AI in a vat: Fundamental limits of efficient world modelling for agent sandboxing and interpretability

Fernando E. Rosas, Alexander Boyd, Manuel Baltieri

Keywords: World models, agent sandboxing, POMDPs, AI interpretability, AI safety

Summary

While traditionally conceived as tools to improve the task performance of model-based reinforcement learning agents, recent work has proposed *world models* as a way to build controlled virtual environments where AI agents can be thoroughly evaluated before deployment. The efficacy of these approaches, however, critically relies on the ability of world models to accurately represent real environments, which can result on high computational costs that may substantially restrict testing capabilities. Drawing inspiration from the 'brain in a vat' thought experiment, here we investigate methods to simplify world models that remain agnostic to the agent under evaluation. Our results reveal a fundamental trade-off inherent to the construction of world models related to their efficiency and interpretability. Building on this trade-off, we develop approaches that either minimise memory usage, establish the limits on what is learnable, or enable retrodictive analyses tracking the causes of undesirable outcomes. Overall, these results sheds light on the fundamental constraints that shape the design space of world modelling for agent sandboxing and interpretability.

Contribution(s)

1. This paper conceptualises and formalises a novel problem: building efficient world models for an operator to sandbox, evaluate, and interpret AI agents before deployment.
   Context: Prior work (e.g. (Ha & Schmidhuber, 2018; Hafner et al., 2020)) focuses on world models from the perspective of the agent using for boosting performance, and has not considered this safety-inspired perspective.
2. We introduce generalised transducers based on quasi-probabilities, leading to a more efficient approach to compress world models at the expense of their interpretability.
   Context: Generalised transducers are an extension of generalised hidden Markov models, which have been thoroughly studied in previous works (Upper, 1997; Vidyasagar, 2011).
3. We provide a unifying framework to investigate and reason about world models of beliefs, and show that all models that can be calculated by an agent in real time can be bisimulated into a canonical world model known as ϵ-*transducer*.
   Context: The minimality of the ϵ-transducer among prescient rival partitions was proven in (Barnett & Crutchfield, 2015), without investigating links with bisimulation or other concepts from reinforcement learning. Relationships between bisimulation and other computational mechanics constructions were investigated by Zhang et al. (2019).
4. We introduce the notion of *reverse* interpretability, which is related to retrodictive analyses that can identify the roots of undesirable outcomes.
   Context: Standard interpretability approaches assess agents with respect to their capabilities to predict and plan with respect to future events (Nanda et al., 2023; Gurnee & Tegmark, 2023; Shai et al., 2025).
5. We introduce the notion of reversible transducer, and identify necessary and sufficient conditions for its construction. We also introduce and explore the notion of retrodictive beliefs.
   Context: Retrodictive and reversible hidden Markov models have been investigated by Ellison et al. (2009; 2011).

AI in a vat: Fundamental limits of efficient world modelling for agent sandboxing and interpretability

Fernando E. Rosas[1-5], **Alexander Boyd**[1,6], **Manuel Baltieri**[7,1]
f.rosas@sussex.ac.uk, alecboy@gmail.com, manuel_baltieri@araya.org

[1]**Department of Informatics, University of Sussex**
[2]**Sussex AI and Sussex Centre for Consciousness Science, University of Sussex**
[3]**Centre for Complexity Science and Center for Psychedelic Research, Department of Brain Sciences, Imperial College London**
[4]**Center for Eudaimonia and Human Flourishing, University of Oxford**
[5]**Principles of Intelligent Behaviour in Biological and Social Systems (PIBBSS)**
[6]**Beyond Institute for Theoretical Science (BITS)**
[7]**Araya Inc.**

Abstract

Recent work proposes to use world models as controlled virtual environments in which AI agents can be tested before deployment to ensure their reliability and safety. However, accurate world models often have high computational demands that can severely restrict the scope and depth of such assessments. Inspired by the classic 'brain in a vat' thought experiment, here we investigate ways of simplifying world models that remain agnostic to the AI agent under evaluation. By following principles from computational mechanics, our approach reveals a fundamental trade-off in world model construction between efficiency and interpretability, demonstrating that no single world model can optimise all desirable characteristics. Building on this trade-off, we identify procedures to build world models that either minimise memory requirements, delineate the boundaries of what is learnable, or allow tracking causes of undesirable outcomes. In doing so, this work establishes fundamental limits in world modelling, leading to actionable guidelines that inform core design choices related to effective agent evaluation.

1 Introduction

Advances in deep reinforcement learning are progressively enabling AI agents capable of mastering complex tasks across a wide array of domains (Arulkumaran et al., 2017; Wang et al., 2022), and a new generation of agents leveraging large language models (Wang et al., 2024) and large multimodal models (Yin et al., 2024) is expected to drive a new wave of technological innovation with the potential to benefit all sectors of the global economy (Larsen et al., 2024). Alongside these benefits, the proliferation of increasingly advanced autonomous AI systems will also bring important new risks regarding their safety, controllability, and alignment with human values (Bengio et al., 2024; Tang et al., 2024). Given these far-reaching prospects, it is imperative to develop frameworks and methodologies to guarantee the safe and beneficial integration of these technologies to our societies.

One path to pursue AI safety and alignment is to use synthetic world models as sandbox environments to evaluate AI agents without real-world consequences (Dalrymple et al., 2024; Díaz-Rodríguez et al., 2023). These simulated environments are ideal for observing how AI agents handle edge cases and respond to novel situations, potentially revealing safety issues or alignment failures before deployment (He et al., 2024). The efficacy of this approach, however, critically relies on the world model accurately representing relevant aspects of real environments, which is key for guaranteeing that the agent's behaviour in simulation may transfer to real-world settings. Thus, a key

challenge lies in dealing with the computational demands of high-fidelity simulations, whose costs can impose heavy restrictions on the breadth and depth of safety and reliability assessments.

Here we address these issues by investigating the fundamental limits that shape the design of world models. By bridging concepts from reinforcement learning, control theory, and computational mechanics, we identify a fundamental trade-off between the computational efficiency of a world model and its interpretability. This approach also leads to the distinction between *forward* and *reverse* interpretability approaches, where the former characterises the predictive capabilities of agents, and the latter enables retrodictive analyses of the origins of undesirable outcomes. Overall, this work establishes foundational groundwork that leads to actionable guidelines for building world models to study AI agents following different desiderata.

2 Scenario and approach

Representation and what is represented belong to two completely different worlds.

Hermann von Helmholtz, *Handbuch der physiologischen Optik* (1867)

Consider the design of a world model to sandbox and test an AI agent (Dalrymple et al., 2024). What should this world model look like? What information should it encode? And for what purpose?

To ensure a reliable assessment of AI agent behaviour from simulations to a real-world setting, world models must faithfully reflect the world's structure and dynamics. This could be seen as suggesting that designing reliable world models is critically bounded by a trade-off between accuracy and computational tractability. Interestingly, this trade-off can be partially circumvented by recognising that effective world models only need to incorporate variables that make a difference for the AI's actions, and these variables only require a granularity that is sufficient to simulate their dynamics.

To illustrate this idea, consider how one could choose to build a world model to sandbox a simple robot. Although one could in principle design a simulation that includes the quantum dynamics of the whole planet, such a simulation would not only be computationally unfeasible but also unnecessary to answer most questions of interest. Indeed, such a world model would likely be too spatially extended (by including regions of the planet that are inaccessible to the robot) and have too much resolution (by including quantum effects for a fundamentally classical robot). To avoid this, one could instead design a more parsimonious world model that factors out indistinguishable properties from the robot's perspective, focusing on the agent's 'interface' including, for instance, sensorimotor contingencies (O'Regan & Noë, 2001; Baltieri & Buckley, 2017; 2019; Tschantz et al., 2020; Mannella et al., 2021) or task-relevant information (Zhang et al., 2021).

Related questions have been extensively investigated in the philosophy of mind and cognitive (neuro)science literature, and more recently in reinforcement learning. These works suggest an important insight: while an agent's actions turn into outcomes through the mediation of the external world, the agent lacks direct access to the world's 'true nature' and only interacts with it via its inputs and outputs (Clark, 2013; Seth & Tsakiris, 2018). This notion is illustrated by the classical *'brain in a vat'* thought experiment, which proposes that if an organism's brain were to be placed in a vat, and a computer used to read the brain's output signals and generate plausible sensory signals, then the brain may not be able to tell it is in fact in a vat.[1] This thought experiment suggests that an ideal world model should depend only on three elements: the set of possible actions of the agent $\mathcal{A}$, the set of possible outcomes affecting the agent $\mathcal{Y}$,[2] and the statistical relationship between action sequences and outcomes. In fact, it should be possible — at least in principle — to build a compressed representation of the 'effective world' of an AI agent that cannot be distinguished from a full simulation, irrespectively of how smart or powerful it may be.

[1] The modern form of this thought experiment is due to Putnam (1981), but has roots in Descartes' 'evil demon' (Descartes, 1641) and Plato's cave allegory (Plato, 375 BC) while serving as inspiration for popular media such as *The Matrix* movies.

[2] The outcome may be a combination of a quantity observable by the agent and a reward signal, so that $\mathcal{Y} = \mathcal{O} \times \mathbb{R}$.

These ideas can be operationalised using principles from computational mechanics (Crutchfield, 1994; 2012), which reveal how observable processes can be generated by multiple data-generating procedures (see App. A for an example). Embracing this multiplicity leads to a perspective that we playfully label as *'AI in a vat'*, which posits that designers should not focus on a single world model, but instead should (i) consider the class of all world models that are indistinguishable from the AI agent's perspective, (ii) characterise their properties, and then (iii) choose one depending on specific priorities. After setting some formal foundations in Sec. 3, the remainder of this paper discusses design choices and procedures that allow the building of world models serving distinct goals related to AI safety and alignment (see Figure 1):

- *Computational efficiency* (Sec. 4): sandboxing agents to evaluate their behaviour or provide formal guarantees about their capabilities (Dalrymple et al., 2024) using a minimal amount of resources.
- *Forward interpretability* (Sec. 5): building models to study which features are learnable by agents, and how representations are encoded inside them (Shai et al., 2025).
- *Reverse interpretability* (Sec. 6): deploying models that can be run backward in time to investigate the origins and tipping points that lead to specific — desirable or undesirable — outcomes.

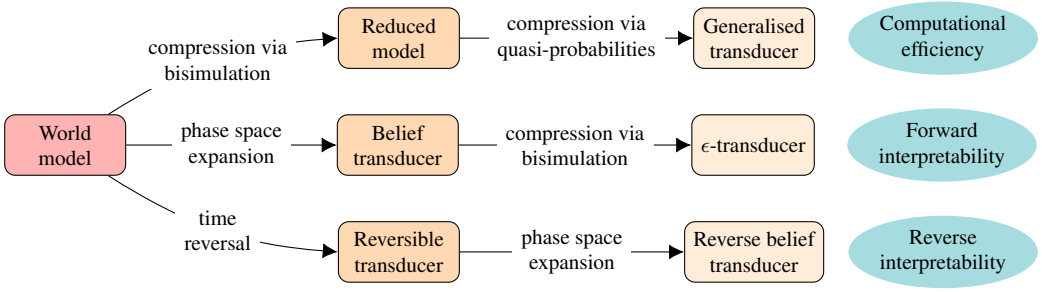

Figure 1: Recommendations for building world optimal models, including implementations (boxes), transformations (arrows), and design criteria (ellipses).

3 Generating interfaces via transducers

This section formalises the notions of 'world model' and 'interface'. In the following, uppercase letters (e.g. X, Y) are used to denote random variables and lowercase (e.g. x, y) their realisations, calligraphic letters (e.g. $\mathcal{X}, \mathcal{Y}$) denote the sets over which they take values, and the symbol Δ (as in $\Delta(\mathcal{X}), \Delta(\mathcal{Y})$) is used to denote the collection of all distributions over those sets. We use the shorthand notation $p(x|y) = \Pr(X = x|Y = y)$ to express probabilities when there is no risk of ambiguity, and assume that equalities of the form $p(x|y, z) = p(x|y)$ hold for all realisations that can take place with non-zero probability. $\mathbb{N} = \{0, 1, 2, \ldots\}$ corresponds to zero-based numbering, and we use the following abbreviations: $\boldsymbol{x}_{a:b} = (x_a, \ldots, x_b)$, $\boldsymbol{x}_{:b} = \boldsymbol{x}_{0:b}$, $\boldsymbol{x}_{a:} = \boldsymbol{x}_{a:\infty}$, and $\boldsymbol{x}_{:} = \boldsymbol{x}_{0:\infty}$.

3.1 World models

We operationalise interfaces as descriptions of how actions turn into outcomes for a particular agent.

Definition 1. *An **interface** $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$ is a collection of distributions $\{p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:}), t \in \mathbb{N}\}$ corresponding to a stochastic process over outcome sequences $\boldsymbol{y}_{:} \in \mathcal{Y}^{\mathbb{N}}$ conditioned on action sequences $\boldsymbol{a}_{:} \in \mathcal{A}^{\mathbb{N}}$. An interface is **anticipation-free** if $p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:}) = p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:t})$ for all $t \in \mathbb{N}$.*

Essentially, an interface describes a semi-infinite stochastic process (Kallenberg, 1997; Loomis & Crutchfield, 2023) for each action sequence $\boldsymbol{a}_{:}$ while being agnostic to the agent's computational capabilities, architecture, or internal functioning. Interfaces can be constructed from an underlying world model that specifies how actions turn into outcomes. Next, we introduce a general notion of a world model in terms of statistical sufficiency (App. B), and use $h_t = (a_t, y_t)$ so that $\boldsymbol{h}_{:t}$ denotes the joint history of the interface up to time t.

Definition 2. *A **world model** for an interface $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$ is a collection of distributions $\{p(\boldsymbol{s}_{:t}|\boldsymbol{h}_{:}), t \in \mathbb{N}\}$ corresponding to a stochastic process over state sequences $\boldsymbol{s}_{:} := (s_0, s_1, \dots) \in \mathcal{S}^{\mathbb{N}}$ that satisfies*

$$(1)\ p(y_t|\boldsymbol{s}_{:t}, \boldsymbol{a}_{:t}, \boldsymbol{y}_{:t-1}) = p(y_t|s_t, a_t) \quad \textit{and} \quad (2)\ p(\boldsymbol{s}_{:t}|\boldsymbol{h}_{:t-1}, \boldsymbol{a}_{t:}) = p(\boldsymbol{s}_{:t}|\boldsymbol{h}_{:t-1}) \quad \forall t \in \mathbb{N}. \tag{1}$$

Thus, world models are candidate mechanisms for implementing the statistical relationships between actions and outcomes, taking the form of auxiliary processes S_t that encapsulate relevant information between past events and present outcomes (condition 1) while guaranteeing time's arrow so future actions cannot affect previous world states or outcomes (condition 2). We may informally denote a world model simply by S_t when it is unambiguous from context. This definition includes models of the 'external world' such as *partially observed Markov decision processes* (POMDPs), as well as their 'epistemic' counterparts, *belief MDPs* (Kaelbling et al., 1998), as explained in Sec. 3.3. The unifying property of all world models is presented next (proof in App. C).

Lemma 1. *A process S_t is a world model for an anticipation-free interface $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$ if and only if*

$$p(\boldsymbol{y}_{:t}, \boldsymbol{s}_{:t+1}|\boldsymbol{a}_{:}) = p(s_0) \prod_{\tau=0}^{t} p(y_\tau|s_\tau, a_\tau) p(s_{\tau+1}|\boldsymbol{s}_{:\tau}, \boldsymbol{h}_{:\tau}), \quad \forall t \in \mathbb{N}. \tag{2}$$

Thus, world models let us express interfaces in terms of probabilistic graphical models (Koller & Friedman, 2009). Among other things, Eq. (2) can be used to generate outcomes for given sequences $\boldsymbol{a}_{:t}$ and $\boldsymbol{s}_{:t}$ by sampling $p(\boldsymbol{y}_{:t}|\boldsymbol{s}_{:t}, \boldsymbol{a}_{:t}) = \prod_{\tau=0}^{t} p(y_\tau|s_\tau, a_\tau)$. In this sense, we say that the world model S_t ***generates*** the interface $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$, and that the graphical model outlined in Eq. (2) establishes a ***presentation*** of the interface. These ideas are illustrated by an example in App. A.

3.2 Transducers

Sampling sequences of world model's states can be highly non-trivial if their dynamics are non-Markovian. This issue can be avoided by restricting ourselves to building world models using *transducers* (Barnett & Crutchfield, 2015), a computational structure that is introduced next.

Definition 3. *A **transducer** is a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{Y}, \kappa, \mathrm{p})$, where $\mathcal{S}$ is a set of memory states, $\mathcal{A}$ and $\mathcal{Y}$ are sets of inputs and outputs, $\kappa : \mathcal{A} \times \mathcal{S} \times \mathbb{N} \to \Delta(\mathcal{Y} \times \mathcal{S})$ is a Markov kernel of the form $\{\kappa_\tau(y, \tilde{s}|a, s) : s, \tilde{s} \in \mathcal{S}, y \in \mathcal{Y}, a \in \mathcal{A}, \tau \in \mathbb{N}\}$, and $\mathrm{p} \in \Delta(\mathcal{S})$ is an initial distribution over states.*

If a transducer's memory can only take $|\mathcal{S}| = n$ different states, then their transitions can be described via substochastic matrices $T_\tau^{(y|a)}$ of the form

$$T_\tau^{(y|a)} := \sum_{i=1}^{n} \sum_{j=1}^{n} \kappa_\tau(y, s_i|a, s_j) \boldsymbol{e}_i \boldsymbol{e}_j^\intercal\ , \tag{3}$$

where $\boldsymbol{e}_k$ is a binary vector with a 1 at the k-th position and zeros elsewhere. These transducers are also known as stochastic automata (Claus, 1971; Cakir et al., 2021), generalising deterministic automata (Minsky, 1967) by using stochastic transitions to generate outputs and update their state. Running a transducer generates an interface $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$ given by its inputs and outputs according to

$$p(\boldsymbol{y}_{:t}\boldsymbol{s}_{:t+1}|\boldsymbol{a}_{:}) = p(s_0) \prod_{\tau=0}^{t} \kappa_\tau(y_\tau, s_{\tau+1}|a_\tau, s_t), \tag{4}$$

providing a graphical model that can be used to simulate the interface (see Figure 2). Comparing Eq. (4) with Lemma 1 let us to formalise this fact as follows.

Lemma 2. *Transducers correspond to world models of anticipation-free interfaces whose dynamics satisfy the Markov condition $p(s_{\tau+1}|\boldsymbol{s}_{:\tau}, \boldsymbol{h}_{:\tau}) = p(s_{\tau+1}|s_\tau, h_\tau)$ for all $\tau \in \mathbb{N}$.*

We may denote a transducer informally as (S_t, A_t, Y_t) when it is unambiguous from the context, and describe its memory state S_t as a world model when appropriate. Complementary characterisations of transducers in terms of sufficient statistics and information properties are provided in App. D.

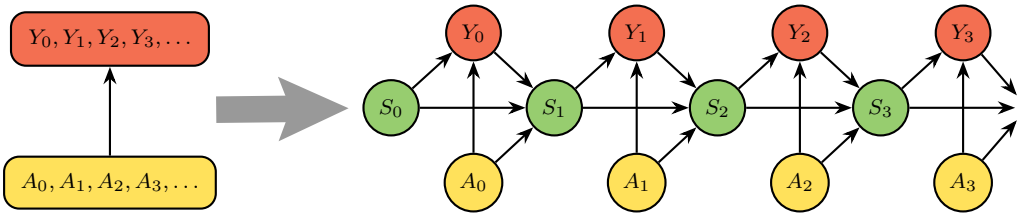

Figure 2: Illustration of an interface (left) and a possible unravelling of it via a presentation with a world model built from the memory states of a transducer (right), as given by Eq. (4).

Transducers can be seen as reflecting the memory structure of interfaces. In particular, an interface $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$ is said to be ***memoryless*** if $S_t = 0$ is a valid transducer, and is ***fully* observable** if $S_t = Y_t$ yields a valid transducer — including Markov decision processes (MDPs) as a main example. We next characterise the statistics of such interfaces, which clarifies that elaborate world models are required by interfaces with non-Markovian dynamics (proof in App. E).

Lemma 3. *An interface is fully observable if and only if* $p(y_{t+1}|\boldsymbol{y}_{:t}, \boldsymbol{a}_{:}) = p(y_{t+1}|y_t, a_t)$*, and is memoryless if and only if* $p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:}) = \prod_{\tau=0}^{t} p(y_\tau|a_\tau)$.

Note that if $p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:}) = p(\boldsymbol{y}_{:t})$, corresponding to 'contemplative' or passive agents that do not act but only sense, then transducers reduce to hidden Markov models (Ephraim & Merhav, 2002).

