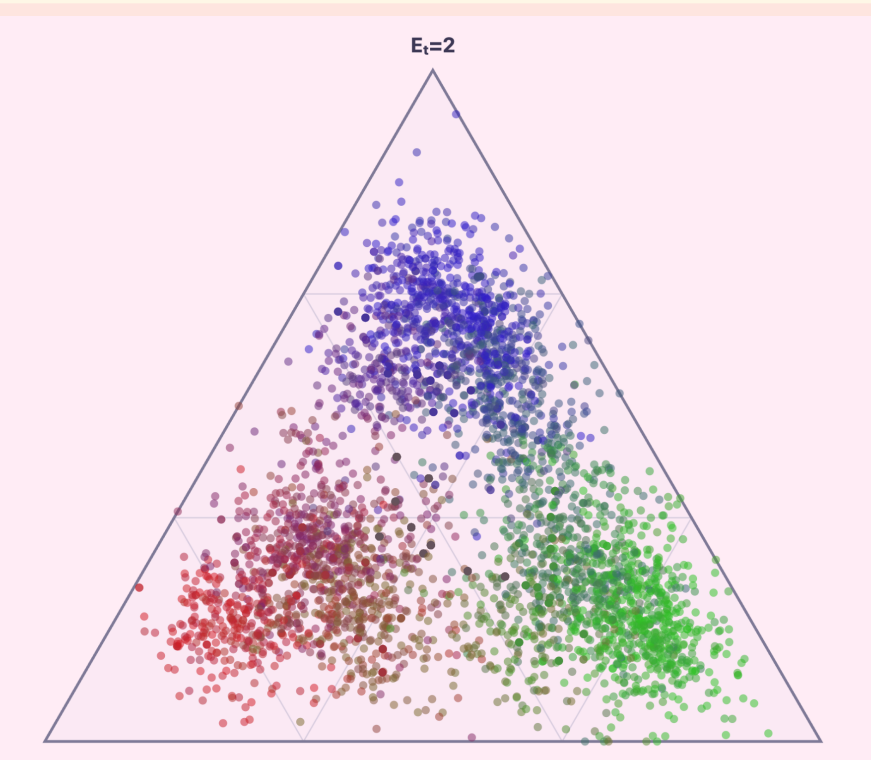
$P(O_{t+1} \mid O_{:t}, A_{:t})P(A_t \mid O_{:t}, A_{:t-1})$



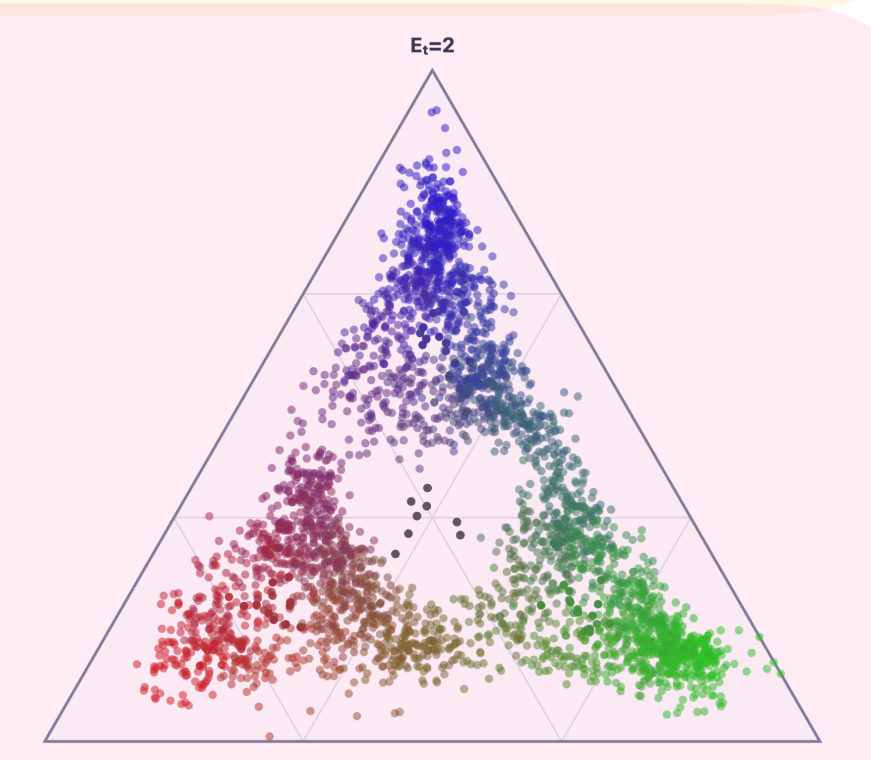
(Training)



