A Bayesian Interpretation of the Internal Model Principle

Manuel Baltieri, Araya Inc.†
Martin Biehl, Cross Labs, Cross Compass Ltd.
Matteo Capucci, University of Strathclyde and Independent Researcher
Nathaniel Virgo, University of Hertfordshire and
Earth-Life Science Institute, Institute of Science Tokyo

***Abstract*—The internal model principle, originally proposed in the theory of control of linear systems, nowadays represents a more general class of results in control theory and cybernetics. The central claim of these results is that, under suitable assumptions, if a system (a controller) can regulate against a class of external inputs (from the environment), it is because the system contains a model of the system causing these inputs, which can be used to generate signals counteracting them. Similar claims on the role of internal models appear also in cognitive science, especially in modern Bayesian treatments of cognitive agents, often suggesting that a system (a human subject, or some other agent) models its environment to adapt against disturbances and perform goal-directed behaviour. It is however unclear whether the Bayesian internal models discussed in cognitive science bear any formal relation to the internal models invoked in standard treatments of control theory. Here, we first review the internal model principle and present a precise formulation of it using concepts inspired by categorical systems theory. This leads to a formal definition of "model" generalising its use in the internal model principle. Although this notion of model is not *a priori* related to the notion of Bayesian reasoning, we show that it can be seen as a special case of *possibilistic* Bayesian filtering. This result is based on a recent line of work formalising, using Markov categories, a notion of *interpretation*, describing when a system can be interpreted as performing Bayesian filtering on an outside world in a consistent way.**

***Index Terms*—Cybernetics, Control Theory, Internal Model Principle, Interpretation Map, Bayesian Inference, Bayesian Filtering.**

I. Introduction

A classic slogan in cybernetics states that "every good regulator of a system must be a model of that system" [1]. This idea, based on the "law of requisite variety" [2] also underpins major developments in fields heavily influenced by cybernetics, such as control theory, biology, artificial intelligence and cognitive science.

In control theory specifically, the "internal model principle" (IMP) [3] refers to a general principle (a "mold", or a guide for a class of results as argued by [4], [5]) that formalises claims made in [1], [2] by defining sufficient conditions for the existence of internal models of the environment in controllers for certain classes of regulation problems.

In biology, internal models are usually invoked based on the IMP and form the basis of the modern understanding of homeostasis and (perfect) adaptation in living organisms at all scales, including microorganisms such as bacteria [6]–[8].

In artificial intelligence, internal models often appear under the name of *world models* [9]–[11], and underlie a research programme with applications to reinforcement learning, robotics and deep learning, focusing on learning how to represent hidden properties of the environment [12].

In cognitive science and neuroscience, internal models are broadly thought to constitute the computational basis of perception, motor control and high-level cognitive reasoning [13]–[16], although there is no shortage of debate about this, e.g. [17]–[20]. In the context of neuroscience, internal models are often, though by no means universally, presented under a Bayesian framework. According to the Bayesian view, brains or agents as whole systems, can be thought of as Bayesian reasoners and their cognitive processes as instances of Bayesian inference [21]–[24].

While the label "internal model," or just "model" is used across different disciplines, it is unclear whether it always refers to the same underlying formal concept. If cognitive scientists propose internal models for the study of cognition, are they referring to the same kind of mathematical objects as control theorists working with internal models for regulation problems? We do not fully answer these questions here, but take some steps towards answering them.

To do so, we structure this work in two main parts. In the first part (Section II), we present the IMP developed by [25]–[29] using concepts inspired by categorical systems theory, a mathematical formalisation of systems and their interactions based on category theory [30]–[32]. We take particular care of spelling out the assumptions that underpin work on the internal model principle, and along the way, we provide a definition of a "model". This definition is inspired by and compatible with the one found in the IMP literature for closed and autonomous systems, but also applies to systems with inputs. We then focus, in the remainder of the paper, on the case of autonomous systems, as treated by the standard IMP, and highlight one of the IMP's central aspects. Its assumptions ensure that although the controller isn't autonomous, its dynamics are effectively described by an autonomous system that we call the "attracting controller". This autonomous system is the one with a model of the environment.

In the second part, we first introduce the notion of a

† Correspondence: manuel_baltieri@araya.org

Bayesian filtering interpretation (Section III). This formally captures what it means for a system to support an interpretation as a Bayesian reasoner, and can also be considered as a formalisation of "model", albeit in a rather different sense than in the IMP. Bayesian filtering interpretations were formulated in previous work [33] in the setting of Markov categories, a recent approach to synthetic probability theory [34], [35] situated in the tradition of applied category theory, and related to ideas on process theories [36], [37] and the graphical language of string diagrams. Here we work in a particular Markov category, $\mathbf{Rel}^{+}$ (of sets and total relations), that captures *possibilistic* rather than probabilistic non-determinism.

In Section IV we then show that, for autonomous systems, whenever there is an internal model in the IMP sense, there is also a corresponding Bayesian filtering interpretation. To the best of our knowledge, this is the first result that establishes a formal connection between the concept of (internal) "model" used in the IMP, and the Bayesian inference/filtering literature. This Bayesian filtering interpretation is however of a particular kind. On the one hand it corresponds to a simplistic case of Bayesian filtering in which the system doing filtering never makes use of its inputs in any non-trivial way, meaning that although the prior changes over time in order to track the changing hidden state, the posterior for a given prior does not depend on the observation or input. Because of this, we consider Bayesian filtering interpretations to be a more general notion of model than the one used in the IMP. On the other hand, the Bayesian filtering interpretation employs a kind of approximation of the modelled system and not the modelled system directly.

II. The Internal Model Principle

The IMP appears in the control theory literature as a series of different results that include, for instance, proposals using linear [3] and nonlinear [38] systems, and a focus on geometric [3], [39] or algebraic approaches [25]. The latter are of particular interest in this work. The algebraic approach proposed by Wonham [25], later refined in [26]–[28], paves the road for a more abstract version of the IMP, capturing its core assumptions for a broad class of system, without requiring explicit assumptions regarding the geometry (or *any* geometry) of a system, see for instance [29], [40].

In this section, we provide a self-contained account and critique of [26]–[28]'s IMP. In the remainder of this work, unless otherwise stated, "IMP" will refer to the one described in this section. We start by first giving a definition of system that will be used throughout this work.

A. Systems and their maps

For the sake of simplicity, we will be working with discrete (in both time and space) dynamical systems, focusing on sets and functions. We note however that, since we adopt a structural approach to our description, most of what follows can be easily generalised to dynamical systems of other kinds (e.g. smooth dynamical systems, as in [26]–[28], or topological/stochastic dynamical systems).[1]

[1] Indeed, all open dynamical systems as above fit in a common structural description, as shown in [30], [31].

Definition II.1. A *system* (or more precisely, a *fully observable system*) X is comprised of a set X of *states*, a set I of *inputs* (or *observations*), and an *update* (or *dynamics*) function:

$$\mathsf{upd}_{\mathsf{X}} : X \times I \to X, \tag{1}$$

The pair $\binom{I}{X}$ is collectively referred to as the *interface* of the system, and we write $\mathsf{X} : \mathbf{Sys}\binom{I}{X}$ to mean X has such an interface.

It might seem odd that the interface of a system includes its state space (as well as its inputs), however this is because we assume that the state space is exposed for other systems to observe. Unless otherwise stated, we thus note that all of our uses of "system" in this work will specifically refer to the definition given above, meaning we will be dealing with fully observable systems. Some of these systems will also have trivial inputs, meaning I is a singleton $1 = \{*\}$. We'll call such systems *autonomous*.