3.3 General classes of transducers

Transducers use their kernels κ_τ to generate (s_{t+1}, y_t) jointly from (s_t, a_t), corresponding to what has been described as a 'Mealy' machines (Virgo, 2023; Bonchi et al., 2024). Simpler computational structures can be obtained by imposing constraints in the kernel as follows:

- ***Input-Moore transducers*** generate outputs ignoring the current input, corresponding to kernels of the form $\kappa_\tau(y, \tilde{s}|a, s) = \mu_\tau(y|s)\nu_\tau(\tilde{s}|y, a, s)$.
- ***Output-Moore transducers*** update their state without considering the current output, corresponding to kernels of the form $\kappa(y, \tilde{s}|a, s) = \mu_\tau(y|a, s)\nu_\tau(\tilde{s}|a, s)$.
- ***I-O Moore transducers*** satisfy both previous conditions, corresponding to kernels of the form $\kappa(y, \tilde{s}|a, s) = \mu_\tau(y|s)\nu_\tau(\tilde{s}|a, s)$.

Based on ideas from automata theory (Lee & Seshia, 2017), input-Moore transducers serve as models for interfaces that satisfy the stronger anticipation-free condition $p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:}) = p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:t-1})$, corresponding to scenarios where Y_t takes place before A_t, thus reflecting a time-indexing convention. In contrast, output-Moore transducers build on the hidden Markov models literature (Riechers, 2016), and are used to represent (non-quantum) physical processes whose evolution is not affected by the observations made by the agent — as opposed to epistemic processes such as belief MPDs, in which the opposite happens (see Sec. 5). POMDPs can be shown to be examples of either output-Moore or I-O Moore transducers depending on the specific definition, as explained in App. F.

4 Minimal world models

After setting the formal foundations of world models, and establishing transducers as a natural way to construct them, we now investigate how to build *minimal* world models.

4.1 Reducing world models

We begin by showing that all interfaces have at least one transducer presentation, and hence one can focus on transducers without loss of generality (see the proof in App. G).

Lemma 4. *The world model* $S_t = \boldsymbol{H}_{:t-1}$ *yields a transducer that generates the interface* $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$.

This world model is far from parsimonious, resembling Borges' character *Funes the memorious* in its inability to forget. This suggests the importance of 'reducing' world models. To formalise this, we extend the notion of MDP homomorphism (Ravindran & Barto, 2003) to transducers as follows.

Definition 4. *A **homomorphism** between transducers (S_t, A_t, Y_t) and (S'_t, A'_t, Y'_t) is given by the mappings $\langle \phi : \mathcal{S} \to \mathcal{S}', f : \mathcal{Y} \to \mathcal{Y}', g : \mathcal{A} \to \mathcal{A}' \rangle$ satisfying two compatibility conditions:*

(i) $\Pr\big(Y'_t = y' | S'_t = \phi(s), A'_t = g(a)\big) = \sum_{y \in f^{-1}(y')} \Pr\big(Y_t = y | S_t = s, A_t = a\big)$.

(ii) $\Pr\big(S'_{t+1} = s' | S'_t = \phi(s), H'_t = (f(y), g(a))\big) = \sum_{s \in \phi^{-1}(s')} \Pr\big(S_{t+1} = s | S_t = \tilde{s}, H_t = (y, a)\big)$ *and* $\Pr\big(S'_0 = s'\big) = \sum_{s \in \phi^{-1}(s')} \Pr\big(S_0 = s\big)$.

*A **reduction** of a world model S_t into a world model S'_t is a homomorphism $\langle \phi, f, g \rangle$ between the transducers (S_t, A_t, Y_t) and (S'_t, A_t, Y_t), where $f : \mathcal{Y} \to \mathcal{Y}$ and $g : \mathcal{A} \to \mathcal{A}$ are identity mappings and ϕ is a surjective map $\phi : S_t \to S'_t$. Two transducers are **isomorphic** if they are reductions of each other, and a transducer is **minimal** if all its reductions are isomorphic to itself.*

A world reduction can be informally described as a coarse-graining ϕ between the memory states of two transducers of the same interface. Condition (i) above ensures that outcomes are generated with the same statistics, and (ii) that the resulting world model is still Markovian — as can be confirmed by relating it with the notion of 'lumpability' of Markov chains (Tian & Kannan, 2006). These properties guarantee that reductions do not distort the corresponding interface (proof in App. H).

Lemma 5. *A world model reduction yields a transducer presentation of the original transducer.*

The next two sections study different approaches to look for minimal world models.[3]

4.2 Reduction via bisimulation

A natural way to reduce a world model is via the notion of bisimulation, which is a way of merging states that behave equivalently (Givan et al., 2003). Here we leverage previous work on bisimulations for hidden Markov models (Jansen et al., 2012) to define bisimulations of transducers.

Definition 5. *For a given transducer with world model S_t and kernel κ_t, a **bisimulation** is an equivalence relation $\mathcal{B}_t \subseteq \mathcal{S} \times \mathcal{S}$ such that $s, s' \in \mathcal{S}$ are equivalent if they satisfy two conditions:*

(i) $p_t(y|s, a) = p_t(y|s', a)$, *where* $p_t(y|s, a) = \sum_{s'' \in \mathcal{S}} \kappa_t(y, s''|s, a)$.

(ii) $p_t(C|s, a) = p_t(C|s', a)$ *for all equivalence classes* $C \subseteq \mathcal{S}$, *where* $p_t(C|s, a) = \sum_{y \in \mathcal{Y}} \sum_{s'' \in C} \kappa_t(y, s''|s, a)$.

There is a direct correspondence between world model reductions (Def. 4) and bisimulations, as shown next by adapting (Taylor et al., 2008, Theorem 3) to our setup (proof in App. I).

Proposition 1. $\phi : S_t \to S'_t$ *is a reduction of world models if and only if the equivalence relation it induces, with equivalence classes given by* $\phi^{-1}(s') = \{s \in \mathcal{S} : \phi(s) = s'\}$, *is a bisimulation.*

Together with Lemma 5, this result confirms that the bisimulation of a transducer yields another transducer presentation for the same interface. This has a simple and yet powerful implication: a full reduction of a given transducer can be attained by coarse-graining all bisimilar states.

There may be cases where bisimulations do not produce the most efficient world model that generates a given interface, since reducing a particular transducer usually does not lead to a global minimum. To investigate this claim, we consider a world model with $|\mathcal{S}| = n$ states and build vectors $\boldsymbol{w}(\boldsymbol{h}_{:t}) \in \mathbb{R}^n$ containing the probabilities of generating $\boldsymbol{y}_{:t}$ given $\boldsymbol{a}_{:t}$ when starting from different world states, so that its k-th coordinate is $[\boldsymbol{w}(\boldsymbol{h}_{:t})]_k = \Pr(\boldsymbol{Y}_{:t} = \boldsymbol{y}_{:t} | \boldsymbol{A}_{:t} = \boldsymbol{a}_{:t}, S_0 = s_k)$. Intuitively, if vectors $\boldsymbol{w}(\boldsymbol{h}_{:t})$ for different $\boldsymbol{h}_{:t}$ are linearly dependent, some of their dimensions (and, hence, the corresponding world states) are still in some sense redundant. Crucially, the coarse-grainings associated to bisimulation can only lump states that have identical components, but cannot

[3] Minimality can also be studied via the entropy of the world's dynamics. Interestingly, minimal entropy models may not coincide with the models with fewer states — although the two coincide for predictive models (Loomis & Crutchfield, 2019).

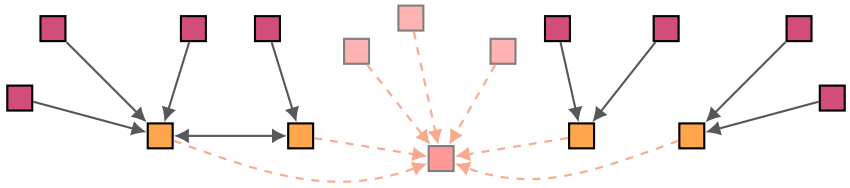

Figure 3: Illustration of the minimisation of world models. Purple boxes represent reducible models and orange boxes represent minimal ones, and arrows correspond to reductions. Red boxes are generalised models following quasi-probabilities, which (if allowed) establish global minima.

reduce linear dependencies between states more generally. Relaxing the criteria for merging states (e.g. via bisimulation metrics (Ferns & Precup, 2014)) does not solve this issue, as this introduces distortion in the interface due to the imprecisions allowed in the state merging procedure.

These ideas can be made concrete by studying the *canonical dimension* of a transducer $\mathcal{T}$, given by[4]

$$d(\mathcal{T}) \coloneqq \lim_{m\to\infty} \dim(U_m), \qquad \text{where} \quad U_m = \text{Span}\{\boldsymbol{w}(\boldsymbol{h}_{:t}) : t \leq m\} \subseteq \mathbb{R}^n. \tag{5}$$

The canonical dimension is an important indicator of the compressibility of a transducer as shown next, whose proof can be found in (Cakir et al., 2021, Cor. 4.9) — also see (Ito et al., 1992; Balasubramanian, 1993) for related results in hidden Markov models.

Theorem 1. *If $\mathcal{T}$ is a transducer with $|\mathcal{S}| = n \in \mathbb{N}$, then $d(\mathcal{T}) = n$ implies that there are no transducers with fewer memory states that can generate the same interface.*

The minimal bisimulation of a transducer $\hat{\mathcal{T}}$ with world states in $\hat{\mathcal{S}}$ could still exhibit $d(\hat{\mathcal{T}}) < |\hat{\mathcal{S}}|$. In fact, there are interfaces for which no transducer reaches $d(\mathcal{T}) = |\mathcal{S}|$. Even if there exists a transducer with $d(\hat{\mathcal{T}}) = |\mathcal{S}|$, we are not aware of any general algorithm that can directly build it.[5]

4.3 Reduction via pseudo-probabilities and generalised transducers

This section focuses on the reduction of world models with a finite number of states $|\mathcal{S}| = n$ but $d(\mathcal{T}) < n$. As discussed in Sec. 3.2, the probabilities of $\boldsymbol{y}_{:t}$ given $\boldsymbol{a}_{:t}$ can then be calculated as

$$p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:t}) = \mathbf{1}^\intercal \cdot \left(\prod_{\tau=0}^{t} T_\tau^{(y_\tau|a_\tau)} \right) \cdot \boldsymbol{p}, \tag{6}$$

where $\mathbf{1}^\intercal$ is a transposed vector with n ones as components. Normally, the substochastic matrices $T_t^{(y|a)}$ and the initial distribution $\boldsymbol{p}$ are assumed to contain only non-negative terms. However, a more general class of transducers can be explored by removing this constraint and considering *quasi-distributions* $\boldsymbol{v} \in \mathbb{R}^n$, which may have negative components but still satisfy $\sum_{i=1}^n v_i = 1$, and quasi-stochastic matrices whose columns are quasi-distributions (Balasubramanian, 1993; Upper, 1997). This leads to the following generalisation of a transducer.

Definition 6. *A **generalised transducer** for an interface $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$ is a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{Y}, \{A^{(y|a)}\}, \boldsymbol{v}, \boldsymbol{u})$ with $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ and $A^{(y|a)} \in \mathbb{R}^{n\times n}$ that satisfy*

$$p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:t}) = \boldsymbol{u}^\intercal \cdot \left(\prod_{\tau=0}^{t} A_\tau^{(y_\tau|a_\tau)} \right) \cdot \boldsymbol{v} \qquad \forall \boldsymbol{y}_{:t} \in \mathcal{Y}^{t+1}, \boldsymbol{a}_{:t} \in \mathcal{A}^{t+1}. \tag{7}$$

Generalised transducers are useful because, in contrast to standard transducers (or POMDPs), they can always be reduced to find representations with a minimal number of states, as shown next.

[4] If a transducer has $|\mathcal{S}| = n$ memory states, then $\lim_{m\to\infty} \dim(U_m) = \dim(U_{n-1})$ (Cakir et al., 2021, Prop. 4.3).

[5] In fact, the relatively simpler case of reducing hidden Markov models is still not fully solved (Vidyasagar, 2011), although algorithms that can address some cases have been developed (Huang et al., 2015; Ohta, 2021).

Theorem 2. *A generalised transducer $\tilde{T}$ with $d(\tilde{T}) < n$ can always be reduced via a linear transformation into another transducer that generates the same interface using fewer states.*

This result follows directly from the proofs provided in (Balasubramanian, 1993, Ch. 3) and related results found in (Upper, 1997; Vidyasagar, 2011), in which reductions correspond to linear projections. These proofs lead to practical algorithms that can be used to effectively reduce transducers with $d(\tilde{T}) < n$ (Huang et al., 2015). In this way, generalised transducers achieve a minimal computational complexity at the cost of introducing world models whose trajectories cannot be sampled (due to the quasi-probabilities), resulting in a substantial lack of interpretability.

5 Forward interpretability via epistemic world models

The previous section shows how computational efficiency can be achieved by either compressing memory state spaces with bisimulations or by allowing memory states of transducers to be encoded by quasi-probabilities. While the latter generally yields higher efficiency, this comes at the cost of making those reduced world models highly uninterpretable due to the possible presence of negative probabilities. This section takes a different route by investigating specific types of world models that focus on interpretability, bringing insights about what agents can learn.

5.1 Beliefs as world models

Let us start by reviewing properties of certain classes of world models that make them learnable by agents in real time. A world model S_t is ***predictive*** if it contains only past information, which in information-theoretic terms corresponds to $I(S_t; \boldsymbol{Y}_{t:}|\boldsymbol{H}_{:t-1}, \boldsymbol{A}_{t:}) = 0$. A world model is ***observable*** if $S_{t+1} = f_t(\boldsymbol{H}_{:t})$, i.e. if it can be estimated from action-output history via mappings f_t. Finally, a world model is ***unifilar*** if $S_{t+1} = \hat{f}_t(S_t, A_t, Y_t)$, so its state can be deterministically updated given inputs and outputs. Thus, observable models are always predictive, and unifilar models are observable if there is no randomness in the world's initial condition. Moreover, unifilar models correspond to transducer whose kernels have the form $\kappa_\tau(y, \tilde{s}|a, s) = \delta^{\tilde{s}}_{\hat{f}_\tau(y,a,s)}\mu_\tau(y|a, s)$.

The literature contains several procedures for building observable world models from non-observable ones (see (Subramanian et al., 2022; Ni et al., 2024) for general reviews and (Virgo et al., 2021; Biehl & Virgo, 2022; Virgo, 2023) for a categorical formulation). These approaches suggest to expand the phase space of world models from elements in $\mathcal{S}$ to distributions over those $\Delta(\mathcal{S})$, henceforth called *belief states*. This idea has been extensively studied for POMDPs via the notion of *belief MDP* (Kaelbling et al., 1998). We extend these ideas to more general transducers.

Definition 7. *A **belief transducer** over a transducer $(\mathcal{S}, \mathcal{A}, \mathcal{Y}, \kappa, \mathrm{p})$ is another transducer $(\mathcal{B}, \mathcal{A}, \mathcal{Y}, \hat{\kappa}, \delta_{\mathrm{b}_0})$ where $\mathcal{B} \subseteq \Delta(\mathcal{S})$ is a set of belief states, $\hat{\kappa} : \mathcal{A} \times \mathcal{B} \times \mathbb{N} \to \Delta(\mathcal{Y} \times \mathcal{B})$ is a Markov kernel of the form $\{\hat{\kappa}_\tau(y, b'|a, b) : b, b' \in \mathcal{B}, a \in \mathcal{A}, y \in \mathcal{Y}, \tau \in \mathbb{N}\}$ such that $\hat{\kappa} = F\{\kappa\}$ for some functional F, and $\mathrm{b}_0 \in \mathcal{B}$ is an initial belief. A belief transducer is said to be **faithful** if it generates the same interface as the original transducer.*

A natural way to define beliefs about an underlying world model S_t is via ***predictive Bayesian beliefs*** corresponding to the posterior distribution $\mathrm{b}_t(s_t) \coloneqq \Pr(S_t = s_t|\boldsymbol{H}_{:t-1} = \boldsymbol{h}_{:t-1})$. The dynamics of the updates of such beliefs are described by Bayesian prediction (Jazwinski, 1970; Särkkä & Svensson, 2023), and their properties have been further studied under the name of 'mixed-states' in computational mechanics (Riechers & Crutchfield, 2018; Jurgens & Crutchfield, 2021b). Building on this literature, we show that the predictive Bayesian beliefs of the memory states of transducers are unifilar and can be used to generate the same interface (proof in App. J).

Proposition 2. *If (S_t, A_t, Y_t) is a transducer and B_t is the predictive Bayesian belief of S_t, then (B_t, A_t, Y_t) is a unifilar belief transducer whose state dynamics are given by*

$$\mathrm{b}_{t+1}(s_{t+1}) = \frac{1}{Z}\sum_{s_t} p(y_t, s_{t+1}|a_t, s_t)\mathrm{b}_t(s_t), \tag{8}$$

with Z a normalisation constant. Moreover, (B_t, A_t, Y_t) is faithful if $\mathrm{b}_0 = p(s_0)$.

Interestingly, I-O Moore transducers (Sec. 3.3) also allow for ***postdictive Bayesian beliefs***[6] of the form $\mathrm{d}_t(s_t) := \Pr(S_t = s_t | \boldsymbol{G}_{:t} = \boldsymbol{g}_{:t})$ with $g_t = (y_t, a_{t-1})$, in which Y_t is used to infer S_t. Our next result explains how predictive and postdictive Bayesian beliefs and MSPs relate, and how they set the bases for Bayesian and Kalman filtering (proof in App. K).

Proposition 3. *If (S_t, A_t, Y_t) is a I-O Moore transducer and D_t is the postdictive Bayesian belief of S_t, then (D_t, A_t, Y_t) is a belief transducer whose state dynamics are given by*

$$\mathrm{d}_{t+1}(s_{t+1}) = \frac{p(y_{t+1}|s_{t+1})}{Z'} \sum_{s_t} p(s_{t+1}|s_t, a_t)\mathrm{d}_t(s_t), \tag{9}$$

with Z' a normalisation constant. Moreover, (D_t, A_t, Y_t) is faithful if $\mathrm{d}_0 = p(s_0)$.

Corollary 1. *Their predictive and postdictive Bayesian beliefs of I-O Moore transducers can be updated as $\mathrm{b}_{t-1} \xrightarrow{\text{predict}} \mathrm{d}_t \xrightarrow{\text{update}} \mathrm{b}_t$ following the 'predict-update' process from Bayesian filtering.*

A faithful belief transducers can be said to provide a ***mixed-state presentation*** (MSP) of the underlying transducer, extending previous work on MSPs of hidden Markov models (Jurgens & Crutchfield, 2021a;b). MSPs are generally not minimal, as they tend to have different mixed-states that are bisimilar. The reduction of these is studied in the next section.

5.2 Minimal predictive world models

Following Barnett & Crutchfield (2015), we now present a method to build an observable world model directly from an interface $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$ without the need to bootstrap from another world model. For this, consider an equivalence relation between histories in which $\boldsymbol{h}_{:t-1} \sim_\epsilon \boldsymbol{h}'_{:t-1}$ when

$$p(\boldsymbol{y}_{t:t+T}|\boldsymbol{h}_{:t-1}, \boldsymbol{a}_{t:t+T}) = p(\boldsymbol{y}_{t:t+T}|\boldsymbol{h}'_{:t-1}, \boldsymbol{a}_{t:t+T}), \quad \forall \boldsymbol{y}_{t:t+L}, \boldsymbol{a}_{t:t+L}, L \in \mathbb{N}. \tag{10}$$

Let's denote by ϵ_t the coarse-graining mapping that assigns each history to its corresponding equivalence class $\epsilon_t(\boldsymbol{h}_{:t-1}) = [\boldsymbol{h}_{:t-1}]_{\sim_\epsilon}$, and define $M_t = \epsilon_t(\boldsymbol{H}_{:t-1})$. This construction is known as *predictive state representations* (Littman & Sutton, 2001; Singh et al., 2004) and *instrumental states* (Kosoy, 2019), and is based on older ideas for stochastic processes (without inputs/actions) from computational mechanics (Crutchfield & Young, 1989). Interestingly, the equivalence classes induced by ϵ are the minimal bisimulation of the world model $S_t = \boldsymbol{H}_{:t-1}$ (Lemma 4), and therefore serve as memory states of a transducer that generates the original interface (first shown in (Barnett & Crutchfield, 2015, Prop 2), alternative proof in App. L).