$P(O_{t+1} \mid O_{:t}, A_{:t})$



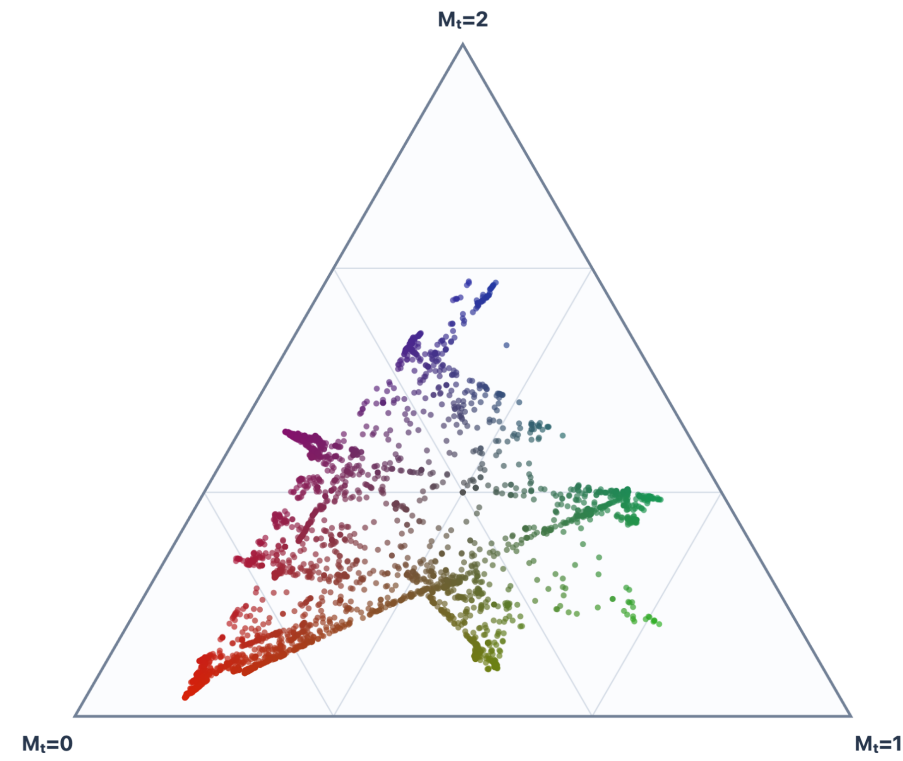
$P(A_t \mid O_{:t}, A_{:t-1})$



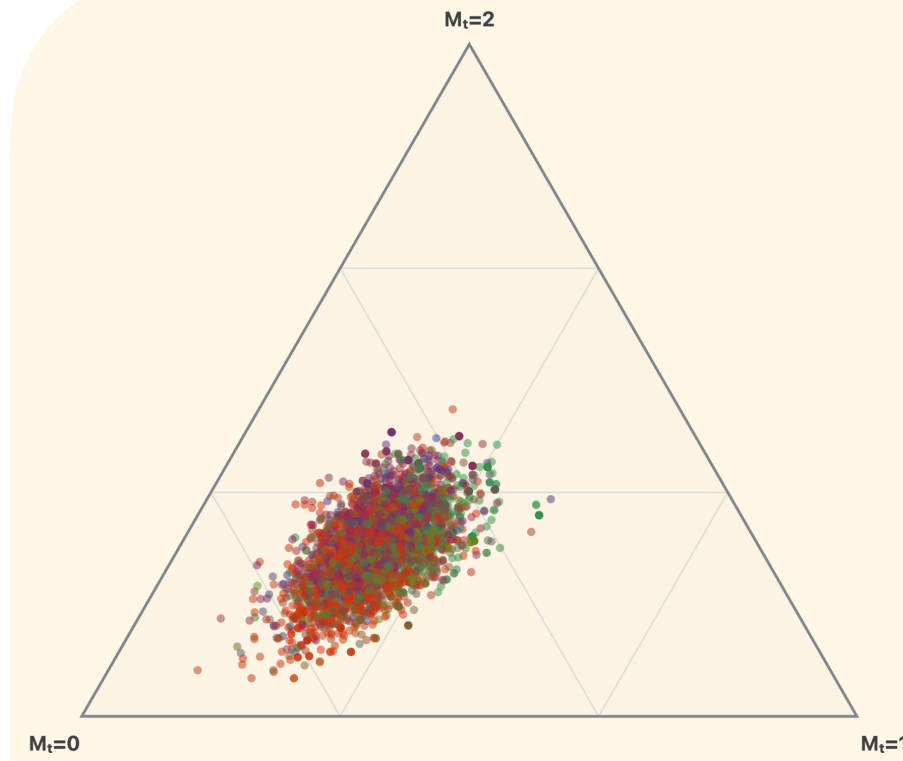
$P(O_{t+1} \mid O_{:t}, A_{:t})P(A_t \mid O_{:t}, A_{:t-1})$

Beliefs about the agent

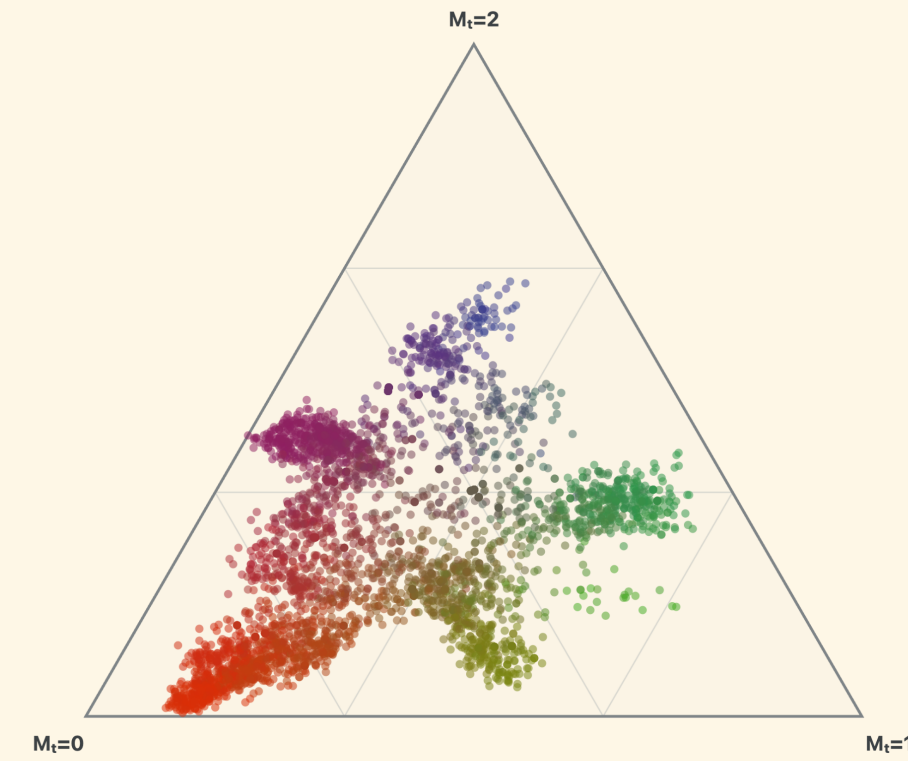
Showing $b_t^M = P(M_t \mid O_{:t}, A_{:t-1})$, ridge reconstruction after B_3