Remark II.2. We will denote a system X in a sans-serif font, and its state space X with the same letter but in regular font. A point we want to make in this section is that it is crucial to distinguish the two.

Next we consider maps between systems. This definition takes into account the fact that a system is comprised not only of a state space, but also of an interface and dynamics on its states. So a map of systems is given by maps between their respective inputs and states which preserves their dynamics. Here and in the following we use the diagrammatic order for composition of functions i.e. if $f : X \to Y$ and $g : Y \to Z$ we write $(f \fatsemi g) : X \to Z$ for the composite function where $(f \fatsemi g)(x) \coloneqq g(f(x))$.

Definition II.3 (Map of systems)**.** Let $\mathsf{X} : \mathbf{Sys}\binom{I}{X}$ and $\mathsf{X}' : \mathbf{Sys}\binom{I'}{X'}$ be systems. A *map of systems* $f : \mathsf{X} \to \mathsf{X}'$ is comprised of two parts:

1) a *map on states*, given by a function
$$f_{\mathsf{s}} : X \to X', \tag{2}$$
2) a *map on inputs*, given by a function
$$f_{\mathsf{i}} : X \times I \to I', \tag{3}$$

such that the following diagram commutes:

$$\begin{array}{ccc} X \times I & \xrightarrow{(\pi_X \fatsemi f_{\mathsf{s}}, f_{\mathsf{i}})} & X' \times I' \\ \downarrow{\scriptstyle \mathsf{upd}_{\mathsf{X}}} & & \downarrow{\scriptstyle \mathsf{upd}_{\mathsf{X}'}} \\ X & \xrightarrow{f_{\mathsf{s}}} & X' \end{array} \tag{4}$$

meaning that, for every $x \in X, i \in I$, the following equation is satisfied:

$$f_{\mathsf{s}}(\mathsf{upd}_X(x,i)) = \mathsf{upd}_{X'}(f_{\mathsf{s}}(x), f_{\mathsf{i}}(x,i)), \tag{5}$$

It's worth commenting on the map on inputs, since one might expect it to have type $I \to I'$ rather than the more general $X \times I \to I'$. This latter choice comes from the notion of *chart* in categorical systems theory [31], and allows a more general class of maps between systems. We make use of this in Remark II.7, since the maps described there would not exist if we used the more restrictive definition. When X and X' have the same set of inputs, a map between them can be given by just specifying the map on states.

Construction II.4. In what follows, we will need to compose maps of systems, so we explain here how that happens. Given maps $f : \mathsf{X} \to \mathsf{X}'$ and $g : \mathsf{X}' \to \mathsf{X}''$, where $\mathsf{X} : \mathbf{Sys}\binom{I}{X}$, $\mathsf{X}' : \mathbf{Sys}\binom{I'}{X'}$ and $\mathsf{X}'' : \mathbf{Sys}\binom{I''}{X''}$, their composition $f \mathbin{⨟} g : \mathsf{X} \to \mathsf{X}''$ is given on states by the composition of the maps on states:

$$(f \mathbin{⨟} g)_{\mathsf{s}} = f_{\mathsf{s}} \mathbin{⨟} g_{\mathsf{s}}, \tag{6}$$

while on inputs $(f \mathbin{⨟} g)_{\mathsf{i}} : X \times I \to I''$ is defined as follows:

$$(f \mathbin{⨟} g)_{\mathsf{i}}(x, i) = g_{\mathsf{i}}(f_{\mathsf{s}}(x), f_{\mathsf{i}}(x, i)). \tag{7}$$

The notions of subsystem of a system and attracting subsystem will be crucial to define the regulation condition for the IMP, hence we introduce them here:

Definition II.5 (Subsystem)**.** A *subsystem* of X is a forward-invariant subset of X together with updates restricted to the subset, thus a map of systems $\gamma : \mathsf{X}' \to \mathsf{X}$ given on states by an inclusion $\gamma_{\mathsf{s}} : X' \to X$ and on inputs by projection $\gamma_{\mathsf{i}} = \pi_I : X \times I \to I$.

Forward-invariant here means that if the state is in X' then the next state will also be in X', no matter what observation is received. The dynamics of a subsystem are induced by the dynamics of its parent system. In fact one can say that a subsystem of X is given by a subset $X' \subseteq X$ such that $\mathsf{upd}_{\mathsf{X}}(x', i) \in X'$ for all $x' \in X'$ and $i \in I$. The update function of the subsystem is then just $\mathsf{upd}_{\mathsf{X}}$ restricted to $X' \times I$.

Note that the term "subsystem" is often used to mean a component, or a part, of a system. This is not the sense in which we use the term here. For example, the controller is not a subsystem of the full system, as defined below.

Definition II.6 (Attracting subsystem)**.** An *attracting subsystem* for X is a *non-empty* subsystem $\mathsf{X}^* \to \mathsf{X}$ such that, for each $x \in X$, there exists $n \in \mathbb{N}$ such that $\mathsf{upd}^t_{\mathsf{X}}(x, \bar{i}) \in X^*$ for all $t \geq n$ and $\bar{i} \in I^t$.

Notice that we do not assume X^* is the smallest subsystem of X meeting this criterion. In part because of this, our definition is rather more general than the usual definition of attractor. For example, X^* could be the basin of attraction of a fixed point, or it could contain multiple distinct orbits.

B. The IMP assumptions

Having fixed what we mean by "systems", let us define the systems of interest for the IMP. The following is a typical control theoretic setup, and is used in particular in the work by Hepburn and Wonham [26]–[28], where it is presented a bit differently. Throughout this section we introduce the assumptions used by Hepburn and Wonham to derive the internal model principle. We note in advance that one of them, Assumption 4, seems to us rather difficult to motivate. We begin with the following:

Assumption 1 (Environment, plant, controller)**.** *The following three components are so defined:*

1) *the environment* $\mathsf{E} : \mathbf{Sys}\binom{1}{E}$ *is an autonomous system*
$$\mathsf{upd}_{\mathsf{E}} : E \to E, \tag{8}$$
2) *the plant* $\mathsf{P} : \mathbf{Sys}\binom{E \times C}{P}$ *is a system*
$$\mathsf{upd}_{\mathsf{P}} : P \times E \times C \to P, \tag{9}$$
3) *the controller* $\mathsf{C} : \mathbf{Sys}\binom{P}{C}$ *is a system*
$$\mathsf{upd}_{\mathsf{C}} : C \times P \to C. \tag{10}$$

The full system $\mathsf{S} : \mathbf{Sys}\binom{1}{E \times P \times C}$ *is the following composite autonomous system:*

$$\begin{aligned} \mathsf{upd}_{\mathsf{S}} : E \times P \times C &\longrightarrow E \times P \times C \\ (s_{\mathsf{E}}, s_{\mathsf{P}}, s_{\mathsf{C}}) &\longmapsto (\mathsf{upd}_{\mathsf{E}}(s_{\mathsf{E}}), \mathsf{upd}_{\mathsf{P}}(s_{\mathsf{P}}, s_{\mathsf{E}}, s_{\mathsf{C}}), \\ &\qquad \mathsf{upd}_{\mathsf{C}}(s_{\mathsf{C}}, s_{\mathsf{P}})). \end{aligned} \tag{11}$$

Let $S = E \times P \times C$ *denote the state space of the full system* S.

We next look at the maps of systems that can be defined between these components thanks to the definition of maps of inputs in Definition II.3.