Proposition 4. *(M_t, A_t, Y_t) with $M_t = \epsilon_t(\boldsymbol{H}_{:t-1})$ is a transducer presentation for $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$.*

The transducer with memory states given by $M_t = \epsilon_t(\boldsymbol{H}_{:t-1})$ resulting from Prop. 4 is known as the ***ϵ*-transducer** of the interface $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$, and is unique up to isomorphism. The link between computational mechanics and other approaches such as predictive state representations was first noticed by Zhang et al. (2019), which explored it using a different computational structure instead of transducers. A salient feature of these approaches is that they can provide observable world models over fewer states than other methods (Littman & Sutton, 2001). Our next result strengthens this intuition by proving that the ϵ-transducer yields the most efficient predictive world model possible (proof in App. M), which takes inspiration from and extends (Barnett & Crutchfield, 2015, Lemma 1).

Theorem 3. *If R_t is a predictive world model of a transducer, then its minimal bisimulation is isomorphic to the ϵ-transducer.*

Corollary 2. *The ϵ-transducer is the minimal predictive model that generates a given interface.*

These findings complement previous studies considering the relationship between bisimulation and predictive state representations (Castro, 2011, Sec. 3). These results imply, for instance, that the

[6] In general, ***postdictive world models*** S_t satisfy $I(S_t; \boldsymbol{Y}_{t+1:}|\boldsymbol{H}_{:t}, \boldsymbol{A}_{t+1:}) = 0$.

bisimulation of predictive beliefs about the states of a world model necessarily converge into the memory states of the corresponding ϵ-transducer. In effect, while bisimulations of arbitrary transducers may not fully reduce world models (Sec. 4.2), bisimulations of predictive transducers necessarily do so (see App. N). An analogous result can be derived for postdictive beliefs, which are reduced into a 'time-shifted' ϵ-transducer. This will be developed in a future publication.

6 Reverse interpretability via retrodictive world models

The results of the last section show that the ϵ-transducer is a universal construction that distils the information that is relevant for predicting future events, which can be used to evaluate the extent to which agents can learn through a given interface. However, prediction alone does not exhaust the possible knowledge-driven activities that can involve an agent. This section investigates reversible and retrodictive world models, exploring new opportunities for agent interpretability.

6.1 Reversible transducers

The kernel of a transducer is usually used to update a world model S_t from s_τ to $s_{\tau+1}$. Interestingly, some transducers can be used to run things 'backwards', so that the world state can be updated from $s_{\tau+1}$ to s_τ while generating the same interface. This is formalised by the next definition.[7]

Definition 8. *A **reversible transducer** is a transducer $(\mathcal{S}, \mathcal{A}, \mathcal{Y}, \kappa, \mathrm{p})$ together with an additional Markov kernel κ^R of the form $\{\kappa^R_t(y, s'|a, s) : a \in \mathcal{A}, y \in \mathcal{Y}, s, s' \in \mathcal{S}, t \in \mathbb{N}\}$ such that*

$$p(\boldsymbol{y}_{:t}, \boldsymbol{s}_{:t+1}|\boldsymbol{a}_{:}) = p(s_0) \prod_{\tau=0}^{t} \kappa_\tau(y_\tau, s_{\tau+1}|s_\tau, a_t) = p(s_{t+1}|\boldsymbol{a}_{:t}) \prod_{\tau=0}^{t} \kappa^{\mathrm{R}}_\tau(y_\tau, s_\tau|a_\tau, s_{\tau+1}). \tag{11}$$

A reversible transducer can be run in reverse to produce the same interface. This can be used to analyse prior events that resulted in an undesirable world state s^*_{t+1} after an agent executed actions $\boldsymbol{a}_{:t}$. This can be investigated via the distribution

$$p(\boldsymbol{s}_{:t}, \boldsymbol{y}_{:t}|\boldsymbol{a}_{:t}, s^*_{t+1}) = \frac{p(\boldsymbol{y}_{:t}, \boldsymbol{s}_{:t}, s^*_t|\boldsymbol{a}_{:})}{p(s^*_{t+1}|\boldsymbol{a}_{:t})} = \kappa^{\mathrm{R}}_\tau(y_t, s_t|a_t, s^*_{t+1}) \prod_{\tau=0}^{t-1} \kappa^{\mathrm{R}}_\tau(y_\tau, s_\tau|a_\tau, s_{\tau+1}), \tag{12}$$

which allows to study how outputs lead to actions and identify tipping points in the world dynamics.

Unfortunately, not all transducers are reversible, as swapping past and future could break the condition of anticipation-free — which is needed for a world model to yield a transducer (see App. O). A necessary and sufficient condition for transducers to be reversed is provided next (proof in App. P).

Theorem 4. *A transducer is reversible for $\tau \leq T$ if and only if the dynamics of its memory state satisfy $p(s_\tau|s_{\tau+1}, \boldsymbol{a}_{:T}) = p(s_\tau|s_{\tau+1}, a_\tau)$ for all $\tau \in \{0, \ldots, T\}$, with a reverse kernel given by*

$$\kappa^{\mathrm{R}}_\tau(y, s|a, \tilde{s}) = \frac{\Pr(S_\tau = s|A_\tau = a)}{\Pr(S_{\tau+1} = \tilde{s}|A_\tau = a)} \kappa_\tau(y, \tilde{s}|a, s). \tag{13}$$

Although Eq. (13) always leads to a valid kernel due to Bayes rule, this may not generate the same interface — in fact, Eq. (11) only holds when the conditions in Theorem 4 are met. Interestingly, those conditions can be attained in a variety of ways. For example, memoryless transducers (see Lemma 3) are always reversible as $p(s_t|s_{t+1}, \boldsymbol{a}_{:t}) = p(s_t|s_{t+1}, a_t) = p(s_t)$. Also, consistent with results by Ellison et al. (2011), *action-agnostic* transducers (i.e. hidden Markov models) can be shown to be always reversible (see App. P). Finally, if the transducer is **action-counifilar** (i.e. if there exists f such that $S_t = f(S_{t+1}, A_t)$ can be deterministically updated)[8] is also sufficient for reversibility, as such transducer satisfies $p(s_\tau|s_{\tau+1}, \boldsymbol{a}_{:T}) = \delta^{s_\tau}_{f(s_{\tau+1}, a_\tau)} = p(s_\tau|s_{\tau+1}, a_\tau)$. Examples of these conditions are illustrated in Figure 4.

[7] This definition differs importantly from thermodynamically reversible transducers (Jurgens & Crutchfield, 2020).

[8] This is a special case of ***counifilar*** transducers, in which $S_t = f_t(S_{t+1}, A_t, Y_t)$ holds.

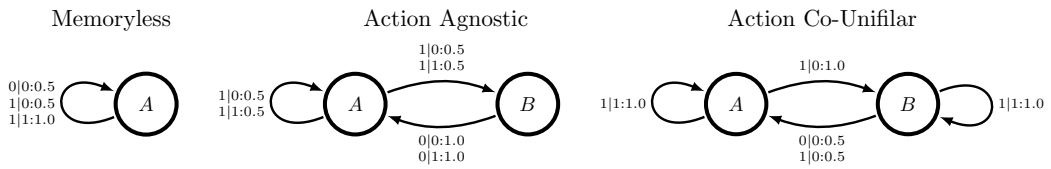

Figure 4: Three examples of reversible transducers. Circles represent world states, and arrows represent transitions and their labels describe the associated actions and outputs. For instance, the label $1|0{:}0.5$ on the edge from s_0 to s_1 indicates that $\Pr(S_{t+1} = s_1, Y_t = 1 | A_t = 0, S_t = s_0) = 0.5$.

6.2 Retrodictive beliefs

The previous subsection showed how there are substantial restrictions on the reversibility of transducers. Even if an interface cannot be generated via a reversible transducer, there are still 'retrodictive' constructions that can be used to investigate their dynamics. Retrodiction uses the future to learn about the past in the same way that prediction uses the past to learn about the future. Formal treatments of retrodiction include classic work in physics (Watanabe, 1955) and filtering theory (Jazwinski, 1970), and in more recent years have been formalised in computational mechanics (Ellison et al., 2009) and category theory (Parzygnat & Buscemi, 2023; Parzygnat, 2024).

Following these ideas, one can build ***retrodictive Bayesian beliefs*** (or mixed states) of a world model S_t as distributions over $\mathcal{S}$ given by $\mathrm{r}_t(s_0) \coloneqq \Pr(S_0 = s_0 | \boldsymbol{H}_{:t} = \boldsymbol{h}_{:t})$. These beliefs provide an analogue of the backward pass of Bayesian smoothing (Jazwinski, 1970), in the same way that the predictive and postdictive beliefs of input-Moore transducers correspond to different steps of Bayesian filtering (Prop. 3). However, in contrast with predictive Bayesian beliefs which can always be faithfull (Prop. 2), retrodictive beliefs may not be able to generate the same interface.

In order to study the dynamics of retrodictive beliefs, we introduce the ***bi-directional mixed-state matrix*** (BDMSM) of an action-outcome sequence $\rho(\boldsymbol{y}_{0:t}, \boldsymbol{a}_{0:t})$ as the $|\mathcal{S}| \times |\mathcal{S}|$ matrix given by

$$\rho(\boldsymbol{y}_{0:t}, \boldsymbol{a}_{0:t}) \coloneqq \sum_{s_0, s_{t+1} \in \mathcal{S}} p(s_0, s_{t+1} | \boldsymbol{y}_{0:t}, \boldsymbol{a}_{0:t}) \boldsymbol{e}_{s_{t+1}} \boldsymbol{e}_{s_0}^{\intercal} \,. \tag{14}$$

The BDMSM allows to calculate retrodictive beliefs and their dynamics (proof in App. Q).

Theorem 5. *Given a world model S_t, its BDMSM, predictive Bayesian beliefs b_τ and retrodictive Bayesian beliefs r_τ can be calculated as*

$$\rho(\boldsymbol{y}_{0:\tau}, \boldsymbol{a}_{0:\tau}) = \frac{T^{(\boldsymbol{y}_{0:\tau} | \boldsymbol{a}_{0:\tau})} \rho_0}{\mathbf{1}^{\intercal} \cdot T^{(\boldsymbol{y}_{0:\tau} | \boldsymbol{a}_{0:\tau})} \rho_0 \cdot \mathbf{1}}, \quad \mathrm{b}_\tau = \rho(\boldsymbol{y}_{0:\tau}, \boldsymbol{a}_{0:\tau}) \cdot \mathbf{1}, \quad \text{and} \quad \mathrm{r}_\tau = \rho(\boldsymbol{y}_{0:\tau}, \boldsymbol{a}_{0:\tau})^{\intercal} \cdot \mathbf{1},$$

where $T^{(y_{\tau:\tau'} | a_{\tau:\tau'})} = \prod_{j=\tau} T_j^{(y_j | a_j)}$ and $\rho_t = \sum_{s_t} p(s_t) \boldsymbol{e}_{s_t} \boldsymbol{e}_{s_t}^{\intercal}$ is a diagonal matrix.

Corollary 3. *The forward-time update of the BDMSM is given by*

$$\rho(\boldsymbol{y}_{0:\tau+1}, \boldsymbol{a}_{0:\tau+1}) = \frac{T^{(y_{\tau+1} | a_{\tau+1})}}{\mathbf{1}^{\intercal} \cdot T^{(y_{\tau+1} | a_{\tau+1})} \rho(\boldsymbol{y}_{0:\tau}, \boldsymbol{a}_{0:\tau}) \cdot \mathbf{1}} \cdot \rho(\boldsymbol{y}_{0:\tau}, \boldsymbol{a}_{0:\tau}), \tag{15}$$

while the reverse-time update is

$$\rho(\boldsymbol{y}_{-1:\tau}, \boldsymbol{a}_{-1:\tau}) = \rho(\boldsymbol{y}_{0:\tau}, \boldsymbol{a}_{0:\tau}) \cdot \frac{\rho_0^{-1} T^{(y_{-1} | a_{-1})} \rho_{-1}}{\mathbf{1}^{\intercal} \cdot \rho(\boldsymbol{y}_{0:\tau}, \boldsymbol{a}_{0:\tau}) \rho_0^{-1} T^{(y_{-1} | a_{-1})} \rho_{-1} \cdot \mathbf{1}}. \tag{16}$$

Retrodictive beliefs can be used to infer the most likely past states of the world given a sequence of future actions and outcomes. This could lead, for instance, to identifying the origins of specific behavioural patterns exhibited by an AI agent, which can in turn be used to characterise favourable or dangerous initial conditions via counterfactual reasoning (Karimi et al., 2021).

7 Conclusion

This paper investigated the fundamental limits that shape the usage of world models as tools to evaluate AI agents. This follows recent proposals to use world models not as tools for the agent (as in standard model-based reinforcement learning), but as tools for the scientist in charge of evaluating its safety and reliability (Dalrymple et al., 2024). By formalising these ideas via principles from computational mechanics, this approach led to a series of proposals for how to assess AI agents that require no assumptions about an agent's policy, architecture, or capabilities, being broadly applicable to systems regardless of how they were designed or trained. This framework revealed fundamental limits, challenges, and opportunities inherent to world modelling, leading to actionable guidelines that can inform core design choices instrumental for effective agent evaluation (see Figure 1).

Our framework revealed a fundamental trade-off between the efficiency and interpretability of world models. Generalised transducers were found to generate the most efficient implementations, but these come at the cost of inducing quasi-probabilities — yielding opaque world models that cannot be sampled.[9] Our results also revealed that the ϵ-transducer, a generalisation of the geometric belief structure recently found in the residual stream of transformers (Shai et al., 2025), yields the unique minimal world model that could be calculated by an agent in real time. The uniqueness of the ϵ-transducer implies that the refinement of the beliefs of any optimal predictive agent must eventually reach this model, regardless of the world model the agent uses. Thus, the ϵ-transducer can be seen as encapsulating all the predictive information that is available for agents, and hence establishes what is learnable about an environments through a particular interface.

We also introduced retrodictive world models as tools to investigate the origins of undesirable events or behaviours. These models allow retrospective analyses that could, for instance, identify 'danger zones' that are likely to lead to undesirable future outcomes. This view complements standard interpretability approaches, which typically assess agents via their capabilities to predict and plan with respect to future events (Nanda et al., 2023; Gurnee & Tegmark, 2023; Shai et al., 2025).

While this work focused on the fundamental limits of world modelling under the dictum of perfect reconstruction, future work may relax this constraint by employing notions such as approximate homomorphisms (Taylor et al., 2008) or bisimulation (Girard & Pappas, 2011), rate-distortion trade-offs (Marzen & Crutchfield, 2016), or other approaches (Subramanian et al., 2022). Another promising direction to enable efficient modelling is to exploit the compositional structure of the world (Lake & Baroni, 2023; Elmoznino et al., 2024; Baek et al., 2025; Fu et al., 2025).

The approach taken here complements the substantial body of work that employs world models to improve the performance of agents in model-based reinforcement learning (Ha & Schmidhuber, 2018; Hafner et al., 2020; 2023; Hansen et al., 2024), and also on representations from the point of view of the agent (see (Ni et al., 2024) and references within). In fact, the formalism presented here provides a unified framework for reasoning about both (i) models that represent physical processes external to the agent and (ii) models that describe knowledge-gathering processes internal to the agent (Kaelbling et al., 1998; Biehl & Virgo, 2022; Virgo, 2023). Furthermore, the relationship found between predictive and postdictive machines in I-O Moore transducers and Bayesian and Kalman filtering sheds new light into the mechanisms supporting these well-established procedures. Moreover, the formalism of belief transducers opens several interesting avenues for future work, including the investigation of more general belief update dynamics on, for example, curved statistical manifolds (Morales & Rosas, 2021; Aguilera et al., 2024).

Overall, the ideas put forward here establish new bridges between related subjects in reinforcement learning, control theory, and computational mechanics, which we hope may serve as a Rosetta stone for navigating across these literatures. These new insights also have interesting implications for cognitive and computational neuroscience (Matsuo et al., 2022), particularly pertaining the formal characterisation of the internal world ('umwelt') of an agent (Von Uexküll, 1909; Ay & Löhr, 2015; Baltieri et al., 2025), which will be explored in future work.

[9] This is reminiscent of the notion of Kantian noumena, which suggests that things-in-themselves are beyond knowledge.

Acknowledgments

The authors thank Lionel Barnett, Martin Biehl, Chris Buckley, Matteo Capucci, James Crutchfield, Alexander Gietelink Oldenziel, Adam Goldstein, Alexandra Jurgens, Vanessa Kosoy, Sarah Marzen, Paul Riechers, Anil Seth, Adam Shai, and Lucas Teixera for inspiring discussions and useful feedback. The work of F.R. and A.B. has been supported by UK ARIA's Safeguarded AI programme. F.R. has also been supported by the PIBBSS Affiliateship programme. M.B. was supported by JST, Moonshot R&D, Grant Number JPMJMS2012.

References

Miguel Aguilera, Pablo Morales, Fernando Rosas, and Hideaki Shimazaki. Explosive neural networks via higher-order interactions in curved statistical manifolds. *arXiv preprint arXiv:2408.02326*, 2024.

Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6):26–38, 2017.

Shahab Asoodeh, Fady Alajaji, and Tamás Linder. Notes on information-theoretic privacy. In *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1272–1278. IEEE, 2014.

Nihat Ay and Wolfgang Löhr. The umwelt of an embodied agent—a measure-theoretic definition. *Theory in Biosciences*, 134(3):105–116, 2015.

Junyeob Baek, Yi-Fu Wu, Gautam Singh, and Sungjin Ahn. Dreamweaver: Learning compositional world representations from pixels. *arXiv preprint arXiv:2501.14174*, 2025.

Vijay Balasubramanian. Equivalence and reduction of hidden Markov models. Technical report, Massachusetts Institute of Technology, 1993.

Manuel Baltieri and Christopher L Buckley. An active inference implementation of phototaxis. In *Artificial Life Conference Proceedings*, pp. 36–43. MIT Press, 2017.

Manuel Baltieri and Christopher L. Buckley. Generative models as parsimonious descriptions of sensorimotor loops. *Behavioral and Brain Sciences*, 42, 2019.

Manuel Baltieri, Martin Biehl, Matteo Capucci, and Nathaniel Virgo. A bayesian interpretation of the internal model principle. *arXiv preprint arXiv:2503.00511*, pp. e218, 2025.

Nix Barnett and James Crutchfield. Computational mechanics of input–output processes: Structured transformations and the ϵ-transducer. *Journal of Statistical Physics*, 161(2):404–451, 2015.

Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Danielle Goldfarb, Hoda Heidari, Leila Khalatbari, et al. International scientific report on the safety of advanced AI (interim report). *arXiv preprint arXiv:2412.05282*, 2024.

Martin Biehl and Nathaniel Virgo. Interpreting systems as solving pomdps: a step towards a formal understanding of agency. In *International Workshop on Active Inference*, pp. 16–31. Springer, 2022.

D Blackwell, RV Ramamoorthi, et al. A bayes but not classically sufficient statistic. *The Annals of Statistics*, 10(3):1025–1026, 1982.

Filippo Bonchi, Elena Di Lavore, and Mario Román. Effectful Mealy machines: Bisimulation and trace. *arXiv preprint arXiv:2410.10627*, 2024.

Merve Nur Cakir, Mehwish Saleemi, and Karl-Heinz Zimmermann. On the theory of stochastic automata. *arXiv preprint arXiv:2103.14423*, 2021.

George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

Pablo Samuel Castro. *On planning, prediction and knowledge transfer in Fully and Partially Observable Markov Decision Processes*. McGill University (Canada), 2011.

Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204, 2013.

Volker Claus. *Stochastische Automaten*. Teubner Studienskripten, 1971.

Thomas Cover and Joy Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

James Crutchfield. The calculi of emergence: Computation, dynamics and induction. *Physica D: Nonlinear Phenomena*, 75(1-3):11–54, 1994.

James Crutchfield. Between order and chaos. *Nature Physics*, 8(1):17–24, 2012.

James Crutchfield and Karl Young. Inferring statistical complexity. *Physical review letters*, 63(2): 105, 1989.

David Dalrymple, Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, et al. Towards guaranteed safe AI: A framework for ensuring robust and reliable AI systems. *arXiv preprint arXiv:2405.06624*, 2024.

René Descartes. *Meditationes de Prima Philosophia*. Michael Soly, 1641.

Natalia Díaz-Rodríguez, Javier Del Ser, Mark Coeckelbergh, Marcos López de Prado, Enrique Herrera-Viedma, and Francisco Herrera. Connecting the dots in trustworthy artificial intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Information Fusion*, 99:101896, 2023.

Christopher Ellison, John Mahoney, and James Crutchfield. Prediction, retrodiction, and the amount of information stored in the present. *Journal of Statistical Physics*, 136:1005–1034, 2009.

Christopher Ellison, John Mahoney, Ryan James, James Crutchfield, and Jörg Reichardt. Information symmetries in irreversible processes. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 21(3):037107, 2011.

Eric Elmoznino, Thomas Jiralerspong, Yoshua Bengio, and Guillaume Lajoie. A complexity-based theory of compositionality. *arXiv preprint arXiv:2410.14817*, 2024.