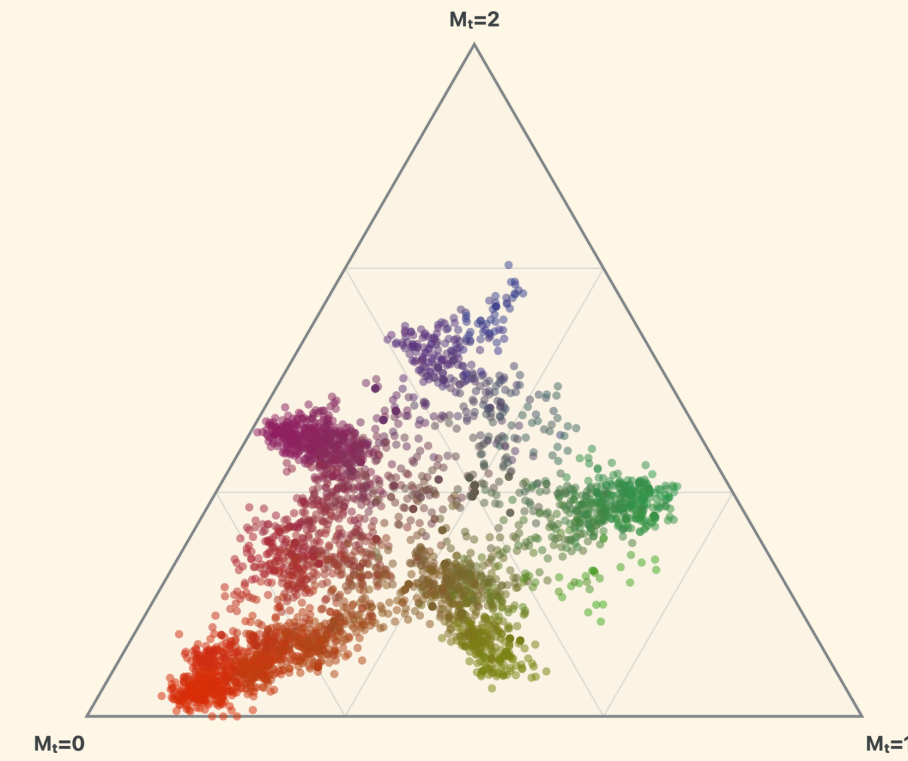
(Training)



