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

Anonymous authors
Paper under double-blind review

Keywords: World modelling, POMDP, agent sandboxing, AI safety, AI interpretability.

Summary

World models provide controlled virtual environments in which AI agents can be tested before deployment to ensure their reliability and safety. Unfortunately, the scope and depth of safety assessments can be severely restricted by the computational demands imposed by high-fidelity simulations. Inspired by the classic 'brain in a vat' thought experiment, here we investigate ways to simplify world models that remain agnostic to the AI agent under evaluation. Our analysis reveals fundamental trade-offs in the construction of world models related to their computational efficiency and interpretability. We identify procedures to build world models that either minimise memory requirements, delineate the limits of what a capable agent could learn about the world, or enable retrospective analyses to reveal the causes of undesirable outcomes. In doing so, we take a first step toward charting the fundamental limits of agent sandboxing, while establishing a common language bridging reinforcement learning, control theory, and computational mechanics.

Contribution(s)

1. This paper conceptualises and formalises a novel problem: building efficient world models to sandbox and evaluate the safety of AI agents before deployment.
   Context: Prior work (e.g. (Ha & Schmidhuber, 2018; Hafner et al., 2020)) has used world models for boosting performance, and has not considered this safety-inspired perspective.
2. We introduce generalised transducers based on quasi-probabilities, which lead to a computationally efficient approach to reduce world models.
   Context: Generalised transducers are an extension of generalised hidden Markov models, which have been thoroughly studied by previous work (Upper, 1997; Vidyasagar, 2011).
3. We provide a unifying formal framework to investigate and reason about world models of beliefs, and show that all such models can be bisimulated into a cannonical world model known as the ϵ-transducer.
   Context: The minimality of the ϵ-transducer among predictive processes was proven in (Barnett & Crutchfield, 2015), without investigating the 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 it. 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 safe agent sandboxing

Anonymous authors
Paper under double-blind review

Abstract

World models provide controlled virtual environments in which AI agents can be tested before deployment to ensure their reliability and safety. Unfortunately, the scope and depth of safety assessments can be severely restricted by the computational demands imposed by high-fidelity simulations. Inspired by the classic 'brain in a vat' thought experiment, here we investigate ways to simplify world models that remain agnostic to the AI agent under evaluation. Our analysis reveals fundamental trade-offs in the construction of world models related to their computational efficiency and interpretability. We identify procedures to build world models that either minimise memory requirements, delineate the limits of what a capable agent could learn about the world, or enable retrospective analyses to reveal the causes of undesirable outcomes. In doing so, we take a first step toward charting the fundamental limits of agent sandboxing, while establishing a common language bridging reinforcement learning, control theory, and computational mechanics.

1 Introduction

Breakthroughs in deep 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) are expected to drive a new wave of technological innovation with the potential to benefit every sector of the global economy (Larsen et al., 2024). Alongside all these benefits, the proliferation of increasingly advanced autonomous AI systems will also bring important new risks regarding their safety, controllability, and alignment to 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 world models as sandbox environments to test and evaluate AI agents without real-world consequences (Dalrymple et al., 2024; Díaz-Rodríguez et al., 2023; EU Council, 2024). These simulated environments are ideal for observing how AI agents handle edge cases and respond to novel situations while pursuing their objectives, potentially revealing safety issues or alignment failures before deployment (He et al., 2024). However, the efficacy of this approach 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 unfortunate restrictions on the breadth and depth of safety assessments.

In this work we address these issues by investigating the fundamental limits that shape the design of world models for AI sandboxing. By bridging concepts from different disciplines, we identify a fundamental trade-off between the computational efficiency of a world model and its interpretability. Moreover, we identify between *forward* and *reverse* interpretability approaches, where the former characterises the predictive capabilities of agents and the latter enables retrodictive analyses that can identify the roots of undesirable outcomes. We provide practical suggestions for building world models that are optimal according to different desiderata, while making no assumptions about the agent's policy or capabilities.

2 Scenario and approach

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

H. von Helmholtz, *Handbuch der physiologischen Optik* (1867)

Consider the task of designing a world model to sandbox and test the safety of 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 real world's structure and dynamics. This could be seen as suggesting that designing reliable world models is critically limited 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 accurately simulate their dynamics.

To illustrate this idea, consider how one could construct a world model to sandbox a small agent such as a bacterium. While one could in principle run a simulation that includes the quantum dynamics of the whole planet, such simulation would be not only computationally unfeasible but also unnecessary to answer most questions of interest at that scale. Indeed, such a world model would likely be too spatially extended (by including regions of the planet that are inaccessible to the agent) and too high-resolution (by including quantum effects for a fundamentally classical agent). To avoid this, the designer could instead choose to build an a more computationally-efficient world model that factor out indistinguishable properties from the bacterium perspective, and instead focuses on sensorimotor contingencies (O'Regan & Noë, 2001; Baltieri & Buckley, 2017; 2019; Tschantz et al., 2020; Mannella et al., 2021), or in the agent's 'interface' that only considers information relevant for an agent and the particular task at end (Zhang et al., 2021).

Related questions have been extensively investigated in the philosophy of mind and cognitive (neuro)science literatures for decades, and more recently in reinforcement learning. These investigations highlight the fact that while an agent's actions turn into outcomes due to the mediation of the external world, the agent has no direct access to the world 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 suggests that if organism's brain were to be placed inside 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]

Following this line of reasoning, an ideal world model should depend only on three key elements: (i) the set of possible actions of the agent $\mathcal{A}$, (ii) the set of possible outcomes affecting the agent $\mathcal{Y}$,[2] and (iii) the statistical relationship between action sequences and outcomes. Crucially, it should be possible to build a compressed representation of the effective world of an AI agent, such that it cannot be distinguished from a full simulation — irrespective of how smart or powerful it may be. This *'AI in a vat'* perspective suggests that designers should not focus on a single world model, but instead consider the class of all world models that are indistinguishable from the AI agent's perspective, characterise their properties, and then use different ones depending on specific priorities. The remainder of this article formalises some of these issues and takes steps towards their resolution, while identifying fundamental trade-offs intrinsic to the design of world models.

[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 $\mathcal{Y} = \mathcal{O} \times \mathbb{R}$.

3 Generating interfaces via transducers

We start by formalising the ideas of 'world model' and 'interface'. In the following, uppercase letters (e.g. X, Y) denote random variables and lowercase (e.g. x, y) their realisations, $\mathbb{N} = \{0, 1, 2, \ldots\}$ corresponds to zero-based numbering. 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 $p(x|y) = p(x)$ hold for all realisations that can take place with non-zero probability. We also 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 $t \in \mathbb{N}$.*

Interfaces can be generated from an underlying world model that describes the transduction of actions into outcomes. Notably, interfaces are agnostic to the computational capabilities of agents, their architecture, and internal functioning. We next introduce a general notion of world model stated in terms of sufficient statistics (App. A), and use the shorthand notation $h_t = (a_t, y_t)$ 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 $p(\boldsymbol{s}_{:t}|\boldsymbol{h}_{:})$ for $t \in \mathbb{N}$ corresponding to a stochastic process over sequences of states $\boldsymbol{s}_{:} := (s_0, s_1, \ldots) \in \mathcal{S}^{\mathbb{N}}$ that satisfies*

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

*A world model is **anticipation-free** if it also satisfies* $(3)\ p(\boldsymbol{s}_{:t}|\boldsymbol{h}_{:t-1}, \boldsymbol{a}_{t':}) = p(\boldsymbol{s}_{:t}|\boldsymbol{h}_{:t-1}) \quad \forall t' \geq t.$

Intuitively, world models are auxiliary stochastic processes that 'unravel' interfaces. More precisely, world models encapsulate the relevant information between the past events and future outcomes (condition 1) and guarantee the arrow of time (conditions 2 & 3). This definition, together with the one of an interface, generalise popular modelling approaches such as partially observed Markov decision processes (POMDPs) (Kaelbling et al., 1998) (see App. B). We may denote a world model informally simply by S_t when it is unambiguous from context.

A key property of anticipation-free world models is that they allow to express interfaces as (App. C)

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

This provides a description of the interface in terms of a probabilistic graphical model (Koller & Friedman, 2009), which can be used to efficiently simulate it. Such graphical model allows, among other things, to generate outcomes for given sequence of actions $\boldsymbol{a}_{:\tau}$ and world states $\boldsymbol{s}_{:\tau}$ by directly sampling the posterior distribution $p(\boldsymbol{y}_{:\tau}|\boldsymbol{s}_{:\tau+1}, \boldsymbol{a}_{:}) = \prod_{t=0}^{\tau} p(y_t|s_t, a_t)$. 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.

3.2 Transducers

Unfortunately, sampling of world trajectories can be highly non-trivial as their dynamics may be non-Markovian. One way to address this problem is to build world models via *transducers* (Barnett & Crutchfield, 2015), a computational structure that we introduce next.

Definition 3. *A **transducer** is a tuple $(\mathcal{S}, \mathcal{Y}, \mathcal{A}, \mathcal{K}, p)$, where $\mathcal{S}$ is the set of memory states, $\mathcal{A}$ and $\mathcal{Y}$ are the sets of inputs and outputs, $\mathcal{K} = \{\kappa_t(y, s'|a, s) : a \in \mathcal{A}, y \in \mathcal{Y}, s, s' \in \mathcal{S}, t \in \mathbb{N}\}$ is a collection of stochastic kernels, and p is an initial distribution for the memory states.*

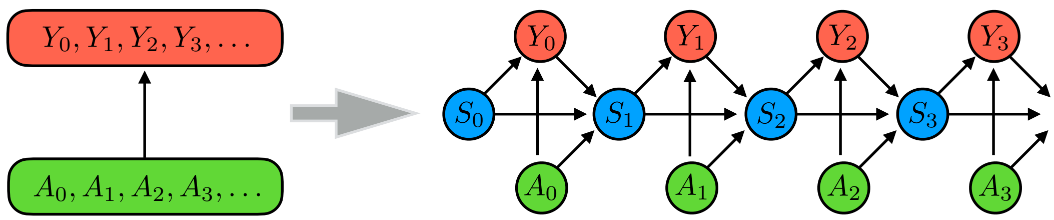

Figure 1: Illustration of an interface (left) and its unravelling via a presentation with world model built from the memory states of a transducer (right), shown in Eq. (4).

We may denote transducers informally as (Y_t, A_t, S_t) when it is unambiguous from the context. If the transducer's memory can only take $|\mathcal{S}| = n$ different states, then the transducer's dynamics can be described via symbol-labelled substochastic matrices $T_t^{(y|a)}$ of the form

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

with $\kappa_t(y, s'|a, s) = \Pr(Y_t = y, S_{t+1} = s' | A_t = a, S_t = s)$ a Markov kernel and $\boldsymbol{e}_k$ a binary vector with a one at the k-th position and zeros elsewhere. Transducers are closely related to stochastic automata (Claus, 1971; Cakir et al., 2021), a generalisation of classic automata (Minsky, 1967) that use stochastic transitions to generate outputs and update their state. In the degenerate case where $p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:}) = p(\boldsymbol{y}_{:t})$, corresponding to 'contemplative' agents that do not act but only sense, transducers reduce to a hidden Markov models (Ephraim & Merhav, 2002).

Our next result provides alternative characterisations of transducers, which clarify under which conditions the memory states can be used as the world model of an interface (proof is given in App. D).

Lemma 1. *The following are alternative characterisations of a transducer:*

1. *S_t is an anticipation-free world model for $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$ whose dynamics satisfy $p(s_{t+1}|\boldsymbol{s}_{:t}, \boldsymbol{h}_{:t}) = p(s_{t+1}|s_t, h_t)$ for all $t \geq 0$.*
2. *S_t satisfies $p(s_0|\boldsymbol{a}_{:}) = p(s_0)$ 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 $t \geq 0$.*
3. *S_t satisfies $I(\boldsymbol{S}_{:t}, \boldsymbol{Y}_{:t-1}; \boldsymbol{A}_{t:} | \boldsymbol{A}_{:t-1}, S_0) = I(\boldsymbol{S}_{t+1:}, \boldsymbol{Y}_{t:}; \boldsymbol{Y}_{:t-1}, \boldsymbol{S}_{:t-1}, \boldsymbol{A}_{:t-1} | \boldsymbol{A}_{t:}, S_t) = 0$.*

Lemma 1 implies that *transducers are world models with Markovian dynamics*. Thanks to this, transducers can be used to conveniently express interfaces as

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

providing a graphical model that can be used to simulate the interface (Figure 1). In this construction, (s_{t+1}, y_t) gets generated jointly out of (s_t, a_t), corresponding to what the literature describes as a 'Mealy' machine (Virgo, 2023; Bonchi et al., 2024). This can be made simpler in several ways. Following the HMM literature (Riechers, 2016), we define an *output-Moore* transducer as the ones satisfying $p(s_{t+1} \mid s_t, h_t) = p(s_{t+1} \mid s_t, a_t)$, so that the future world state do not depend on the current output conditioned on the present state. Alternatively, following the automata literature (Lee & Seshia, 2017), we define an *input-Moore* transducer as the ones satisfying $p(y_t | s_t, s_{t+1}, a_t) = p(y_t | s_t, s_{t+1})$, so that the output does not depend on the current action.[3] Finally, both conditions can be combined to form ***I-O Moore*** transducers that satisfy $p(y_t | s_t, s_{t+1}, a_t) = p(y_t | s_t)$, which correspond to partially observed Markov decision processes (POMDPs) (Kaelbling et al., 1998) as shown in App. B.

[3] Output-Moore systems can be used to represent physical processes whose evolution is not affected by observation, contrasting with models reflecting epistemic processes (see Sec. 5). The input-Moore condition is typically used as a modelling choice to determine the temporal ordering between A_t and Y_t.

Some interfaces admit a very simple transducer. For example, an interface corresponding to a memoryless input-output processes with $p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:}) = \prod_{\tau=0}^{t} p(y_\tau|a_\tau)$ can be generated by a trivial world model $S_t = 0$. This is a degenerate case of a broader family of interfaces that afford simple world models, which we define next — including Markov decision processes (MDPs) as a main example.

Definition 4. *An interface $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$ is **fully observable** if $S_t = Y_t$ yields a valid transducer.*

Interestingly, non-trivial world models are required by interfaces with non-Markovian dynamics.

Lemma 2. *An interface is fully observable if and only if $p(y_{t+1}|\boldsymbol{y}_{:t}, \boldsymbol{a}_{:}) = p(y_{t+1}|y_t, a_t)$.*

Proof. This follows directly from using condition (2) from Lemma 1, an noticing that $S_t = Y_t$ yields a transducer if and only if $p(y_{t+1}, y_t|\boldsymbol{y}_{:t}, \boldsymbol{a}_{:}) = p(y_{t+1}, y_t|y_t, a_t) = p(y_{t+1}|y_t, a_t)$. □

4 Reducing world models

After setting the formal foundations of world models, and transducers as a way to construct computationally efficient ones, we now investigate minimal world models.

4.1 Minimal world models

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

Lemma 3. *The world model $S_t = \boldsymbol{H}_{:t-1}$ yields a valid transducer for any interface $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$.*

Unfortunately, the world model highlighted in Lemma 3 is far from parsimonious: resembling Borges' character *Funes the memorious*, it does not forget anything and hence its implementation would require an unbounded amount of memory. Thus, from here onwards we focus on the following question: *how can one reduce/simplify a given transducer presentation of an interface?*.

To address this question, we first establish what it means to 'reduce' a transducer. For this, we build on the idea of MDP homomorphism (Ravindran, 2003), which we extend to transducers as follows.