Yariv Ephraim and Neri Merhav. Hidden Markov processes. *IEEE Transactions on information theory*, 48(6):1518–1569, 2002.

Norman Ferns and Doina Precup. Bisimulation metrics are optimal value functions. In *UAI*, pp. 210–219, 2014.

Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368, 1922.

Shuhao Fu, Andrew Jun Lee, Anna Wang, Ida Momennejad, Trevor Bihl, Hongjing Lu, and Taylor W Webb. Evaluating compositional scene understanding in multimodal generative models. *arXiv preprint arXiv:2503.23125*, 2025.

Antoine Girard and George J Pappas. Approximate bisimulation: A bridge between computer science and control theory. *European Journal of Control*, 17(5-6):568–578, 2011.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in Markov decision processes. *Artificial intelligence*, 147(1-2):163–223, 2003.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *Proceedings of the International Conference on Learning Representations (ICLR'23)*, 2023.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *Proceedings of the International Conference on Learning Representations (ICLR'20)*, 2020.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *Proceedings of the International Conference on Learning Representations (ICLR'24)*, 2024.

Yifeng He, Ethan Wang, Yuyang Rong, Zifei Cheng, and Hao Chen. Security of AI agents. *arXiv preprint arXiv:2406.08689*, 2024.

Qingqing Huang, Rong Ge, Sham Kakade, and Munther Dahleh. Minimal realization problems for hidden Markov models. *IEEE Transactions on Signal Processing*, 64(7):1896–1904, 2015.

Hisashi Ito, S-I Amari, and Kingo Kobayashi. Identifiability of hidden Markov information sources and their minimum degrees of freedom. *IEEE transactions on information theory*, 38(2):324–333, 1992.

David N Jansen, Flemming Nielson, and Lijun Zhang. Belief bisimulation for hidden Markov models: Logical characterisation and decision algorithm. In *Proceedings for NASA Formal Methods: 4th International Symposium, NFM 2012*, pp. 326–340, 2012.

Andrew H. Jazwinski. *Stochastic Processes and Filtering Theory*, volume 64 of *Mathematics in Science and Engineering*. Academic Press, New York, 1970. ISBN 0123815509.

Alexandra Jurgens and James Crutchfield. Functional thermodynamics of Maxwellian ratchets: Constructing and deconstructing patterns, randomizing and derandomizing behaviors. *Physical Review Research*, 2(3):033334, 2020.

Alexandra Jurgens and James Crutchfield. Divergent predictive states: The statistical complexity dimension of stationary, ergodic hidden markov processes. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31(8), 2021a.

Alexandra Jurgens and James Crutchfield. Shannon entropy rate of hidden Markov processes. *Journal of Statistical Physics*, 183(2):32, 2021b.

Leslie Pack Kaelbling, Michael Littman, and Anthony Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Olav Kallenberg. *Foundations of modern probability*, volume 2. Springer, 1997.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 353–362, 2021.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Andrei Nikolaevitch Kolmogorov. Determination of the centre of dispersion and degree of accuracy for a limited number of observation. *Izv. Akad. Nauk, USSR Ser. Mat*, 6:3–32, 1942.

Vanessa Kosoy. Reinforcement learning with imperceptible rewards, 2019. URL https://www.alignmentforum.org/posts/aAzApjEpdYwAxnsAS/reinforcement-learning-with-imperceptible-rewards-1.

Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

Brenden Lake and Marco Baroni. Human-like systematic generalization through a meta-learning neural network. *Nature*, 623(7985):115–121, 2023.

Benjamin Larsen, Cathy Li, Stephanie Teeuwen, Olivier Denti, Jason DePerro, and Efi Raili. Navigating the AI frontier: A primer on the evolution and impact of AI agents. Technical report, World Economic Forum, December 2024.

Edward Ashford Lee and Sanjit Arunkumar Seshia. *Introduction to embedded systems: A cyber-physical systems approach*. MIT press, 2017.

Erich Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

Erich Leo Lehmann and Henry Scheffé. Completeness, similar regions, and unbiased estimation-part i. In *Selected Works of EL Lehmann*, pp. 233–268. Springer, 2012.

Michael Littman and Richard Sutton. Predictive representations of state. *Advances in neural information processing systems*, 14, 2001.

Samuel Loomis and James Crutchfield. Strong and weak optimizations in classical and quantum models of stochastic processes. *Journal of Statistical Physics*, 176(6):1317–1342, 2019.

Samuel Loomis and James Crutchfield. Topology, convergence, and reconstruction of predictive states. *Physica D: Nonlinear Phenomena*, 445:133621, 2023.

Francesco Mannella, Federico Maggiore, Manuel Baltieri, and Giovanni Pezzulo. Active inference through whiskers. *Neural Networks*, 144:428–437, 2021.

Sarah Marzen and James Crutchfield. Predictive rate-distortion for infinite-order Markov processes. *Journal of Statistical Physics*, 163:1312–1338, 2016.

Yutaka Matsuo, Yann LeCun, Maneesh Sahani, Doina Precup, David Silver, Masashi Sugiyama, Eiji Uchibe, and Jun Morimoto. Deep learning, reinforcement learning, and world models. *Neural Networks*, 152:267–275, 2022.

Marvin Lee Minsky. *Computation: Finite and Infinite Machines*. Prentice-Hall Englewood Cliffs, 1967.

Pablo A Morales and Fernando E Rosas. Generalization of the maximum entropy principle for curved statistical manifolds. *Physical Review Research*, 3(3):033216, 2021.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 16–30, 2023.

Tianwei Ni, Benjamin Eysenbach, Erfan Seyedsalehi, Michel Ma, Clement Gehring, Aditya Mahajan, and Pierre-Luc Bacon. Bridging state and history representations: Understanding self-predictive rl. *arXiv preprint arXiv:2401.08898*, 2024.

Yoshito Ohta. On the realization of hidden Markov models and tensor decomposition. *IFAC-PapersOnLine*, 54(9):725–730, 2021.

J Kevin O'Regan and Alva Noë. A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, 24(5):939–973, 2001.

Arthur J Parzygnat. Reversing information flow: retrodiction in semicartesian categories. *arXiv preprint arXiv:2401.17447*, 2024.

Arthur J Parzygnat and Francesco Buscemi. Axioms for retrodiction: achieving time-reversal symmetry with a prior. *Quantum*, 7:1013, 2023.

Plato. *Republic*. The Academy, 375 BC.

Hilary Putnam. *Reason, truth and history*. Cambridge University Press Cambridge, 1981.

Balaraman Ravindran and Andrew G. Barto. SMDP homomorphisms: An algebraic approach to abstraction in semi-Markov decision processes. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, Aug 2003.

Paul Riechers and James Crutchfield. Spectral simplicity of apparent complexity. I. the nondiagonalizable metadynamics of prediction. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(3), 2018.

Paul Michael Riechers. *Exact results regarding the physics of complex systems via linear algebra, hidden Markov models, and information theory*. University of California, Davis, 2016.

Simo Särkkä and Lennart Svensson. *Bayesian filtering and smoothing*, volume 17. Cambridge University Press, 2023.

Anil K Seth and Manos Tsakiris. Being a beast machine: The somatic basis of selfhood. *Trends in Cognitive Sciences*, 22(11):969–981, 2018.

Adam Shai, Lucas Teixeira, Alexander Oldenziel, Sarah Marzen, and Paul Riechers. Transformers represent belief state geometry in their residual stream. *Advances in Neural Information Processing Systems*, 37:75012–75034, 2025.

Satinder Singh, Michael James, and Matthew Rudary. Predictive state representations: A new theory for modeling dynamical systems. In *Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence*, pp. 512–518, 01 2004.

Jayakumar Subramanian, Amit Sinha, Raihan Seraj, and Aditya Mahajan. Approximate information state for approximate planning and reinforcement learning in partially observed systems. *Journal of Machine Learning Research*, 23(12):1–83, 2022.

Xiangru Tang, Qiao Jin, Kunlun Zhu, Tongxin Yuan, Yichi Zhang, Wangchunshu Zhou, Meng Qu, Yilun Zhao, Jian Tang, Zhuosheng Zhang, et al. Prioritizing safeguarding over autonomy: Risks of LLM agents for science. *arXiv preprint arXiv:2402.04247*, 2024.

Jonathan Taylor, Doina Precup, and Prakash Panagaden. Bounding performance loss in approximate MDP homomorphisms. *Advances in Neural Information Processing Systems*, 21, 2008.

Jianjun Tian and Dan Kannan. Lumpability and commutativity of Markov processes. *Stochastic analysis and Applications*, 24(3):685–702, 2006.

Alexander Tschantz, Anil K Seth, and Christopher L Buckley. Learning action-oriented models through active inference. *PLOS Computational Biology*, 16(4):e1007805, 2020.

Daniel Ray Upper. *Theory and algorithms for hidden Markov models and generalized hidden Markov models*. PhD thesis, University of California, Berkeley, 1997.

Mathukumalli Vidyasagar. The complete realization problem for hidden Markov models: A survey and some new results. *Mathematics of Control, Signals, and Systems*, 23(1):1–65, 2011.

Nathaniel Virgo. Unifilar machines and the adjoint structure of bayesian filtering. *arXiv preprint arXiv:2305.02826*, 2023.

Nathaniel Virgo, Martin Biehl, and Simon McGregor. Interpreting dynamical systems as bayesian reasoners. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 726–762. Springer, 2021.

Jakob Von Uexküll. *Umwelt und innenwelt der tiere*. Springer, 1909.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

Xu Wang, Sen Wang, Xingxing Liang, Dawei Zhao, Jincai Huang, Xin Xu, Bin Dai, and Qiguang Miao. Deep reinforcement learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4):5064–5078, 2022.

Satosi Watanabe. Symmetry of physical laws. part iii. prediction and retrodiction. *Reviews of Modern Physics*, 27(2):179–186, 1955.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):1–20, 2024.

Amy Zhang, Zachary C Lipton, Luis Pineda, Kamyar Azizzadenesheli, Anima Anandkumar, Laurent Itti, Joelle Pineau, and Tommaso Furlanello. Learning causal state representations of partially observable environments. *arXiv preprint arXiv:1906.10437*, 2019.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Supplementary Materials

The following content was not necessarily subject to peer review.

A Example of an interface and multiple world models

Let us present an example to illustrate the notions of interface and world models. This example showcases how a single interface can be generated by various world models with different properties.

Consider a robot manipulating a deck of cards. The state of the deck can be described via a world model that can adopt $|\mathcal{S}| = 13! \approx 6 \times 10^{10}$ possible states. At every time point, the robot can take two possible actions: either it puts the front card in the back (α_1), or it shuffles the deck (α_2). Thus, the set of actions available to the robot is $\mathcal{A} = \{\alpha_1, \alpha_2\}$. Additionally, at every time point the robot observes the card that is on top of the deck. However, we will assume that the sensory apparatus of the robot is not capable of reading the number or the suit of the card, but only its colour. Hence, the possible outcomes of this scenario for the robot are $\mathcal{Y} = \{\texttt{black}, \texttt{red}\}$.

In this scenario, the interface of the robot is constituted by a collection of probability distributions of the form $p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:t})$ relating sequences of actions with sequences of colours. Furthermore, if one tracks the state of the deck at time t via the variable S_t, this results in a transformer (S_t, A_t, Y_t) (see Def. 3). However, if we wanted to implement this transducer on a computer, specifying its kernel would require a substantial amount of memory due to the large number of possible world states.

Following the considerations made in Sec. 2, one could instead forget about the fact that there is an underlying deck of cards and focus just on the sequences of colours that the agent records. By doing this, one may notice that, given a sequence of actions and outcomes $(\boldsymbol{a}_{:t}, \boldsymbol{y}_{:t})$, the only information that is relevant to predict the next outcome y_{t+1} is the number of red and black cards observed since the last time the agent took action α_2 (shuffling the deck). If one tracks this information at time t via the variable M_t, then Prop. 4 guarantees that (M_t, A_t, Y_t) provides an alternative transducer presentation of the original interface. Note that this is an 'epistemic' world model that reflects the agent's state of knowledge (as described in Sec. 5), contrasting with S_t which reflects an objective physical process taking place 'out there'. Interestingly, M_t uses only roughly 13^2 states instead of $13!$ states, requiring substantially fewer memory resources. Furthermore, M_t reflects all the relevant information that an agent with this particular interface (i.e. limited to recognising colours) could ever want to take into account in order to make informed actions in this scenario. Therefore, this new world model is not only more memory efficient, but also reveals what information an agent of this kind could and should learn.

Before concluding, let us add some considerations related to the ideas explored in the latest part of the manuscript. Sec. 6 studies world models that 'run backwards' — i.e. can be updated in reverse time. The interface chosen in this example does not allow for such a reversible presentation, as the combinations of shuffling (α_2) and card flipping (α_1) lead to world dynamics that violate the conditions outlined in Theorem 4. One could attain a reversible interface if we considered a different set of actions, for example $\mathcal{A}' = \{\alpha_1, \alpha_3\}$ with α_3 corresponding to the robot taking the card in the back and putting that on top of the deck. Indeed, the world dynamics resulting from such a set of actions do satisfy the sufficient condition for reversibility discussed at the end of Sec. 6.1.

B Sufficient statistics

Given the importance of the notion of sufficient statistics in this work, we use this appendix to provide an account of its origins and significance.

Consider a random vector $\boldsymbol{X} = (X_1, \ldots, X_n) \in \mathcal{X}^n$ that follows a distribution with parameter $\theta \in \Theta$, and a 'statistic' $T(\cdot)$ (that is, a mapping $T : \mathcal{X}^n \to \mathbb{R}$). Following Fisher (1922), $Y = T(\boldsymbol{X})$ is a *classical/frequentist sufficient statistic* for $\boldsymbol{X}$ w.r.t. θ if the value of $\Pr_\theta(\boldsymbol{X} = \boldsymbol{x}|Y = y)$ is the

same $\forall\theta \in \Theta$ (Casella & Berger, 2002). This means that the information given by $\boldsymbol{X}$ that is not in Y is irrelevant to estimating the value of θ.

Another approach to statistical sufficiency due to Kolmogorov (1942), which can be called *strong bayesian statistical sufficiency*, states that Y is sufficient for $\boldsymbol{X}$ w.r.t. θ if $\boldsymbol{X}$ is statistically independent of θ given Y for any prior distribution over θ. Strong Bayesian sufficiency can be shown to imply classical sufficiency, but the converse does not necessarily hold (Blackwell et al., 1982).

A useful generalisation of the above condition, which we simply call *(weak) Bayesian statistical sufficiency*, follows Kolmogorov's condition just for a given distribution of θ (Cover & Thomas, 2012). In particular, given two random variables X and Y, a statistic $T = f(X)$ is said to be a *Bayesian sufficient statistic for X w.r.t. Y* if X is statistically independent of Y given T. In terms of Shannon's mutual information, this corresponds to the condition $\Pr(X = x|Y = y, T = t) = \Pr(X = x|T = t)$. This is equivalent to the information-theoretic condition $I(X; Y|T) = 0$, which states that X and Y share no information that is not given by T (Cover & Thomas, 2012). This is the definition of sufficient statistics that we use in this work.

Another way to think about sufficient statistics is to notice that if $T = f(X)$ for some mapping f then $T - X - Y$ is a Markov chain. Then, thanks to the data processing inequality, $I(Y; X) \geq I(Y; T)$ as 'processing' X into T cannot increase its information about Y (Cover & Thomas, 2012). Interestingly, the equality $I(Y; X) = I(Y; T)$ is attained if an only if $X - T - Y$ is also a Markov chain, which corresponds to when T is a sufficient statistic. In summary, sufficient statistics are related to optimal (i.e. lossless) data processing (Kullback, 1997).

Sufficient statistics always exist — in particular, X is always sufficient for itself. The search for optimal but also efficient statistics leads to the idea of minimal sufficiency: a sufficient statistic S is minimal if for all other sufficient statistic T exists a function $f(\cdot)$ such that $S = f(T)$ (Lehmann & Scheffé, 2012), or equivalently, the following Markov chain holds: $S - T - X - Y$. From an information-theoretic point of view, a minimal sufficient statistic is the sufficient statistic of minimal entropy, hence providing the most parsimonious representation of the relevant information. Minimal sufficient statistics exist for a wide range of settings (Lehmann & Casella, 2006, Sec. 1.6), and are unique up to isomorphisms (i.e. re-labelling). Moreover, the minimal sufficient statistics of X w.r.t. Y can be built explicitly, built as the partition induced by the following equivalence relation (Asoodeh et al., 2014, Def. 2):

$$x \sim x' \quad \text{iff} \quad \forall y \in \mathcal{Y}: \; p(y|x) = p(y|\boldsymbol{x}'). \tag{17}$$

Note the similarities between this way of building minimal sufficient statistics, bisimulation (Def. 5), and the construction of the ϵ-transducer via the equivalence relation in Eq. (10).

C Proof of Lemma 1

Proof. Let us first prove that if S_t is a world model for the interface $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$, then Eq. (2) holds. Using property (2) of world models, together with the fact that the interface is anticipation-free, one can show that

$$\begin{aligned} p(\boldsymbol{y}_{:\tau}, \boldsymbol{s}_{:\tau+1}|\boldsymbol{a}_{:}) &= p(\boldsymbol{y}_{:\tau}|\boldsymbol{a}_{:})p(\boldsymbol{s}_{:\tau+1}|\boldsymbol{y}_{:\tau}, \boldsymbol{a}_{:}) \\ &= p(\boldsymbol{y}_{:\tau}|\boldsymbol{a}_{:\tau})p(\boldsymbol{s}_{:\tau+1}|\boldsymbol{y}_{:\tau}, \boldsymbol{a}_{:\tau}) \\ &= p(\boldsymbol{y}_{:\tau}, \boldsymbol{s}_{:\tau+1}|\boldsymbol{a}_{:\tau}), \end{aligned} \tag{18}$$

which holds for all $\tau \in \mathbb{N}$. Then, one can use this equality recursively to derive the following:

$$\begin{aligned}
p(\boldsymbol{y}_{:t}, \boldsymbol{s}_{:t+1}|\boldsymbol{a}_{:}) &= p(\boldsymbol{y}_{:t}, \boldsymbol{s}_{:t+1}|\boldsymbol{a}_{:t}) \\
&= p(\boldsymbol{y}_{:t-1}, \boldsymbol{s}_{:t}|\boldsymbol{a}_{:t})p(y_t, s_{t+1}|\boldsymbol{y}_{:t-1}, \boldsymbol{s}_{:t}, \boldsymbol{a}_{:t}) \\
&= p(\boldsymbol{y}_{:t-1}, \boldsymbol{s}_{:t}|\boldsymbol{a}_{:t-1})p(y_t, s_{t+1}|\boldsymbol{y}_{:t-1}, \boldsymbol{s}_{:t}, \boldsymbol{a}_{:t}) \\
&= p(\boldsymbol{y}_{:t-2}, \boldsymbol{s}_{:t-1}|\boldsymbol{a}_{:t-1}) \prod_{\tau=t-1}^{t} p(y_\tau, s_{\tau+1}|\boldsymbol{y}_{:\tau-1}, \boldsymbol{s}_{:\tau}, \boldsymbol{a}_{:\tau}) \\
&= \dots \\
&= p(s_0) \prod_{\tau=0}^{t} p(y_\tau, s_{\tau+1}|\boldsymbol{y}_{:\tau-1}, \boldsymbol{s}_{:\tau}, \boldsymbol{a}_{:\tau}). \qquad (19)
\end{aligned}$$

Note that, in the last step, $p(s_0|a_0) = p(s_0)$ follows by applying property (2) for $t = 0$. Now, using property (1) one can show that

$$\begin{aligned}
p(y_\tau, s_{\tau+1}|\boldsymbol{y}_{:\tau-1}, \boldsymbol{s}_{:\tau}, \boldsymbol{a}_{:\tau}) &= p(y_\tau|\boldsymbol{y}_{:\tau-1}, \boldsymbol{s}_{:\tau}, \boldsymbol{a}_{:\tau})p(s_{\tau+1}|\boldsymbol{y}_{:\tau}, \boldsymbol{s}_{:\tau}, \boldsymbol{a}_{:\tau}) \\
&= p(y_\tau|s_\tau, a_\tau)p(s_{\tau+1}|\boldsymbol{y}_{:\tau}, \boldsymbol{s}_{:\tau}, \boldsymbol{a}_{:\tau}). \qquad (20)
\end{aligned}$$

The desired result follows from putting together Eq. (19) and Eq. (20).