Remark II.7. There are maps of systems $\pi_{\mathsf{E}} : \mathsf{S} \to \mathsf{E}$, $\pi_{\mathsf{P}} : \mathsf{S} \to \mathsf{P}$ and $\pi_{\mathsf{C}} : \mathsf{S} \to \mathsf{C}$ induced by projecting out states of S:

$$\begin{array}{ccc} S \times 1 & \xrightarrow{(\pi_E, \mathsf{id}_1)} & E \times 1 \\ {\scriptstyle \mathsf{upd}_{\mathsf{S}}}\downarrow & & \downarrow{\scriptstyle \mathsf{upd}_{\mathsf{E}}} \\ S & \xrightarrow[\pi_E]{} & E \end{array} \qquad \begin{array}{ccc} S \times 1 & \xrightarrow{(\pi_C, \pi_P)} & C \times P \\ {\scriptstyle \mathsf{upd}_{\mathsf{S}}}\downarrow & & \downarrow{\scriptstyle \mathsf{upd}_{\mathsf{C}}} \\ S & \xrightarrow[\pi_C]{} & C \end{array}$$
$$\begin{array}{ccc} S & \xrightarrow{(\pi_P, \pi_{E \times C})} & P \times E \times C \\ {\scriptstyle \mathsf{upd}_{\mathsf{S}}}\downarrow & & \downarrow{\scriptstyle \mathsf{upd}_{\mathsf{P}}} \\ S & \xrightarrow[\pi_P]{} & P \end{array} \tag{12}$$

In this setting, the environment's role is to produce undesirable signals, or disturbances, the controller and plant must adapt to. The plant-controller subsystem has no control over it, since E is an autonomous system, i.e. it does not receive inputs from controller, plant or any other system. In [25] Wonham calls the environment a "convenient fiction" which can be used as a placeholder for disturbances that C is designed to compensate. We will comment more on this point later.

The controller's job is to *regulate* plant and environment, according to some notion of regulation. Formally, we define:

Definition II.8 (Regulation problem)**.** A *regulation problem* (or *reguland* [1]) is a triple $(\mathsf{E}, \mathsf{P}, \tilde{K} \subseteq E \times P)$.

$\tilde{K}$ is a set of *target states* (also known as goals, or references). This set specifies what the controller is supposed to achieve. We define it as a subset of $E \times P$ rather than just P, which means that the desired states of the plant can also depend on the state of the environment. We then define the set $K = C \times \tilde{K}$ of states of the complete system in which the plant-environment components are in a target state. By slight abuse of terminology, we also refer to the elements of K as target states. A regulation problem is thus the complete collection of *things* we need to formulate the IMP. From now on, the rest of the assumptions will be *properties* we impose on these things.

We should note at point that the dynamics of S might exit and enter target states K. Thus, in order to say that C actually regulates S, we need to make assumptions regarding the way K fits in the dynamics of S:

Assumption 2 (Regulation condition)**.** *C solves its regulation problem, meaning there exists an attracting subsystem (Definition II.6) $\mathsf{S}^* \to \mathsf{S}$ such that, on states, $S^* \subseteq K$.*

The idea of this assumption is that no matter what state the whole system starts in, there eventually will be a time at which the environment-plant system is in K *and* it will remain there indefinitely. The only non-trivial condition imposed by Assumption 2 is that S^* is non-empty: in fact there is always a maximal subsystem of S contained in K (just take the union of all such subsystems), but it might possibly be empty. In particular, C might be *small* compared to the rest of S. Therefore, the next assumptions buy some extra structure so that we can ultimately find a way to compare C and E.

Looking again at Assumption 2 we see that, even though K contains all states of C, this might not be true anymore for S^*. We can define the set of *attracting control states* as

$$C^* := \{s_{\mathsf{C}} \in C \mid \exists s_{\mathsf{E}} \in E, s_{\mathsf{P}} \in P, (s_{\mathsf{E}}, s_{\mathsf{P}}, s_{\mathsf{C}}) \in S^*\}.$$

It is obtained along with a surjection $\pi_{C^*} : S^* \to C^*$ by taking the image of S^* under the projection map $\pi_C : S \to C$:

$$\begin{array}{ccc} S^* & \xrightarrow{\pi_{C^*}} & C^* \\ \downarrow & & \downarrow \\ S & \xrightarrow[\pi_C]{} & C \end{array} \tag{13}$$

We would like to show next that there is a system with state-space C^* that tracks E (or a subsystem of E whose states are contained in K), but (1) C^* isn't necessarily forward-invariant as a subset of C and (2) E is autonomous, while C isn't.

To resolve this issue, following [25]–[28], we assume that the controller C operates in a particular way. When the system is outside the target states, the controller is in a *closed loop* regime in which its evolution depends on the state of the plant, by virtue of the inputs received from it. It is however allowed to switch to an *open loop* strategy, and in particular when the system is in a target state, we assume that the controller operates purely in an open loop setting, in which its outputs and state changes don't depend on the state of the plant. Plenty of control strategies fall under this assumption, namely those that employ *error feedback*: the controller measures how far the plant-environment system is from a target state, and counteracts accordingly, so that when inside the target state the controller essentially operates in an open loop regime. Still, it's important to remember that, while this assumption plays an important role in certain areas of control theory, for instance in the "disturbance decoupling problem" [41], and cognitive science, where it corresponds to the idea of "decouplability" or "detachability" [42], it rules out controllers that use closed loop feedback to maintain a target regime; plenty of these also exist.

Assumption 3 (Error feedback structure)**.** *C^* supports an autonomous dynamics $\mathsf{upd}_{\mathsf{C}^*} : C^* \to C^*$, thus defining a system $\mathsf{C}^* : \mathbf{Sys}\binom{1}{C^*}$, that we call attracting controller, and making π_{C^*} a full-fledged map of systems:*

$$\pi_{\mathsf{C}^*} : \mathsf{S}^* \longrightarrow \mathsf{C}^* \tag{14}$$

We now give the following definition which we believe to be novel, though in spirit very much like that proposed in [28] as well as in [1].

Definition II.9 (Model)**.** A *model* of a system $\mathsf{X} \in \mathbf{Sys}\binom{I}{X}$ is:

- a system $\mathsf{M} \in \mathbf{Sys}\binom{J}{M}$ (the *archetype*), and
- a map of systems (the *model per se*)

$$\mathsf{X} \xrightarrow{\mu} \mathsf{M} \tag{15}$$

such that

1) its part on states $\mu_{\mathsf{s}} : X \to M$ is surjective, and
2) its part on inputs $\mu_{\mathsf{i}}(x, -) : I \to J$ is surjective for each $x \in X$.

Often we will just say "M models X", leaving μ implicit. The idea of such a definition is that the map μ forgets some of the complexity of the system X, which is mapped (surjectively) into the simpler system M. It is also connected to "coarse-grainings" [43], "variable aggregation" [44], "state aggregation" [45], "lumpability" [46], "model reduction" [47], "dynamical consistency" [48] or other similar concepts, and reflected in standard ideas of "supervisory control" [40], for readers already familiar with any of these. Importantly however, this is not the same as any of the standard definitions mentioned above since the map on inputs corresponds to a chart [31] for fully observable systems, as alluded to in the definition of maps of systems (Definition II.3): it is of the form $X \times I \to J$ and not simply $I \to J$.