Definition 5. *A **homomorphim** between transducers (Y_t, A_t, S_t) and (Y'_t, A'_t, S'_t) is given by the mappings $\phi : \mathcal{S} \to \mathcal{S}'$, $f : \mathcal{Y} \to \mathcal{Y}'$, and $g : \mathcal{A} \to \mathcal{A}'$ satisfying two compatibility conditions:*

(i) $\Pr\big(Y'_t = f(y)|S'_t = \phi(s), A'_t = g(a)\big) = \Pr\big(Y_t = y|S_t = s, A_t = a\big)$.

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

*A **reduction** of a world model $S_t \xrightarrow{\phi} S'_t$ is a homomorphism between transducers with the same inputs and outputs (Y_t, A_t, S_t) and (Y_t, A_t, S'_t) in which f and g are identity mappings and ϕ is surjective. Two worlds are **isomorphic** if they are reductions of each other. Finally, a world model S_t is **minimal** if all its reductions are isomorphic to itself.*

An homomorphism is a structure-preserving map between transducers, and a world reduction is a coarse-graining 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 Markovian — as can be confirmed by relating it with the notion of 'lumpability' of Markov chains (Tian & Kannan, 2006). These properties let reductions of transducers to generate the same interfaces as the transducers they reduce, as shown next (see App. F for a proof).

Lemma 4. *A transducer and all its reductions generate the same interface.*

The next two sections study different approaches to look for minimal world models.[4]

[4]Minimality can also be studied via the entropy of the world dynamics, which better accounts for encoding cost. Interestingly, minimal entropy models may not coincide with the models with fewer states — although the two coincide for predictive models (Loomis & Crutchfield, 2019)

4.2 Reduction via bisimulation

A natural way to reduce a world model is via the notion of bisimulation, which is typically studied in the context of MDPs as a way of merging states that have an equivalent role in generation and dynamics (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 6. *For a given transducer with world model S_t and kernel κ_t, a* ***bisimulation*** *is an equivalence relationship $\mathcal{B}_t \subseteq \mathcal{S} \times \mathcal{S}$ such that $s \sim s'$ if they satisfy the following conditions*

(i) $p_t(y|s,a) = p_t(y|s',a)$, where $p_t(y|s,a) = \sum_{s'' \in \mathcal{S}} \kappa_t(y, s''|s,a)$ is the probability of generating y given s and a, and

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

World model reductions (Def. 5) and bisimulations are two faces of the same coin, as shown next by extending a standard result from Taylor et al. (2008) to our world models (see proof in App. G).

Proposition 1. *$S_t \xrightarrow{\phi} S'_t$ if and only if the equivalence relationship with classes given by $\phi^{-1}(s') = \{s \in \mathcal{S} : \phi(s) = s'\}$ is a bisimulation.*

This proposition has a simple yet powerful implication: it shows that the optimal way to reduce a given world model is to coarse-grain its states with a bisimulation.

Unfortunately, bisimulation is often not able to deliver the smallest world model capable of generating a given interface. To investigate this, let us consider a world model with $|\mathcal{S}| = n$ states and build 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 $[\boldsymbol{w}(\boldsymbol{h}_{:t})]_k = \Pr(\boldsymbol{Y}_{:t} = \boldsymbol{y}_{:t} | \boldsymbol{A}_{:t} = \boldsymbol{a}_{:t}, S_0 = s_k)$. Intuitively, if the vectors $\boldsymbol{w}(\boldsymbol{h}_{:t})$ are linearly dependent, that suggests that some of their dimensions — and, hence, their corresponding world states — are not being exploited. Crucially, the coarse-grainings related to bisimulation can only remove states that have identical components, but cannot reduce more general linear dependencies between states (see also Sec. 5.2 and App. M). Note that relaxing the criteria for merging states — e.g. via bisimulation metrics (Ferns & Precup, 2014) — does not solve this issue, as this would necessarily introduce changes in the resulting interface.

These ideas can be made concrete by studying the so-called *canonical dimension* of a transducer $\mathcal{T}$, which is defined as

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

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). The canonical dimension is an important index of a transducer, as shown by the next result, whose proof can be found in (Cakir et al., 2021, Th. 4.8), and related results can be found in (Ito et al., 1992; Balasubramanian, 1993).

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.*

Unfortunately, it is often the case that the minimal bisimulation of a given transducer $\hat{\mathcal{T}}$ with world states in $\hat{\mathcal{S}}$ still exhibits $d(\hat{\mathcal{T}}) < |\hat{\mathcal{S}}|$. In fact, there are interfaces for which no transducer reaches $d(\mathcal{T}) = |\mathcal{S}|$. Furthermore, 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. 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).

4.3 Pseudo-probabilities and generalised transducers

In this section we focus on the reduction of world models with a finite number of states $|\mathcal{S}| = n$. As discussed in Sec. 3.2, if a transducer has a world that can take $n < \infty$ number of states, then the probabilities of $\boldsymbol{y}_{:t}$ given $\boldsymbol{a}_{:t}$ can be calculated via

$$p(\boldsymbol{y}_{:t}|\boldsymbol{y}_{:t}) = \mathbf{1}^\intercal \cdot \left(\prod_{i=0}^{t} T_i^{(y_i|a_i)} \right) \cdot \boldsymbol{p}, \tag{6}$$

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. A more general class of transducers can be explored by reducing 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 a generalised notion of transducer, which we introduce next.

Definition 7. *A* ***generalised transducer*** *for an interface* $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$ *is a tuple* $(\mathcal{A}, \mathcal{Y}, \mathcal{S}, \{A^{(y|a)}\}, \boldsymbol{v}, \boldsymbol{u})$ *with* $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ *and* $A^{(y|a)} \in \mathbb{R}^{n\times n}$ *that satisfy*

$$\Pr(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:t}) = \boldsymbol{u}^\intercal \cdot \left(\prod_{i=0}^{t} A_i^{(y_i|a_i)} \right) \cdot \boldsymbol{v} \qquad \forall \boldsymbol{y}_{:t} \in \mathcal{Y}^{t+1}, \boldsymbol{a}_{:t} \in \mathcal{A}^{t+1}. \tag{7}$$

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.

Theorem 2. *A generalised transducer* $\tilde{T}$ *with* $d(\tilde{T}) < n$ *can always be reduced to another transducer that generate the same interface using fewer states.*

This result follows directly from the proofs provided in (Balasubramanian, 1993, Ch. 3), and related results can be found in (Upper, 1997; Vidyasagar, 2011). Notably, these proofs lead to practical algorithms that can be used to efficiently reduce transducers with $d(\tilde{T}) < n$ (see App. H). In this way, generalised transducers achieve a minimal computational complexity at the cost of introducing an opaque world model whose trajectories cannot be sampled (due to the quasi-probabilities), which results in a substantial lack of interpretability.

5 Forward interpretability via epistemic world models

The previous section shows how maximal computational efficiency can be achieved by either compressing memory state spaces with bisimulations, or by allowing memory states of transducers to follow 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. In this section we take a different route by investigating specific types of world models that focus on interpretability, bringing insights about what AI agents can learn.

5.1 World models of beliefs

Let us start by highlighting properties that can make world models more interpretable.

Definition 8. *A world model* S_t *is* ***predictive*** *if* $I(S_t; \boldsymbol{Y}_{t:}|\boldsymbol{H}_{:t-1}, \boldsymbol{A}_{t:}) = 0$*, so that the present world state contains no present or future information (given the actions). A world model is* ***observable*** *if there is a mapping* $f : \mathcal{Y} \times \mathcal{A} \to \mathcal{S}$ *such that* $S_{t+1} = f(\boldsymbol{Y}_{:t}, \boldsymbol{A}_{:t})$*. A world model is* ***unifilar*** *if there is a function* f *such that* $S_{t+1} = f(Y_t, A_t, S_t)$*.*[5]

These classes of models are linked in interesting ways: observable world models are always predictive, and unifilar models are observable if there is no randomness in the world's initial condition.

The literature contains various procedures that expand world models that trade computational complexity for observability. Many of these approaches model processes of inference and accumulation

[5] Or $S_{t+1} = f(Y_{t+1}, A_t, S_t)$, depending on time indexing conventions.

of knowledge (Virgo et al., 2021; Biehl & Virgo, 2022). Following Bayesian principles (Jaynes, 2003), these approaches shift the world configurations from elements in a set $\mathcal{S}$ to distributions over $\mathcal{S}$ — henceforth called *belief states* — that reflect different states of knowledge of agents. Moreover, by focusing on processes of optimal reasoning, one can assume that these belief states are updated via unifilar dynamics (Virgo, 2023). These ideas are captured in our next definition.

Definition 9. *A* ***predictive belief transducer*** *on the states a world model $\mathcal{S}$ is a tuple $\big(\mathcal{B}, \mathcal{Y}, \mathcal{A}, \hat{\mathcal{K}}, b_0\big)$, where $\mathcal{B} \subseteq \Delta(\mathcal{S})$ is a set of belief states, $\hat{\mathcal{K}} = \{\hat{\kappa}_t(y, d'|a, d) : a \in \mathcal{A}, y \in \mathcal{Y},\ d, d' \in \mathcal{B}, t \in \mathbb{N}\}$ are stochastic kernels of the form $\hat{\kappa}_t(y, d'|a, d) = p(y|a, d)\delta^{d'}_{f(y,a,d)}$ with $f : \mathcal{Y} \times \mathcal{A} \times \Delta(\mathcal{S}) \to \Delta(\mathcal{S})$, and $b_0 \in \Delta(\mathcal{S})$ is an initial belief.*

Predictive belief transducers are predictive, observable since their initialisation is always deterministic, and unifilar by construction. In general, predictive belief transducers may not generate the same interface as the world model over which they are built.

Various types of belief transducer can be built by choosing different update functions f from the literature. A well-known update rule in the reinforcement literature comes from the notion of *belief MDP* (Kaelbling et al., 1998), which we extend to input-Moore transducers.

Definition 10. *An* ***update transducer*** *is a belief transducer determined by memory states of the form $\mathcal{B} = \{b_t = p(s_t|\boldsymbol{y}_{:t}, \boldsymbol{a}_{:t-1}) : s_t \in \mathcal{S}, \boldsymbol{y}_{:t} \in \mathcal{Y}^{t+1}, \boldsymbol{a}_{:t-1} \in \mathcal{A}^t, t \in \mathbb{N}\}$ and an update rule given by*

$$b_t(s_t) = \frac{p(y_t|s_t)}{Z} \sum_{s_{t-1}} p(s_t|s_{t-1}, a_{t-1}) b_{t-1}(s_{t-1}), \tag{8}$$

with Z a normalising constant that does not depend on s_t.

Above, Eq. (8) is the natural Bayes updating procedure that arises from the functional form of b_t when S_t is a Moore transducer (for a derivation, see App. I). Update transducers are important as they enable policies that reach optimal control in partially observable settings (Sawaki & Ichikawa, 1978; Åström, 1965; Yang et al., 2023).

Another way to build belief states from a world model from 'mixed-states' (Riechers & Crutchfield, 2018; Jurgens & Crutchfield, 2021), which we now generalise to world models.

Definition 11. *A* ***mixed-state transducer*** *is a belief transducer determined by memory states of the form $\mathcal{B} = \{d_t = p(s_t|\boldsymbol{h}_{:t-1}) : s_t \in \mathcal{S}, \boldsymbol{h}_{:t-1} \in \mathcal{Y}^t \times \mathcal{A}^t, t \in \mathbb{N}\}$ and an update rule given by*

$$d_{t+1}(s_{t+1}) = \frac{1}{Z'(a_t, y_t)} \sum_{s_t} p(y_t, s_{t+1}|s_t, a_t) d_t(s_t), \tag{9}$$

with Z' a normalising constant that does not depend on s_t.

A useful fact about mixed-state transducers is that they generate the same interface as the original transducer when their initial condition matches the one of the latter (proof in App. J). Interestingly, update and mixed-state transducers can be seen as two facets of Bayesian updating, corresponding to alternating phases of Bayesian filtering (Chen, 2003) as shown next (proof in App. I).

Lemma 5. *If S_t is the memory state of an input-Moore transducer, then the dynamics between update and mixed states follow the 'predict-update' process from Bayesian filtering:*

$$b_{t-1} = p(s_{t-1}|\boldsymbol{h}_{:t-1}) \xrightarrow{\text{predict}} d_t = p(s_t|\boldsymbol{h}_{:t-1}) \xrightarrow{\text{update}} b_t = p(s_t|\boldsymbol{h}_{:t}). \tag{10}$$

5.2 Minimal predictive world models

Following Barnett & Crutchfield (2015), let us now present a method from computational mechanics to build an observable world model directly from an interface $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$ without the need to bootstrap from a world model. This will tell us, in some sense, what is reasonable to assume about a world model given only its interface.

For this, let us first consider the equivalence relationship of histories given by

$$\boldsymbol{h}_{:t} \sim_\epsilon \boldsymbol{h}'_{:t} \quad \text{iff} \quad p(\boldsymbol{y}_{t+1:}|\boldsymbol{h}_{:t}, \boldsymbol{a}_{t+1:}) = p(\boldsymbol{y}_{t+1:}|\boldsymbol{h}'_{:t}, \boldsymbol{a}_{t+1:}), \quad \forall \boldsymbol{y}_{t+1:}, \boldsymbol{a}_{t+1:}. \tag{11}$$

Let's denote by ϵ the coarse-graining mapping that assigns each history to its corresponding equivalence class $\epsilon(\boldsymbol{h}_{:t}) = [\boldsymbol{h}_{:t}]_{\sim_\epsilon}$, and define $M_t = \epsilon(\boldsymbol{H}_{:t})$. This construction is known to be an effective way to build belief states without relying on a world model, known as *predictive state representations* (Littman & Sutton, 2001; Singh et al., 2004) in reinforcement learning, which are based on older ideas for stochastic processes (without inputs/actions) from computational mechanics (Crutchfield & Young, 1989). This construction is also closely related to the notion of *instrumental states* presented by Kosoy (2019). We now show that these equivalence classes serve as memory states of a transducer that generates the original interface (proof in App. K).

Proposition 2. *The triplet (Y_t, A_t, M_t) yields a valid transducer that is isomorphic to the minimal bisimulation of the world model $S_t = \boldsymbol{H}_{:t-1}$.*

The link between computational mechanics methods and predictive state representations was first noticed by Zhang et al. (2019), which addressed it using a different computational structure instead of transducers. Following Barnett & Crutchfield (2015), we now formally define the transducer that results from the ϵ coarse-graining.

Definition 12. *The **ϵ-transducer** of the interface $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$ is the transducer with memory state given by $M_t = \epsilon(\boldsymbol{H}_{:t-1})$, where ϵ is defined as in Eq. (11).*

Every interface has a unique (up to isomorphism) ϵ-transducer. The next result shows that the ϵ-transducer generates its interface, which was first proven in (Barnett & Crutchfield, 2015, Prop. 2). We provide an alternative proof that leverages links with the reinforcement learning literature.

Lemma 6. *The ϵ-transducer of an interface $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$ always generates the same interface.*

Proof. For a given interface $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$, Prop. 2 shows that the ϵ-transducer is a bisimulation of $S_t = \boldsymbol{H}_{:t-1}$. Given that S_t generates the interface (as shown in Lemma 3), Lemma 4 and Prop. 1 guarantee that the ϵ-transducer also does so. □

A salient feature of the ϵ-transducer (or, equivalently, predictive state representation) is that it provides belief dynamics over much fewer states than regular belief MDPs or equivalent methods (Littman & Sutton, 2001). Our next result further sediments this by showing that it yields the most efficient predictive world model possible.