For the converse, let's show that if Eq. (2) holds, then S_t satisfies the two properties of world models as given in Def. 2. The first property can be proven directly as follows:

$$\begin{aligned}
p(y_t|\boldsymbol{s}_{:t}, \boldsymbol{a}_{:t}, \boldsymbol{y}_{:t-1}) &= \frac{\sum_{s_{t+1}} p(\boldsymbol{y}_{:t}, \boldsymbol{s}_{:t+1}|\boldsymbol{a}_{:t})}{\sum_{s_{t+1}, y_t} p(\boldsymbol{y}_{:t}, \boldsymbol{s}_{:t+1}|\boldsymbol{a}_{:t})} \\
&= \frac{\sum_{s_{t+1}} p(s_0) \prod_{\tau=}^{t} p(y_\tau|s_\tau, a_\tau)p(s_{\tau+1}|\boldsymbol{s}_{:\tau}, \boldsymbol{h}_{:\tau})}{\sum_{s_{t+1}, y_t} p(s_0) \prod_{\tau=0}^{t} p(y_\tau|s_\tau, a_\tau)p(s_{\tau+1}|\boldsymbol{s}_{:\tau}, \boldsymbol{h}_{:\tau})} \\
&= \frac{\sum_{s_{t+1}} p(y_t|s_t, a_t)p(s_{t+1}|\boldsymbol{s}_{:t}, \boldsymbol{h}_{:t})}{\sum_{s_{t+1}, y_t} p(y_t|s_\tau, a_t)p(s_{t+1}|\boldsymbol{s}_{:t}, \boldsymbol{h}_{:t})} \\
&= p(y_t|s_t, a_t). \qquad (21)
\end{aligned}$$

Similarly, the second property can be proven as follows:

$$\begin{aligned}
p(\boldsymbol{s}_{:t}|\boldsymbol{h}_{:t-1}, \boldsymbol{a}_{t:}) &= \frac{p(\boldsymbol{s}_{:t}, \boldsymbol{y}_{:t-1}|\boldsymbol{a}_{:})}{p(\boldsymbol{y}_{:t-1}|\boldsymbol{a}_{:})} \\
&= \frac{p(s_0) \prod_{\tau=}^{t-1} p(y_\tau|s_\tau, a_\tau)p(s_{\tau+1}|\boldsymbol{s}_{:\tau}, \boldsymbol{h}_{:\tau})}{\sum_{\boldsymbol{s}_{:t}} p(s_0) \prod_{\tau=}^{t-1} p(y_\tau|s_\tau, a_\tau)p(s_{\tau+1}|\boldsymbol{s}_{:\tau}, \boldsymbol{h}_{:\tau})} \\
&\overset{(a)}{=} p(\boldsymbol{s}_{:t}|\boldsymbol{h}_{:t-1}). \qquad (22)
\end{aligned}$$

Above, (a) follows from the fact that the variables $\boldsymbol{a}_{t:}$ do not appear in the previous expression. □

D Alternative characterisations of a transducer

Lemma 2 characterises transducers as world models. However, this characterisation is not the most useful when investigating if a given process S_t qualifies as the memory state of a transducer. Here we provide two alternative characterisations of a transducer that are better suited to those tasks.

Lemma 6. *The process S_t provides a memory state for a transducer presentation of an anticipation-free interface $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$ if and only if one of the following conditions hold:*

1. $p(s_0|\boldsymbol{a}_{:}) = p(s_0)$ *and* $p(s_{t+1}, y_t|\boldsymbol{s}_{:t}, \boldsymbol{h}_{:t-1}, \boldsymbol{a}_{t:}) = p(s_{t+1}, y_t|s_t, a_t)$ *for all* $t \in \mathbb{N}$.

2. $I(\boldsymbol{S}_{t_i+1:t}, \boldsymbol{Y}_{t_i:t-1}; \boldsymbol{A}_{t:}|\boldsymbol{A}_{t_i:t-1}, S_{t_i}) = I(\boldsymbol{S}_{t+1:}, \boldsymbol{Y}_{t:}; \boldsymbol{Y}_{:t-1}, \boldsymbol{S}_{:t-1}, \boldsymbol{A}_{:t-1}|\boldsymbol{A}_{t:}, S_t) = 0$ *for all* $t_i, t \in \mathbb{N}$ *with* $t_i \leq t$.

Proof. We prove these equivalences in two steps.

Step 1: Equivalence between condition (1) and Lemma 2

Let's first prove that Lemma 2 imply condition (1). By Lemma 1, if S_t is a world model then $p(s_{t+1}, y_t|\boldsymbol{s}_{:t}, \boldsymbol{h}_{:t-1}, \boldsymbol{a}_{t:}) = p(y_t|s_t, a_t)p(s_{t+1}|\boldsymbol{s}_{:t}, \boldsymbol{h}_{:t})$. Combining this with $p(s_{t+1}|\boldsymbol{s}_{:t}, \boldsymbol{h}_{:t}) = p(s_{t+1}|s_t, h_t)$ (Lemma 2), it is clear that $p(s_{t+1}, y_t|\boldsymbol{s}_{:t}, \boldsymbol{h}_{:t-1}, \boldsymbol{a}_{t:}) = p(s_{t+1}, y_t|s_t, a_t)$.

To prove the converse, let us now show that condition (1) guarantees that S_t is a world model that satisfies $p(s_{t+1}|\boldsymbol{s}_{:t}, \boldsymbol{h}_{:t}) = p(s_{t+1}|s_t, h_t)$. Property (1) of world models can be proven as follows:

$$p(y_t|\boldsymbol{s}_{:t}, \boldsymbol{a}_{:t}, \boldsymbol{y}_{:t-1}) = \sum_{s_{t+1}} p(s_{t+1}, y_t|\boldsymbol{s}_{:t}, \boldsymbol{h}_{:t-1}, a_t) \overset{(a)}{=} \sum_{s_{t+1}} p(s_{t+1}, y_t|s_t, a_t) = p(y_t|s_t, a_t), \quad (23)$$

where (a) uses condition (1). Property (2) follows from

$$\begin{aligned} p(\boldsymbol{s}_{:t}|\boldsymbol{h}_{:t-1}, \boldsymbol{a}_{t:}) &= \frac{p(\boldsymbol{s}_{:t}, \boldsymbol{y}_{:t-1}|\boldsymbol{a}_{:})}{p(\boldsymbol{y}_{:t-1}|\boldsymbol{a}_{:})} \\ &= \frac{\prod_{\tau=0}^{t-1} p(s_{\tau+1}, y_\tau|\boldsymbol{s}_{:\tau}, \boldsymbol{y}_{:\tau-1}, \boldsymbol{a}_{:})}{p(\boldsymbol{y}_{:t-1}|\boldsymbol{a}_{:})} \\ &\overset{(b)}{=} \frac{\prod_{\tau=0}^{t-1} p(s_{\tau+1}, y_\tau|\boldsymbol{s}_{:\tau}, \boldsymbol{y}_{:\tau-1}, \boldsymbol{a}_{:t-1})}{p(\boldsymbol{y}_{:t-1}|\boldsymbol{a}_{:t-1})} \\ &= \frac{p(\boldsymbol{s}_{:t}, \boldsymbol{y}_{:t-1}|\boldsymbol{a}_{:t-1})}{p(\boldsymbol{y}_{:t-1}|\boldsymbol{a}_{:t-1})} \\ &= p(\boldsymbol{s}_{:t}|\boldsymbol{h}_{:t-1}), \end{aligned} \quad (24)$$

where (b) is using condition (1) and the fact that the interface is anticipation free. Finally, the Markovianity of state dynamics can be proven as follows:

$$p(s_{t+1}|\boldsymbol{s}_{:t}, \boldsymbol{h}_{:t}) = \frac{p(s_{t+1}, y_t|\boldsymbol{s}_{:t}, a_t, \boldsymbol{h}_{:t-1})}{\sum_{s_{t+1}} p(s_{t+1}, y_t|\boldsymbol{s}_{:t}, a_t, \boldsymbol{h}_{:t-1})} = \frac{p(s_{t+1}, y_t|s_t, a_t)}{\sum_{s_{t+1}} p(s_{t+1}, y_t|s_t, a_t)} = p(s_{t+1}|s_t, h_t). \quad (25)$$

Part 2: Equivalence between conditions (1) and (2)

Let's first show that condition (2) implies condition (1). For this, let's first note that in general if $I(A; B|C) = 0$ holds for some variables A, B, and C, then $p(a|c) = p(a|b, c)$. Thus, the condition $I(\boldsymbol{S}_{t_i+1:t}, \boldsymbol{Y}_{t_i:t-1}; \boldsymbol{A}_{t:}|\boldsymbol{A}_{t_i:t-1}, S_{t_i}) = 0$ implies that

$$p(\boldsymbol{s}_{t_i+1:\tau}, \boldsymbol{y}_{t_i:\tau-1}|\boldsymbol{a}_{t_i:}, s_{t_i}) = p(\boldsymbol{s}_{t_i+1:\tau}, \boldsymbol{y}_{t_i:\tau-1}|\boldsymbol{a}_{t_i:\tau-1}, s_{t_i}), \quad (26)$$

holding for all $\tau \in \mathbb{N}$ with $t_i \leq \tau \leq t$. Similarly, $I(\boldsymbol{S}_{t+1:}, \boldsymbol{Y}_{t:}; \boldsymbol{Y}_{:t-1}, \boldsymbol{S}_{:t-1}, \boldsymbol{A}_{:t-1}|\boldsymbol{A}_{t:}, S_t) = 0$ implies that

$$p(\boldsymbol{s}_{\tau+1:t+1}, \boldsymbol{y}_{\tau:t}|\boldsymbol{y}_{:\tau-1}, \boldsymbol{s}_{:\tau}, \boldsymbol{a}_{:}) = p(\boldsymbol{s}_{\tau+1:t+1}, \boldsymbol{y}_{\tau:t}|\boldsymbol{a}_{\tau:}, s_\tau), \quad (27)$$

holding for all $\tau \in \mathbb{N}$ with $t_i \leq \tau \leq t$. Note that S_0 is an element of the past $\boldsymbol{S}_{:t-1}$, so we can multiply these together to obtain

$$\begin{aligned} &p(\boldsymbol{s}_{t_i+1:\tau}, \boldsymbol{y}_{t_i:\tau-1}|\boldsymbol{a}_{t_i:\tau-1}, s_{t_i})p(\boldsymbol{s}_{\tau+1:t+1}, \boldsymbol{y}_{\tau:t}|\boldsymbol{a}_{\tau:}, s_\tau) \\ &\quad = p(\boldsymbol{s}_{t_i+1:\tau}, \boldsymbol{y}_{t_i:\tau-1}|\boldsymbol{a}_{t_i:}, s_{t_i})p(\boldsymbol{s}_{\tau+1:t+1}, \boldsymbol{y}_{\tau:t}|\boldsymbol{y}_{t_i:\tau-1}, \boldsymbol{s}_{t_i:\tau}, \boldsymbol{a}_{t_i:}) \\ &\quad = p(\boldsymbol{s}_{t_i+1:t+1}, \boldsymbol{y}_{t_i:t}|\boldsymbol{a}_{t_i:}, s_{t_i}). \end{aligned} \quad (28)$$

Using this relation recursively, one can find that

$$\begin{aligned} p(\boldsymbol{s}_{:t+1}, \boldsymbol{y}_{:t}, |\boldsymbol{a}_{:}, s_0) &\overset{(c)}{=} p(\boldsymbol{s}_{:1}, y_0|a_0, s_0)p(\boldsymbol{s}_{2:t+1}, \boldsymbol{y}_{1:t}|\boldsymbol{a}_{1:}, s_1) \\ &\overset{(d)}{=} p(s_1, y_0|a_0, s_0)p(s_2, y_1|s_1, a_1)p(\boldsymbol{s}_{3:t+1}, \boldsymbol{y}_{2:t}|s_2, \boldsymbol{a}_{2:}) \\ &= \dots \\ &= \prod_{\tau=0}^{t} p(s_{\tau+1}, y_\tau|s_\tau, a_\tau), \end{aligned} \tag{29}$$

where (c) is obtained by using $t_i = 0$ and $\tau = 1$, (d) by using $t_i = 1$ and $\tau = 2$, and so on. By comparing with Lemma 2, this means that condition (2) implies condition (1).

Let us now show that condition (1) implies condition (2). Condition (1) implies that

$$\begin{aligned} p(\boldsymbol{s}_{t_i+1:t_f}, \boldsymbol{y}_{t_i:t_f-1}, |\boldsymbol{a}_{t_i:}, s_{t_i}) &= \prod_{\tau=t_i}^{t_f} p(s_{t+1}, y_t|s_t, a_t) \\ &= \left(\prod_{j=t}^{t_f} p(s_{j+1}, y_j|s_j, a_j)\right)\left(\prod_{k=t_i}^{t} p(s_{k+1}, y_k|s_k, a_k)\right) \\ &= p(\boldsymbol{s}_{t+1:t_f}, \boldsymbol{y}_{t:t_f-1}|\boldsymbol{a}_{t:t_f-1}, s_t)p(\boldsymbol{s}_{t_i+1:t}, \boldsymbol{y}_{t_i:t-1}|\boldsymbol{a}_{t_i:t-1}, s_{t_i}). \end{aligned} \tag{30}$$

This can be used to show that

$$\begin{aligned} p(\boldsymbol{s}_{t_i+1:t}, \boldsymbol{y}_{t_i:t-1}|\boldsymbol{a}_{t_i:t-1}, s_{t_i}) &= \sum_{\substack{\boldsymbol{s}_{t+1:t_f} \\ \boldsymbol{y}_{t:t_f-1}}} p(\boldsymbol{s}_{t+1:t_f}, \boldsymbol{y}_{t:t_f-1}|\boldsymbol{a}_{t:t_f-1}, s_t)p(\boldsymbol{s}_{t_i+1:t}, \boldsymbol{y}_{t_i:t-1}|\boldsymbol{a}_{t_i:t-1}, s_{t_i}) \\ &= \sum_{\substack{\boldsymbol{s}_{t+1:t_f} \\ \boldsymbol{y}_{t:t_f-1}}} p(\boldsymbol{s}_{t_i+1:t_f}, \boldsymbol{y}_{t_i:t_f-1}, |\boldsymbol{a}_{t_i:}, s_{t_i}) \\ &= p(\boldsymbol{s}_{t_i+1:t}, \boldsymbol{y}_{t_i:t-1}, |\boldsymbol{a}_{t_i:}, s_{t_i}), \end{aligned} \tag{31}$$

which implies $I(\boldsymbol{S}_{t_i+1:t}, \boldsymbol{Y}_{t_i:t-1}; \boldsymbol{A}_{t:}|\boldsymbol{A}_{t_i:t-1}, S_{t_i}) = 0$. To prove the second information equality, one can divide both sides of Eq. (30) by $p(\boldsymbol{s}_{t_i+1:t}, \boldsymbol{y}_{t_i:t-1}|\boldsymbol{a}_{t_i:t-1}, s_{t_i})$ to obtain

$$\begin{aligned} p(\boldsymbol{s}_{t+1:t_f}, \boldsymbol{y}_{t:t_f-1}|\boldsymbol{a}_{t:t_f-1}, s_t) &= \frac{p(\boldsymbol{s}_{t_i+1:t_f}, \boldsymbol{y}_{t_i:t_f-1}, |\boldsymbol{a}_{t_i:}, s_{t_i})}{p(\boldsymbol{s}_{t_i+1:t}, \boldsymbol{y}_{t_i:t-1}|\boldsymbol{a}_{t_i:t-1}, s_{t_i})} \\ &= p(\boldsymbol{s}_{t+1:t_f}, \boldsymbol{y}_{t:t_f-1}, |\boldsymbol{y}_{t_i:t-1}, \boldsymbol{s}_{t_i:t}, a_{t_i:}). \end{aligned} \tag{32}$$

Given that t_i and t_f are arbitrary, this implies that $I(\boldsymbol{S}_{t+1:}, \boldsymbol{Y}_{t:}; \boldsymbol{Y}_{:t-1}, \boldsymbol{S}_{:t}, \boldsymbol{A}_{:t-1}|\boldsymbol{A}_{t:}, S_t) = 0$. □

E Proof of Lemma 3

Lemma 7. *An interface is fully observable if and only if $p(y_{t+1}|\boldsymbol{y}_{:t}, \boldsymbol{a}_{:}) = p(y_{t+1}|y_t, a_t)$, and is memoryless if and only if $p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:}) = \prod_{\tau=0}^{t} p(y_\tau|a_\tau)$.*

Proof. To prove the first part of the lemma, one can use condition (1) in Lemma 6 which implies that $S_t = Y_t$ yields a transducer if and only if $p(y_\tau, y_{\tau+1}|\boldsymbol{y}_{:\tau}, \boldsymbol{a}_{:\tau}) = p(y_\tau, y_{\tau+1}|y_\tau, a_\tau) = p(y_{\tau+1}|y_\tau, a_\tau)$. To prove the second part of the lemma, note that an interface satisfies $p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:}) = \prod_{\tau=0}^{t} p(y_\tau|a_\tau)$ if and only if $S_t = 0$ yields a factorisation of $p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:})$ as in Eq. (4). This shows that $S_t = 0$ is the state of a transducer presentation of $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$ if and only if the interface is memoryless.

□

F Relationship between transducers and POMDPs

A POMDP is a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \tau, \mu, \rho)$ in which $\mathcal{S}$ correspond to states of the world, $\mathcal{A}$ the action space, $\mathcal{O}$ the observation space, and the probability kernels $\tau : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, $\mu : \mathcal{S} \to \Delta(\mathcal{O})$, and $\rho : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$ specify the world dynamics, observation map, and reward function (Kaelbling et al., 1998). Under a POMDP, the joint dynamics satisfy Eq. (4), which — thanks to condition (1) in Lemma 6 — is sufficient to show that the POMDP induces a transducer. This, together with Lemma 2, implies that the process S_t in a POMDP is a world model, in the sense that it satisfies the conditions in Def. 2.

Note also that the standard presentaiton of POMDPs correspond to a transducer whose kernel allows the following factorisation:

$$p(s_{t+1}, y_t | s_t, a_t) = \tau(s_{t+1}|s_t, a_t)\mu(o_t|s_t)\rho(r_t|s_t, a_t). \tag{33}$$

This corresponds to a I-O Moore transducer, as defined in Sec. 3.3. The different types of transducers are illustrated in Figure 5.

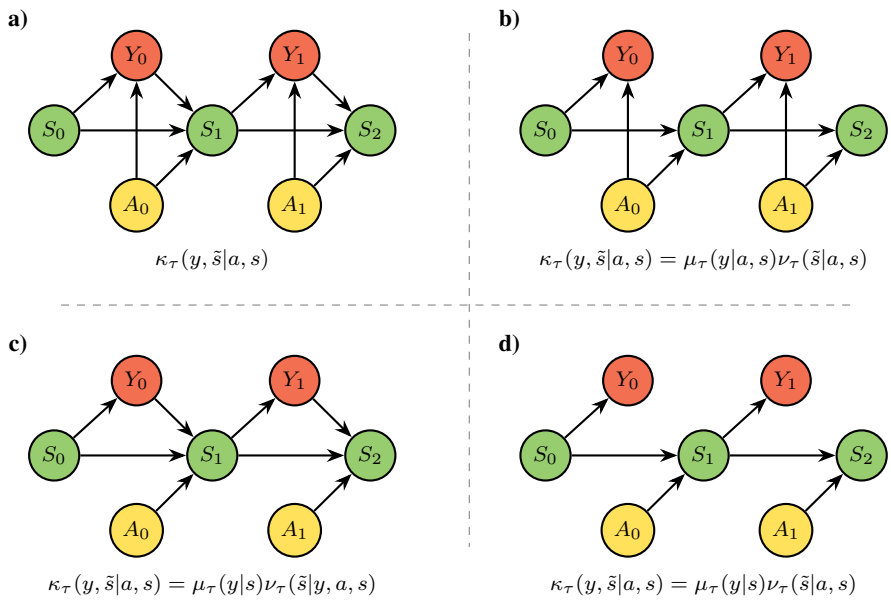

Figure 5: Illustration of different types of transducers: Mealy transducers (a), output-Moore transducers (b), input-Moore transducer (c), and I-O Moore transducer.