Having a model $\mu : \mathsf{X} \to \mathsf{M}$ means that for each state $m \in M$ we have a set of states $\mu^{-1}(m) \in X$, called the *fibre* of m and which represents a subset of elements of X that are *indistinguishable* from the perspective of the simpler system M as they all map to the same element $m \in M$ via the surjective function μ_{s}. As $m \in M$ varies along the function $\mathsf{upd}_{\mathsf{M}}$, this variation is consistent with the variation described by the function $\mathsf{upd}_{\mathsf{X}}$ for each element x of the fibre $\mu^{-1}(m)$ of m. This will be further unpacked in Construction IV.1.

Remark II.10. When applied to autonomous systems, a model reduces to the definition implicit in [27], [28]. It is implicit there because it only appears in cases where X is either the attracting full system S^*, or the attracting environment E^* defined below. It is also mixed with other assumptions, particularly Assumption 4, which we believe are not strictly necessary (and potentially problematic) for a notion of model. Our definition is also in terms of sets as in [29], [40] (rather than in terms of manifolds as in [27], [28]), and corresponds to a surjective map of states commuting with the dynamics:

$$\begin{array}{ccc} X & \xrightarrow{\mu_{\mathsf{s}}} & M \\ {\scriptstyle \mathsf{upd}_{\mathsf{X}}}\downarrow & & \downarrow{\scriptstyle \mathsf{upd}_{\mathsf{M}}} \\ X & \xrightarrow[\mu_{\mathsf{s}}]{} & M \end{array} \tag{16}$$

This is because the map on inputs of a model is necessarily of the form $X \times 1 \to 1$ for autonomous systems, due to the surjectivity condition. Another consequence of this is that systems that model autonomous systems must also be autonomous.

Remark II.11. Like any good definition, Definition II.9 admits a trivial instance. A *trivial model* is one where $\mu : X \to M$ is a product projection, which means there exists a system $\mathsf{F} \in \mathbf{Sys}\binom{H}{F}$ such that the state space and inputs of the system X factor as $X = M \times F$ and $I = J \times H$ respectively, *and* its dynamics decompose in that of M and F:

$$\mathsf{upd}_{\mathsf{X}}((m,f),(j,h)) = (\mathsf{upd}_{\mathsf{M}}(m,j),\, \mathsf{upd}_{\mathsf{F}}(f,h)). \tag{17}$$

Thus, in a trivial model, knowledge of M does not afford knowledge about the rest of X, because M and F are uncoupled.

We stress that what makes a trivial model *trivial* is Eq. (17), rather than μ_{s} and, fibrewise, μ_{i}, being product projections. As an instructive and relevant example of this fact, observe that $\pi_{\mathsf{E}} : \mathsf{S} \to \mathsf{E}$ from Remark II.7 satisfies the definition of model and it is a product projection *on states*, but it is *not* a trivial model. It thus means that *we could meaningfully interpret* E as describing a coarse-grained version of S: even if a state of E always represents all possible states of the plant-controller component (so it is hardly an informative belief), the dynamics is non-trivial since knowing the state of E does constrain the possible evolution of S. In particular, π_{E} does not satisfy Eq. (17) unless P does not make use of its input from the environment.

To obtain what Hepburn and Wonham called the IMP, we need another assumption. It is however quite strong, and hard to motivate in practice as far as we can tell. To state it, we first introduce the following lemma.

Lemma II.12. *Let $\gamma : \mathsf{X}^* \to \mathsf{X}$ be an attracting system for X, and let $p : \mathsf{X} \to \mathsf{Y}$ be a map of autonomous systems surjective on states. Then the image of X^* under p defines an attracting system for Y.*

Proof. Let $y \in \mathsf{Y}$. Since the map on states p_{s} is surjective, there is an $x \in X$ such that $p_{\mathsf{s}}(x) = y$. By assumption, X^* is an attracting system so there exists an $n(x) \in \mathbb{N}$ such that $\mathsf{upd}_{\mathsf{X}}^m(x) \in \mathsf{X}^*$ for all $m \geq n(x)$. Now since p_{s} commutes with $\mathsf{upd}_{\mathsf{X}}$ and $\mathsf{upd}_{\mathsf{Y}}$, $p_{\mathsf{s}}(\mathsf{X}^*) \ni p_{\mathsf{s}}(\mathsf{upd}_{\mathsf{X}}^m(x)) = \mathsf{upd}_{\mathsf{Y}}^m(p_{\mathsf{s}}(x)) = \mathsf{upd}_{\mathsf{Y}}^m(y)$, thus proving that $p(\mathsf{X}^*)$ is an attracting system for Y. □

The projection $\pi_{\mathsf{E}} : \mathsf{S} \to \mathsf{E}$ described in Remark II.7 satisfies the hypotheses of Lemma II.12, meaning the attracting full system S^* induces an analogous global attractor E^* in E (again, obtained by restricting E to those states which are part of at least one state of S^*). We will call the system E^* the *attracting environment*, by analogy with the attracting controller in Assumption 3. The next assumption is arguably the core of the IMP as described in [26]–[28].

Assumption 4. *There is an isomorphism of systems $\mathsf{S}^* \cong \mathsf{E}^*$, meaning that for each environment state $s_{\mathsf{E}} \in E^*$, there is* exactly one *$s \in S^*$ such that $\pi_E(s) = s_{\mathsf{E}}$.*

This assumption can be used to obtain a map of systems between the attracting controller C^* and the attracting environment E^*, which leads to a complete reformulation of [28]:

Theorem II.13 (*Internal Model Principle (attracting environment E^*)*)**.** *Let S be a system subject to Assumptions 1 to 4. Then C^* models the attracting environment E^* via the dashed map below:*[2]

$$\begin{array}{ccccc} \mathsf{E}^* & & \overset{\nu}{\dashrightarrow} & & \mathsf{C}^* \\ & \underset{\text{Assumption 4}}{\searrow^{\sim}} & & \underset{\text{Assumption 3}}{\nearrow^{\pi_{\mathsf{C}^*}}} & \\ & & \mathsf{S}^* & & \end{array} \tag{18}$$

In the next section, we will link this notion of model to that of Bayesian interpretation put forward by [33], [49]. This is perhaps surprising, because models in Bayesian statistics and models according to definition Definition II.9 are quite different things and it's not a priori obvious that they would be connected at all.

III. Bayesian updates and interpretations

The notion of interpretation map was introduced in [33] as a map between the state of a system and probability distributions on the states of the external world describing the former's beliefs about the latter. The concept has since then been developed further in a category theory context in [50]. The goal was to understand what it means to *interpret* a physical system as performing Bayesian inference; that is, what properties must a physical system have in order to be able to make the claim that it *has a Bayesian prior* over some hidden variable? This question, and the approach taken to it, is similar in spirit to [51], with the main difference being the focus on Bayesian reasoning rather than computation.

This notion has also been applied to partially observable Markov decision processes [49], sharing some mathematical background with the state-space control theoretic approach behind the internal model principle. In this section we briefly summarise some of the results in [33], [49] that will help us later establish connections between the internal model principle and Bayesian reasoning ideas.

A. Markov categories

Markov categories are a synthetic approach to probability theory and formalise the compositional structure of non-deterministic processes that behave like Markov kernels. The canonical reference for Markov categories is [35]. We begin with the definition of a category.

Definition III.1 (Category)**.** A category $\mathbf{C}$ consists of:

[2] Such a composite is defined by Construction II.4.