Theorem 3. *If R_t is a predictive world model of a transducer (such as, e.g., belief MDPs or mixed-states), then its minimal bisimulation is isomorphic to the ϵ-transducer.*

Proof. For a given transducer with memory R_t, one can build an equivalence relationship via

$$\epsilon(r) = \epsilon(r') \quad \text{iff} \quad \Pr(\boldsymbol{Y}_{t:}|\boldsymbol{A}_{t:}, R_t = r) = \Pr(\boldsymbol{Y}_{t:}|\boldsymbol{A}_{t:}, R_t = r'). \tag{12}$$

Then, one can show that if R_t is a predictive world model, then $\epsilon(R_t)$ are isomorphic to the memory states of the ϵ-transducer. A proof of this can be found in App. L. □

This result leads to an important corollary relalated to the bisimulation of beliefs (Castro et al., 2009): while the bisimulation of general transducers may not fully reduce world models (as discussed in Sec. 4.2), bisimulations of beliefs necessarily lead to the ϵ-transducer.

Corollary 1. *The ϵ-transducer is the minimal predictive model that generates a given interface.*

A discussion between the minimality of predictive vs general transducers is provided in App. M.

6 Backwards interpretability via retrodictive world models

The previous section highlights the ϵ-transducer as the universal solution for scenarios where one needs a minimal predictive model. This model particularly useful to evaluate the capabilities of agents to distil information that is relevant to predict future events. Despite this being the prevalent approach to agent interpretability, it is crucial to note that prediction does not exhaust the possible knowledge-based activities in which an agent can be involved. In this section we explore retrodictive world models, which opens a new dimension of agent interpretability.

6.1 Retrodictive transducers

World models can in general contain information that the agent can only have access to in the future, without this violating the arrow of time. For example, a world model could be such that its state at $t = 0$ could already contain outcomes for any possible sequence of future actions (see App. N). In fact, in some scenarios the present state of a world models can be more strongly correlated with future observations rather than past ones (Ellison et al., 2009), and while this architecture may appear counterintuitive, it has been shown to be maximally efficient for processes that generate structure (Boyd et al., 2018).

Building on these ideas, we now consider 'retrodictive' world models that only contain future information, being duals to predictive models as introduced in Def. 13. Following Riechers et al. (2016), we also introduce a dual notion to unifiliarity.

Definition 13. *A world model S_t is **retrodictive** if $I(S_t; \boldsymbol{Y}_{:t-1}|\boldsymbol{H}_{t:}, \boldsymbol{A}_{:t-1}) = 0$. A world model is **counifilar** if there is a function f such that $S_t = f(S_{t+1}, A_t, Y_t)$.*

This notion makes one wonder if transducers could be made to 'run backwards', and what conditions would be necessary for this to happen. Functionally, Eq. (4) suggests that this could be done if rather than employing a forward-time kernel $\kappa(y_\tau, s_{\tau+1}|s_\tau, a_\tau) = p(y_\tau, s_{\tau+1}|s_\tau, a_\tau)$ that updates the memory state from s_τ to $s_{\tau+1}$, one could build a reverse-time kernel $\kappa^R(y_\tau, s_\tau|s_{\tau+1}, a_\tau) = p(y_\tau, s_\tau|s_{\tau+1}, a_\tau)$ that updates the memory from $s_{\tau+1}$ to s_τ.

Definition 14. *A **reversible transducer** is a transducer $\big(\mathcal{S}, \mathcal{Y}, \mathcal{A}, \mathcal{K}, p\big)$ together with an additional stochastic kernels $\mathcal{K}^R = \{\kappa_t^R(y, s'|a, s) : a \in \mathcal{A}, y \in \mathcal{Y}, s, s' \in \mathcal{S}, t \in \mathbb{N}\}$ such that*

$$p(\boldsymbol{y}_{:t}, \boldsymbol{s}_{:t+1}|\boldsymbol{a}_{:}) = p(s_0) \prod_{\tau=0}^{t} \kappa_t(y_\tau, s_{\tau+1}|s_\tau, a_t) = p(s_{t+1}|\boldsymbol{a}_{:t}) \prod_{\tau=0}^{t} \kappa_t^R(y_\tau, s_\tau|s_{\tau+1}, a_t). \tag{13}$$

While previous work has shown that all input-agnostic transducers (i.e. HMMs) can be time-reversed (Ellison et al., 2011), some transducers cannot. The key issue is that when swapping past and future one may break the condition of anticipation-free, which — according to Lemma 1 — is necessary for a world model to yield a transducer (an illustration of this is provided in App. O). Our next result provides necessary and sufficient conditions for a transducer to be reversed (see proof in App. P).

Theorem 4. *A transducer is reversible if an only if the dynamics of its memory state satisfy $p(s_t|s_{t+1}, \boldsymbol{a}_{:t}) = p(s_t|s_{t+1}, a_t)$.*

Theorem 4 shows that if $p(s_t|s_{t+1}\boldsymbol{a}_{:t}) \neq p(s_t|s_{t+1}, a_t)$ then S_t does not yield a transducer that can be run backwards. This result also reveals that reversible transducers can be achieved in a variety of ways (see Figure 2). For example, if the transducer is memoryless then the condition is satisfied trivially since $p(s_t|s_{t+1}, \boldsymbol{a}_{:t}) = p(s_t|s_{t+1}, a_t) = p(s_t)$. Also, if the transducer is *action-agnostic* (i.e. it is an HMM), then it is reversible as argued in Sec. P.2. Finally, being action counifiliar (i.e. if there exists f such that $S_t = f(S_{t+1}, A_t)$) is also sufficient for reversibility, as shown next.

Lemma 7. *If a transducer is action counifilar, then it is reversible.*

Proof. An action unifilar transducer satisfies $p(s_t|s_{t+1}, \boldsymbol{a}_{:t}) = \delta_{s_t, f(s_{t+1}, a_t)} = p(s_t|s_{t+1}, a_t)$. $\square$

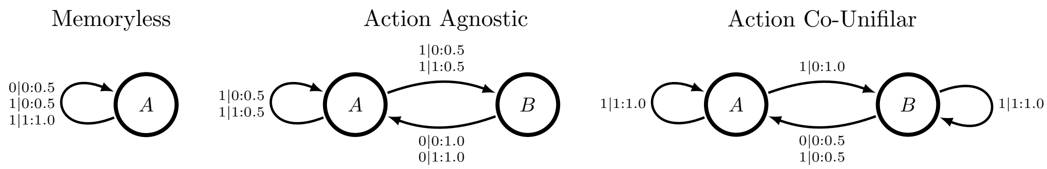

Figure 2: Three examples of reversible transducers. Circles represent world states, and arrows represent transitions and their labels describe the associated actions and outputs. For instance, the label 1|0:0.5 on the edge from s_0 to s_1 indicates that $\Pr(S_{t+1} = s_1, Y_t = 1|A_t = 0, S_t = s_0) = 0.5$.

6.2 Retrodictive beliefs and reverse interpretability

The previous section showed how, unlike for HMMs, there are strong restrictions on the reversibility of transducers. However, even if an interface cannot be generated via a reversed transducer, there still are retrodictive constructions that can be used to investigate those dynamics.

Definition 15. *A* ***retrodictive belief transducer*** *of a world model $S_t \in \mathcal{S}$ is a belief transducer $\big(\mathcal{B}, \mathcal{Y}, \mathcal{A}, \tilde{\mathcal{K}}, r_{t^*}\big)$ where the initial condition r_{t^*} may depend on $\boldsymbol{a}_{:t^*}$, and the stochastic kernels update the states following the mapping $S_t = g(Y_t, A_y, S_{t+1})$.*

Using this as a foundation, let us construct retrodictive mixed-states — which provide an analogue to the backward pass of Bayesian smoothing (Särkkä & Svensson, 2023), in the same way that update beliefs and mixed-states correspond to different steps of Bayesian filtering (Lemma 5).

Definition 16. *The* ***retrodictive mixed-states*** *of a world model S_t are given by the collection of distribution over $\mathcal{S}$ given by $r_{0,t}(s_0) = p(s_0|\boldsymbol{y}_{0:t-1}, \boldsymbol{a}_{0:t-1})$ for all $\boldsymbol{y}_{0:t-1} \in \mathcal{Y}^t, \boldsymbol{a}_{0:t-1} \in \mathcal{A}^t$.*

In contrast with predictive mixed-state beliefs (Def. 11), which always yield a presentation of the interface (as shown in App. J), retrodictive mixed-states may not do this. Nevertheless, one can still evaluate their dynamics and use them for useful analyses via linear operators, as shown next.

Definition 17. *The* ***bi-directional mixed-state matrix*** *(BDMSM) of an action-outcome sequence $\rho(\boldsymbol{y}_{0:t}, \boldsymbol{a}_{0:t})$ is a $|\mathcal{S}| \times |\mathcal{S}|$ matrix given by*

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

The BDMSM is directly linked with predictive and retrodictive mixed-states (proof in App. Q).

Theorem 5. *The predictive mixed-states d_t, retrodictive mixed-states $r_{0,t}$, and the BDMSM can be calculated as*

$$\rho(\boldsymbol{y}_{0:\tau}, \boldsymbol{a}_{0:\tau}) = \frac{T^{(\boldsymbol{y}_{0:\tau}|\boldsymbol{a}_{0:\tau})}\rho_0}{\boldsymbol{1}^\intercal \cdot T^{(\boldsymbol{y}_{0:\tau}|\boldsymbol{a}_{0:\tau})}\rho_0 \cdot \boldsymbol{1}}, \quad d_t = \rho(\boldsymbol{y}_{0:\tau}, \boldsymbol{a}_{0:\tau}) \cdot \boldsymbol{1}, \quad \text{and} \quad e_{0,t} = \rho(\boldsymbol{y}_{0:\tau}, \boldsymbol{a}_{0:\tau})^\intercal \cdot \boldsymbol{1},$$

where $\boldsymbol{1}$ is a $|\mathcal{S}|$-dimensional vector of 1's, $T^{(y_{0:t}|a_{0:t})} \equiv \prod_{\tau=0}^{t} T^{(y_\tau|a_\tau)}$, and $\rho_t = \sum_{s_t} p(s_t)\boldsymbol{e}_{s_t}\boldsymbol{e}_{s_t}^\intercal$ is a diagonal matrix.

Corollary 2. *The forward-time update of the BDMSM is given by*

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

while the reverse-time update is

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

Given that not every transducer is reversible, the operation $\rho_t^{-1}T^{(y|a)}\rho_{t-1}$ do not generally yield the action of a transducer. It 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.

7 Discussion

This paper explores the benefits of designing world models for sandboxing and testing AI systems by focusing on the agent's interface, which characterises the viewpoint of the agent in consideration. This leads to a policy-agnostic approach that require no assumptions about the agent's architecture and capabilities, being applicable to systems irrespective of how they were designed or trained. This allowed us to identify fundamental limits and trade-offs inherent to world modelling, whose demarcation leads to a number of practical recommendations to guide designers when constructing world models (Figure 3).

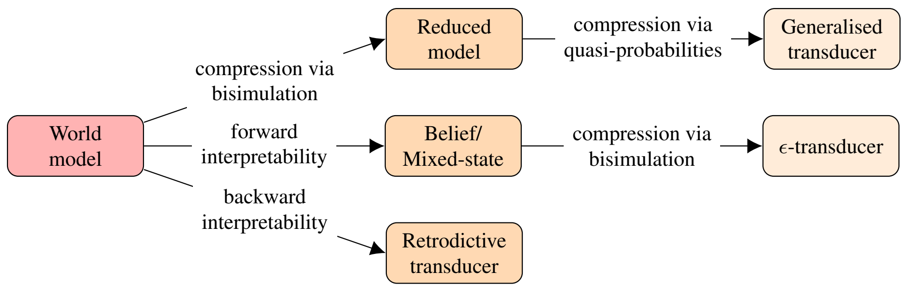

Figure 3: Recommendations for building world optimal models according to different desiderata.

Our analyses revealed a fundamental trade-off between the computational efficiency and interpretability of world models. Generalised transducers were found to yield the most efficient implementations at the cost of having to employ quasi-probabilities, resulting on opaque world models that cannot be sampled — remaining unknowable, akin to the Kantian noumena. In contrast, the ϵ-transducer, a generalisation of the geometric belief structure recently found in the residual stream of transformers (Shai et al., 2025), was found to yield the unique minimal predictive world model. 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 to learn about their environments.

We also introduced the notion of retrodictive world models for facilitating retrospective analyses to study the origins of undesirable events or behaviours. These models allow to, for instance, identify 'danger zones' that are likely to lead to undesirable future states. 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 yield efficient modelling is to explode the compositional structure of the world (Lake & Baroni, 2023; Elmoznino et al., 2024; Baek et al., 2025).

Overall, the approach taken in this work complements the substantial body of work that uses world models to boost the performance of agents (Ha & Schmidhuber, 2018; Hafner et al., 2020; 2023; Hansen et al., 2024), and work on representations from the point of view of the agent (Ni et al., 2024). Additionally, the ideas put forward here establish new bridges between related subjects in reinforcement learning, control theory, and theoretical physics, and may serve as a rosetta stone for navigating across these literatures. Finally, the new insights related to world models revealed in this work also have significant 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), which will be developed in a separate publication.

References

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.

Karl Johan Åström. Optimal control of markov processes with incomplete state information. *Journal of mathematical analysis and applications*, 10:174–205, 1965.

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, 01 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:e218, 2019. DOI: 10.1017/S0140525X19001353.

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.

Alexander Boyd, Dibyendu Mandal, and James Crutchfield. Thermodynamics of modularity: Structural costs beyond the landauer bound. *Physical Review X*, 8(3):031036, 2018.

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, Prakash Panangaden, and Doina Precup. Equivalence relations in fully and partially observable markov decision processes. In *IJCAI*, volume 9, pp. 1653–1658, 2009.

Zhe Chen. Bayesian filtering: From kalman filters to particle filters, and beyond. *Statistics*, 182(1): 1–69, 2003.

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 M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, 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 J Ellison, John R Mahoney, Ryan G James, James P Crutchfield, and Jörg Reichardt. Information symmetries in irreversible processes. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 21(3), 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.

EU Council. The Artificial Intelligence Act, Chapter 6, 2024. URL https://artificialintelligenceact.eu/chapter/6.

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.

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 *NASA Formal Methods: 4th International Symposium, NFM 2012, Norfolk, VA, USA, April 3-5, 2012. Proceedings 4*, pp. 326–340. Springer, 2012.

Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

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

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

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 M 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 L 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 S Sutton. Predictive representations of state. *Advances in neural information processing systems*, 14, 2001.

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

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

Sarah E Marzen and James P 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.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In Yonatan Belinkov, Sophie Hao, Jaap Jumelet, Najoung Kim, Arya McCarthy, and Hosein Mohebbi (eds.), *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 16–30, Singapore, December 2023. Association for Computational Linguistics. DOI: 10.18653/v1/2023.blackboxnlp-1.2. URL https://aclanthology.org/2023.blackboxnlp-1.2/.

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.

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

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

Balaraman Ravindran. 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 M Riechers, John R Mahoney, Cina Aghamohammadi, and James P Crutchfield. Minimized state complexity of quantum-encoded cryptic processes. *Physical Review A*, 93(5):052317, 2016.

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.

Katsushige Sawaki and Akira Ichikawa. Optimal control for partially observable markov decision processes over an infinite horizon. *Journal of the Operations Research Society of Japan*, 21(1): 1–16, 1978.

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.