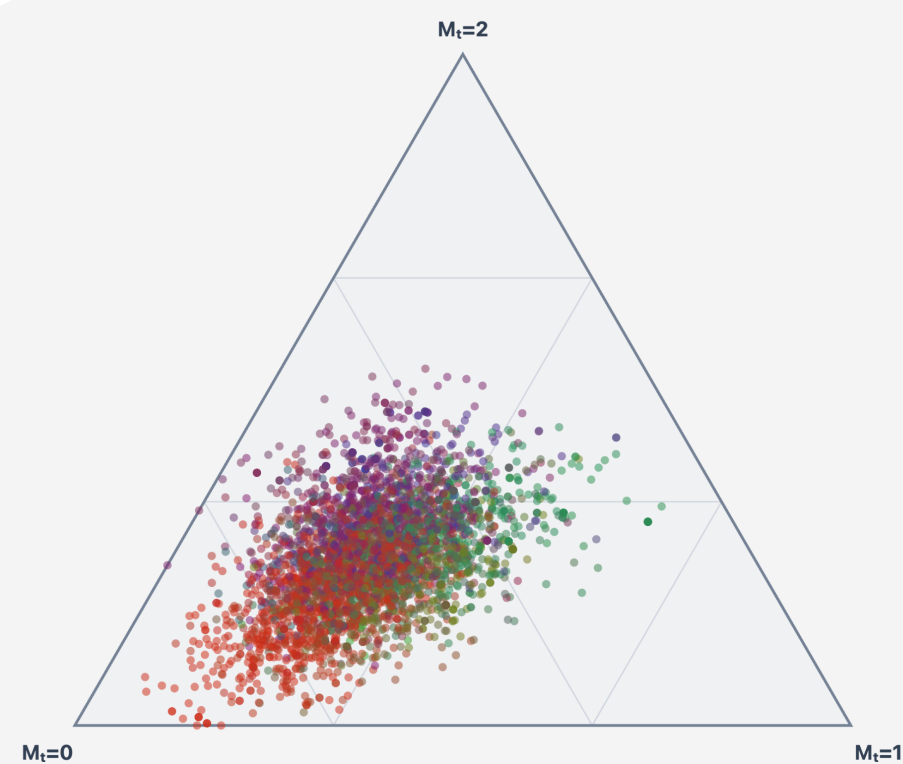
$P(O_{t+1} \mid O_{:t}, A_{:t})$



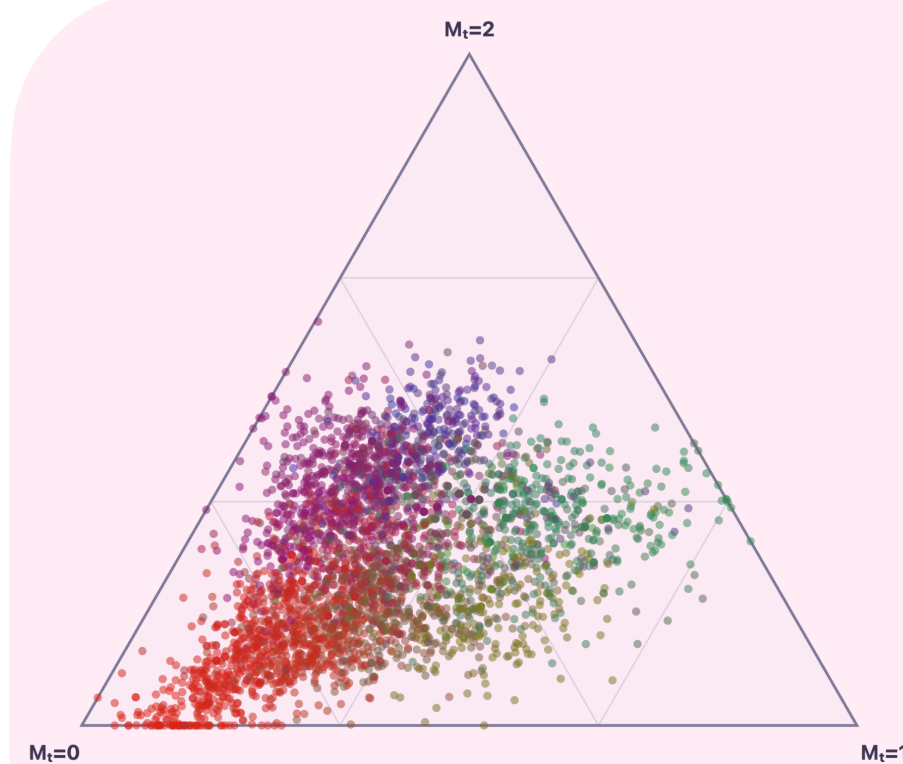
$P(A_t \mid O_{:t}, A_{:t-1})$



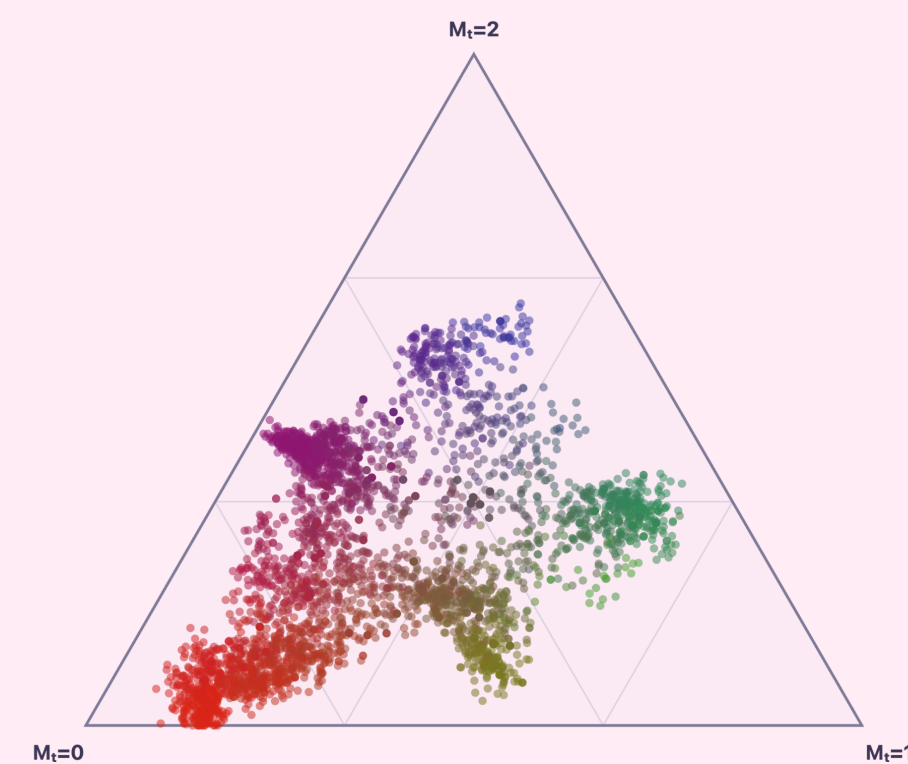
$P(O_{t+1} \mid O_{:t}, A_{:t})P(A_t \mid O_{:t}, A_{:t-1})$