- a class of *objects*, $\mathrm{ob}(\mathbf{C})$, e.g. $A, B, C, \ldots$,
- a class of *maps*, *arrows* or *morphisms*, $\mathrm{arrow}(\mathbf{C})$ (these three terms are interchangeable),
- for each arrow in $\mathrm{arrow}(\mathbf{C})$, a *source* and a *target*, which are objects, i.e. elements of $\mathrm{ob}(\mathbf{C})$, if an arrow f has source A and target B then we often write it as $f : A \to B$, and we say that $A \to B$ is the arrow's *type*,
- for each object of $\mathrm{ob}(\mathbf{C})$, an *identity morphism* $\mathsf{id}_A : A \to A$,
- a binary operation $⨟$ on arrows called the *composition rule*, such that given morphisms $f : A \to B$ and $g : B \to C$, their composite $f ⨟ g$ is an arrow with type $A \to C$; composition is defined when (and only when) the target of one arrow equals the source of another, and must obey the following laws:
  - *associativity*: given morphisms $f : A \to B$, $g : B \to C$ and $h : C \to D$, we must have $f ⨟ (g ⨟ h) = (f ⨟ g) ⨟ h$,
  - left and right *unit laws*: for every pair of objects A, B and morphism $f : A \to B$, we must have $\mathsf{id}_A ⨟ f = f = f ⨟ \mathsf{id}_B$.

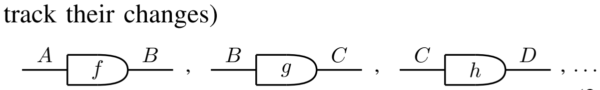
Next we introduce a graphical notation, string diagrams, for a special class of categories with extra structure, known as symmetric monoidal categories, that will be used throughout this work. We give an abridged definition that gives enough information to understand the rest of the work, skipping over some important details (the so-called coherence laws) that are needed for the graphical language to be formally valid [52]. For a concise formal treatment of symmetric monoidal categories and their graphical language, a good reference is [53]; see also [54] for more on string diagrams and [37] for a much more comprehensive beginner-friendly introduction.

Definition III.2 (Symmetric monoidal category)**.** A *symmetric monoidal category* (also known as a *process theory* [37]) is a category $\mathbf{C}$ with the following additional structure:

- objects as wires, or strings, $A, B, C, \ldots$

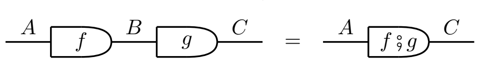
(19)

- morphisms as boxes, $f : A \to B$, $g : B \to C$, $h : C \to D, \ldots$, converting objects to new objects and forming processes (combinations of objects and morphisms to track their changes)



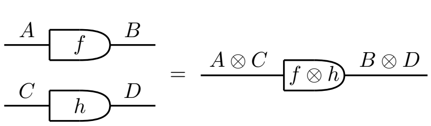
(20)

- composition appears in two forms, sequential and parallel; sequential composition, denoted by $⨟$, corresponds to the connection of wires with the appropriate type (A's with A's, B's with B's, etc.) such that



(21)

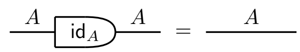
while parallel composition, denoted by $\otimes$, is simply depicted as



(22)

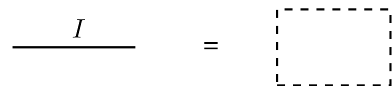
The composition operations must obey:

- associativity, for sequential composition given by (equalities) $f ⨟ (g ⨟ h) = (f ⨟ g) ⨟ h = f ⨟ g ⨟ h$ and for parallel composition by (isomorphism) $f \otimes (g \otimes h) \cong (f \otimes g) \otimes h \cong f \otimes g \otimes h$,
- identity, for sequential composition in the form of an identity box (usually not drawn, as on the right hand side below)



(23)

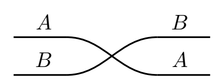
and for parallel composition in the form of a distinguished wire I, often drawn as no wire,



(24)

such that for every object $A \in \mathbf{C}$, $I \otimes A \cong A \cong A \otimes I$, i.e. in string diagrammatic terms, any wire can be seen as having infinitely many not-drawn identity wires in parallel,

- and symmetry, in the form of a morphism $\sigma_{A,B} : A \otimes B \to B \otimes A$, drawn as two wires crossing:

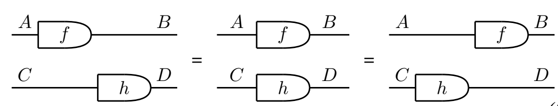


(25)

such that $\sigma_{A,B} ⨟ \sigma_{B,A} = \mathrm{id}_{A \otimes B}$.

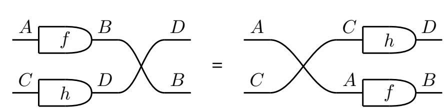
The isomorphisms witnessing associativity, unitality and symmetry satisfy coherence laws [52, Definition 6.1.1], as mentioned before, will not be of concern here. We will however state the following properties in the graphical notation, as they will be regularly used in string diagrams manipulations later on:

- interchange law



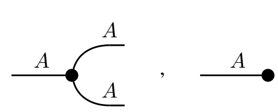
(26)

- naturality of the swap map



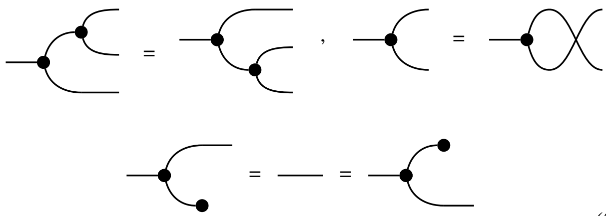
(27)

Definition III.3 (Markov category)**.** Markov categories (or affine cd-categories [34]) are symmetric monoidal categories where every object is equipped with extra structure that reproduces, synthetically, the algebra of *non-deterministic* processes resembling (normalised) Markov kernels. This structure includes *copy* and *delete* operations doing what their names explicitly suggest: the former creates two copies of its input, while the latter deletes its input:



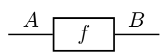
(28)

Copy and delete obey the following laws (for readability, strings aren't labelled here, but imagine them all having the same label):



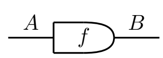
(29)

For extra clarity, we adopt the convention introduced by [33] to denote deterministic maps as boxes:



(30)

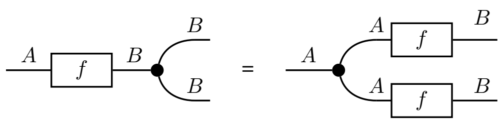
and keep general (possibly non-deterministic) ones as boxes with a curved right edge:



(31)

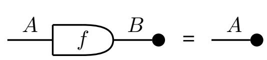
In text, we write the type of a deterministic morphism as $f : A \to B$, whereas for a general (possibly non-deterministic) morphism we use the special notation $f : A \twoheadrightarrow B$.

The defining property of deterministic maps in Markov categories is



(32)

which is sometimes called the *naturality of copy*. Naturality doesn't apply to non-deterministic maps, intuitively because copying the result of a die roll is not the same as rolling two dice. On the other hand, all maps in a Markov category satisfy the normalisation law, or *naturality of delete*, stating that mapping an input to an output and then deleting the output, is the same as deleting the input:



(33)

For further technical details, and discussions, we refer the reader to [35], [55].