Yujie Yang, Yuxuan Jiang, Jianyu Chen, Shengbo Eben Li, Ziqing Gu, Yuming Yin, Qian Zhang, and Kai Yu. Belief state actor-critic algorithm from separation principle for pomdp. In *2023 American Control Conference (ACC)*, pp. 2560–2567, 2023. DOI: 10.23919/ACC55779.2023.10155792.

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, 11 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 Sufficient statistics

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

Consider a random vector $\boldsymbol{X} = (X_1, \dots, X_n) \in \mathcal{X}^n$ 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 when estimating the value of θ via, e.g., maximum likelihood, the information given by $\boldsymbol{X}$ after Y has been fixed is irrelevant.

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 $\boldsymbol{X} \perp\!\!\!\perp \theta|Y$ for any prior distribution over θ. It can been shown that strong Bayesian sufficiency 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, an statistic $T = f(X)$ is said to be a *Bayesian sufficient statistic for X w.r.t. Y* if $X \perp\!\!\!\perp Y|T$, i.e. if $\Pr(X = x|Y = y, T = t) = \Pr(X = x|T = t)$. This is equivalent to the information-theoretic condition $I(X;Y|T) = 0$, which state 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 through this work.

Another way to think of sufficient statistics is by noticing that, if $X - T - Y$ is a Markov chain, which implies that all the information shared between X and Y necessarily "goes through" T. Interestingly, for all mappings f, if $T = f(X)$ then the following Markov chain hold: $T - X - Y$. Moreover, the data processing inequality says that for any such Markov chains then $I(Y;X) \geq I(Y;T)$; therefore "processing" X cannot increase its information about Y. Moreover, following Cover & Thomas (2012), the equality $I(Y;X) = I(Y;T)$ is attained if an only if $X - T - Y$ is also a Markov chain; i.e. if T is a sufficient statistic. In summary, sufficient statistics are related to optimal (i.e. lossless) data processing (Kullback, 1997).

Sufficient statistics always exists — in particular, X is always sufficient for itself. The search for optimal but also efficient statistics lead 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 $\boldsymbol{X}$ w.r.t. Y can be build explicitly, built as the partition induced by the follwoing equivalence relationship (Asoodeh et al., 2014, Def. 2):

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

It is worth noticing the similarities between this way to build minimal sufficient statistics, Def. 6, and Eq. (11).

B Relationship between transducers and POMDPs

A POMDP is a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \tau, \mu, \rho)$, where $\mathcal{S}$ are the states of the world, $\mathcal{A}$ the action space, $\mathcal{O}$ the observation space, and the probability kernels $\tau : \mathcal{S} \times \mathcal{A} \to P(\mathcal{S})$, $\mu : \mathcal{S} \to P(\mathcal{O})$, and

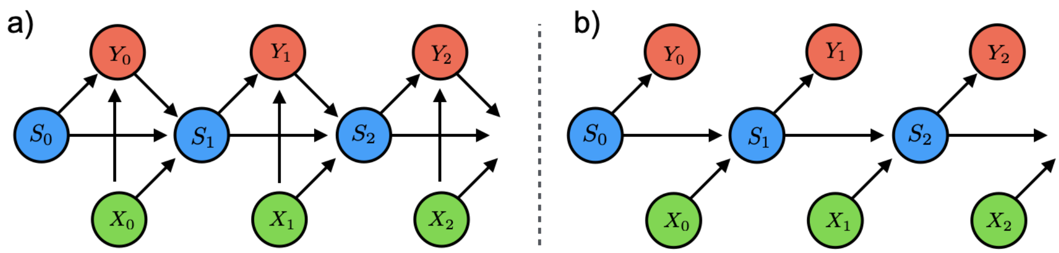

Figure 4: Mealy transducers (left) exhibit all connections, I-O Moore ones (right) restrict some.

$\rho : \mathcal{S} \times \mathcal{A} \to P(\mathbb{R})$ specify the world dynamics, observation map, and reward function (Kaelbling et al., 1998).

From this definition one can see, under a POMDP, the joint dynamics satisfy Eq. (4), which — thanks to the second alternative definition in Lemma 1 — is sufficient to show that the POMDP induces a transducer. This, together with the first alternative definition in Lemma 1, imply that the process S_t in a POMDP is a world model, in the sense that it satisfies the conditions in Def. 2. Finally, one can observe that the kernel of the correspodning transducer allows the following factorisation:

$$p(s_{t+1}, y_t | s_t, a_t) = \tau(s_{t+1}|s_t, a_t)\mu(o_t|s_t)\rho(r_t|s_t, a_t). \tag{18}$$

This shows that POMDP is a I-O Moore transducer, as defined in Sec. 3.2 and illustrated in Figure 4.

C Derivation of Eq. (2)

The properties of anticipation-free world models allows to factorise interfaces as follows:

$$p(\boldsymbol{y}_{:\tau}, \boldsymbol{s}_{:\tau+1}|\boldsymbol{a}_{:}) = p(s_0|\boldsymbol{a}_{:}) \prod_{t=0}^{\tau-1} p(s_{t+1}, y_t|\boldsymbol{h}_{:t-1}, \boldsymbol{s}_{:t}, \boldsymbol{a}_{t:}) \tag{19}$$

$$= p(s_0|\boldsymbol{a}_{:}) \prod_{t=0}^{\tau-1} p(y_t|\boldsymbol{h}_{:t-1}, \boldsymbol{s}_{:t}, \boldsymbol{a}_{t:}) p(s_{t+1}|, \boldsymbol{h}_{:t}, \boldsymbol{s}_{:t}, \boldsymbol{a}_{t+1:}). \tag{20}$$

Now, using the properties of world models, one can find that

$$p(y_t|\boldsymbol{h}_{:t-1}, \boldsymbol{s}_{:t}, \boldsymbol{a}_{t:}) = p(y_t|s_t, \boldsymbol{a}_{t:}) = p(y_t|s_t, a_t), \tag{21}$$

where the first equality uses the first property in Def. 2, and the second equality the second property. Similarly, assuming that the world model is anticipation-free, then the expression for the dynamics of the world model can be simplified as follows:

$$p(s_{t+1}|\boldsymbol{h}_{:t}, \boldsymbol{s}_{:t}, \boldsymbol{a}_{t+1:}) = \frac{p(\boldsymbol{s}_{:t+1}|\boldsymbol{h}_{:t}, \boldsymbol{a}_{t+1:})}{\sum_{s_{t+1}} p(\boldsymbol{s}_{:t+1}|\boldsymbol{h}_{:t}, \boldsymbol{a}_{t+1:})} = \frac{p(\boldsymbol{s}_{:t+1}|\boldsymbol{h}_{:t})}{\sum_{s_{t+1}} p(\boldsymbol{s}_{:t+1}|\boldsymbol{h}_{:t})} = p(s_{t+1}|\boldsymbol{s}_{:t}, \boldsymbol{h}_{:t}). \tag{22}$$

Also, the anticipation-free property also guarantees that $p(s_0|\boldsymbol{a}_{:}) = p(s_0)$.

Putting this together, we find that

$$p(\boldsymbol{y}_{:\tau}, \boldsymbol{s}_{:\tau+1}|\boldsymbol{a}_{:}) = p(s_0) \prod_{t=0}^{\tau-1} p(y_t|s_t, a_t) p(s_{t+1}|\boldsymbol{h}_{:t}, \boldsymbol{s}_{:t}). \tag{23}$$

D Proof of Lemma 1

For clarity, let us divide the proof into sub-parts.

Part 1: Being a transducer is equivalent to Condition (2)

Proof. Let us first show that being a transducer is equivalent to condition (2) of Lemma 1. One direction of the implication is trivial, as a transducer satisfies condition (2) by construction. To prove the converse, let's assume that condition (2) holds. Then, one can define the kernel $\kappa_t(y, s|a, s') = \Pr(Y_t = y, S_{t+1} = s|A_t = a, S_t = s)$. In virtue of condition (2), it is direct to see that Eq. (4) holds, which implies that this kernel gives rise to the dynamics. □

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

Proof. Let's first prove that condition (1) implies condition (2). Using the derivations presented in Eq. (21) and Eq. (22), one finds that if S_t is a world model then $p(s_{t+1}, y_t|\boldsymbol{s}_{:t}, \boldsymbol{h}_{:t-1}, \boldsymbol{a}_{t:}) = p(y_t|s_t, a_t)p(s_{t+1}|\boldsymbol{s}_{:t}, \boldsymbol{h}_{:t})$. Then, using the definition of being a transducer $p(s_{t+1}|\boldsymbol{s}_{:t}, \boldsymbol{h}_{:t}) = p(s_{t+1}|s_t, h_t)$, which in turn implies that $p(s_{t+1}, y_t|\boldsymbol{s}_{:t}, \boldsymbol{h}_{:t-1}, \boldsymbol{a}_{t:}) = p(s_{t+1}, y_t|s_t, a_t)$.

Let us now prove that condition (2) implies condition (1). The first property of world models can be proven as follows:

$$\begin{aligned}
p(\boldsymbol{y}_{t:t'}|\boldsymbol{h}_{:t-1}, \boldsymbol{s}_{:t}, \boldsymbol{a}_{t:}) &= \sum_{\boldsymbol{s}_{t+1:t'+1}} p(\boldsymbol{s}_{t+1:t'+1}, \boldsymbol{y}_{t:t'}|\boldsymbol{h}_{:t-1}, \boldsymbol{s}_{:t}, \boldsymbol{a}_{t:}) && (24)\\
&= \sum_{\boldsymbol{s}_{t+1:t'+1}} \prod_{\tau=t}^{t'} p(s_{\tau+1}, y_\tau|\boldsymbol{h}_{:\tau-1}, \boldsymbol{s}_{:\tau}, \boldsymbol{a}_{\tau:}) && (25)\\
&= \sum_{\boldsymbol{s}_{t+1:t'+1}} \prod_{\tau=t}^{t'} p(s_{\tau+1}, y_\tau|\boldsymbol{y}_{t:\tau-1}, \boldsymbol{s}_{t:\tau}, \boldsymbol{a}_{t:}) && (26)\\
&= \sum_{\boldsymbol{s}_{t+1:t'+1}} p(\boldsymbol{s}_{t+1:t'+1}, \boldsymbol{y}_{t:t'}|s_t, \boldsymbol{a}_{t:}) && (27)\\
&= p(\boldsymbol{y}_{t:t'}|s_t, \boldsymbol{a}_{t:}). && (28)
\end{aligned}$$

Above, note that the third equality uses condition (2) to drop some of the conditioning elements. The second property of world models follows from this calculation:

$$p(y_t|\boldsymbol{a}_{t:}, s_t) = \sum_{s_{t+1}} p(s_{t+1}, y_t|\boldsymbol{a}_{t:}, s_t) = \sum_{s_{t+1}} p(s_{t+1}, y_t|a_t, s_t) = p(y_t|a_t, s_t). \quad (29)$$

The condition of anticipation-free world model is satisfied as follows:

$$\begin{aligned}
p(\boldsymbol{s}_{:t}|\boldsymbol{h}_{:t-1}, \boldsymbol{a}_{t':}) &= \frac{p(\boldsymbol{s}_{:t}, \boldsymbol{y}_{:t-1}|\boldsymbol{a}_{:t-1}, \boldsymbol{a}_{t':})}{p(\boldsymbol{y}_{:t-1}|\boldsymbol{a}_{:t-1}, \boldsymbol{a}_{t':})} = \frac{\prod_{\tau=0}^{t} p(s_\tau, y_{\tau-1}|\boldsymbol{s}_{:\tau-1}, \boldsymbol{y}_{:\tau-2}, \boldsymbol{a}_{:t-1}, \boldsymbol{a}_{t':})}{p(\boldsymbol{y}_{:t-1}|\boldsymbol{a}_{:t-1})} && (30)\\
&= \frac{\prod_{\tau=0}^{t} p(s_\tau, y_{\tau-1}|\boldsymbol{s}_{:\tau-1}, \boldsymbol{y}_{:\tau-2}, \boldsymbol{a}_{:t-1})}{p(\boldsymbol{y}_{:t-1}|\boldsymbol{a}_{:t-1})} = \frac{p(\boldsymbol{s}_{:t}, \boldsymbol{y}_{:t-1}|\boldsymbol{a}_{:t-1})}{p(\boldsymbol{y}_{:t-1}|\boldsymbol{a}_{:t-1})} = p(\boldsymbol{s}_{:t}|\boldsymbol{h}_{:t-1}). && (31)
\end{aligned}$$

Finally, the Markovianity of state dynamics can be proven as follows:

$$p(s_{t+1}|\boldsymbol{h}_{:t}, \boldsymbol{s}_{:t}) = \frac{p(s_{t+1}, y_t|a_t, \boldsymbol{h}_{:t-1}, \boldsymbol{s}_{:t})}{\sum_{s_{t+1}} p(s_{t+1}, y_t|a_t, \boldsymbol{h}_{:t-1}, \boldsymbol{s}_{:t})} = \frac{p(s_{t+1}, y_t|a_t, s_t)}{\sum_{s_{t+1}} p(s_{t+1}, y_t|a_t, s_t)} = p(s_{t+1}|s_t, h_t). \quad (32)$$

□

Part 3: Being a transducer is equivalent to condition (3)

Proof. We have two conditions that we claim are equivalent:

1) The following equalities hold

$$I[\boldsymbol{S}_{:t-1}, S_t, \boldsymbol{Y}_{:t-1}; \boldsymbol{A}_{t:}|\boldsymbol{A}_{:t-1}, S_{t_i}] = 0 \tag{33}$$
$$I[\boldsymbol{S}_{t+1:}, \boldsymbol{Y}_{t:}; \boldsymbol{Y}_{:t-1}, \boldsymbol{S}_{:t-1}, \boldsymbol{A}_{:t-1}|\boldsymbol{A}_{t:}, S_t] = 0. \tag{34}$$

2) The joint distribution can be implemented by a transducer.