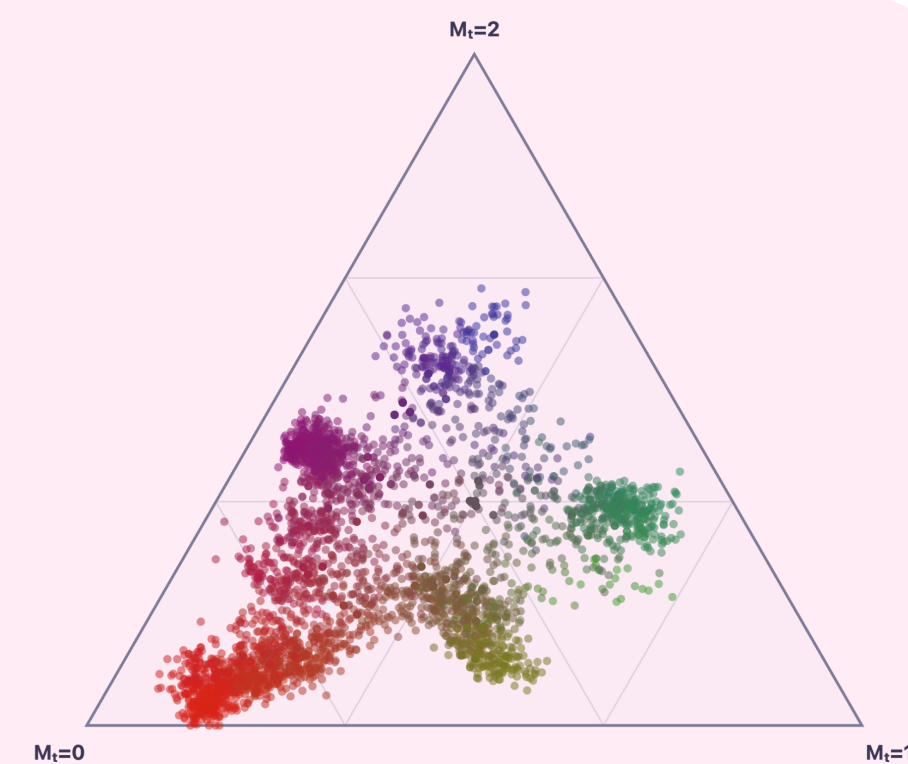
(Training)