G Proof of Lemma 4

Proof. Let consider $S_t = \boldsymbol{H}_{t-1}$ and $S_0 = 0$. To prove that S_t yields a transducer presentation of $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$, we will use condition (1) from Lemma 6. For this, note that

$$p(s_{t+1}, y_t | \boldsymbol{s}_{:t}, \boldsymbol{h}_{:t-1}, \boldsymbol{a}_{t:}) = p(s_{t+1}, y_t | \boldsymbol{h}_{:t-1}, \boldsymbol{a}_{t:}) = p(s_{t+1}|\boldsymbol{h}_{:t}, \boldsymbol{a}_{t+1:})p(y_t|\boldsymbol{h}_{:t-1}, \boldsymbol{a}_{t:}). \tag{34}$$

Let us develop each of those terms separately. First, one can find that

$$p(s_{t+1}|\boldsymbol{h}_{:t}, \boldsymbol{a}_{t+1:}) = p(s_{t+1}|s_t, h_t, \boldsymbol{a}_{t+1:}) = \delta_{s_{t+1}}^{(s_t, h_t)} = p(s_{t+1}|s_t, h_t), \tag{35}$$

where δ_a^b is the Kroneker delta. Similarly, the second term can be developed as follows:

$$p(y_t|\boldsymbol{h}_{:t-1}, \boldsymbol{a}_{t:}) = \frac{p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:})}{p(\boldsymbol{y}_{:t-1}|\boldsymbol{a}_{:})} \overset{(a)}{=} \frac{p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:t})}{p(\boldsymbol{y}_{:t-1}|\boldsymbol{a}_{:t})} = p(y_t|\boldsymbol{y}_{:t-1}, \boldsymbol{a}_{:t}) = p(y_t|s_t, a_t), \tag{36}$$

where (a) uses the fact that the interface is anticipation-free. Finally, by combining Eq. (34), Eq. (35), and Eq. (36) one finds that

$$p(s_{t+1}, y_t|\boldsymbol{s}_{:t}, \boldsymbol{h}_{:t-1}, \boldsymbol{a}_{t:}) = p(s_{t+1}|s_t, h_t)p(y_t|s_t, a_t) = p(s_{t+1}, y_t|s_t, a_t), \tag{37}$$

which shows that condition (1) from Lemma 6 is satisfied. □

H Proof of Lemma 5

Proof. Consider $S'_t = \phi(S_t)$ a reduction of the memory state S_t of a transducer. Then

$$\begin{aligned}
p(\boldsymbol{y}_{:t}\boldsymbol{s}'_{:t+1}|\boldsymbol{a}_{:}) &= \sum_{\tau=0}^{t} \sum_{\substack{s_\tau \in \mathcal{S} \\ \phi(s_\tau)=s'_\tau}} p(\boldsymbol{y}_{:t}\boldsymbol{s}_{:t+1}|\boldsymbol{a}_{:}) \\
&\overset{(a)}{=} \sum_{\tau=0}^{t} \sum_{\substack{s_\tau \in \mathcal{S} \\ \phi(s_\tau)=s'_\tau}} p(s_0) \prod_{\tau=0}^{t} p(y_\tau|s_\tau, a_\tau)p(s_{\tau+1}|s_\tau, h_\tau) \\
&\overset{(b)}{=} \sum_{\tau=0}^{t} \sum_{\substack{s_\tau \in \mathcal{S} \\ \phi(s_\tau)=s'_\tau}} p(s_0) \prod_{\tau=0}^{t} p(y_\tau|s'_\tau, a_\tau)p(s_{\tau+1}|s_\tau, h_\tau) \\
&= \sum_{\tau=0}^{t-1} \sum_{\substack{s_\tau \in \mathcal{S} \\ \phi(s_\tau)=s'_\tau}} p(s_0) \prod_{\tau=0}^{t-1} p(y_\tau|s'_\tau, a_\tau)p(s_{\tau+1}|s_\tau, h_\tau)p(y_t|s'_t, a_t) \sum_{\substack{s_{t+1} \in \mathcal{S} \\ \phi(s_{t+1})=s'_{t+1}}} p(s_{t+1}|s_t, h_\tau) \\
&\overset{(c)}{=} \sum_{\tau=0}^{t-1} \sum_{\substack{s_\tau \in \mathcal{S} \\ \phi(s_\tau)=s'_\tau}} p(s_0) \prod_{\tau=0}^{t-1} p(y_\tau|s'_\tau, a_\tau)p(s_{\tau+1}|s_\tau, h_\tau)p(y_t|s'_t, a_t)p(s'_{t+1}|s'_t, h_\tau) \\
&\overset{(d)}{=} \dots \\
&= \left[\sum_{\substack{s_0 \in \mathcal{S} \\ \phi(s_0)=s'_0}} p(s_0) \right] \prod_{\tau=0}^{t} p(y_\tau|s'_\tau, a_\tau)p(s'_{\tau+1}|s'_\tau, h_\tau) \\
&\overset{(e)}{=} p(s'_0) \prod_{\tau=0}^{t} p(y_\tau|s'_\tau, a_\tau)p(s'_{\tau+1}|s'_\tau, h_\tau).
\end{aligned} \tag{38}$$

Above, (a) uses that S_t is the memory state of a transducer, (b) uses property (i) of homomorphisms (see Def. 4), (c) and (e) uses property (ii) of homomorphisms, and (d) denotes that the same steps of previous equations are done iteratively. Finally, Eq. (38) together with Lemma 1 confirm that S'_t yields a valid transducer for the same interface. □

I Proof of Prop. 1

Proof. Let's first assume that the mapping ϕ induces a reduction of the world model S_t into S'_t, and define the equivalence relation B such that $s \sim s'$ when $\phi(s) = \phi(s')$. In this setting, let's prove that B is a bisimulation. For this, one can note that if $s \sim s'$ then one can use the first property of homomorphims to find that

$$\begin{aligned}
\Pr(Y_t = y|S_t = s, A_t = a) &= \Pr(Y_t = y|S'_t = \phi(s), A_t = a) \\
&= \Pr(Y_t = y|S'_t = \phi(s'), A_t = a) \\
&= \Pr(Y_t = y|S_t = s', A_t = a).
\end{aligned} \tag{39}$$

Additionally, using the second property one finds that

$$\begin{aligned}\sum_{s''\in[\tilde{s}]} \Pr\left(S_{t+1}=s''|S_t=s,H_t=(y,a)\right) &= \Pr\left(S'_{t+1}=\tilde{s}|S'_t=\phi(s),H_t=(y,a)\right)\\ &= \Pr\left(S'_{t+1}=\tilde{s}|S'_t=\phi(s'),H_t=(y,a)\right)\\ &= \sum_{s''\in[\tilde{s}]} \Pr\left(S_{t+1}=s''|S_t=s',H_t=(y,a)\right), \quad (40)\end{aligned}$$

where $[\tilde{s}] = \{s \in \mathcal{S} : \phi(s) = \tilde{s}\}$. Together, these two results show that B is a bisimulation in the sense of Def. 5.

For proving the converse statement, let's assume that $B \subseteq \mathcal{S} \times \mathcal{S}$ is a bisimulation, and define $\phi(s) = [s]$ as a function that maps each state $s \in \mathcal{S}$ into its equivalence class according to B. Let's prove that $S_t \xrightarrow{\phi} \phi(S_t) = [S_t]$ is a reduction. First, for B being a bisimulation implies that $\Pr\left(Y_t = y|S_t = s, A_t = a\right) = \Pr\left(Y_t = y|S_t = s', A_t = a\right)$ for any $(s, s') \in B$, which in turn implies that

$$\Pr\left(Y_t=y|\phi(S_t)=[s],A_t=a\right)=\Pr\left(Y_t=y|S_t=s,A_t=a\right), \quad (41)$$

showing that ϕ satisfies the first property of homomorphisms. Furthermore, if $(s, s') \in B$ then

$$\begin{aligned}\Pr\left(\phi(S_{t+1})=[\tilde{s}]|S_t=s,H_t=(y,a)\right) &= \sum_{s''\in[\tilde{s}]} \Pr\left(S_{t+1}=s''|S_t=s,H_t=(y,a)\right)\\ &= \sum_{s''\in[\tilde{s}]} \Pr\left(S_{t+1}=s''|S_t=s',H_t=(y,a)\right)\\ &= \Pr\left(\phi(S_{t+1})=[\tilde{s}]|S_t=s',H_t=(y,a)\right), \quad (42)\end{aligned}$$

which implies that

$$\Pr\left(\phi(S_{t+1})=[\tilde{s}]|S_t=s,H_t=(y,a)\right)=\Pr\left(\phi(S_{t+1})=[\tilde{s}]|\phi(S_t)=[s],H_t=(y,a)\right). \quad (43)$$

Using this, one can finally show that

$$\begin{aligned}\Pr\left(\phi(S_{t+1})=[\tilde{s}]|\phi(S_t)=[s],H_t=(y,a)\right) &= \sum_{s''\in[\tilde{s}]} \Pr\left(S_{t+1}=s''|\phi(S_t)=[s],H_t=(y,a)\right)\\ &= \sum_{s''\in[s]} \Pr\left(S_{t+1}=\tilde{s}|S_t=s,H_t=(y,a)\right) \quad (44)\end{aligned}$$

□

J Proof of Prop. 2

Proof. To prove the first part of the proposition, let's consider the Bayesian beliefs of the memory state S_t of a transducer (S_t, Y_t, A_t) as given by $\mathrm{b}_t(s_t) = p(s_t|\boldsymbol{h}_{:t-1})$. Let's also define the postdictive beliefs as $\mathrm{d}_t = p(s_t|\boldsymbol{h}_{:t})$. Then, their update can be calculated as follows:

$$\mathrm{b}_{t+1}(s_{t+1}) = \sum_{s_t} p(s_{t+1}|s_t,\boldsymbol{h}_{:t})p(s_t|\boldsymbol{h}_{:t}) \overset{(i)}{=} \sum_{s_t} p(s_{t+1}|s_t,h_t)\mathrm{d}_t(s_t), \quad (45)$$

where (i) uses the Markovianity of the memory states following Lemma 2. In a similar way, one can find that

$$\mathrm{d}_t(s_t) = \frac{p(s_t,\boldsymbol{h}_{:t-1},a_t,y_t)}{p(\boldsymbol{h}_{:t-1},a_t,y_t)} = \frac{p(y_t|s_t,\boldsymbol{h}_{:t-1},a_t)p(s_t|\boldsymbol{h}_{:t-1},a_t)}{p(y_t|\boldsymbol{h}_{:t-1},a_t)} \overset{(ii)}{=} \frac{p(y_t|s_t,a_t)}{Z'}\mathrm{b}_t(s_t) \quad (46)$$

with Z' a normalising constant. Above, (ii) uses that S_t is a world model together with the fact that that

$$p(s_t|\boldsymbol{h}_{:t-1}, a_t) = \frac{p(s_t, \boldsymbol{h}_{:t-1}, a_t)}{p(\boldsymbol{h}_{:t-1}, a_t)} = \frac{p(a_t|s_t, \boldsymbol{h}_{:t-1})p(s_t|\boldsymbol{h}_{:t-1})}{p(a_t|\boldsymbol{h}_{:t-1})} = p(s_t|\boldsymbol{h}_{:t-1}), \tag{47}$$

where the last equality holds due to the fact that actions depend on histories and not on states, and hence $p(a_t|s_t, \boldsymbol{h}_{:t-1}) = p(a_t|\boldsymbol{h}_{:t-1})$. Then, combining Eq. (45) and Eq. (46) one can find that

$$\mathrm{b}_{t+1}(s_{t+1}) = \hat{f}(\mathrm{b}_t, y_t, a_t) := \frac{1}{Z'} \sum_{s_t} p(s_{t+1}|s_t, h_t)p(y_t|s_t, a_t)\mathrm{b}_t(s_t), \tag{48}$$

proving the first part of the proposition.

To prove the second part of the proposition, first note that Eq. (48) combined with condition (1) in Lemma 6 imply that (B_t, A_t, Y_t) is a belief transducer with unifilar kernel given by

$$\hat{\kappa}_\tau(y, \tilde{\mathrm{b}}|a, \mathrm{b}) = \delta_{\tilde{\mathrm{b}}}^{\hat{f}_\tau(\mathrm{b},y,a)} \sum_{\tilde{s},s\in\mathcal{S}} \kappa_\tau(y, \tilde{s}|a, s)\mathrm{b}(s) = \delta_{\tilde{\mathrm{b}}}^{\hat{f}_\tau(\mathrm{b},y,a)} \Big(\mathbf{1}^\intercal \cdot T_\tau^{(y|a)} \cdot \mathrm{b}\Big), \tag{49}$$

where $T_\tau^{(y|a)}$ is the linear operator defined as in Eq. (3), and the second term corresponds to the probability of emitting y given b, i.e.

$$\mathbf{1}^\intercal \cdot T_\tau^{(y_\tau|a_\tau)} \cdot \mathrm{b}_\tau = \sum_{s_\tau\in\mathcal{S}} p(y_\tau|s_\tau, a_\tau)\mathrm{b}_\tau(s_\tau). \tag{50}$$

Now, let's consider a transducer (S_t, A_t, Y_t) and the belief transducer of predictive Bayesian beliefs (B_t, A_t, Y_t). Given that (S_t, A_t, Y_t) is a transducer, then the update rule given by Eq. (48) can be re-written as

$$\hat{f}_\tau(\mathrm{b}, y, a) = \frac{T_\tau^{(y|a)} \cdot \mathrm{b}}{\mathbf{1}^\intercal \cdot T_\tau^{(y|a)} \cdot \mathrm{b}}. \tag{51}$$

Moreover, for a given sequences $\boldsymbol{y}_{:\tau}, \boldsymbol{a}_{:\tau}$ and beliefs $\mathrm{b}_0, \ldots, \mathrm{b}_\tau$ following this updating rule, this can be applied recursively yielding

$$\begin{aligned}
\hat{f}_\tau(\mathrm{b}_\tau, y_\tau, a_\tau) &= \frac{T_\tau^{(y_\tau|a_\tau)} \cdot \mathrm{b}_\tau}{\mathbf{1}^\intercal \cdot T_\tau^{(y_\tau|a_\tau)} \cdot \mathrm{b}_\tau} \\
&= \frac{T_{\tau-1}^{(\boldsymbol{y}_{\tau-1:\tau}|\boldsymbol{a}_{\tau-1:\tau})} \cdot \mathrm{b}_{\tau-1}}{\mathbf{1}^\intercal \cdot T_{\tau-1}^{(\boldsymbol{y}_{\tau-1:\tau}|\boldsymbol{a}_{\tau-1:\tau})} \cdot \mathrm{b}_{\tau-1}} \\
&= \ldots \\
&= \frac{T_0^{(\boldsymbol{y}_{:\tau}|\boldsymbol{a}_{:\tau})} \cdot \mathrm{b}_0}{\mathbf{1}^\intercal \cdot T_0^{(\boldsymbol{y}_{:\tau}|\boldsymbol{a}_{:\tau})} \cdot \mathrm{b}_0},
\end{aligned} \tag{52}$$

where we are using $T_t^{(\boldsymbol{y}_{t:t'}|\boldsymbol{a}_{t:t'})} := \prod_{\tau=t}^{t'} T_\tau^{(y_\tau|a_\tau)}$ as a shorthand notation. Note that, similarly as Eq. (49), the denominator in Eq. (52) corresponds to

$$1^\intercal \cdot T_0^{(\boldsymbol{y}_{:\tau}|\boldsymbol{a}_{:\tau})} \cdot \mathrm{b}_0 = p(\boldsymbol{y}_{:\tau}|\boldsymbol{a}_{:\tau}, s_0)\mathrm{b}_0(s_0). \tag{53}$$

Finally, combining all these expressions one can directly calculate what is the result of successive applications of the kernel of a predictive Bayesian belief transducer:

$$\begin{aligned}\sum_{\mathrm{b}_{:t+1}} \prod_{\tau=0}^{t} \hat{\kappa}_\tau(y_\tau, \mathrm{b}_{\tau+1}|a_\tau, \mathrm{b}_\tau) &\overset{(a)}{=} \sum_{\mathrm{b}_{:t+1}} \prod_{\tau=0}^{t} \delta_{\mathrm{b}_{\tau+1}}^{\hat{f}(\mathrm{b}_\tau, y_\tau, a_\tau)} \left(\mathbf{1}^\intercal \cdot T_\tau^{(y_\tau|a_\tau)} \cdot \mathrm{b}_\tau\right) \\ &\overset{(b)}{=} \prod_{\tau=0}^{t} \mathbf{1}^\intercal \cdot T_\tau^{(y_\tau|a_\tau)} \cdot \frac{T_0^{(\boldsymbol{y}_{:\tau-1}|\boldsymbol{a}_{:\tau-1})} \cdot \mathrm{b}_0}{\mathbf{1}^\intercal \cdot T_0^{(\boldsymbol{y}_{:\tau-1}|\boldsymbol{a}_{:\tau-1})} \cdot \mathrm{b}_0} \\ &= \mathbf{1}^\intercal \cdot T_0^{(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:t})} \cdot \mathrm{b}_0 \\ &\overset{(c)}{=} \sum_{s_0} p(\boldsymbol{y}_{:\tau}|\boldsymbol{a}_{:\tau}, s_0)\mathrm{b}_0(s_0). \end{aligned} \tag{54}$$

Above, (a) uses Eq. (49), (b) resolves the Dirac delta with the summation and uses Eq. (52), and (c) uses Eq. (53). To conclude, one can notice that if $\mathrm{b}_0(s_0 = \Pr(S_0 = s_0)$, then this shows that (S_t, A_t, Y_t) and (B_t, A_t, Y_t) generate the same interface.

□

K Proof of Prop. 3

Before the proof, let us note that the kernel of I-O Moore transducers can be re-organised as

$$\begin{aligned} p(\boldsymbol{y}_{:t}, \boldsymbol{s}_{:t+1}|\boldsymbol{a}_{:t}) &= p(s_0) \prod_{\tau=0}^{t} \kappa_\tau(y_\tau, s_{\tau+1}|s_\tau, a_\tau) \\ &= p(s_0) \prod_{\tau=0}^{t} \mu_\tau(y_\tau|s_\tau)\nu_\tau(s_{\tau+1}|s_\tau, a_\tau) \\ &= p(s_0) \prod_{\tau=0}^{t} \kappa_\tau^S(y_\tau, s_\tau|s_{\tau-1}, a_{\tau-1}), \end{aligned} \tag{55}$$

where $\kappa_\tau^S(y_{t+1}, s_{t+1}|s_t, a_t)$ is a 'time-shifted' kernel defined as

$$\kappa_\tau^S(y, \tilde{s}|s, a) := \begin{cases} \mu_\tau(y|\tilde{s})\nu_{\tau-1}(\tilde{s}|s, a) & \text{if } \tau \geq 1, \\ \mu_0(y|s) & \text{if } \tau = 0. \end{cases}$$

This means that I-O Moore transducers yield two associated kernels: the standard one κ and the time-shifted one κ^S, and — crucially — both generate the interface. Following Eq. (3), let's defined the time-shifted linear operators

$$T_\tau^{'(y|a)} := \sum_{i=1}^{n} \sum_{j=1}^{n} \kappa_\tau^S(y, s_i|a, s_j)\boldsymbol{e}_i\boldsymbol{e}_j^\intercal \ , \tag{56}$$

with the understanding that for $\tau = 0$ then

$$T_0^{'(y|a)} := \sum_{i=1}^{n} \sum_{j=1}^{n} \mu_0(y|s_j)\boldsymbol{e}_i\boldsymbol{e}_j^\intercal \ . \tag{57}$$

Proof. To derive the update rule of postdictive beliefs, one can combine Eq. (45) and Eq. (46) to find that

$$\mathrm{d}_{t+1}(s_{t+1}) = \frac{p(y_{t+1}|s_{t+1}, a_{t+1})}{Z'} \sum_{s_t} p(s_{t+1}|s_t, h_t)\mathrm{d}_t(s_t). \tag{58}$$

Furthermore, if the transformer is I-O Moore, then

$$p(y_{t+1}|s_{t+1}, a_{t+1}) = p(y_{t+1}|s_{t+1}) \quad \text{and} \quad p(s_{t+1}|s_t, h_t) = p(s_{t+1}|s_t, a_t). \tag{59}$$

Using these relationships in Eq. (58) directly yields the desired update rule.