- 1) $\Rightarrow$ 2): If the equality $I[A; B|C]$ holds, then we have the equality of probabilities $p(A|C) = p(A|BC)$. Thus, the two information equalities imply the probability equalities

$$\Pr(\boldsymbol{S}_{t+1:}, \boldsymbol{Y}_{t:}|\boldsymbol{Y}_{:t-1}, \boldsymbol{S}_{:t-1}, \boldsymbol{A}_{:t-1}, \boldsymbol{A}_{t:}, S_t) = \Pr(\boldsymbol{S}_{t+1:}, \boldsymbol{Y}_{t:}|\boldsymbol{A}_{t:}, S_t) \tag{35}$$
$$\Pr(\boldsymbol{S}_{:t-1}, \boldsymbol{Y}_{:t-1}, S_t|\boldsymbol{A}_{:t-1}, \boldsymbol{A}_{t:}, S_{t_i}) = \Pr(\boldsymbol{S}_{:t-1}, \boldsymbol{Y}_{:t-1}, S_t|\boldsymbol{A}_{:t-1}, S_{t_i}). \tag{36}$$

  Note that S_{t_i} is an element of the past $\boldsymbol{S}_{:t-1}$, so we can multiply these together to obtain

$$\Pr(\boldsymbol{S}_{:t-1}, \boldsymbol{Y}_{:t-1}, S_t|\boldsymbol{A}_{:t-1}, S_{t_i}) \Pr(\boldsymbol{S}_{t+1:}, \boldsymbol{Y}_{t:}|\boldsymbol{A}_{t:}, S_t) \tag{37}$$
$$= \Pr(\boldsymbol{S}_{:t-1}, \boldsymbol{Y}_{:t-1}, S_t|\boldsymbol{A}_{:t-1}, \boldsymbol{A}_{t:}, S_{t_i}) \Pr(\boldsymbol{S}_{t+1:}, \boldsymbol{Y}_{t:}|\boldsymbol{Y}_{:t-1}, \boldsymbol{S}_{:t-1}, \boldsymbol{A}_{:t-1}, \boldsymbol{A}_{t:}, S_t, S_{t_i}) \tag{38}$$
$$= \Pr(S_t, \boldsymbol{S}_{t+1:}, \boldsymbol{Y}_{t:}, \boldsymbol{Y}_{:t-1}, \boldsymbol{S}_{:t-1}|\boldsymbol{A}_{:t-1}, \boldsymbol{A}_{t:}, S_{t_i}) \tag{39}$$
$$= \Pr(\boldsymbol{S}_{:t-1}\boldsymbol{S}_{t:}, \boldsymbol{Y}_{:t-1}, \boldsymbol{Y}_{t:}|\boldsymbol{A}_{:t-1}, \boldsymbol{A}_{t:}, S_{t_i}) \tag{40}$$
$$= \Pr(\boldsymbol{S}_{t_i:}, \boldsymbol{Y}_{t_i:}, |\boldsymbol{A}_{t_i:}, S_{t_i}), \tag{41}$$

  which uses the fact that the whole trajectory of $\boldsymbol{X}_{:t-1}\boldsymbol{X}_{t:}$ is the same as the forward trajectory from the initial time $\boldsymbol{X}_{t_i:}$ We can apply this recursively by first considering $t = t_i + 1$:

$$\Pr(\boldsymbol{S}_{t_i:}, \boldsymbol{Y}_{t_i:}, |\boldsymbol{A}_{t_i:}, S_{t_i}) = \Pr(\boldsymbol{S}_{:t_i}, \boldsymbol{Y}_{:t_i}, S_{t_i+1}|\boldsymbol{A}_{:t_i}, S_{t_i}) \Pr(\boldsymbol{S}_{t_i+2:}, \boldsymbol{Y}_{t_i+1:}|\boldsymbol{A}_{t_i+1:}, S_{t_i+1}) \tag{42}$$
$$= \Pr(S_{t_i}, Y_{t_i}, S_{t_i+1}|A_{t_i}, S_{t_i}) \Pr(\boldsymbol{S}_{t_i+2:}, \boldsymbol{Y}_{t_i+1:}|\boldsymbol{A}_{t_i+1:}, S_{t_i+1}) \tag{43}$$

  Note that $\Pr(A, B|A, C) = \Pr(B|A, C)$, so we can simplify to the recursive relation:

$$\Pr(\boldsymbol{S}_{t_i:}, \boldsymbol{Y}_{t_i:}, |\boldsymbol{A}_{t_i:}, S_{t_i}) = \Pr(S_{t_i}, Y_{t_i}|A_{t_i}, S_{t_i}) \Pr(\boldsymbol{S}_{t_i+1:}, \boldsymbol{Y}_{t_i+1:}|\boldsymbol{A}_{t_i+1:}, S_{t_i+1}), \tag{44}$$

  where we have isolated the kernel κ_t as $\Pr(S_{t+1}, Y_t|A_t, S_t)$.

Through recursion, we see that the joint probability can be constructed from the kernel

$$\Pr(\boldsymbol{S}_{t_i:}, \boldsymbol{Y}_{t_i:}, |\boldsymbol{A}_{t_i:}, S_{t_i}) = \prod_{t=t_i}^{t_f-1} \Pr(S_{t+1}, Y_t|A_t, S_t), \tag{45}$$

meaning that this channel and world model can indeed be expressed as a transducer.

2) $\Rightarrow$ 1): If the world model can be expressed as a transducer, then the joint probability of hidden state, action, output trajectories can be broken into the product of terms

$$\Pr(\boldsymbol{S}_{t_i:t_f}, \boldsymbol{Y}_{t_i:t_f-1}, |\boldsymbol{A}_{t_i:t_f-1}, S_{t_i}) = \prod_{t=t_i}^{t_f} \Pr(S_{t+1}, Y_t|A_t, S_t). \tag{46}$$

This can be split into the product of two terms

$$\Pr(\boldsymbol{S}_{t_i:t_f}, \boldsymbol{Y}_{t_i:t_f-1}|\boldsymbol{A}_{t_i:t_f-1}, S_t) = \left(\prod_{j=t}^{t_f-1} \Pr(S_{j+1}, Y_j|A_j, S_j)\right) \left(\prod_{j=t_i}^{t-1} \Pr(S_{j+1}, Y_j|A_j, S_j)\right) \tag{47}$$
$$= \Pr(\boldsymbol{S}_{t:t_f}, \boldsymbol{Y}_{t:t_f-1}|\boldsymbol{A}_{t:t_f-1}, S_t) \Pr(\boldsymbol{S}_{t_i+1:t}, \boldsymbol{Y}_{t_i:t-1}|\boldsymbol{A}_{t_i:t-1}, S_{t_i}). \tag{48}$$

Applying the definitions of past and futures of t, we have

$$\begin{aligned}\Pr(\boldsymbol{S}_{:t-1}S_t\boldsymbol{S}_{t+1:}, \boldsymbol{Y}_{:t-1}\boldsymbol{Y}_{t:}|\boldsymbol{A}_{:t-1}\boldsymbol{A}_{t:}, S_{t_i}) &= \Pr(S_t\boldsymbol{S}_{t+1:}, \boldsymbol{Y}_{t:}|\boldsymbol{A}_{t:}, S_t)\Pr(\boldsymbol{S}_{:t-1}, S_t, \boldsymbol{Y}_{:t-1}|\boldsymbol{A}_{:t-1}, S_{t_i}) \quad (49)\\ &= \Pr(\boldsymbol{S}_{t+1:}, \boldsymbol{Y}_{t:}|\boldsymbol{A}_{t:}, S_t)\Pr(\boldsymbol{S}_{:t-1}, S_t, \boldsymbol{Y}_{:t-1}|\boldsymbol{A}_{:t-1}, S_{t_i}), \quad (50)\end{aligned}$$

using the fact that $\Pr(A, B|A) = \Pr(B|A)$. If we sum over output/hidden-state futures, we get the relation:

$$\Pr(\boldsymbol{S}_{:t-1}, S_t, \boldsymbol{Y}_{:t-1}|\boldsymbol{A}_{:t-1}\boldsymbol{A}_{t:}, S_{t_i}) = \Pr(\boldsymbol{S}_{:t-1}, S_t, \boldsymbol{Y}_{:t-1}|\boldsymbol{A}_{:t-1}, S_{t_i}), \quad (51)$$

which implies our first information equality

$$I[\boldsymbol{A}_{t:}; \boldsymbol{S}_{:t-1}S_t, \boldsymbol{Y}_{:t-1}|\boldsymbol{A}_{:t-1}, S_{t_i}] = 0. \quad (52)$$

Then, divide both sides of Eq. (50) by $\Pr(\boldsymbol{S}_{:t-1}S_t, \boldsymbol{Y}_{:t-1}|\boldsymbol{A}_{:t-1}\boldsymbol{A}_{t:}, S_{t_i})$ to obtain

$$\begin{aligned}\Pr(\boldsymbol{S}_{t+1:}, \boldsymbol{Y}_{t:}|\boldsymbol{S}_{:t-1}, S_t, \boldsymbol{Y}_{:t-1}, \boldsymbol{A}_{:t-1}\boldsymbol{A}_{t:}, S_{t_i}) &= \Pr(\boldsymbol{S}_{t+1:}, \boldsymbol{Y}_{t:}|\boldsymbol{A}_{t:}, S_t) \quad (53)\\ S_i &\text{ is part of} \quad (54)\\ \boldsymbol{S}_{:t-1}\Pr(\boldsymbol{S}_{t+1:}, \boldsymbol{Y}_{t:}|\boldsymbol{S}_{:t-1}, \boldsymbol{Y}_{:t-1}, \boldsymbol{A}_{:t-1}\boldsymbol{A}_{t:}, S_t) &= \Pr(\boldsymbol{S}_{t+1:}, \boldsymbol{Y}_{t:}|\boldsymbol{A}_{t:}, S_t), \quad (55)\end{aligned}$$

that implies our second equality

$$I[\boldsymbol{S}_{t+1:}, \boldsymbol{Y}_{t:}; \boldsymbol{S}_{:t-1}, \boldsymbol{Y}_{:t-1}, \boldsymbol{A}_{:t-1}|\boldsymbol{A}_{t:}, S_t] = 0. \quad (56)$$

□

E Proof of Lemma 3

Proof. Let consider $S_t = \boldsymbol{H}_{t-1}$. The two conditions for being a world model, stated in Eq. (1), can be proved as follows. The first property follows directly by noticing that $\boldsymbol{s}_{:t} = \boldsymbol{h}_{:t-1} = s_t$, and the second one from the following calculation:

$$\begin{aligned}p(y_t|s_t, \boldsymbol{a}_{t:}) =& p(y_t|\boldsymbol{h}_{:t-1}, \boldsymbol{a}_{t:}) = p(y_t|\boldsymbol{y}_{:t-1}, \boldsymbol{a}_{:}) = \frac{p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:})}{p(\boldsymbol{y}_{:t-1}|\boldsymbol{a}_{:})} = \frac{p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:t})}{p(\boldsymbol{y}_{:t-1}|\boldsymbol{a}_{:t})}\\ =& p(y_t|\boldsymbol{y}_{:t-1}, \boldsymbol{a}_{:t}) = p(y_t|\boldsymbol{h}_{:t-1}, a_t) = p(y_t|s_t, a_t), \quad (57)\end{aligned}$$

where we are using the fact that the interface is anticipation-free. Finally, the condition for being a transducer from Def. 14 can be proven by

$$p(s_{t+1}|\boldsymbol{s}_{:t}, \boldsymbol{h}_{:t}, \boldsymbol{a}_{t+1:}) = p(s_{t+1}|s_t, h_t, \boldsymbol{a}_{t+1:}) = \delta_{s_{t+1}}^{(s_t, h_t)} = p(s_{t+1}|s_t, h_t), \quad (58)$$

where δ_a^b is the Kroneker delta that is one if $a = b$. □

F Proof of Lemma 4

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

$$
\begin{aligned}
p(\boldsymbol{y}_{:t}\boldsymbol{s}'_{:t+1}|\boldsymbol{a}_{:}) &= p(s'_0|\boldsymbol{a}_{:}) \prod_{\tau=0}^{t} p(y_\tau, s'_{\tau+1}|\boldsymbol{h}_{:\tau-1}, \boldsymbol{s}'_{:\tau}, \boldsymbol{a}_{\tau:}) && (59)\\
&= \sum_{\tau=0}^{t} \sum_{\substack{s_\tau\in\mathcal{S}\\ \phi(s_\tau)=s'_\tau}} p(s_0|\boldsymbol{a}_{:}) \prod_{\tau=0}^{t} p(y_\tau, s_{\tau+1}|\boldsymbol{h}_{:\tau-1}, \boldsymbol{s}_{:\tau}, \boldsymbol{a}_{\tau:}) && (60)\\
&\overset{(a)}{=} \sum_{\tau=0}^{t} \sum_{\substack{s_\tau\in\mathcal{S}\\ \phi(s_\tau)=s'_\tau}} p(s_0) \prod_{\tau=0}^{t} p(y_\tau|s_\tau, a_\tau) p(s_{\tau+1}|s_\tau, h_\tau) && (61)\\
&\overset{(b)}{=} \sum_{\tau=0}^{t} \sum_{\substack{s_\tau\in\mathcal{S}\\ \phi(s_\tau)=s'_\tau}} p(s_0) \prod_{\tau=0}^{t} p(y_\tau|s'_\tau, a_\tau) p(s_{\tau+1}|s_\tau, h_\tau) && (62)\\
&= \sum_{\tau=0}^{t-1} \sum_{\substack{s_\tau\in\mathcal{S}\\ \phi(s_\tau)=s'_\tau}} p(s_0) \prod_{\tau=0}^{t-1} p(y_\tau|s'_\tau, a_\tau) p(s_{\tau+1}|s_\tau, h_\tau) p(y_t|s'_t, a_t) \sum_{\substack{s_{t+1}\in\mathcal{S}\\ \phi(s_{t+1})=s'_{t+1}}} p(s_{t+1}|s_t, h_\tau) && (63)\\
&\overset{(c)}{=} \sum_{\tau=0}^{t-1} \sum_{\substack{s_\tau\in\mathcal{S}\\ \phi(s_\tau)=s'_\tau}} p(s_0) \prod_{\tau=0}^{t-1} p(y_\tau|s'_\tau, a_\tau) p(s_{\tau+1}|s_\tau, h_\tau) p(y_t|s'_t, a_t) p(s'_{t+1}|s'_t, h_\tau) && (64)\\
&\overset{(d)}{=} \dots && (65)\\
&= \left[\sum_{\substack{s_0\in\mathcal{S}\\ \phi(s_0)=s'_0}} p(s_0) \right] \prod_{\tau=0}^{t} p(y_\tau|s'_\tau, a_\tau) p(s'_{\tau+1}|s'_\tau, h_\tau) && (66)\\
&\overset{(e)}{=} p(s'_0) \prod_{\tau=0}^{t} p(y_\tau|s'_\tau, a_\tau) p(s'_{\tau+1}|s'_\tau, h_\tau). && (67)
\end{aligned}
$$

Above, (a) uses that S_t is the memory state of a transducer, (b) and (c/e) use the first and second properties of homomorphisms, respectively, and (d) assumes the same steps of previous equations are done iteratively. This result shows that S'_t yields a transducer for the same interface, given that

$$p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:}) = \sum_{\boldsymbol{s}_{:t+1}} p(\boldsymbol{y}_{:t}\boldsymbol{s}_{:t+1}|\boldsymbol{a}_{:}) = \sum_{\boldsymbol{s}'_{:t+1}} p(\boldsymbol{y}_{:t}\boldsymbol{s}'_{:t+1}|\boldsymbol{a}_{:}). \tag{68}$$

□

G 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

$$
\begin{aligned}
\Pr(Y_t = y|S_t = s, A_t = a) &= \Pr(Y_t = y|S'_t = \phi(s), A_t = a) && (69)\\
&= \Pr(Y_t = y|S'_t = \phi(s'), A_t = a) && (70)\\
&= \Pr(Y_t = y|S_t = s', A_t = a). && (71)
\end{aligned}
$$

Additionally, using the second property one finds that

$$\sum_{s'' \in [\tilde{s}]} \Pr\big(S_{t+1} = s'' | S_t = s, H_t = (y,a)\big) = \Pr\big(S'_{t+1} = \tilde{s} | S'_t = \phi(s), H_t = (y,a)\big) \tag{72}$$

$$= \Pr\big(S'_{t+1} = \tilde{s} | S'_t = \phi(s'), H_t = (y,a)\big) \tag{73}$$

$$= \sum_{s'' \in [\tilde{s}]} \Pr\big(S_{t+1} = s'' | S_t = s', H_t = (y,a)\big), \tag{74}$$

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

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 $\Pr\big(Y_t = y | S_t = s, A_t = a\big) = \Pr\big(Y_t = y | S_t = s', A_t = a\big)$ for any $(s, s') \in B$, which in turn implies that

$$\Pr\big(Y_t = y | \phi(S_t) = [s], A'_t = a\big) = \Pr\big(Y_t = y | S_t = s, A_t = a\big), \tag{75}$$

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

$$\Pr\big(\phi(S_{t+1}) = [\tilde{s}] | S_t = s, H_t = (y,a)\big) = \sum_{s'' \in [\tilde{s}]} \Pr\big(S_{t+1} = s'' | S_t = s, H_t = (y,a)\big) \tag{76}$$

$$= \sum_{s'' \in [\tilde{s}]} \Pr\big(S_{t+1} = s'' | S_t = s', H_t = (y,a)\big) \tag{77}$$

$$= \Pr\big(\phi(S_{t+1}) = [\tilde{s}] | S_t = s', H_t = (y,a)\big), \tag{78}$$

which implies that

$$\Pr\big(\phi(S_{t+1}) = [\tilde{s}] | S_t = s, H_t = (y,a)\big) = \Pr\big(\phi(S_{t+1}) = [\tilde{s}] | \phi(S_t) = [s], H_t = (y,a)\big). \tag{79}$$

Using this, one can finally show that

$$\Pr\big(\phi(S_{t+1}) = [\tilde{s}] | \phi(S_t) = [s], H_t = (y,a)\big) = \sum_{s'' \in [\tilde{s}]} \Pr\big(S_{t+1} = s'' | \phi(S_t) = [s], H_t = (y,a)\big) \tag{80}$$

$$= \sum_{s'' \in [s]} \Pr\big(S_{t+1} = \tilde{s} | S_t = s, H_t = (y,a)\big) \tag{81}$$

□

H Algorithms to reduce a transducer

, then one can reduce the world model as follows:

1. Compute a singular value decomposition $U_m = U \Lambda V^\intercal$, where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are unitary matrices of singular vectors and $\Lambda \in \mathcal{R}^{m \times n}$ is a diagonal matrix with $\text{Rank}(V_m) = r$ non-zero elements.
2. Collect the r left singular vectors associated with non-zero singular values, and create the matrix $C = [\boldsymbol{v}_1, \dots, \boldsymbol{v}_n] \in \mathbb{R}^{n \times r}$.
3. Use C as a transformation matrix to define the new world states, and calculate the resulting quasi-stochastic matrices.