$P(O_{t+1} \mid O_{:t}, A_{:t})$



$P(A_t \mid O_{:t}, A_{:t-1})$



$P(O_{t+1} \mid O_{:t}, A_{:t})P(A_t \mid O_{:t}, A_{:t-1})$

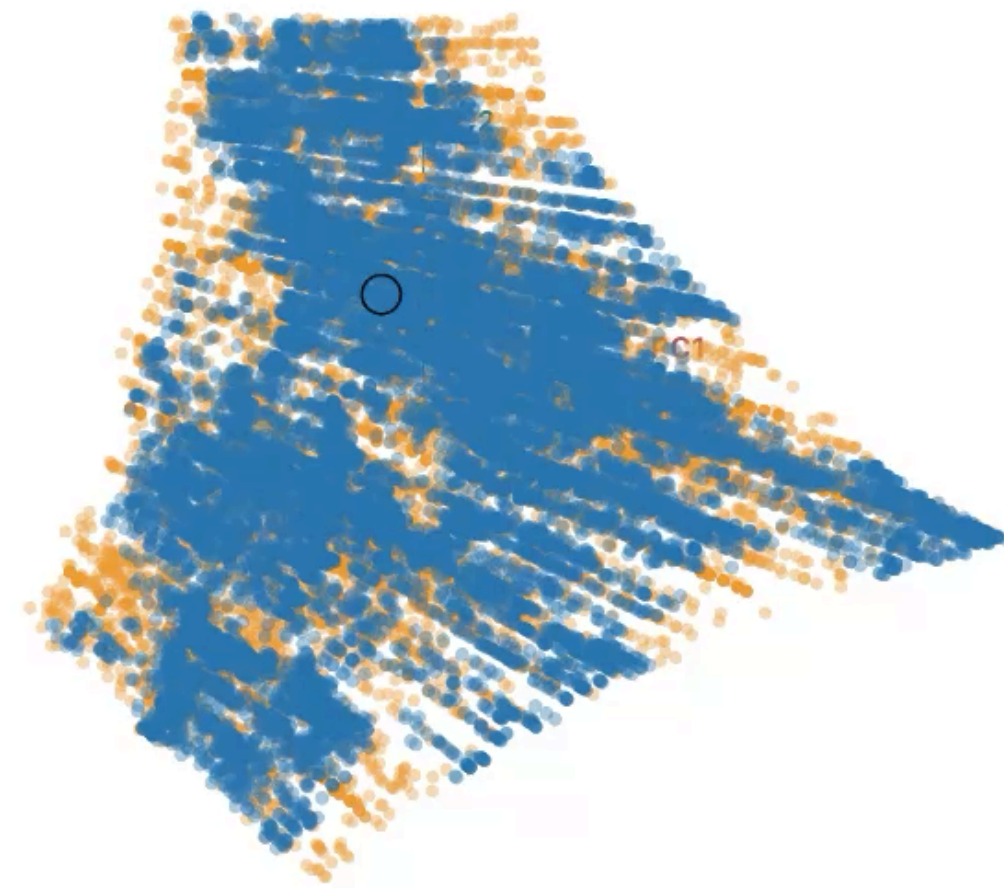
Support restriction, intuitively

Orange: open loop/full support, blue: closed loop/support restricted

Full joint belief · T6 · closed

Affine joint-belief readout · full-theory PCA 1–3

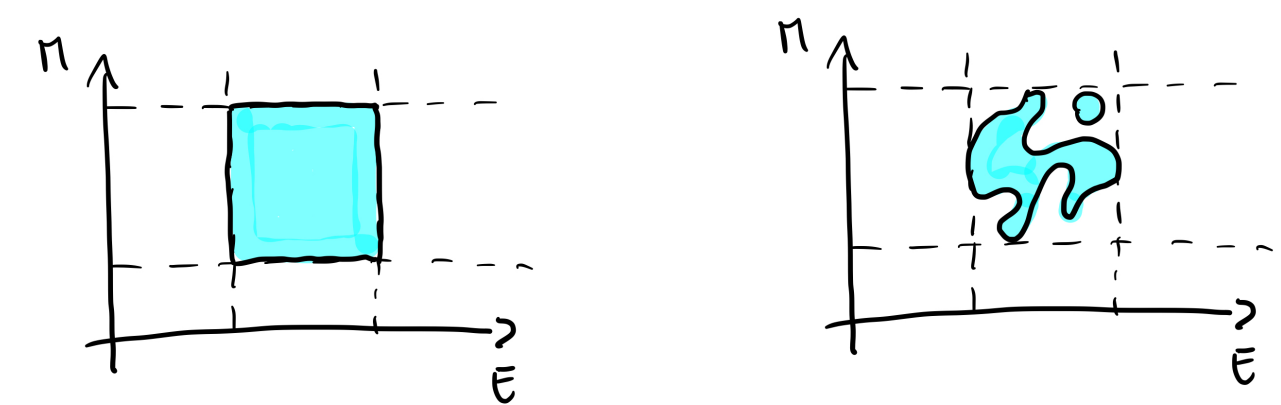
LOSSY THREE-AXIS PROJECTION



LOSSY THREE-AXIS PROJECTION

Correlation on factored beliefs

Reconstruction by a single affine map: not ok vs ok-ish

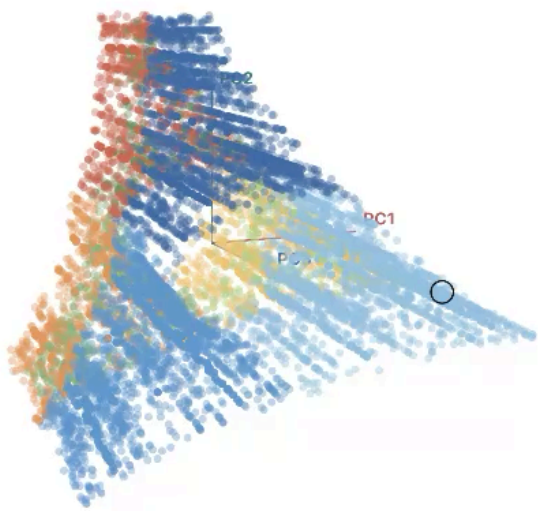


Full joint belief · T3 · open

Affine joint-belief readout · full-theory PCA 1–3

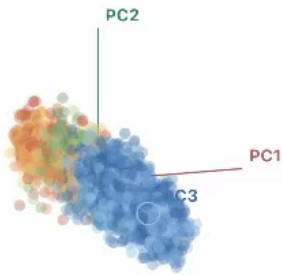
Exact theory

LOSSY THREE-AXIS PROJECTION



Affine joint-belief readout

LOSSY THREE-AXIS PROJECTION

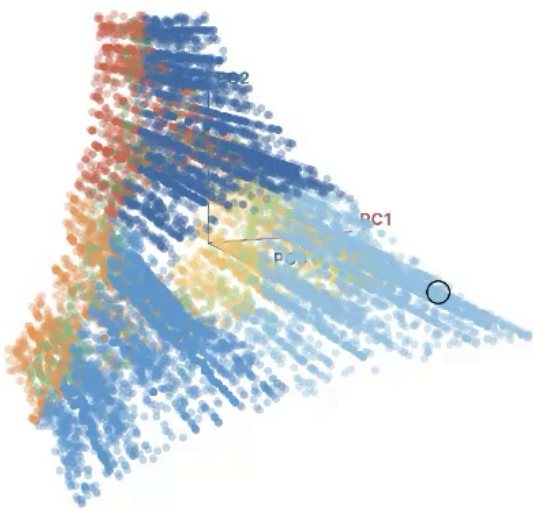


Full joint belief · T3 · open

Product of phase-aligned marginal readouts · full-theory PCA 1–3