To prove the second part of the proposition, let us first note for I-O Moore transducers then $p(y_{t+1}|s_{t+1}) = \mu_{t+1}(y_{t+1}|s_{t+1})$ and $p(s_{t+1}|s_t, a_t) = \nu_t(s_{t+1}|s_t, a_t)$. Using this, the update rule can be re-written as $\mathrm{d}_\tau = f'_\tau(y_\tau, a_{\tau-1}, \mathrm{d}_{\tau-1})$ with f' given by

$$\Big[f'_\tau(y, a, \mathrm{d})\Big](s) := \frac{\mu_\tau(y|s)}{Z'} \sum_{s'} \nu_{\tau-1}(s|s', a)\mathrm{d}(s') = \frac{1}{Z'} \sum_{s'} \kappa^S_\tau(y, s|a, s')\mathrm{d}(s'). \tag{60}$$

Moreover, comparing this with Eq. (48), one can find that

$$f'_\tau(y, a, \mathrm{d}) = \frac{T'^{(y|a)}_\tau \cdot \mathrm{d}}{\mathbf{1}^\intercal \cdot T'^{(y|a)}_\tau \cdot \mathrm{d}}, \tag{61}$$

and following a similar derivation one finds that for sequences $\boldsymbol{y}_{:\tau}, \boldsymbol{a}_{:\tau}$ and beliefs $\mathrm{d}_0, \dots, \mathrm{d}_\tau$

$$\begin{aligned} f'_\tau(y_\tau, a_{\tau-1}, \mathrm{d}_{\tau-1}) &= \frac{T'^{(y_\tau|a_{\tau-1})}_\tau \cdot \mathrm{d}_{\tau-1}}{\mathbf{1}^\intercal \cdot T'^{(y_\tau|a_{\tau-1})}_\tau \cdot \mathrm{d}_{\tau-1}} \\ &= \dots \\ &= \frac{T_0^{(\boldsymbol{y}_{:\tau}|\boldsymbol{a}_{:\tau-1})} \cdot \mathrm{d}_0}{\mathbf{1}^\intercal \cdot T_0^{(\boldsymbol{y}_{:\tau}|\boldsymbol{a}_{:\tau-1})} \cdot \mathrm{d}_0}, \end{aligned} \tag{62}$$

where we use $T'^{(\boldsymbol{y}_{t:t'}|\boldsymbol{a}_{t-1:t'-1})}_t := \prod_{\tau=t}^{t'} T'^{(y_\tau|a_{\tau-1})}_\tau$ as a shorthand notation.

With these tools at hand, let's note that Eq. (9) combined with condition (1) in Lemma 6 imply that (D_t, A_t, Y_t) is a belief transducer with unifilar kernel given by

$$\hat{\kappa}^S_\tau(y, \tilde{\mathrm{d}}|a, \mathrm{d}) = \delta_{\tilde{\mathrm{d}}}^{f'_\tau(y,a,\mathrm{d})} \sum_{s\in\mathcal{S}} \kappa^S_\tau(y, \tilde{s}|a, s)\mathrm{d}(s) = \delta_{\tilde{\mathrm{d}}}^{f'_\tau(y,a,\mathrm{d})} \Big(\mathbf{1}^\intercal \cdot T'^{(y|a)}_\tau \cdot \mathrm{d}\Big), \tag{63}$$

being analogous to the result found in Eq. (49). Then, combining all these expressions one can directly calculate what is the result of successive applications of the kernel of a predictive Bayesian belief transducer:

$$\begin{aligned} \sum_{\mathrm{d}_{:t}} \prod_{\tau=0}^{t} \hat{\kappa}^S_\tau(y_\tau, \mathrm{d}_\tau|a_{\tau-1}, \mathrm{d}_{\tau-1}) &\overset{(a)}{=} \sum_{\mathrm{d}_{:t}} \prod_{\tau=0}^{t} \delta_{\mathrm{d}_\tau}^{f'(y_\tau, a_{\tau-1}, \mathrm{d}_{\tau-1})} \Big(\mathbf{1}^\intercal \cdot T'^{(y_\tau|a_{\tau-1})}_\tau \cdot \mathrm{d}_{\tau-1}\Big) \\ &\overset{(b)}{=} \prod_{\tau=0}^{t} \mathbf{1}^\intercal \cdot T^{(y_\tau|a_{\tau-1})}_\tau \cdot \frac{T_0^{(\boldsymbol{y}_{:\tau-1}|\boldsymbol{a}_{:\tau-2})} \cdot \mathrm{d}_0}{\mathbf{1}^\intercal \cdot T_0^{(\boldsymbol{y}_{:\tau-1}|\boldsymbol{a}_{:\tau-2})} \cdot \mathrm{d}_0} \\ &= \mathbf{1}^\intercal \cdot T_0^{(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:t-1})} \cdot \mathrm{d}_0 \\ &= \sum_{s_0} p(\boldsymbol{y}_{:\tau}|\boldsymbol{a}_{:\tau-1}, s_0)\mathrm{d}_0(s_0). \end{aligned} \tag{64}$$

Above, (a) uses Eq. (63), (b) resolves the Dirac delta with the summation and uses Eq. (62). To conclude, note that this shows that (S_t, A_t, Y_t) and (D_t, A_t, Y_t) generate the same interface when $\mathrm{d}_0(s_0) = \Pr(S_0 = s_0)$.

Before finishing, note that Cor. 1 follows directly from recognising Eq. (45) and Eq. (46) as the equations related to 'predict' and 'update' steps in Kalman and Bayesian filtering (Särkkä & Svensson, 2023) corresponding to

$$b_{t-1} = p(s_{t-1}|\boldsymbol{h}_{:t-1}) \xrightarrow{\text{predict}} d_t = p(s_t|\boldsymbol{h}_{:t-1}) \xrightarrow{\text{update}} b_t = p(s_t|\boldsymbol{h}_{:t}). \tag{65}$$

□

L Proof of Prop. 4

This result was originally proven in (Barnett & Crutchfield, 2015, Prop. 2). Here we provide an alternative proof that leverages the link established between transducer homomorphisms and bisimulations.

Proof. We first show that the equivalence class defined by Eq. (10) is a bisimulation of the world model $S_t = \boldsymbol{H}_{:t-1}$. For this, we first show that the coarse-graining defined by Eq. (10) is a bisimulation — i.e., it it satisfies the two properties outlined in Def. 5. Condition (i) follows from Eq. (10) directly, since it only considers futures of length $L = 1$. A proof that Condition (ii) follows from the fact that the dynamics of the equivalence classes are conditionally Markovian given the actions, which has been shown in (Barnett & Crutchfield, 2015, Prop. 6).

With this, the desired result follows directly from noting that $S_t = \boldsymbol{H}_{:t-1}$ is always a valid transducer of $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$ (Lemma 4), and that the bisimulation of a transducer always yields a valid transducer (Prop. 1) that generates the same interface (Lemma 5). □

M Proof of Theorem 3

This proof is a direct extension of (Barnett & Crutchfield, 2015, Lemma 1), which focuses on 'rival partitions' rather than predictive transducers. The core idea of the proof is that, for a given predictive transducer with memory $R_t \in \mathcal{R}$, one can build an equivalence relation in $\mathcal{R} \times \mathcal{R}$ induced by a coarse-graining map ϵ defined as

$$\epsilon'_t(r) = \epsilon'_t(r') \quad \text{iff} \quad \Pr(\boldsymbol{Y}_{t:t'}|\boldsymbol{A}_{t:t'}, R_t = r) = \Pr(\boldsymbol{Y}_{t:t'}|\boldsymbol{A}_{t:t'}, R_t = r') \quad \forall t' \geq t. \tag{66}$$

Then, one can show that if R_t is a predictive world model, then $\epsilon'_t(R_t)$ is isomorphic to the memory states of the ϵ-transducer. The full proof of this is given next, after presenting the following lemma.

Lemma 8. *A predictive transducer* (S_t, A_t, Y_t) *satisfies*

$$p(\boldsymbol{y}_{t:t'}|\boldsymbol{a}_{t:t'}, s_t) = p(\boldsymbol{y}_{t:t'}|\boldsymbol{a}_{t:t'}, s_t, \boldsymbol{h}_{:t-1}) = p(\boldsymbol{y}_{t:t'}|\boldsymbol{a}_{t:t'}, \boldsymbol{h}_{:t-1}) \quad \forall t' \geq t. \tag{67}$$

Proof. By definition (see Sec. 5.1) a predictive transducer has memory states S_t that satisfy the condition $I(\boldsymbol{Y}_{t:}, S_t|\boldsymbol{H}_{:t-1}, \boldsymbol{A}_{t:}) = 0$, which implies that for all $t' \geq t$

$$\begin{aligned} p(\boldsymbol{y}_{t:t'}|\boldsymbol{a}_{t:t'}, s_t, \boldsymbol{h}_{:t-1}) &\overset{(i)}{=} p(\boldsymbol{y}_{t:t'}|\boldsymbol{a}_{t:}, s_t, \boldsymbol{h}_{:t-1}) \\ &= p(\boldsymbol{y}_{t:t'}|\boldsymbol{a}_{t:}, \boldsymbol{h}_{:t-1}) \\ &\overset{(ii)}{=} p(\boldsymbol{y}_{t:t'}|\boldsymbol{a}_{t:t'}, \boldsymbol{h}_{:t-1}), \end{aligned} \tag{68}$$

holding whenever $p(\boldsymbol{y}_{t:t'}, s_t|\boldsymbol{a}_{t:t'}, \boldsymbol{h}_{:t-1}) \neq 0$, which means that these events are compatible. Above, (i) and (ii) use the fact that the interface and world models are non anticipatory. Additionally, using the properties of transducers one can show that

$$\begin{aligned} p(\boldsymbol{y}_{t:t'}|\boldsymbol{a}_{t:t'}, s_t, \boldsymbol{h}_{:t-1}) &= \sum_{\boldsymbol{s}_{t+1:t'+1}} p(\boldsymbol{y}_{t:t'}, \boldsymbol{s}_{t+1:t'+1}|\boldsymbol{a}_{t:t'}, s_t, \boldsymbol{h}_{:t-1}) \\ &= \sum_{\boldsymbol{s}_{t+1:t'+1}} \prod_{\tau=t}^{t'} p(y_\tau, s_{\tau+1}|\boldsymbol{a}_{t:t'}, \boldsymbol{s}_{t:\tau}, \boldsymbol{y}_{t:\tau-1}, \boldsymbol{h}_{:t-1}) \\ &\overset{(iii)}{=} \sum_{\boldsymbol{s}_{t+1:t'+1}} \prod_{\tau=t}^{t'} p(y_\tau, s_{\tau+1}|\boldsymbol{a}_{t:t'}, \boldsymbol{s}_{t:\tau}, \boldsymbol{y}_{t:\tau-1}) \\ &= \sum_{\boldsymbol{s}_{t+1:t'+1}} p(\boldsymbol{y}_{t:t'}, \boldsymbol{s}_{t+1:t'+1}|\boldsymbol{a}_{t:t'}, s_t) \\ &= p(\boldsymbol{y}_{t:t'}|s_t, \boldsymbol{a}_{t:t'}) \end{aligned} \tag{69}$$

for all $t' > t$, where (iii) uses condition (1) from Lemma 6. Finally, combining Eq. (68) and Eq. (69) gives the desired result. □

Proof of Theorem 3. Consider S_t the memory state of a transducer. Let us first note that

$$p(\boldsymbol{y}_{t:t'}|\boldsymbol{a}_{t:t'}, \boldsymbol{h}_{:t-1}) = \sum_{j\in\mathcal{J}} \alpha_j p(\boldsymbol{y}_{t:t'}|\boldsymbol{a}_{t:t'}, s_t, \boldsymbol{h}_{:t-1}) \tag{70}$$

where $\alpha_j = p(s_t|\boldsymbol{h}_{:t-1}, \boldsymbol{a}_{t:t'}) = p(s_t|\boldsymbol{h}_{:t-1})$, where the second equality holds because S_t is a world model (property (2) in Def. 2). Above, we are using a suitable set of indices $\mathcal{J}$ corresponding to the possible values of s_t, which satisfy $\sum_{j\in\mathcal{J}} \alpha_j = \sum_{s_t\in\mathcal{S}} p(s_t|\boldsymbol{h}_{:t-1}) = 1$. Given that the Shannon entropy is concave, Eq. (70) implies

$$H\Big[p(\boldsymbol{y}_{t:t'}|\boldsymbol{a}_{t:t'}, s_t)\Big] \geq \sum_{j\in\mathcal{J}} \alpha_j H\Big[p(\boldsymbol{y}_{t:t'}|\boldsymbol{a}_{t:t'}, s_t, \boldsymbol{h}_{:t-1})\Big], \tag{71}$$

where $H[p]$ is a shorthand notation for the entropy of a variable with distribution p. Moreover, given that H is strictly concave, Eq. (71) turns into an equality if and only if all α_j's are either 0 or 1 for all $j \in \mathcal{J}$.

Now, consider S_t the memory state of a predictive transducer and $M_t = \epsilon(\boldsymbol{H}_{:t-1})$ the memory state of the ϵ-transducer, as determined by the coarse-graining mapping defined in Eq. (10). Note that (M_t, A_t, Y_t) is a predictive transducer, and hence Lemma 8 can be used yielding

$$p\big(\boldsymbol{y}_{t:t'}|\boldsymbol{a}_{t:t'}, \epsilon(\boldsymbol{h}_{:t-1})\big) = p(\boldsymbol{y}_{t:t'}|\boldsymbol{a}_{t:t'}, \epsilon(\boldsymbol{h}_{:t-1}), \boldsymbol{h}_{:t-1}) = p(\boldsymbol{y}_{t:t'}|\boldsymbol{a}_{t:t'}, \boldsymbol{h}_{:t-1}) \tag{72}$$

for all $t' > t$. Then, using Lemma 8 this time on (S_t, A_t, Y_t), one can find that

$$\begin{aligned}\Pr\big(\boldsymbol{Y}_{t:t'}|\boldsymbol{A}_{t:t'}, M_{t-1} = \epsilon(\boldsymbol{h}_{:t-1})\big) &= \Pr(\boldsymbol{Y}_{t:t'}|\boldsymbol{A}_{t:t'}, \boldsymbol{H}_{:t-1} = \boldsymbol{h}_{:t-1}) \\ &= \Pr(\boldsymbol{Y}_{t:t'}|\boldsymbol{A}_{t:t'}, S_t = s_t, \boldsymbol{H}_{:t-1} = \boldsymbol{h}_{:t-1}) \\ &= \Pr(\boldsymbol{Y}_{t:t'}|\boldsymbol{A}_{t:t'}, S_t = s_t).\end{aligned} \tag{73}$$

This implies that for each equivalence class $\epsilon_t(\boldsymbol{h}_{:t-1}) = [\boldsymbol{h}_{:t-1}]_{\sim_\epsilon}$ there exists at least one $s_t \in \mathcal{S}$ such that $\Pr\big(\boldsymbol{Y}_{t:t'}|\boldsymbol{A}_{t:t'}, M_{t-1} = \epsilon(\boldsymbol{h}_{:t-1})\big) = \Pr(\boldsymbol{Y}_{t:t'}|\boldsymbol{A}_{t:t'}, S_t = s_t)$. Moreover, the previous equalities imply that if S_t is a predictive transducer, then Eq. (71) necessarily becomes an equality. This, in turn, implies that $\alpha_j = p(s_t|\boldsymbol{h}_{:t-1})$ is 1 for all $s_t \in \mathcal{S}$ for which

$$\begin{aligned}\Pr(\boldsymbol{Y}_{t:t'}|\boldsymbol{A}_{t:t'}, S_t = s_t) &= \Pr(\boldsymbol{Y}_{t:t'}|\boldsymbol{A}_{t:t'}, S_t = s_t, \boldsymbol{H}_{:t-1} = \boldsymbol{h}_{:t-1}) \\ &= \Pr(\boldsymbol{Y}_{t:t'}|\boldsymbol{A}_{t:t'}, \boldsymbol{H}_{:t-1} = \boldsymbol{h}_{:t-1})\end{aligned} \tag{74}$$

holds, or 0 otherwise. This implies that the mapping ϵ' given by

$$\epsilon'(s_t) = \epsilon'(s'_t) \quad \Leftrightarrow \quad \Pr(\boldsymbol{Y}_{t:t'}|\boldsymbol{A}_{t:t'}, S_t = s_t) = \Pr(\boldsymbol{Y}_{t:t'}|\boldsymbol{A}_{t:t'}, S_t = s'_t) \quad \forall t' \geq t \tag{75}$$

satisfies $\epsilon'_t(S_t) = \epsilon_t(\boldsymbol{H}_{:t-1}) = M_t$. □

N Comparing the reduction of general vs predictive transducers

Let us explain why reducing a predictive transducer is different from reducing an arbitrary transducer. For this, leveraging the discussion about the canonical dimension of a transducer (see Eq. (5)), let us consider on transducers with finite memory states (i.e. $|\mathcal{S}| = n$), and consider the matrix W whose columns given by the vectors $\boldsymbol{w}(\boldsymbol{h}_{:t}) \in \mathbb{R}^n$ of probabilities of generating $\boldsymbol{y}_{:t}$ given $\boldsymbol{a}_{:t}$ when starting from different world states, so that its k-th coordinate is $[\boldsymbol{w}(\boldsymbol{h}_{:t})]_k = \Pr(\boldsymbol{Y}_{:t} = \boldsymbol{y}_{:t}|\boldsymbol{A}_{:t} = \boldsymbol{a}_{:t}, S_0 = s_k)$ for all possible sequences when $t = n-1$ (see (Cakir et al., 2021, Prop. 4.3)). Then, the coarse-graining ϵ defined by Eq. (10) correspond to merging together all rows of W_t that are equal. In contrast, the canonical dimension $d(\mathcal{T})$ defined in Eq. (5) corresponds to the number of

linearly independent rows. The crucial point is that, if a transducer with memory states S_t is predictive, then any coarse-graining $\epsilon(S_t)$ will also be predictive. However, reductions via more general procedures to trim linearly dependent components may not be attainable via coarse-grainings. In particular, the matrix W_t of an ϵ-transducer may have linearly dependent rows, and reducing those would — due to Cor. 2 — necessary make the transducer to stop being predictive.

It is interesting to note that the predictive beliefs of the ϵ-transducer are isomorphic to the states of the ϵ-transducer. However, the ϵ-transducer is not the only machine whose MSP produces the ϵ-transducer — in fact, the MSP of any transducer without redundant states will produce the same.

Finally, it is worth noting that a non-predictive finite world model may have an associate ϵ-transducer that has infinite states filling the simplex of belief states with intricate fractal patterns (Jurgens & Crutchfield, 2021a). This fact makes techniques to study transducers at various degrees of resolution (e.g., via approximate homomorphisms (Taylor et al., 2008) or bisimulation (Girard & Pappas, 2011), or using rate-distortion trade-offs (Marzen & Crutchfield, 2016)) particularly important.

O Example of a non-reversible transducer

Let us provide an example of how reversing a transducer can lead to a violation of the anticipation free condition. Let's consider the so-called delay channel, for which the output Y_{t+1} is equal to the previous action A_t (Barnett & Crutchfield, 2015). This channel displays acausal behaviour when time reversed: somehow the action A_t determines the outcome at the previous time step Y_{t-1}, meaning that

$$I(\boldsymbol{Y}_{:t-1}; \boldsymbol{A}_{t:}|\boldsymbol{A}_{t-1:}) = I(Y_{t-1}; A_t|\boldsymbol{A}_{t-1:}) = H(A_t|\boldsymbol{A}_{t-1:}), \tag{76}$$

which is nonzero if the entropy rate of the actions is nonzero. This leads to an interface that does not satisfy anticipation free, violating the conditions of Def. 1.

P Reversing processes and proof of Theorem 4

Here we present an extended exposition of the conditions for reversing various types of stochastic processes.

P.1 Reversing Markov processes

Let's start by considering a Markov process X_t, so that $p(x_t|\boldsymbol{x}_{:t-1}) = p(x_t|x_{t-1})$ for all $t \in \mathbb{N}$. Then the reverse process (given by $X_t, X_{t-1}, \dots$) is also Markov, as

$$\begin{aligned} p(x_t|\boldsymbol{x}_{t+1:t'}) &= \frac{p(\boldsymbol{x}_{t:t'})}{p(\boldsymbol{x}_{t+1:t'})} = \frac{p(x_t)\prod_{k=t+1}^{t'} p(x_k|\boldsymbol{x}_{t:k-1})}{p(x_{t+1})\prod_{j=t+2}^{t'} p(x_j|\boldsymbol{x}_{t+1:j-1})} \\ &= \frac{p(x_t)\prod_{k=t+1}^{t'} p(x_k|x_{k-1})}{p(x_{t+1})\prod_{j=t+2}^{t'} p(x_j|x_{j-1})} = \frac{p(x_t)p(x_{t+1}|x_t)}{p(x_{t+1})} = p(x_t|x_{t+1}). \end{aligned} \tag{77}$$

P.2 Reversing hidden Markov models

Let's now consider a general (i.e. Mealy (Riechers, 2016)) hidden Markov model, in which $p(s_{t+1}, y_t|\boldsymbol{s}_{:t}, \boldsymbol{y}_{:t-1}) = p(s_{t+1}, y_t|s_t)$ holds for all $t \in \mathbb{N}$. As for Markov chains, one can show

that the reverse process is also an hidden Markov model, as

$$\begin{aligned}
p(s_t, y_t|\boldsymbol{s}_{t+1:t'+1}, \boldsymbol{y}_{t+1:t'}) &= \frac{p(\boldsymbol{s}_{t:t'+1}, \boldsymbol{y}_{t:t'})}{p(\boldsymbol{s}_{t+1:t'+1}, \boldsymbol{y}_{t+1:t'})} \\
&= \frac{p(s_t, y_t, s_{t+1}) \prod_{k=t+1}^{t'} p(s_{k+1}, y_k|\boldsymbol{s}_{t:k}, \boldsymbol{y}_{t:k-1})}{p(s_{t+1}, y_{t+1}, s_{t+2}) \prod_{j=t+2}^{t'} p(s_{j+1}, y_j|\boldsymbol{s}_{t:j}, \boldsymbol{y}_{t:j-1})} \\
&= \frac{p(s_t, y_t, s_{t+1}) \prod_{k=t+1}^{t'} p(s_{k+1}, y_k|s_k)}{p(s_{t+1}, y_{t+1}, s_{t+2}) \prod_{j=t+2}^{t'} p(s_{j+1}, y_j|s_j)} \\
&= \frac{p(s_t) \prod_{k=t}^{t'} p(s_{k+1}, y_k|s_k)}{p(s_{t+1}) \prod_{j=t+1}^{t'} p(s_{j+1}, y_j|s_j)} \\
&= \frac{p(s_t)p(s_{t+1}, y_t|s_t)}{p(s_{t+1})} \\
&= p(s_t, y_t|s_{t+1}).
\end{aligned} \tag{78}$$

Note that in this case the structure is not perfectly time-symmetric, but could be described as 'co-Mealy' structure — as the time indices of the world are shifted.