It can be shown that the resulting representation is minimal as in Def. 5. For more details on this procedure, see (Balasubramanian, 1993, Sec. 3) and also (Huang et al., 2015, Algorithm 1).

I Proof of Lemma 5

Proof. The predict step is given by

$$d_t = \sum_{s_{t-1}} p(s_t|s_{t-1}, \boldsymbol{h}^{t-1}) p(s_{t-1}|\boldsymbol{h}^{t-1}) = \sum_{s_{t-1}} p(s_t|s_{t-1}, h_{t-1}) b_{t-1}(s_{t-1}), \tag{82}$$

and the update step is given by

$$b_t = \frac{p(s_t, \boldsymbol{h}^{t-1}, a_t, y_t)}{p(\boldsymbol{h}^{t-1}, a_t, y_t)} = \frac{p(y_t|s_t, \boldsymbol{h}^{t-1}, a_t) p(s_t|\boldsymbol{h}^{t-1}, a_t)}{p(y_t|\boldsymbol{h}^{t-1}, a_t)} = \frac{p(y_t|s_t, a_t)}{Z'} d_t(s_t) \tag{83}$$

with Z' a normalising constant, where the last equality uses the fact that that

$$p(s_t|\boldsymbol{h}^{t-1}, a_t) = \frac{p(s_t, \boldsymbol{h}^{t-1}, a_t)}{p(\boldsymbol{h}^{t-1}, a_t)} = \frac{p(a_t|s_t, \boldsymbol{h}^{t-1}) p(s_t|\boldsymbol{h}^{t-1})}{p(a_t|\boldsymbol{h}^{t-1})} = p(s_t|\boldsymbol{h}^{t-1}), \tag{84}$$

thanks to the fact that actions depend on histories and not on states, and hence $p(a_t|s_t, \boldsymbol{h}^{t-1}) = p(a_t|\boldsymbol{h}^{t-1})$. Direct updates between b's and d's can be calculated from these equations directly, giving

$$b_t(s_t) = \frac{p(y_t|s_t, a_t)}{Z'} \sum_{s_{t-1}} p(s_t|s_{t-1}, h_{t-1}) b_{t-1}(s_{t-1}), \tag{85}$$

$$d_t(s_t) = \frac{1}{Z'} \sum_{s_{t-1}} p(s_t|s_{t-1}, h_{t-1}) p(y_{t-1}|s_{t-1}, a_{t-1}) d_{t-1}(s_{t-1}), \tag{86}$$

corresponding to the updates of beliefs and mixed-states. Finally, notice that if S_t belongs to a input-Moore transducer, then $p(y_t|s_t, a_t) = p(y_t|s_t)$ and hence one arrives to Eq. (8). □

J Mixed-states are transducers and generate the same interface

As mentioned earlier, the predictive mixed-state presentation (MSP) of a transducer is determined by the probability of hidden state s_t given the action-outcome history

$$p(s_t|a_{0:t-1}, y_{0:t-1}). \tag{87}$$

As discussed in App. R, the belief states represent points in the Hilbert space of $\mathcal{S}$:

$$|\rho^P(y_{0:t-1}, a_{0:t-1})\rangle \equiv \sum_{s_t} |s_t\rangle p(s_t|a_{0:t-1}, y_{0:t-1}). \tag{88}$$

This can be exactly calculated from the vector that represents the initial state distribution

$$|\rho_0\rangle \equiv \sum_{s_0} |s_0\rangle p(s_0), \tag{89}$$

and applying the linear operators of the transducer in sequence

$$|\rho^P(\boldsymbol{y}_{0:t-1}, \boldsymbol{a}_{0:t-1}))\rangle = \frac{T^{(\boldsymbol{y}_{0:t-1}|\boldsymbol{a}_{0:t-1})}|\rho_0\rangle}{\langle 1|T^{(\boldsymbol{y}_{0:t-1}|\boldsymbol{a}_{0:t-1})}|\rho_0\rangle}, \tag{90}$$

where $T^{(\boldsymbol{y}_{0:t-1}|\boldsymbol{a}_{0:t-1})} \equiv \prod_{\tau=0}^{t-1} T^{(y_\tau|a_\tau)}$ and $\langle 1| \equiv \sum_s \langle s|$.

The mixed states are themselves predictive memory states of the transducer. They are functions of the past, and they store the relevant information necessary to produce the future action-outcome mapping:

$$p(y_{t:z}|a_{t:z}, y_{0:t}, a_{0:t}) = \sum_s \langle s| \prod_{\tau=t}^{z} T^{(y_\tau|a_\tau)} |\rho^P(y_{0:t-1}, a_{0:t-1})\rangle, \tag{91}$$

which uses the fact that

$$p(y_{0:t}) = \sum_s \langle s| \prod_{\tau=0}^{t} T^{(y_\tau|a_\tau)}|\rho_0\rangle\rangle. \tag{92}$$

Because the MSP is a predictive transducer, it can be coarse-grained to the ϵ-transducer. The exact form of the transducer of the MSP states is

$$M^{(y|a)}_{|\rho\rangle\to|\rho'\rangle} = \langle 1|T^{(y|a)}|\rho\rangle\delta_{|\rho'\rangle,\frac{T^{(y|a)}|\rho\rangle}{\langle 1|T^{(y|a)}|\rho\rangle}}. \tag{93}$$

Thus, if we can calculate the behavior of actions and outcomes alongside the mixed-state trajectory $|\rho\rangle_{0:t}$:

$$p(|\rho\rangle_{0:t}, \boldsymbol{y}_{0:t-1}|\boldsymbol{a}_{0:t-1}, |\rho_0\rangle) = \prod_{\tau=0}^{t-1} M^{(y_\tau|a_\tau)}_{|\rho_\tau\rangle\to|\rho_{\tau+1}\rangle} \tag{94}$$

$$= \prod_{\tau=0}^{t-1} \langle 1|T^{(y_\tau|a_\tau)}|\rho_\tau\rangle\delta_{|\rho_{\tau+1}\rangle,\frac{T^{(y_\tau|a_\tau)}|\rho_\tau\rangle}{\langle 1|T^{(y_\tau|a_\tau)}|\rho_\tau\rangle}} \tag{95}$$

We can then sum over all mixed-state trajectories to obtain the original interface, using the fact that the only nonzero terms in the sum are those for which $|\rho_\tau\rangle = |\rho(\boldsymbol{y}_{0:\tau-1}, \boldsymbol{a}_{0:\tau-1})\rangle$:

$$\sum_{|\rho\rangle_{0:t}} p(|\rho\rangle_{0:t}, \boldsymbol{y}_{0:t-1}|\boldsymbol{a}_{0:t-1}, |\rho_0\rangle) = \sum_{|\rho\rangle_{0:t}} \prod_{\tau=0}^{t-1} \langle 1|T^{(y_\tau|a_\tau)}|\rho_\tau\rangle\delta_{|\rho_{\tau+1}\rangle,\frac{T^{(y_\tau|a_\tau)}|\rho_\tau\rangle}{\langle 1|T^{(y_\tau|a_\tau)}|\rho_\tau\rangle}} \tag{96}$$

$$= \prod_{\tau=0}^{t-1} \langle 1|T^{(y_\tau|a_\tau)}|\rho(\boldsymbol{y}_{0:\tau-1}, \boldsymbol{a}_{0:\tau-1})\rangle \tag{97}$$

$$= \prod_{\tau=0}^{t-1} \frac{\langle 1|T^{(\boldsymbol{y}_{0:\tau}|\boldsymbol{a}_{0:\tau})}|\rho_0\rangle}{\langle 1|T^{(\boldsymbol{y}_{0:\tau-1}|\boldsymbol{a}_{0:\tau-1})}|\rho_0\rangle} \tag{98}$$

$$= \frac{\langle 1|T^{(\boldsymbol{y}_{0:t-1}|\boldsymbol{a}_{0:t-1})}|\rho_0\rangle}{\langle 1|\rho_0\rangle} \tag{99}$$

$$= \langle 1|T^{(\boldsymbol{y}_{0:t-1}|\boldsymbol{a}_{0:t-1})}|\rho_0\rangle. \tag{100}$$

This is precisely the probability of outcomes given by the original transducer T. Note that with this notation $x_{t:t} = x_t$ and $x_{t:t-1}$ is null, meaning with applying no actions or outcomes. Thus we have constructed a transducer M that uses the mixed-states to generate the original interface, meaning that the corresponding belief transducer is a presentation of that interface. Therefore, we call it the Mixed-State Presentation (MSP) of that particular transducer.

The causal states of the ϵ-transducer are a function of the past $s_t = \epsilon(y_{0:t-1}, a_{0:t-1})$. Therefore, the MSP states are isomorphic to the causal states

$$|\rho^P(y_{0:t-1}, a_{0:t-1})\rangle = \sum_s |s\rangle\delta_{s,\epsilon(y_{0:t-1},a_{0:t-1})}. \tag{101}$$

As a result, the MSP is also the ϵ-transducer.

The ϵ-transducer is not the only machine whose MSP produces the ϵ-transducer. The MSP of any transducer without redundant states will produce be the ϵ-machine. In this case, a redundant state s_t has a future distribution $p(y_{t:}|a_{t:}, s_t)$ that is a linear combination of other states

$$p(y_{t:}|a_{t:}, s_t) = \sum_{s'_t\neq s_t} q(s'_t)p(y_{t:}|a_{t:}, s'_t). \tag{102}$$

If all these states have linearly independent futures, then every linear combination of states produces a distinct future distribution. Thus, it is impossible to coarse-grain further while preserving the functionality of the transducer, and the MSP must be the ϵ-transducer.

K Proof of Prop. 2

Proof. Lemma 3 shows that $S_t = \boldsymbol{H}_{:t-1}$ is always a valid transducer. Also, from Def. 5 and Prop. 1 one can see that a bisimulation of a transducer always yields a valid transducer. Thus, the only thing that remains is to prove that the coarse-graing defined by Eq. (11) has the two properties of a bisimulation (Def. 6). Condition (i) follows from Eq. (11) directly, since it only considers futures of length 1. A proof that Condition (ii) follows from Eq. (11), i.e. that the dynamics of the equivalence classes are conditionally Markovian on the actions, can be found in (Barnett & Crutchfield, 2015, Prop. 5).

□

L Proof of Theorem 3

Proof. A predictive transducer has memory states S_t that satisfy the condition

$$I[S_t, Y_{t:}|A_{t:}, Y_{:t-1}A_{:t-1}] = 0, \tag{103}$$

for all t. In combination of the property of being non-anticipatory

$$I[A_{:t-1}, Y_{:t-1}; Y_{t:}|A_{t:}, S_t] = 0, \tag{104}$$

this is equivalent to the tripartite equality

$$\Pr(Y_{t:}|A_{t:}, S_t, A_{:t-1}, Y_{:t-1}) = \Pr(Y_{t:}|A_{t:}, S_t) = \Pr(Y_{t:}|X_{t:}, A_{:t-1}, Y_{:t-1}), \tag{105}$$

holding whenever $\Pr(Y_{t:}, A_{t:}, S_t, A_{:t-1}, Y_{:t-1}) \neq 0$. In general, $A_{:t-1}$ and $Y_{:t-1}$ are the actions that the agent has already interacted with when it is in configuration S_t, so we can express them to them as the past at time t $\boldsymbol{Y}_{:t-1} \equiv Y_{:t-1}$ $\boldsymbol{A}_{:t-1} \equiv A_{:t-1}$. This allows us to rewrite the condition for an agent being predictive

$$\Pr(\boldsymbol{Y}_{t:}|\boldsymbol{A}_{t:}, S_t, \boldsymbol{A}_{:t-1}, \boldsymbol{Y}_{:t-1}) = \Pr(\boldsymbol{Y}_{t:}|\boldsymbol{A}_{t:}, S_t) = \Pr(\boldsymbol{Y}_{t:}|\boldsymbol{A}_{t:}, \boldsymbol{A}_{:t-1}, \boldsymbol{Y}_{:t-1}), \tag{106}$$

when $\Pr(\boldsymbol{Y}_{t:}, \boldsymbol{A}_{t:}, S_t, \boldsymbol{A}_{:t-1}, \boldsymbol{Y}_{:t-1}) \neq 0$. This condition means that the memory S_t and history $\boldsymbol{A}_{:t-1}, \boldsymbol{Y}_{:t-1}$ are mutually compatible and can coexist.

For comparison, consider the causal equivalence relation that leads to the ϵ-transducer for the same interface:

$$\epsilon(\boldsymbol{a}_{:t-1}, \boldsymbol{y}_{:t-1}) = \epsilon(\boldsymbol{a}'_{:t-1}, \boldsymbol{y}'_{:t-1}) \tag{107}$$

$$\Leftrightarrow \tag{108}$$

$$\Pr(\boldsymbol{Y}_{t:}|\boldsymbol{A}_{t:}, \boldsymbol{A}_{:t-1} = \boldsymbol{a}_{:t-1}, \boldsymbol{Y}_{:t-1} = \boldsymbol{y}_{:t-1}) = \Pr(\boldsymbol{Y}_{t:}|\boldsymbol{A}_{t:}, \boldsymbol{A}_{:t-1} = \boldsymbol{a}'_{:t-1}, \boldsymbol{Y}_{:t-1} = \boldsymbol{y}'_{:t-1}). \tag{109}$$

By construction $S_t = \epsilon(\boldsymbol{A}_{:t-1}, \boldsymbol{Y}_{:t-1})$, which means that

$$\Pr(\boldsymbol{Y}_{t:}|\boldsymbol{A}_{t:}, S_t, \boldsymbol{A}_{:t-1}, \boldsymbol{Y}_{:t-1}) = \Pr(\boldsymbol{Y}_{t:}|\boldsymbol{A}_{t:}, \epsilon(\boldsymbol{A}_{:t-1}, \boldsymbol{Y}_{:t-1}), \boldsymbol{A}_{:t-1}, \boldsymbol{Y}_{:t-1}) \tag{110}$$

$$= \Pr(\boldsymbol{Y}_{t:}|\boldsymbol{A}_{t:}, \boldsymbol{A}_{:t-1}, \boldsymbol{Y}_{:t-1}), \tag{111}$$

using the fact that $\Pr(A|f(B), B) = \frac{\Pr(A, f(B)|B)}{\Pr(f(B)|B)} = \frac{\Pr(f(B)|A,B)\Pr(A|B)}{\Pr(f(B)|B)} = \Pr(A|B)$. Furthermore, the equivalence condition implies that the memory shields that future from the past

$$\Pr(\boldsymbol{Y}_{t:}|\overrightarrow{X}_t, \epsilon(\boldsymbol{A}_{:t-1}, \boldsymbol{Y}_{:t-1}), \boldsymbol{A}_{:t-1}, \boldsymbol{Y}_{:t-1}) = \Pr(\boldsymbol{Y}_{t:}|\boldsymbol{A}_{t:}, \epsilon(\boldsymbol{A}_{:t-1}, \boldsymbol{Y}_{:t-1})), \tag{112}$$

so the ϵ-transducer is predictive.