Example III.4 (**FinStoch** [35])**.** *For a given set Y, denote by $D(Y)$ the set of all finitely-supported probability distributions over it. There is a Markov category* **FinStoch** *whose objects are finite sets and whose morphisms from X to Y, $X \twoheadrightarrow Y$, called Markov kernels, are defined as functions $X \to D(Y)$. For a Markov kernel $f : X \twoheadrightarrow Y$ we write $f(y \mid x)$ for $f(x)(y)$, the probability assigned to y when the kernel is given x as input.*

Sequential composition of Markov kernels $f : X \twoheadrightarrow Y$, $g : Y \twoheadrightarrow Z$ is given by the Markov kernel so defined (i.e. the Chapman–Kolmogorov equation):

$$\begin{aligned} f \fatsemi g &: X \twoheadrightarrow Z \\ (f \fatsemi g)(z \mid x) &:= \sum_{y \in Y} g(z \mid y) f(y \mid x) \end{aligned} \tag{34}$$

while parallel composition is, on objects, the Cartesian product of sets $X \otimes Y := X \times Y$, and for morphisms $f : X \twoheadrightarrow Y$ and $f' : M \twoheadrightarrow Y'$ given by the Markov kernel so defined:

$$\begin{aligned} f \otimes f' &: X \times X' \twoheadrightarrow Y \times Y' \\ (f \otimes f')(y, y' \mid x, x') &:= f(y \mid x) f'(y' \mid x'). \end{aligned} \tag{35}$$

The unit for parallel composition is the singleton $I = 1 = \{*\}$. The copy and delete morphisms for a set X are given, respectively, by the Markov kernel $\Delta_X : X \to X \times X$ mapping $x \in X$ to the Dirac distribution $\delta_{(x,x)}$, and the unique map $\mathsf{del}_X : X \to 1$, mapping $x \in X$ to the only normalised probability distribution that exists over $\{*\}$.

Deterministic morphisms in this Markov category are those that are indeed deterministic, i.e. they map every element of their domain to a Dirac distribution. In other words, the deterministic maps of **FinStoch** *are all and only the usual set-theoretic functions, see [35] for details.*

Example III.5 (**Rel**$^+$ (or **SetMulti**) [35], [56])**.** *For a set Y, denote by $P^+(Y)$ the set of all non-empty subsets of Y. There is a Markov category* **Rel**$^+$ *whose objects are sets and whose morphisms from X to Y are functions $X \to P^+(Y)$, corresponding to left-total relations. Sequential composition of left-total relations $f : X \twoheadrightarrow Y$, $g : Y \twoheadrightarrow Z$ is given by the left-total relation so defined:*

$$\begin{aligned} f \fatsemi g : X &\longrightarrow P^+(Z) \\ x &\longmapsto \{z \in Z \mid \exists y \in Y, y \in f(x) \text{ and } z \in g(y)\}, \end{aligned} \tag{36}$$

while parallel composition is, on objects, the Cartesian product of sets $X \otimes Y := X \times Y$, and for morphisms $f : X \twoheadrightarrow Y$ and $f' : M \twoheadrightarrow Y'$ given by the left-total relation so defined:

$$\begin{aligned} f \otimes f' : X \times M &\longrightarrow P^+(Y \times Y') \\ (x, x') &\longmapsto \{(y, y') \in Y \times Y' \mid \\ &\qquad y \in f(x) \text{ and } y' \in f'(x')\}. \end{aligned} \tag{37}$$

The unit for parallel composition is the singleton $I = 1 = \{*\}$.

The copy and delete morphisms for a set X are given, respectively, by the left-total relation $\Delta_X : X \to X \times X$ mapping $x \in X$ to the singleton $\{(x, x)\}$, and the unique map $\mathsf{del}_X : X \to P^+(1)$.

Deterministic morphisms in this Markov category are those that map every element of their domain to a singleton. In other words, the deterministic maps of **Rel**$^+$ *are all and only the usual set-theoretic functions [35].*

The morphisms in **Rel**$^+$ can be seen as *possibilistic* Markov kernels. For each element in the codomain we specify a set of possible elements of the domain that it might map to, but we don't assign probabilities to these possibilities. Possibilistic representations of uncertainty are less common than probabilistic ones. Nonetheless they are well established and have a long history in different fields, including control theory [57]–[61], artificial intelligence [62] and automata theory [63].

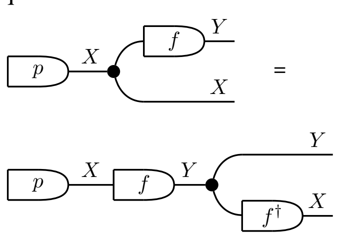
B. Bayesian reasoning and interpretations in Markov categories

We now introduce the main ideas involved in the notion of Bayesian interpretations. To aid intuition, we refer to maps of type $p : I \rightsquigarrow X$ for any object X as *distributions* and to deterministic maps $x : I \to X$ of that type as *elements* of X. We also consider maps $\psi : \Theta \rightsquigarrow X$ to correspond to *parametrised families of distributions* (also called "channels" in [34], [64]), with each element $\theta : I \to \Theta$ a parameter determining a distribution $\psi(X \mid \theta)$ or $\psi(\theta)$ over X.

Furthermore, given a map $f : X \rightsquigarrow Y$ and an element $x : I \to Y$ we call $x \fatsemi f : I \rightsquigarrow Y$ the distribution over Y assigned to the element $x \in X$ by f. We write it as $f(x)$ or $f(Y \mid x)$. In **FinStoch** this terminology and the notation coincide with its common usage in probability theory.

Next we recall the definitions of Bayesian inversion and conjugate prior for Markov categories with conditionals given, e.g. in [35], [64]. We then generalise these two concepts, which brings us closer to the definition of a Bayesian interpretation.

Definition III.6 (Bayesian inversion)**.** In a Markov category, a Bayesian inversion [34] of a map $f : X \rightsquigarrow Y$ with respect to a distribution $p : I \rightsquigarrow X$ is a map $f^\dagger : Y \rightsquigarrow X$ satisfying the following equation:



(38)

In a general Markov category with f and p as above, a Bayesian inverse $f^\dagger$ is not guaranteed to exist. However, many Markov of interest have a property known as "having (all) conditionals" [35, Definition 11.5], which guarantees that they always exist, for every f and p. This is the case in both **FinStoch** and **Rel**$^+$. However, although they always exist in our categories of interest, Bayesian inverses are not generally unique, meaning that for given p and f there might be multiple morphisms $f^\dagger$ satisfying Eq. (38). We discuss the reasons for this shortly.

To get a more concrete feeling for Bayesian inverses, we can examine the form Eq. (38) takes when our Markov category is **FinStoch**, where it corresponds to the equation

$$\begin{aligned} p(x)f(y \mid x) &= f(y \mid p(x))f^\dagger(x \mid y) \\ &= \sum_{x' \in X} p(x')f(y \mid x')f^\dagger(x \mid y), \end{aligned} \tag{39}$$

from which we can derive the standard Bayes rule by dividing by $f(y \mid p(x))$, assuming it's positive ($f(y \mid p(x)) > 0$):

$$\begin{aligned} f^\dagger(x \mid y) &= \frac{p(x)f(y \mid x)}{f(y \mid p(x))} \\ &= \frac{p(x)f(y \mid x)}{\sum_{x' \in X} p(x')f(y \mid x')}. \end{aligned} \tag{40}$$

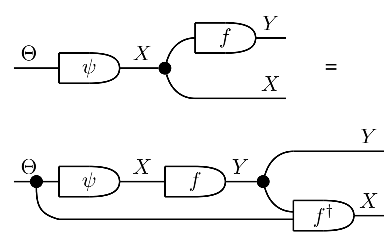
As mentioned, the Bayesian inverse $f^\dagger$ is generally not unique. In **FinStoch** this is because Eq. (39) doesn't constrain the value of $f^\dagger(x \mid y)$ in cases where $f(y \mid p(x)) = 0$.