If the hidden Markov model is Moore (Riechers, 2016), so that $p(s_{t+1}, y_t|\boldsymbol{s}_{:t}, \boldsymbol{y}_{:t-1}) = p(s_{t+1}|s_t)p(y_t|s_t)$, then a similar calculation leads to

$$p(s_t, y_t|\boldsymbol{s}_{t+1:t'+1}, \boldsymbol{y}_{t+1:t'}) = \frac{p(s_t)p(s_{t+1}, y_t|s_t)}{p(s_{t+1})} = \frac{p(s_t)p(s_{t+1}|s_t)p(y_t|s_t)}{p(s_{t+1})} = p(s_t|s_{t+1})p(y_t|s_t), \tag{79}$$

yielding another Moore hidden Markov model.

P.3 Reversing transducers

Using the previous calculations as a foundation, let's now explore the reverse properties of a transducer, where $p(s_{t+1}, y_t|\boldsymbol{s}_{:t}, \boldsymbol{y}_{:t-1}, \boldsymbol{a}_{:}) = p(s_{t+1}, y_t|s_t, a_t)$ holds (see App. D). Using this property, it is direct to see that

$$\begin{aligned}
p(\boldsymbol{y}_{:t}, \boldsymbol{s}_{:t+1}|\boldsymbol{a}_{:}) &= p(s_0) \prod_{\tau=0}^{t} p(y_\tau, s_{\tau+1}|\boldsymbol{y}_{:\tau-1}, \boldsymbol{s}_{:\tau}, \boldsymbol{a}_{:}) \\
&= p(s_0) \prod_{\tau=0}^{t} p(y_\tau, s_{\tau+1}|s_\tau, \boldsymbol{a}_{:t}) \\
&= p(\boldsymbol{y}_{:t}, \boldsymbol{s}_{:t+1}|\boldsymbol{a}_{:t}),
\end{aligned} \tag{80}$$

showing that transducers naturally impose some arrow of time over actions.[10] Now, let's consider expressing $p(\boldsymbol{y}_{:t}, \boldsymbol{s}_{:t+1}|\boldsymbol{a}_{:})$ factor backwards as follows:

$$p(\boldsymbol{y}_{:t}, \boldsymbol{s}_{:t+1}|\boldsymbol{a}_{:}) = p(\boldsymbol{y}_{:t}, \boldsymbol{s}_{:t+1}|\boldsymbol{a}_{:t}) = p(s_{t+1}|\boldsymbol{a}_{:t}) \prod_{\tau=0}^{t} p(y_\tau, s_\tau|\boldsymbol{y}_{\tau+1:t}, \boldsymbol{s}_{\tau+1:t+1}, \boldsymbol{a}_{:t}). \tag{81}$$

[10] Note that the derivation uses the fact that $p(s_0|\boldsymbol{a}_{:}) = p(s_0)$, and it wouldn't work for other initial point where this does not hold.

This shows that we need to looks for ways of simplifying expressions of the form $p(y_\tau, s_\tau | \boldsymbol{y}_{\tau+1:t}, \boldsymbol{s}_{\tau+1:t+1}, \boldsymbol{a}_{:t})$. Using the properties of transducers, one can show that

$$\begin{aligned}
p(s_\tau, y_\tau | \boldsymbol{s}_{\tau+1:t+1}, \boldsymbol{y}_{\tau+1:t}, \boldsymbol{a}_{:t}) &= \frac{p(\boldsymbol{s}_{\tau:t+1}, \boldsymbol{y}_{\tau:t}, \boldsymbol{a}_{:t})}{p(\boldsymbol{s}_{\tau+1:t+1}, \boldsymbol{y}_{\tau+1:t}, \boldsymbol{a}_{:t})} \\
&= \frac{p(s_\tau, y_\tau, s_{\tau+1}, \boldsymbol{a}_{:t}) \prod_{k=\tau+1}^{t} p(s_{k+1}, y_k | \boldsymbol{s}_{\tau:k}, \boldsymbol{y}_{\tau:k-1}, \boldsymbol{a}_{:t})}{p(s_{\tau+1}, y_{\tau+1}, s_{\tau+2}, \boldsymbol{a}_{:t}) \prod_{j=\tau+2}^{t} p(s_{j+1}, y_j | \boldsymbol{s}_{\tau:j}, \boldsymbol{y}_{\tau:j-1}, \boldsymbol{a}_{:t})} \\
&= \frac{p(s_\tau, y_\tau, s_{\tau+1}, \boldsymbol{a}_{:t}) \prod_{k=\tau+1}^{t} p(s_{k+1}, y_k | s_k, a_k)}{p(s_{\tau+1}, y_{\tau+1}, s_{\tau+2}, \boldsymbol{a}_{:t}) \prod_{j=\tau+2}^{t} p(s_{j+1}, y_j | s_j, a_j)} \\
&= \frac{p(s_\tau, \boldsymbol{a}_{:t}) \prod_{k=\tau}^{t} p(s_{k+1}, y_k | s_k, a_k)}{p(s_{\tau+1}, \boldsymbol{a}_{:t}) \prod_{j=\tau+1}^{t} p(s_{j+1}, y_j | s_j, a_j)} \\
&= \frac{p(s_\tau | \boldsymbol{a}_{:t}) p(s_{\tau+1}, y_\tau | s_\tau, a_\tau)}{p(s_{\tau+1} | \boldsymbol{a}_{:t})} \\
&= \frac{p(s_\tau | \boldsymbol{a}_{:t}) p(s_{\tau+1} | s_\tau, a_\tau)}{p(s_{\tau+1} | \boldsymbol{a}_{:t})} p(y_\tau | s_{\tau+1}, s_\tau, a_\tau) \\
&= p(s_\tau | s_{\tau+1}, \boldsymbol{a}_{:t}) p(y_\tau | s_\tau, s_{\tau+1}, a_\tau). \qquad (82)
\end{aligned}$$

This shows that, for any transducer (S_t, A_t, Y_t), we can always 'run it back' using the whole sequence of actions to reproduce the interface, as shown by the factorisation

$$p(\boldsymbol{y}_{:t}, \boldsymbol{s}_{:t+1} | \boldsymbol{a}_{:}) = p(s_{t+1} | \boldsymbol{a}_{:t}) \prod_{\tau=0}^{t} p(s_\tau | s_{\tau+1}, \boldsymbol{a}_{:t}) p(y_\tau |, s_\tau, s_{\tau+1}, a_\tau). \qquad (83)$$

If the transducer satisfies the additional condition $p(s_\tau | s_{\tau+1}, \boldsymbol{a}_{:t}) = p(s_\tau | s_{\tau+1}, a_\tau)$, or equivalently the information relation $I(\boldsymbol{S}_\tau; \boldsymbol{A}_{0:\tau-1} \boldsymbol{A}_{\tau+1:t} | S_{\tau+1}, A_\tau) = 0$, then one obtains a reverse factorisation of the form

$$p(\boldsymbol{y}_{:t}, \boldsymbol{s}_{:t+1} | \boldsymbol{a}_{:}) = p(s_{t+1} | \boldsymbol{a}_{:t}) \prod_{\tau=0}^{t} p(y_\tau, s_\tau | s_{\tau+1}, a_t). \qquad (84)$$

The reverse kernel κ^R can be expressed in terms of the forward one via the following derivation:

$$\kappa_\tau^R(y_\tau, s_\tau | a_\tau, s_{\tau+1}) = p(y_\tau, s_\tau | a_{:\tau}, s_{\tau+1}) = \frac{p(s_\tau | a_\tau)}{p(s_{\tau+1} | a_\tau)} p(y_\tau, s_{\tau+1} | a_{:\tau}, s_\tau), \qquad (85)$$

which implies that for reversible transducers then

$$\kappa_\tau^R(y, s | a, \tilde{s}) = \frac{\Pr(S_\tau = s | A_\tau = a)}{\Pr(S_{\tau+1} = \tilde{s} | A_\tau = a)} \kappa_\tau(y, \tilde{s} | a, s). \qquad (86)$$

It is interesting to note that that replacing Eq. (86) in Eq. (84) could give the impression of not recovering Eq. (4); however, under the assumption of transducer reversibility it does. To confirm this, let us first check that

$$p(s_\tau | s_{\tau+1}, \boldsymbol{a}_{:t}) = \frac{p(s_\tau, s_{\tau+1} | \boldsymbol{a}_{:t})}{p(s_{\tau+1} | \boldsymbol{a}_{:t})} = \frac{p(s_\tau | \boldsymbol{a}_{:t}) p(s_{\tau+1} | s_\tau, \boldsymbol{a}_{:t})}{p(s_{\tau+1} | \boldsymbol{a}_{:t})}, \qquad (87)$$

and similarly

$$p(s_\tau | s_{\tau+1}, a_\tau) = \frac{p(s_\tau, s_{\tau+1} | a_\tau)}{p(s_{\tau+1} | a_\tau)} = \frac{p(s_\tau | a_\tau) p(s_{\tau+1} | s_\tau, a_\tau)}{p(s_{\tau+1} | a_\tau)}. \qquad (88)$$

Now, as s_τ is the memory state of a transducer then $p(s_{\tau+1}|s_\tau, \boldsymbol{a}_{:t}) = p(s_{\tau+1}|s_\tau, a_t)$. Putting all together, the condition $p(s_t|s_{t+1}, \boldsymbol{a}_{:t}) = p(s_t|s_{t+1}, a_t)$ can be seen to imply the following:

$$\frac{p(s_\tau|\boldsymbol{a}_{:t})}{p(s_{\tau+1}|\boldsymbol{a}_{:t})} = \frac{p(s_\tau|a_\tau)}{p(s_{\tau+1}|a_\tau)}. \tag{89}$$

This equality allows us to confirm that replacing Eq. (13) in Eq. (84) indeed recovers Eq. (4).

In conclusion, if $p(s_\tau|s_{\tau+1}, \boldsymbol{a}_{:t}) = p(s_\tau|s_{\tau+1}, a_\tau)$ holds then one can generate the interface via the following procedure:

1. Initialise the state of the world at time $t+1$ by sampling $p(s_{t+1}|\boldsymbol{a}_{:t})$. Alternatively, for counterfactual analyses pick an arbitrary world state $s \in \mathcal{S}$ and set $S_{t+1} = s$.
2. Run the transducer backwards using the kernel $\kappa^R(y_\tau, s_\tau|a_\tau, s_{\tau+1}) = p(y_\tau, s_\tau|s_{\tau+1}, a_t)$.

P.4 Effect of action-unifiliarity

A transducer is action-unifilar if $p(s_{\tau+1}|s_\tau, a_\tau) = \delta_{s_{\tau+1}}^{f(s_\tau, a_\tau)}$ with $S_{\tau+1} = f(S_\tau, A_\tau)$ a function. If the dynamics of the transducer is action-counifilar, meaning that $p(s_\tau|s_{\tau+1}, a_\tau) = \delta_{s_\tau}^{r(s_{\tau+1}, a_\tau)}$ where $S_\tau = r(S_{\tau+1}, A_\tau)$, then we necessarily safisfy the condition of being reversible $p(s_\tau|s_{\tau+1}, a_{:\tau}) = p(s_\tau|s_{\tau+1}, a_\tau)$. However, this is much more restrictive if than action-unifilarity if we insist that every world-state can accept every action $\sum_{s_{\tau+1}} p(s_{\tau+1}|s_\tau, a_\tau) = 1$. Using Bayes rule

$$\begin{aligned} p(s_{\tau+1}|s_\tau, a_\tau) &= p(s_\tau|s_{\tau+1}, a_\tau)\frac{p(s_{\tau+1}|a_\tau)}{p(s_\tau|a_\tau)} \\ &= \delta_{s_\tau}^{r(s_{\tau+1}, a_\tau)}\frac{p(s_{\tau+1}|a_\tau)}{p(s_\tau|a_\tau)}, \end{aligned} \tag{90}$$

we see that there is one nonzero transition for every combination of state $s_{\tau+1}$ and action a_τ. We can think of each transition as an edge betweens states labeled with the action, like a driven transition. This means that there are $|\mathcal{A}|$ transitions per state s_τ. The condition that every world-state can accept every action means that every state has at least one outgoing edge for every action. If this were a non-unifilar model, this would mean that there an action that had two or more outgoing edges. However, that would mean that the total number of edges in the automata is larger than $|\mathcal{A}||\mathcal{S}|$, which is a contradiction. Thus, each state s_τ has exactly one outgoing edge for each action a_τ, meaning that the next state is a function of these states

$$S_{\tau+1} = f(S_\tau, A_\tau). \tag{91}$$

Therefore, every action-counifilar transducer is also action-unifilar, meaning that it obeys a type of reversibility.

Q Proof of Theorem 5

For convenience, in this proof we use Dirac's notation, which uses bras like $\langle v|$ and kets like $|v\rangle$ to express row and column vectors respectively. If we are describing vectors and matrices over states $\mathcal{S}$, then we can use an orthonormal basis $(\{|s\rangle\}_{s\in\mathcal{S}}$ such that $\langle s|s'\rangle = \delta_{s,s'})$ in the Hilbert space $\mathcal{H}_\mathcal{S}$ to express the vector

$$|v\rangle = \sum_s v(s)|s\rangle. \tag{92}$$

Here, $v(s)$ represents the sth element of the vector. Similarly, for a linear operator in this Hilbert space, we can think of

$$\langle s'|M|s\rangle, \tag{93}$$

as the element in the sth row and s'th column, and we can translate a matrix A with elements $A_{ss'}$ into a linear operator in this space by using the outer-product

$$A = \sum_{ss'} |s'\rangle A_{ss'} \langle s|. \tag{94}$$

Using this notation, we can consider a vector space $\mathbb{R}^{|\mathcal{S}|}$ using the orthonormal basis of states $\{|s\rangle\}_{s\in\mathcal{S}}$ such that $\langle s|s'\rangle = \delta_{s,s'}$. Then, the predictive Bayesian belief can be described as

$$|\rho^P(\boldsymbol{y}_{:t}, \boldsymbol{a}_{:t})\rangle = \sum_{s_{t+1}} |s_{t+1}\rangle p(s_{t+1}|\boldsymbol{h}_{:t}) \ , \tag{95}$$

and the retrodictive Bayesian belief as

$$\langle \rho^R(\boldsymbol{y}_{:t}, \boldsymbol{a}_{:t})| = \sum_{s_0} p(s_0|\boldsymbol{h}_{:t}) \langle s_0| \ . \tag{96}$$

Similarly, the matrix corresponding a sequence of actions $\boldsymbol{a}_{:t}$ and outputs $\boldsymbol{y}_{:t}$ can be described as

$$T^{(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:t})} = \prod_{\tau=0}^{t} T^{(y_\tau|a_\tau)} = \sum_{s_0, s_{\tau+1}} |s_{\tau+1}\rangle p(s_{\tau+1}, \boldsymbol{y}_{:\tau}|\boldsymbol{a}_{:\tau}, s_0) \langle s_0|, \tag{97}$$

If we define the initial diagonal operator $\rho_t \equiv \sum_{s_t} |s_t\rangle p(s_t) \langle s_t|$, then we can calculate the probability of joint start and end state as follows

$$T^{(\boldsymbol{y}_{:\tau}|\boldsymbol{a}_{:\tau})} \rho_0 = \sum_{s_0, s_{\tau+1}} |s_{\tau+1}\rangle p(s_{\tau+1}, s_0, \boldsymbol{y}_{:\tau}|\boldsymbol{a}_{:\tau}) \langle s_0|. \tag{98}$$

With this, one can calculate the interface via expressions of the form

$$p(\boldsymbol{y}_{:\tau}|\boldsymbol{a}_{:\tau}) = \langle 1| T^{(\boldsymbol{y}_{:\tau}|\boldsymbol{a}_{:\tau})} \rho_0 |1\rangle, \tag{99}$$

where $|1\rangle \equiv \sum_s |s\rangle$.

Proof of Theorem 5. Using this notation, the BMSM (Sec. 6.2) can be expressed as

$$\rho(\boldsymbol{y}_{:\tau}, \boldsymbol{a}_{:\tau}) = \sum_{s_0, s_{\tau+1}} |s_{\tau+1}\rangle p(s_{\tau+1}, s_0|\boldsymbol{y}_{:\tau}, \boldsymbol{a}_{:\tau}) \langle s_0| \ . \tag{100}$$

Moreover, by using Eq. (98) and Eq. (99) one can find that

$$\rho(\boldsymbol{y}_{:\tau}, \boldsymbol{a}_{:\tau}) = \sum_{s_0, s_{\tau+1}} \frac{|s_{\tau+1}\rangle p(s_{\tau+1}, s_0, \boldsymbol{y}_{:\tau}|\boldsymbol{a}_{:\tau}) \langle s_0|}{p(\boldsymbol{y}_{:\tau}|\boldsymbol{a}_{:\tau})} = \sum_{s_0, s_{\tau+1}} \frac{|s_{\tau+1}\rangle T^{(\boldsymbol{y}_{:\tau}|\boldsymbol{a}_{:\tau})} \rho_0 \langle s_0|}{\langle 1| T^{(\boldsymbol{y}_{:\tau}|\boldsymbol{a}_{:\tau})} \rho_0 |1\rangle} \ , \tag{101}$$

which proves the first part of the theorem. Additionally, by comparing this with Eq. (95) and Eq. (96) one finds that

$$\rho(\boldsymbol{y}_{:\tau}, \boldsymbol{a}_{:\tau})|1\rangle = |\rho^P(\boldsymbol{y}_{:t}, \boldsymbol{a}_{:t})\rangle \quad \text{and} \quad \langle 1|\rho(\boldsymbol{y}_{:\tau}, \boldsymbol{a}_{:\tau}) = \langle \rho^R(\boldsymbol{y}_{:t}, \boldsymbol{a}_{:t})| \ , \tag{102}$$

which proves the second part of the theorem.

The corollary can be proven by noticing that

$$\rho(\boldsymbol{y}_{:\tau+1}, \boldsymbol{a}_{:\tau+1}) = \frac{T^{(y_{\tau+1}|a_{\tau+1})} \rho(\boldsymbol{y}_{:\tau}, \boldsymbol{a}_{:\tau})}{\langle 1| T^{(y_{\tau+1}|a_{\tau+1})} \rho(\boldsymbol{y}_{:\tau}, \boldsymbol{a}_{:\tau})|1\rangle}, \tag{103}$$

where the denominator is a normalisation term. By contrast, the reverse-time update requires applying a modified version of the transducer operator $\rho_0^{-1}T^{(y|a)}\rho_0$ and normalizing:

$$\rho(\boldsymbol{y}_{-1:\tau}, \boldsymbol{a}_{-1:\tau}) = \frac{\rho(\boldsymbol{y}_{:\tau}, \boldsymbol{a}_{:\tau})\rho_0^{-1}T^{(y_{-1}|a_{-1})}\rho_{-1}}{\langle 1|\rho(\boldsymbol{y}_{:\tau}, \boldsymbol{a}_{:\tau})\rho_0^{-1}T^{(y_{-1}|a_{-1})}\rho_{-1}|1\rangle}. \tag{104}$$

Reflecting the fact that not every transducer is reversible, the operation of $\rho_0^{-1}T^{(y|a)}\rho_0$ cannot necessarily be interpreted as the action of a transducer. However, it is nevertheless a valid method for retrodicting the state distribution of the world. □

It is important to note that, since not every transducer is reversible, the operation $\rho_t^{-1}T^{(y|a)}\rho_{t-1}$ generally does not yield the action of a transducer. This operation is, nevertheless, a valid method for retrodicting the state distribution of a world model if its initial state is assumed to be uncorrelated with future action sequences.