Furthermore, any predictive transducer can be coarse-grained to the ϵ-transducer. This can be seen by setting the equivalence relation

$$\epsilon'(s_t) = \epsilon'(s'_t) \tag{113}$$
$$\Leftrightarrow \tag{114}$$
$$\Pr(\boldsymbol{Y}_{t:}|\boldsymbol{A}_{t:}, S_t = s_t) = \Pr(\boldsymbol{Y}_{t:}|\boldsymbol{A}_{t:}, S_t = s'_t). \tag{115}$$

This coarse-graining achieves the ϵ-transducer, because the equality condition can be re-expressed for predictive transducers

$$\Pr(\boldsymbol{Y}_{t:}|\boldsymbol{A}_{t:}, S_t = s_t) = \Pr(\boldsymbol{Y}_{t:}|\boldsymbol{A}_{t:}, S_t = s'_t) \tag{116}$$
$$\Pr(\boldsymbol{Y}_{t:}|\boldsymbol{A}_{t:}, S_t = s_t, \boldsymbol{A}_{:t-1} = \boldsymbol{a}_{:t-1}, \boldsymbol{Y}_{:t-1} = \boldsymbol{y}_{:t-1}) = \Pr(\boldsymbol{Y}_{t:}|\boldsymbol{A}_{t:}, S_t = s'_t, \boldsymbol{A}_{:t-1} = \boldsymbol{a}'_{:t-1}, \boldsymbol{Y}_{:t-1} = \boldsymbol{y}'_{:t-1}) \tag{117}$$
$$\Pr(\boldsymbol{Y}_{t:}|\boldsymbol{A}_{t:}, \boldsymbol{A}_{:t-1} = \boldsymbol{a}_{:t-1}, \boldsymbol{Y}_{:t-1} = \boldsymbol{y}_{:t-1}) = \Pr(\boldsymbol{Y}_{t:}|\boldsymbol{A}_{t:}, \boldsymbol{A}_{:t-1} = \boldsymbol{a}'_{:t-1}, \boldsymbol{Y}_{:t-1} = \boldsymbol{y}'_{:t-1}), \tag{118}$$

when the memory and history are mutually compatible. Thus, we have the causal equivalence relation is satisfied for all histories that are consistent the coarse-grained with the memory states that map to the same state $\epsilon'(s) = \epsilon'(s')$. This logic can be performed in both directions: from memory to past and past to memory. Thus the causal equivalence relation and ϵ' are identical, meaning that the equivalence relation ϵ' yields the ϵ-transducer. □

M Comparing the reduction of general vs predictive transducers

Building upon the discussion about the canonical dimension of a transducer (see Eq. (5)), let us focus 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. (11) correspond to merging together all rows of W_t that are equal. In contrast, the cannonical 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. 1 — necessary make the transducer to stop being predictive.

A mixed-state construction of a transducer is guaranteed to produce the ϵ-transducer when there is no linear dependency in its future distributions for each state. We will use Dirac notation as discussed in App. R for the future vector of each state s_0

$$|\overrightarrow{p}_t(s_0)\rangle \equiv \sum_{\boldsymbol{y}_{:t},\boldsymbol{y}_{:t}} |\boldsymbol{y}_{:t}, \boldsymbol{a}_{:t}\rangle p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:t}, s_0), \tag{119}$$

where the joint ket is defined

$$|\boldsymbol{y}_{:t}, \boldsymbol{a}_{:t}\rangle \equiv \bigotimes_{\tau=0}^{t} |y_\tau\rangle \otimes |a_\tau\rangle \tag{120}$$

Redundancy appears as linear dependence between future distributions, meaning that we can express the future of one state as a linear combination of the others

$$|\overrightarrow{p}_t(s_0)\rangle = \sum_{s'_0 \neq s_0} k(s'_0)|\overrightarrow{p}_t(s'_0)\rangle, \tag{121}$$

where $k(s'_t)$ is some real function of the memory states.

Lemma 8. *If there no linear dependence between the future distributions of a transducers memory states, then the MSP is the ϵ-transducer.*

Proof. The future distribution of belief state d_t is given by

$$|\overrightarrow{p}_t(d_0)\rangle = \sum_{s_0} d_0(s_0)|\overrightarrow{p}_t(s_0)\rangle. \tag{122}$$

If there is another state $d'$of the MSP with the same future distribution, then it must be true that

$$\sum_{s_t}(d(s_t) - d'(s_t))|\overrightarrow{p}_t(s_0)\rangle = 0. \tag{123}$$

However, this contradicts linear independence of the futures, so it must be true that $d = d'$ if they have the same future distribution. Therefore, all states of the MSP have distinct future distributions, meaning that they satisfy they are the states of the ϵ-transducer. □

If there is linear dependence, then there is the possibility that different mixed-states have the same future distribution, in which case the dimensionality MSP can be reduced.

N Some generic retrodictive world models

N.1 A cannonical retrodictive world model

For a given interface $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$, the process $S_t = \boldsymbol{Y}_{t:}$ is a retrodictive world model but is not anticipatory-free, and hence it doesn't lead to a transducer (see Sec. N.2). This world model can be described as a 'transducer with insider information', which knows what decisions are going to be made beforehand.

One can further show that all anticipation-free transducer have a retrodictive transducer, which can be described as 'the profet' as it has an answer to all possible sequence of future actions. To build the world model of this transducer, let us first denote as $\mathcal{T}_\mathcal{A}$ the regular tree with one root and where each node has one branch per elements in $\mathcal{A}$. Let's denote by $\mathcal{N}(\mathcal{T}_\mathcal{A})$ the nodes of the tree, and establish some operations:

- $\mu : \mathcal{N}(\mathcal{T}_\mathcal{A}) \to \mathcal{A}^*$ and $\nu : \mathcal{N}(\mathcal{T}_\mathcal{A}) \to \mathcal{N}(\mathcal{T}_\mathcal{A})^*$, where $\mu(v)$ and $\nu(v)$ returns a vector with all the branches and nodes in the path leading back from v to the root, respectively, with $()^*$ being the Kleene operator.
- $\pi : \mathcal{N}(\mathcal{T}_\mathcal{A}) \times \mathcal{A} \to \mathcal{N}(\mathcal{T}_\mathcal{A})$, where $\pi(v, a)$ gives the descendent of v connected via branch a.
- $\tau : \mathcal{N}(\mathcal{T}_\mathcal{A}) \to \mathbb{N}$, where $\tau(v)$ is the depth of v in the tree.

With all this, we are ready to define our world model. In general, $S_t \in \mathcal{Y}^{\mathcal{T}_\mathcal{A}}$ are random varibles that take values on $\mathcal{T}_\mathcal{A}$-shaped sequences of symbols in $\mathcal{Y}$. Concretely, $S_0 = \big(Z_v : v \in \mathcal{N}(\mathcal{T}_\mathcal{A})\big)$ with $Z_v \in \mathcal{Y}$ being random variables, whose joint distribution is given by

$$\Pr\big(S_0 = (Z_v : v \in \mathcal{T}_\mathcal{A})\big) := \prod_{v \in \mathcal{T}_\mathcal{A}} \Pr(Z_v|\boldsymbol{Z}_{\nu(v)}) \tag{124}$$

with $\boldsymbol{Z}_{\nu(v)}$ the vector of variables corresponding to nodes in $\nu(v)$ and

$$\Pr(Z_v = y|\boldsymbol{Z}_{\nu(v)} = \boldsymbol{y}_{:\nu(v)-1}) := \Pr\big(Y_{\tau(v)} = y|\boldsymbol{Y}_{:\tau(v)-1} = \boldsymbol{y}_{:\nu(v)-1}, \boldsymbol{A}_{:\tau(v)} = \mu(v)\big) \tag{125}$$

Then, the world's dynamics are established recursively by $p(s_{t+1}|\boldsymbol{s}_{:t}, \boldsymbol{h}_:) := \delta_{s_{t+1}}^{f(s_t,a_t)}$ so that $S_{t+1} = f(S_t, A_t)$ a.s., with the unifilar update established by

$$S_{t+1} = (Z_v^{t+1} : v \in \mathcal{T}_\mathcal{A}) \quad \text{with} \quad Z_v^{t+1} = Z_{\pi(v,A_t)}^t. \tag{126}$$

In summary, the world is first initialised at time zero by sampling S_0, i.e. by sampling Z_v for all $v \in \mathcal{T}_\mathcal{A}$ — which stands to sample $\boldsymbol{Y}_:$ for all possible sequences of actions $\boldsymbol{a}_:$. After this, the world evolves deterministically by following the update rule given by f.

N.2 Naive retrodictive model is a world model but cannot be run

Here we prove that taking $R_t = \boldsymbol{Y}_{t:}$ is a valid world model, but is not anticipation-free and hence is not a transducer — as it cannot be properly run without future information.

Proof. For a given interface $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$, let's define a stochastic process $R_t \in \mathcal{Y}^{\mathbb{N}}$ conditional on the semi-infinite history $\boldsymbol{H}_:$ as the coarse-graining $R_t = g(\boldsymbol{Y}_:, \boldsymbol{A}_:) = \boldsymbol{Y}_{t:}$. Let's show that R_t is a valid world model. For this, let's first introduce operations $\psi_0(r_t)$ and $\psi(r_t)$ that are such that $r_t = \big(\psi_0(r_t), \psi(r_t)\big)$, so that ψ_0 is a projection that gives the first component of r_t and ψ gives all the rest without the first component. Then, let us first notice that R_t induces a simple yet valid conditional distributions of the form

$$p(\boldsymbol{r}_{:t}|\boldsymbol{h}_:) = p(r_0|\boldsymbol{y}_:, \boldsymbol{a}_:) \prod_{\tau=0}^{t-1} \delta^{r_{\tau+1}}_{\psi(r_\tau)} = \delta^{\boldsymbol{y}_:}_{r_0} \prod_{\tau=0}^{t-1} \delta^{r_{\tau+1}}_{\psi(r_\tau)}, \tag{127}$$

which is the type of object specified by Def. 2. Furthermore, direct calculations show that

$$p(\boldsymbol{y}_{t:}|\boldsymbol{h}_{:t-1}, \boldsymbol{r}_{:t}, \boldsymbol{a}_{t:}) = \delta^{r_t}_{\boldsymbol{y}_{t:}} = p(\boldsymbol{y}_{t:}|r_t, \boldsymbol{a}_{t:}) \qquad \text{(a.s.)} \tag{128}$$

and also

$$p(y_t|\boldsymbol{a}_{t:}, s_t) = \delta^{\psi_0(r_t)}_{y_t} = p(y_t|a_t, s_t). \tag{129}$$

This proves that R_t is a valid world model. However, it is not anticipation-free given that

$$p(r_0|\boldsymbol{a}_:) = p(\boldsymbol{y}_:|\boldsymbol{a}_:) \neq p(\boldsymbol{y}_:) = p(r_0). \tag{130}$$

Therefore, this world model cannot be ran, as it cannot be properly initialised unless having information about future actions. □

O About non-reversible transducers

Let us consider the delay channel, for which the output Y_{t+1} is equal to the previous action A_t (Barnett & Crutchfield, 2015). This channel displays paradoxically acausal behaviour when time reversed. Now, somehow the action A_t determines the outcome at the previous time step Y_{t-1}, meaning that

$$I[Y_{:t-1}; A_{t:}|A_{t-1:}] = I[Y_{t-1}; A_t|A_{t-1:}] \tag{131}$$
$$= H[A_t|A_{t-1:}], \tag{132}$$

which is nonzero if the entropy rate of the actions is nonzero. Moreover, even if the time-reversed interface is anticipation-free, it may be possible that the dynamics of the memory cannot be implemented causaly in reverse time.

P Reversing processes and proof of Theorem 4

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

P.1 Reversing Markov processes

Let's say X_t is a Markov process X_t, so that $p(x_t|\boldsymbol{x}_{:t-1}) = p(x_t|x_{t-1})$. Then, one can show the reverse process is also Markov, as

$$p(x_t|\boldsymbol{x}_{t+1:t'}) = \frac{p(\boldsymbol{x}_{t:t'})}{p(\boldsymbol{x}_{t+1:t'})} = \frac{p(x_t)\prod_{k=t+1}^{t'} p(x_k|\boldsymbol{x}_{t:k-1})}{p(x_{t+1})\prod_{j=t+2}^{t'} p(x_j|\boldsymbol{x}_{t+1:j-1})} \tag{133}$$

$$= \frac{p(x_t)\prod_{k=t+1}^{t'} p(x_k|x_{k-1})}{p(x_{t+1})\prod_{j=t+2}^{t'} p(x_j|x_{j-1})} = \frac{p(x_t)p(x_{t+1}|x_t)}{p(x_{t+1})} = p(x_t|x_{t+1}). \tag{134}$$

P.2 Reversing HMMs

Let's now consider a general (Mealy) HMM, where $p(s_{t+1}, y_t | \boldsymbol{s}_{:t}, \boldsymbol{y}_{:t-1}) = p(s_{t+1}, y_t | s_t)$. Then, one can show the reverse process is also an HMM, as

$$\begin{aligned}
p(s_t, y_t | \boldsymbol{s}_{t+1:t'+1}, \boldsymbol{y}_{t+1:t'}) &= \frac{p(\boldsymbol{s}_{t:t'+1}, \boldsymbol{y}_{t:t'})}{p(\boldsymbol{s}_{t+1:t'+1}, \boldsymbol{y}_{t+1:t'})} &(135)\\
&= \frac{p(s_t, y_t, s_{t+1}) \prod_{k=t+1}^{t'} p(s_{k+1}, y_k | \boldsymbol{s}_{t:k}, \boldsymbol{y}_{t:k-1})}{p(s_{t+1}, y_{t+1}, s_{t+2}) \prod_{j=t+2}^{t'} p(s_{j+1}, y_j | \boldsymbol{s}_{t:j}, \boldsymbol{y}_{t:j-1})} &(136)\\
&= \frac{p(s_t, y_t, s_{t+1}) \prod_{k=t+1}^{t'} p(s_{k+1}, y_k | s_k)}{p(s_{t+1}, y_{t+1}, s_{t+2}) \prod_{j=t+2}^{t'} p(s_{j+1}, y_j | s_j)} &(137)\\
&= \frac{p(s_t) \prod_{k=t}^{t'} p(s_{k+1}, y_k | s_k)}{p(s_{t+1}) \prod_{j=t+1}^{t'} p(s_{j+1}, y_j | s_j)} &(138)\\
&= \frac{p(s_t) p(s_{t+1}, y_t | s_t)}{p(s_{t+1})} &(139)\\
&= p(s_t, y_t | s_{t+1}). &(140)
\end{aligned}$$

Note that this is not time-symmetric, but a 'co-Mealy' structure — as the time indices of the world are shifted.