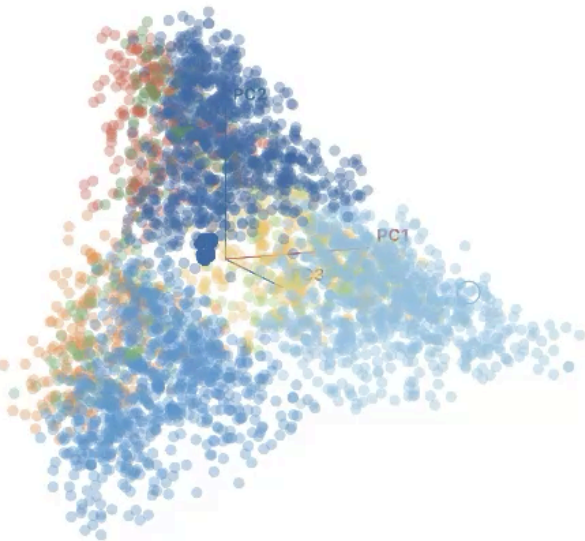
Exact theory

LOSSY THREE-AXIS PROJECTION



Product of phase-aligned marginal readouts

LOSSY THREE-AXIS PROJECTION



LOSSY THREE-AXIS PROJECTION

$$\mathcal{B}_{\text{open}} \approx \mathcal{B}^M \times \mathcal{B}^E$$

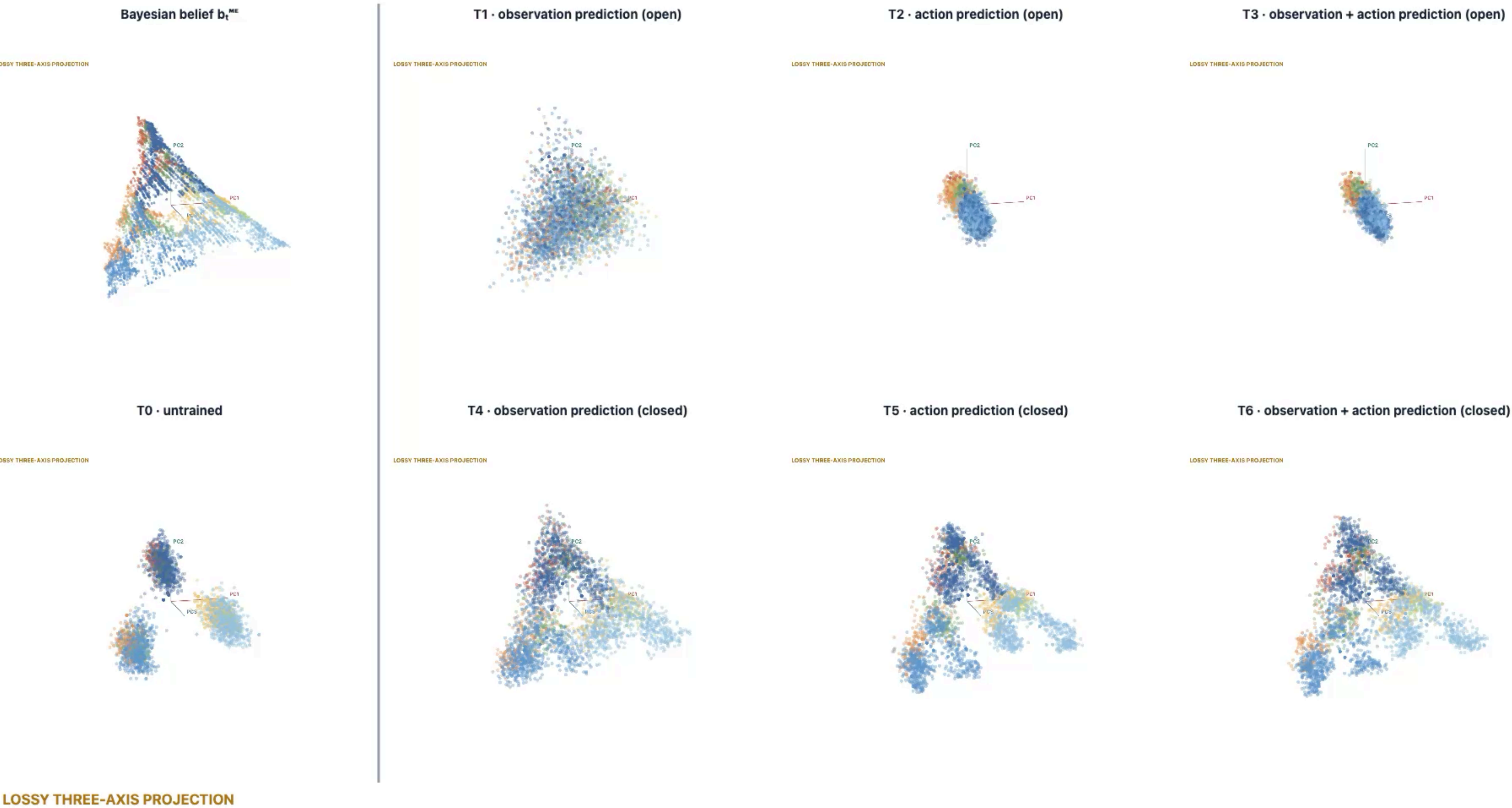
LOSSY THREE-AXIS PROJECTION

$$\mathcal{B}_{\text{closed}} \subsetneq \mathcal{B}^M \times \mathcal{B}^E$$

Reconstruction of closed-loop setup

Full joint belief · Bayesian theory and T0–T6 readouts · closed histories

Affine joint-belief readout · full-theory PCA 1–3



Environment beliefs without explicit training?

Model trained on $P(A_t \mid O_{\cdot t}, A_{\cdot t-1})$

Full joint belief · T5 · three reconstruction methods · open histories

Three reconstruction methods · full-theory PCA 1–3

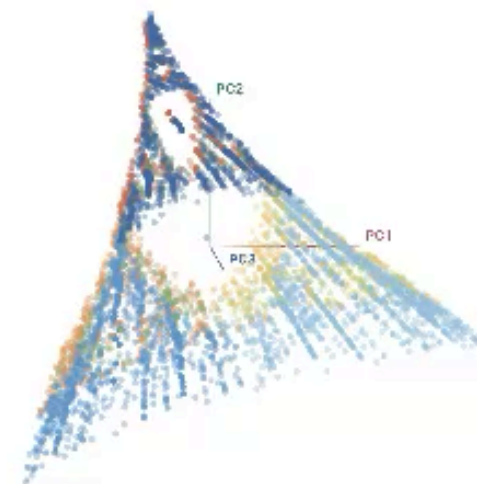
Bayesian belief b_t^{ME}

Affine joint-belief readout

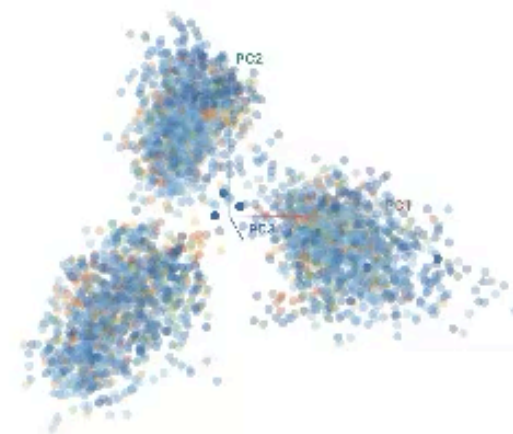
Product of action-position marginal readouts