Bayesian inversions can be used to inspire a notion of *updating* of distributions over a (hidden) variable in response to observations [65]. For this, we consider Y as observations generated from hidden variable X by $f : X \rightsquigarrow Y$ (often called a *statistical model* in a probabilistic context, including for instance in **FinStoch**), the distribution $p : I \rightsquigarrow X$ above as a *prior* distribution before an observation, and the map $f^\dagger : Y \rightsquigarrow X$ as assigning new, *posterior* distributions to the observations Y generated via f. In the case of **FinStoch**, the map $f^\dagger$ can be seen as multiplying the prior by the *likelihood* (i.e. the values of $f(y \mid x)$ for the given *data* y) and then dividing by the *evidence* $\sum_{x' \in X} p(x')f(y \mid x')$ to obtain the posterior.

It is common to refer to distributions that are updated according to this process as *(Bayesian) beliefs*, and to the process itself as *(Bayesian) belief updating*.

In some cases, in place of the prior $p : I \rightsquigarrow X$ we might want a *parametrised family* of priors $\psi : \Theta \rightsquigarrow X$. The idea is that the prior is assumed to come from some family of distributions, say for instance Gaussians, and so we consider one prior for each value of the parameters Θ. In this case, the Bayesian inverse $f^\dagger$ must depend on the parameter as well, because in general the Bayesian inverse depends on the prior. This gives rise to the following definition.

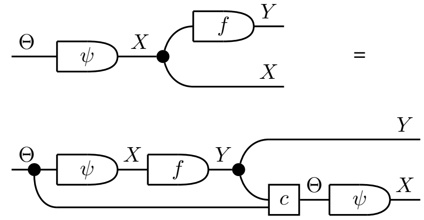
Definition III.7. In a Markov category, we say that a map $f^\dagger : Y \otimes \Theta \rightsquigarrow X$ is a Bayesian inversion of a parametrised family $f : \Theta \rightsquigarrow X$ if



(41)

The next concept, called a *conjugate prior*, expresses an important property that such a parametrised Bayesian inverse might have, namely that it factors through the original family of distributions ψ. This is of particular importance to us because the map c below can be seen as explicitly implementing the belief update, by updating the parameters such that the prior's parameter is updated to the posterior's parameter. Conjugate priors were first expressed in string diagrams in this form in [64].

Definition III.8 (Conjugate prior for Bayesian inference)**.** We call ψ a conjugate prior [64] to $f : X \rightsquigarrow Y$ if there exists a deterministic map $c : Y \otimes \Theta \to \Theta$ that satisfies the following:



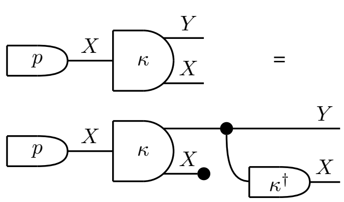
(42)

Note that the map $c \fatsemi \psi$ is a Bayesian inverse of f with respect to ψ, in the sense of Definition III.7. In Bayesian statistics, $\theta \in \Theta$ is sometimes called a *hyperparameter* and $x \in X$ a *parameter*, but we will generally avoid this usage.

If we have a conjugate prior $\psi : \Theta \to X$ to $f : X \to Y$, then every parameter $\theta : I \to \Theta$ determines a Bayesian inversion $\psi(c(-, \theta)) : Y \to X$ of $f : X \to Y$ for $p(\theta)$ [64, Theorem 6.3]. The role of c is to turn parameters of prior distributions into parameters of posterior distributions by taking into account observations generated by f according to the prior ψ. In this way, the map c implements belief updating. The usual informal notion of a prior being conjugate to f if the posterior distribution is in the same family as the prior's is captured by having the map ψ appear twice: once when mapping the initial parameter to the prior and again when mapping the updated parameter to the posterior.

The version of belief updating introduced so far can be seen as a version of *Bayesian inference*, which means that the updated beliefs are about a constant hidden variable. To see this, note that X gets copied, but never gets changed in Eqs. (38), (41) and (42).

We need then a form of belief updating that corresponds to *Bayesian filtering*, where the hidden variable also changes. For this we can replace $\Delta_X \mathbin{⨾} f \otimes \mathsf{id}_X : X \to Y \otimes X$ with a map $\kappa : X \to X \otimes Y$ that produces an observation Y and may change X instead of just copying it. The analogue of the Bayesian inversion (Definition III.6) with respect to a distribution $p : I \to X$ is thus a map $\kappa^\dagger : Y \to X$ satisfying



(43)

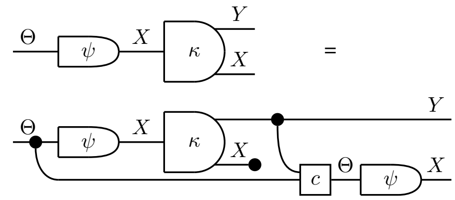
As with ordinary Bayesian inversions (Definition III.6), the map $\kappa^\dagger$ always exists in any Markov category with conditionals, including thus **FinStoch** and $\mathbf{Rel}^+$. Note that setting $\kappa = \Delta_X \mathbin{⨾} f \otimes \mathsf{id}_X : X \to Y \otimes X$ recovers Definition III.6.

In Bayesian filtering, the idea is that κ updates X and generates an observation Y. Similarly to the Bayesian inference inversion, we can view p as a prior distribution before the update of X and observation of Y, and $\kappa^\dagger : Y \to X$ as assigning posterior distributions over the updated X to the observations Y generated via κ. In **FinStoch**, this corresponds to the following equation:

$$\kappa^\dagger(x \mid y) = \frac{\sum_{x' \in X} p(x')\kappa(y, x \mid x')}{\sum_{x', x'' \in X} p(x')\kappa(y, x'' \mid x')}. \tag{44}$$

The analogue to a conjugate prior to $\kappa : X \to X \otimes Y$ in the filtering case is then the following [33].

Definition III.9 (Conjugate prior for Bayesian filtering)**.** We call ψ a conjugate prior for Bayesian filtering to $\kappa : X \to X \otimes Y$ if there exists a deterministic map $c : Y \otimes \Theta \to \Theta$ that satisfies the following:



(45)

The map $c : Y \otimes \Theta \to \Theta$ maps parameters of prior distributions over X to parameters of posterior distributions over *updated* X (cf. Definition III.8 where X is instead fixed), while taking into account observations generated by κ. In a probabilistic context (including in **FinStoch**), the map κ where X is updated is often called a *hidden Markov model* (or a *discrete state-space model*), as opposed to the statistical model f where X is fixed.

Reference [33] inverts this account, by starting with a map $c : Y \otimes \Theta \to \Theta$ and asking whether ψ and κ exist such that Eq. (45) holds. This leads to the following definition.

Definition III.10 (Bayesian filtering interpretation [33])**.** Given a deterministic map $c : Y \otimes \Theta \to \Theta$, a *Bayesian filtering interpretation* of c consists of a map $\psi : \Theta \to X$ called the *interpretation map*, together with a map $\kappa : X \to X \otimes Y$ called *hidden Markov model*, such that Eq. (45) holds. In this context, Eq. (45) is called the *consistency equation*. A map $c : Y \otimes \Theta \to \Theta$ together with an interpretation (ψ, κ) is called a *reasoner*.