If the HMM is Moore, 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

$$p(s_t, y_t | \boldsymbol{s}_{t+1:t'+1}, \boldsymbol{y}_{t+1:t'}) = \frac{p(s_t) p(s_{t+1}, y_t | s_t)}{p(s_{t+1})} = \frac{p(s_t) p(s_{t+1}|s_t) p(y_t|s_t)}{p(s_{t+1})} = p(s_t|s_{t+1}) p(y_t|s_t), \tag{141}$$

yielding another Moore HMM.

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. Using this property, it is direct to see that

$$\begin{aligned}
p(\boldsymbol{y}_{:t}, \boldsymbol{s}_{:t+1} | \boldsymbol{a}_:) &= p(s_0) \prod_{\tau=0}^{t} p(y_\tau, s_{\tau+1} | \boldsymbol{y}_{:\tau-1}, \boldsymbol{s}_{:\tau}, \boldsymbol{a}_:) &(142)\\
&= p(s_0) \prod_{\tau=0}^{t} p(y_\tau, s_{\tau+1} | s_\tau, \boldsymbol{a}_{:t}) &(143)\\
&= p(\boldsymbol{y}_{:t}, \boldsymbol{s}_{:t+1} | \boldsymbol{a}_{:t}), &(144)
\end{aligned}$$

showing that transducers naturally impose some arrow of time over actions. Note that for this to work we are using the fact that $p(s_0|\boldsymbol{a}_:) = p(s_0)$, and it would not work for other initial point where this doesn't hold.

Now, let's consider expressing $p(\boldsymbol{y}_{:t}, \boldsymbol{s}_{:t+1} | \boldsymbol{a}_:)$ factor backwards as follows

$$p(\boldsymbol{y}_{:t}, \boldsymbol{s}_{:t+1} | \boldsymbol{a}_:) = p(\boldsymbol{y}_{:t}, \boldsymbol{s}_{:t+1} | \boldsymbol{a}_{:t}) = p(s_{t+1} | \boldsymbol{a}_{:t}) \prod_{\tau=0}^{t} p(y_\tau, s_\tau | \boldsymbol{y}_{\tau+1:t}, \boldsymbol{s}_{\tau+1:t+1}, \boldsymbol{a}_{:t}). \tag{145}$$

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, we can show that

$$\begin{aligned} p(s_\tau, y_\tau|\boldsymbol{s}_{\tau+1:t+1}, \boldsymbol{y}_{\tau+1:t}, \boldsymbol{a}_{:t}) &= \frac{p(\boldsymbol{s}_{\tau:t+1}, \boldsymbol{y}_{\tau:t}, \boldsymbol{a}_{:t})}{p(\boldsymbol{s}_{\tau+1:t+1}, \boldsymbol{y}_{\tau+1:t}, \boldsymbol{a}_{:t})} && (146) \\ &= \frac{p(s_\tau, y_\tau, s_{\tau+1}, \boldsymbol{a}_{:t}) \prod_{k=\tau+1}^{t} p(s_{k+1}, y_k|\boldsymbol{s}_{\tau:k}, \boldsymbol{y}_{\tau:k-1}, \boldsymbol{a}_{:t})}{p(s_{\tau+1}, y_{\tau+1}, s_{\tau+2}, \boldsymbol{a}_{:t}) \prod_{j=\tau+2}^{t} p(s_{j+1}, y_j|\boldsymbol{s}_{\tau:j}, \boldsymbol{y}_{\tau:j-1}, \boldsymbol{a}_{:t})} \\ &= \frac{p(s_\tau, y_\tau, s_{\tau+1}, \boldsymbol{a}_{:t}) \prod_{k=\tau+1}^{t} p(s_{k+1}, y_k|s_k, a_k)}{p(s_{\tau+1}, y_{\tau+1}, s_{\tau+2}, \boldsymbol{a}_{:t}) \prod_{j=\tau+2}^{t} p(s_{j+1}, y_j|s_j, a_j)} && (147) \\ &= \frac{p(s_\tau, \boldsymbol{a}_{:t}) \prod_{k=\tau}^{t} p(s_{k+1}, y_k|s_k, a_k)}{p(s_{\tau+1}, \boldsymbol{a}_{:t}) \prod_{j=\tau+1}^{t} p(s_{j+1}, y_j|s_j, a_j)} && (148) \\ &= \frac{p(s_\tau|\boldsymbol{a}_{:t}) p(s_{\tau+1}, y_\tau|s_\tau, a_\tau)}{p(s_{\tau+1}|\boldsymbol{a}_{:t})} && (149) \end{aligned}$$

From this point, there are different ways forward. One possibility is to define

$$\Delta_\tau := \frac{p(s_\tau|\boldsymbol{a}_{:t})}{p(s_\tau|a_\tau)} \quad \text{and} \quad \Delta'_\tau := \frac{p(s_\tau|\boldsymbol{a}_{:t})}{p(s_\tau|a_{\tau-1})} \tag{150}$$

as measures of discrepancy, which allow us to express the reverse transducer as follows:

$$\begin{aligned} p(s_\tau, y_\tau|\boldsymbol{s}_{\tau+1:t+1}, \boldsymbol{y}_{\tau+1:t}, \boldsymbol{a}_{:t}) &= \frac{p(s_\tau|\boldsymbol{a}_{:t})}{p(s_{\tau+1}|\boldsymbol{a}_{:t})} p(s_{\tau+1}, y_\tau|s_\tau, a_\tau) && (151) \\ &= \frac{\Delta_\tau}{\Delta'_{\tau+1}} \frac{p(s_\tau|a_\tau)}{p(s_{\tau+1}|a_\tau)} p(s_{\tau+1}, y_\tau|s_\tau, a_\tau) && (152) \\ &= \frac{\Delta_\tau}{\Delta'_{\tau+1}} p(s_\tau, y_\tau|s_{\tau+1}, a_\tau). && (153) \end{aligned}$$

Another option is to try a different algebraic route, and do as follows:

$$\begin{aligned} p(s_\tau, y_\tau|\boldsymbol{s}_{\tau+1:t+1}, \boldsymbol{y}_{\tau+1:t}, \boldsymbol{a}_{:t}) &= \frac{p(s_\tau|\boldsymbol{a}_{:t})}{p(s_{\tau+1}|\boldsymbol{a}_{:t})} p(s_{\tau+1}, y_\tau|s_\tau, a_\tau) && (154) \\ &= \frac{p(s_\tau|\boldsymbol{a}_{:t})}{p(s_{\tau+1}|\boldsymbol{a}_{:t})} p(s_{\tau+1}, y_\tau|s_\tau, \boldsymbol{a}_{:t}) && (155) \\ &= \frac{p(s_{\tau+1}, y_\tau, s_\tau|\boldsymbol{a}_{:t})}{p(s_{\tau+1}|\boldsymbol{a}_{:t})} && (156) \\ &= p(s_\tau, y_\tau|s_{\tau+1}, \boldsymbol{a}_{:t}) && (157) \\ &= p(y_\tau|s_\tau, s_{\tau+1}, a_\tau) p(s_\tau|s_{\tau+1}, \boldsymbol{a}_{:t}). && (158) \end{aligned}$$

In both cases, these calculations reveal what is the problem with running transducers back! This usually break down because generally $p(s_\tau|s_{\tau+1}, \boldsymbol{a}_{:t}) \neq p(s_\tau|s_{\tau+1}, a_\tau)$, or equivalently that $\Delta_\tau \neq 1$ or $\Delta'_\tau \neq 1$.

In summary, for any transducer S_t, we can always run it back to reproduce the interface but this needs the whole sequence of actions, as shown by the factorisation given by

$$p(\boldsymbol{y}_{:t}, \boldsymbol{s}_{:t+1}|\boldsymbol{a}_{:}) = p(s_{t+1}|\boldsymbol{a}_{:t}) \prod_{\tau=0}^{t} p(y_\tau|, s_\tau, s_{\tau+1}, a_\tau) p(y_\tau, s_\tau|s_{\tau+1}, \boldsymbol{a}_{:t}). \tag{159}$$

If the transducer satisfies the additional condition

$$p(s_\tau|s_{\tau+1}, \boldsymbol{a}_{:t}) = p(s_\tau|s_{\tau+1}, a_\tau), \tag{160}$$

or equivalently, the information relation

$$I[S_\tau; A_{0:\tau-1}A_{\tau+1:\infty}|S_{\tau+1}, A_\tau] = 0, \tag{161}$$

or the condition

$$\Delta_t = \Delta'_{t+1} = 1, \tag{162}$$

then one could run all back yielding

$$p(\boldsymbol{y}_{:t}, \boldsymbol{s}_{:t+1}|\boldsymbol{a}_{:}) = p(s_{t+1}|\boldsymbol{a}_{:t}) \prod_{\tau=0}^{t} p(y_\tau, s_\tau|s_{\tau+1}, a_t). \tag{163}$$

So, if the above conditions are satisfied, one could generate the interface by the following procedure:

(1) Initialise the world at $p(s_{t+1}|\boldsymbol{a}_{:t})$. Or, for counterfactual analysis, pick a world state $S_{t+1} = s$ that one want to evaluate.

(2) Then run things backward using $p(y_\tau, s_\tau|s_{\tau+1}, a_t)$.

Notice the difference between the kernel of a transducer,

$$p(s_{\tau+1}, y_\tau|\boldsymbol{s}_{:\tau}, \boldsymbol{y}_{:\tau-1}, \boldsymbol{a}_{:t}) = p(s_{\tau+1}, y_\tau|s_\tau, a_t), \tag{164}$$

and the kernel of a transducer running backwards, Co-transducer:

$$p(s_\tau, y_\tau|\boldsymbol{s}_{\tau+1:t+1}, \boldsymbol{y}_{\tau+1:t}, \boldsymbol{a}_{:t}) = p(s_\tau, y_\tau|s_{\tau+1}, a_t). \tag{165}$$

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

$$p(s_{\tau+1}|s_\tau, a_\tau) = p(s_\tau|s_{\tau+1}, a_\tau)\frac{p(s_{\tau+1}|a_\tau)}{p(s_\tau|a_\tau)} \tag{166}$$

$$= \delta_{s_\tau}^{r(s_{\tau+1}, a_\tau)} \frac{p(s_{\tau+1}|a_\tau)}{p(s_\tau|a_\tau)}, \tag{167}$$

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

$$S_{\tau+1} = f(S_\tau, A_\tau). \tag{168}$$

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

Q Proof of Theorem 5

This appendix uses notation introduced in App. R.

We will represent both in the larger vector space $\mathbb{R}^{|\mathcal{S}|}$ using the orthonormal basis of states $\{|s\rangle\}_{s\in\mathcal{S}}$ such that $\langle s|s'\rangle = \delta_{s,s'}$: The predictive mixed-state belief (MSB) of an action outcome sequence is

$$|\rho^P(\boldsymbol{y}_{0:t}, \boldsymbol{a}_{0:t})\rangle = \sum_{s_{t+1}} |s_{t+1}\rangle p(s_{t+1}|\boldsymbol{y}_{0:t}, \boldsymbol{a}_{0:t}), \tag{169}$$

and the retrodictive MSB is

$$\langle\rho^R(\boldsymbol{y}_{0:t}, \boldsymbol{a}_{0:t})| = \sum_{s_0} p(s_0|\boldsymbol{y}_{0:t}, \boldsymbol{a}_{0:t})\langle s_0|. \tag{170}$$

The matrix corresponding a sequence of actions $a_{0:\tau}$ and outputs $y_{0:\tau}$ has a direct probabilistic interpretation

$$T^{(y_{0:\tau}|a_{0:\tau})} \equiv \prod_{t=0}^{\tau} T^{(y_t|a_t)} \tag{171}$$

$$= \sum_{s_0, s_{\tau+1}} |s_{\tau+1}\rangle p(s_{\tau+1}, y_{0:\tau}|a_{0:\tau}, s_0)\langle s_0|, \tag{172}$$

If we define the initial diagonal state $\rho_t \equiv \sum_{s_t} |s_t\rangle p(s_t)\langle s_t|$ and assume the initial state is uncorrelated with the action sequence, then we can also calculate the probability of joint start and end state

$$T^{(y_{0:\tau}|a_{0:\tau})}\rho_0 = \sum_{s_0, s_{\tau+1}} |s_{\tau+1}\rangle p(s_{\tau+1}, s_0, y_{0:\tau}|a_{0:\tau})\langle s_0|. \tag{173}$$

Therefore, we can exactly calculate the word probability via linear algebraic expression

$$p(y_{0:\tau}|a_{0:\tau}) = \langle 1|T^{(y_{0:\tau}|a_{0:\tau})}\rho_0|1\rangle, \tag{174}$$

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

Definition 18 (Bidirectional Mixed State Matrix)**.** *The joint probability of initial and final density given the intermediate action-observation sequence determines the bidirectional mixed state matrix (BMSM)*

$$\rho(y_{0:\tau}, a_{0:\tau}) \equiv \sum_{s_0, s_{\tau+1}} |s_{\tau+1}\rangle p(s_{\tau+1}, s_0|y_{0:\tau}, a_{0:\tau})\langle s_0|. \tag{175}$$

Lemma 9. *The BMSM can be exactly calculated from the product of the linear operators of the transducer*

$$\rho(y_{0:\tau}, a_{0:\tau}) = \frac{T^{(y_{0:\tau}|a_{0:\tau})}\rho_0}{\langle 1|T^{(y_{0:\tau}|a_{0:\tau})}\rho_0|1\rangle}. \tag{176}$$

Lemma 10. *The BMSM exactly determines both the predictive and retrodictive MSBs*

$$|\rho^P(y_{0:\tau}, a_{0:\tau})\rangle = \rho(y_{0:\tau}, a_{0:\tau})|1\rangle \tag{177}$$

$$\langle\rho^R(y_{0:\tau}, a_{0:\tau})| = \langle 1|\rho(y_{0:\tau}, a_{0:\tau}). \tag{178}$$

From this we see that there are recursive relations that allow us to exactly determine the forward-time and reverse-time update steps for both. The forward-time update is to apply the transducer operator $T^{(y|a)}$ and normalize

$$\rho(y_{0:\tau+1}, a_{0:\tau+1}) = \frac{T^{(y_{\tau+1}|a_{\tau+1})}\rho(y_{0:\tau}, a_{0:\tau})}{\langle 1|T^{(y_{\tau+1}|a_{\tau+1})}\rho(y_{0:\tau}, a_{0:\tau})|1\rangle}. \tag{179}$$

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:

$$\rho(y_{-1:\tau}, a_{-1:\tau}) = \frac{\rho(y_{0:\tau}, a_{0:\tau}) \rho_0^{-1} T^{(y_{-1}|a_{-1})} \rho_{-1}}{\langle 1 | \rho(y_{0:\tau}, a_{0:\tau}) \rho_0^{-1} T^{(y_{-1}|a_{-1})} \rho_{-1} | 1 \rangle}. \tag{180}$$

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.

R Dirac Notation

For notational simplicity, we turn to quantum mechanics for a large portion of our proofs with linear algebra. This notation 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

$$|v\rangle = \sum_s v(s) |s\rangle. \tag{181}$$

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

$$\langle s'|M|s\rangle, \tag{182}$$

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

$$A = \sum_{ss'} |s'\rangle A_{ss'} \langle s|. \tag{183}$$