Product of phase-aligned marginal readouts

LOSSY THREE-AXIS PROJECTION

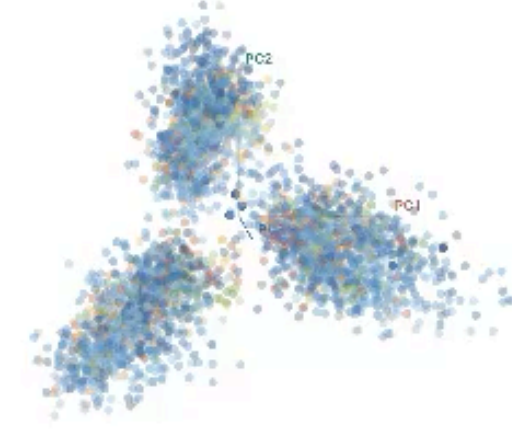


LOSSY THREE-AXIS PROJECTION



Linear regression
from post-B3 to
the joint belief

LOSSY THREE-AXIS PROJECTION



Product of linear
regressions from
post-B2 (env)
and post-final-
normalised-B3
(agent)

LOSSY THREE-AXIS PROJECTION



Product of linear
regressions from
post-B3 (env)
and post-B3
(agent)

Take home messages & Open questions

Learning environment, agent and joint models

Joints made of coupled models have more information
(the coupling)

- Are these agent models self models, in some meaningful sense?
- Does learning about predicting actions (observations) make a model better at predicting observations (actions)?
- Theory: WIP
 - No, if the models are uncoupled
 - ?, if the models are coupled

A “Good Regulator Theorem” for Embodied Agents

Nathaniel Virgo^{1,2}, Martin Biehl³, Manuel Baltieri⁴, and Matteo Capucci⁵

¹University of Hertfordshire, UK ²Earth-Life Science Institute (ELSI), Institute of Science Tokyo, Japan

³Cross Labs, Cross Compass Ltd., Japan ⁴Araya, Inc., Japan ⁵Independent researcher

Abstract

In a classic paper, Conant and Ashby claimed that “every good regulator of a system must be a model of that system.” Artificial Life has produced many examples of systems that perform tasks with apparently no model in sight; these suggest Conant and Ashby’s theorem doesn’t easily generalise beyond its restricted setup. Nevertheless, here we show that a similar intuition can be fleshed out in a different way: whenever an agent is able to perform a regulation task, it is possible for an observer to interpret it as having “beliefs” about its environment, which it “updates” in response to sensory input. This notion of belief updating provides a notion of model that is more sophisticated than Conant and Ashby’s, as well as a theorem that is more broadly applicable. However, it necessitates a change in perspective, in that the observer plays an essential role in the theory: models are not a mere property of the system but are imposed on it from outside. Our theorem holds regardless of whether the system is regulating its environment in a classic control theory setup, or whether it’s regulating its own internal state; the model is of its environment either way. The model might be trivial, however, and this is how the apparent counterexamples are resolved.

work descending from them appear to have reached disparate conclusions: Artificial Life has produced many examples of agents that apparently do not need an internal model in order to achieve their tasks (e.g. Braitenberg, 1986; Beer, 2003), apparently contradicting (C&A).

This is in part because (C&A) doesn’t strictly succeed, even in its own terms, in showing what its title claims. (This has been noted previously in blog posts: Baez, 2016; Wentworth, 2021.) Its theorem doesn’t establish that *every* good regulator is a model but only that some of them are (those that are not “unnecessarily complex” in C&A’s terminology). Although there have been several attempts to generalise it beyond its simple setup, they often seem to need strong assumptions that limit their validity. (See Baltieri et al., 2025 for a recent exposition of the assumptions behind the internal model principle in control theory, for example.)

We claim, however, that these issues can be fixed in a way that removes the need for extra assumptions while also accounting for the apparent counterexamples. To do this, we

A Bayesian Interpretation of the Internal Model Principle

Manuel Baltieri, Araya Inc.[†]

Martin Biehl, Cross Labs, Cross Compass Ltd.

Matteo Capucci, University of Strathclyde and Independent Researcher

Nathaniel Virgo, University of Hertfordshire and
Earth-Life Science Institute, Institute of Science Tokyo

Abstract—The internal model principle, originally proposed in the theory of control of linear systems, nowadays represents a more general class of results in control theory and cybernetics. The central claim of these results is that, under suitable assumptions, if a system (a controller) can regulate against a class of external inputs (from the environment), it is because the system contains a model of the system causing these inputs, which can be used to generate signals counteracting them. Similar claims on the role of internal models appear also in cognitive science, especially in modern Bayesian treatments of cognitive agents, often suggesting that a system (a human subject, or some other agent) models its environment to adapt against disturbances and perform goal-directed behaviour. It is however unclear whether the Bayesian internal models discussed in cognitive science bear any formal relation to the internal models invoked in standard treatments of control theory. Here, we first review the internal model principle and present a precise formulation of it using concepts inspired by categorical systems theory. This leads to a formal definition of “model” generalising its use in the internal model principle. Although the notion of model is not necessary

organisms at all scales, including microorganisms such as bacteria [6]–[8].

In artificial intelligence, the concept of *world model* [9]–[11], closely related to the idea of an internal model, underlies a research programme with applications to reinforcement learning, robotics and deep learning, focusing on learning how to represent hidden properties of the environment [12].

In cognitive science and neuroscience, internal models are broadly thought to constitute the computational basis of perception, motor control and high-level cognitive reasoning [13]–[16], although there is no shortage of debate about this, e.g. [17]–[20]. In the context of neuroscience, internal models are often, though by no means universally, presented under a Bayesian framework. According to the Bayesian view, brains or agents as whole systems, can be thought of as Bayesian reasoners and their cognitive processes as instances