Technically, this is a slight simplification of the main definition of [33], since that paper allows c to be stochastic rather than restricting it to be deterministic. This terminology stems from the proposal to look at a the map c as a physical system, whose states parametrise a Bayesian prior. The interpretation map specifies this parametrisation. The consistency equation, Eq. (45), guarantees that when the prior updates, it does so in a way that is consistent with Bayesian filtering. We note also that it is possible for the same map c to admit many different Bayesian filtering interpretations.

Although the hidden Markov "model" κ is a different kind of thing from the "model" μ in the internal model principle, they are nevertheless related. In the next section we will formally show this connection, providing a Bayesian interpretation of the IMP and showing that interpretation maps can be seen as a generalisation of models in Definition II.9.

IV. Bayesian filtering interpretations from models

To explain the connection between the notion of models in Definition II.9 and Bayesian filtering interpretations, we will now show that every model induces a Bayesian filtering interpretation. This means, in turn, that in the setting of Section II (under the hypotheses of Theorem II.13), the attracting controller C^* has (among possibly many others) an interpretation involving the attracting environment E^*.

A. A possibilistic perspective on the IMP

The existence of power sets allows to turn models (Definition II.9) *upside down*, bringing us closer, as we shall briefly see, to the Bayesian filtering interpretation we are after. We start from the following construction:

Construction IV.1. Given any surjective function $f : A \to B$ we can define a left-total relation $f^{-1} : B \to P^+(A)$ landing in the set $P^+(A) = P(A) \setminus \{\varnothing\}$ of inhabited subsets of A:

$$f^{-1}(b) \coloneqq \{a \in A \mid f(a) = b\} \quad \text{for } b \in B. \tag{46}$$

For a given $b \in B$ the subset $f^{-1}(b)$ is called the *fibre of* $f : A \to B$ *over* $b \in B$. Both $f : A \to B$ and $f^{-1} : B \to P^+(A)$ can be seen as maps in $\mathbf{Rel}^+$ with types $f : A \to B$ and $f^{-1} : B \nrightarrow A$. Composing these two yields a closure map $\diamondsuit_f : A \nrightarrow A$, which *closes* an element $a \in A$ by mapping it to the set of all $a' \in A$ that generate the same image:

$$\diamondsuit_f(a) \coloneqq (f^{-1} \fatsemi f)(a) = \{a' \in A \mid f(a') = f(a)\}. \tag{47}$$

Instantiated for the case of a model μ between autonomous systems, with map of states $\mu_\mathsf{s} : X \to M$, we can thus interpret the objects constructed above as follows:

1) $\mu_\mathsf{s}^{-1} : M \nrightarrow X$ explicitly assigns to each state $m \in M$ the (inhabited) set of states of the system X it models. We can think of this map as encoding a *belief*: $\mu_\mathsf{s}^{-1}(m)$ is the belief (in the form of a set of possibilities) someone using the model μ would have about the state of the system X knowing just m.
2) $\diamondsuit_{\mu_\mathsf{s}}$ completes an element $x \in X$ with all the other states in X which can't be distinguished from x by the model μ, i.e. states *equicredible* with x.

We now state the following proposition, which will be used in Theorem IV.4.

Proposition IV.2. *For autonomous systems* M *and* X*, if* M *models* X *with respect to the map of systems* μ*, the following diagram commutes in* $\mathbf{Rel}^+$*:*

$$\begin{array}{ccc} M & \xrightarrow{\mu_\mathsf{s}^{-1}} & X \\ {\scriptstyle \mathsf{upd}_\mathsf{M}}\downarrow & & \downarrow{\scriptstyle \overline{\mathsf{upd}_\mathsf{X}}} \\ M & \xrightarrow[\mu_\mathsf{s}^{-1}]{} & X \end{array} \tag{48}$$

where

$$\overline{\mathsf{upd}_\mathsf{X}} \coloneqq \mathsf{upd}_\mathsf{X} \fatsemi \diamondsuit_{\mu_\mathsf{s}}. \tag{49}$$

Proof. See Appendix A. ☐

The function $\overline{\mathsf{upd}_\mathsf{X}}$ can be thought of as *the dynamics of* X *from the point of view of the map* $\mu : \mathsf{X} \to \mathsf{M}$ of a system M modelling a system X. We will see shortly that the map μ_s^{-1} can be seen as an interpretation map and $\overline{\mathsf{upd}_\mathsf{X}}$ as the according model of a Bayesian filtering interpretation. Before proceeding with the main theorem, we briefly note the following as a relevant example of Proposition IV.2.

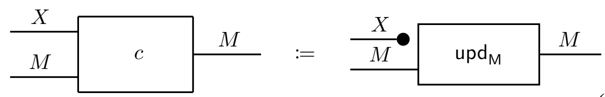
Example IV.3. C^* *models* E^* *with the map* ν*, according to Theorem II.13. Thus Proposition IV.2 tells us that the following diagram commutes:*

$$\begin{array}{ccc} C^* & \xrightarrow{\nu_\mathsf{s}^{-1}} & E^* \\ {\scriptstyle \mathsf{upd}_{\mathsf{C}^*}}\downarrow & & \downarrow{\scriptstyle \overline{\mathsf{upd}_{\mathsf{E}^*}}} \\ C^* & \xrightarrow[\nu_\mathsf{s}^{-1}]{} & E^* \end{array} \tag{50}$$

B. Models imply Bayesian filtering interpretations

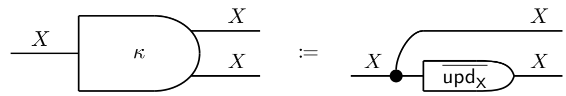
Having obtained a possibilistic version of the IMP with a notion of beliefs given by the map μ_s^{-1}, we now show how a model from Definition II.9 induces a Bayesian filtering interpretation.

Theorem IV.4. *Let* M *model* X *with* $\mu : \mathsf{X} \to \mathsf{M}$*, and assume* M *and* X *are autonomous. Define* $c : X \otimes M \to M$ *as*



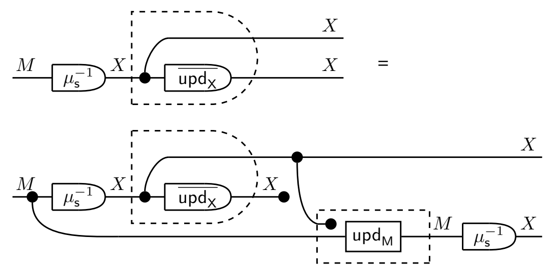
(51)

and $\kappa : X \nrightarrow X \otimes X$ *as:*



(52)

Then κ *is the hidden Markov model, and* $\mu_\mathsf{s}^{-1} : M \nrightarrow X$ *the interpretation map of a Bayesian filtering interpretation of* c*, i.e. we have:*



(53)

where the dashed lines show, informally, where we replaced the definitions above in Eq. (45)*.*

Proof. See Appendix B. ☐

We highlight that the hidden Markov model κ in Theorem IV.4 is possibilistic as opposed to probabilistic since we are now in $\mathbf{Rel}^+$ rather than **FinStoch** or another Markov category. This means, once again, that κ includes transitions to sets of possible states without assigning numerical probabilities, and is related to, among other, standard work on (non-deterministic) labelled transition systems in automata theory, as previously highlighted by, for instance, [33], [66].

We then note that updates $\overline{\mathsf{upd}_\mathsf{X}}$ constitute a kind of approximation of upd_X. For any state $x \in X$, the deterministic result of updates $\mathsf{upd}_\mathsf{X}(x)$ is in fact replaced by an approximate, possibilistic one: the set $\diamondsuit_{\mu_\mathsf{s}}(\mathsf{upd}_\mathsf{X}(x))$ of all states that are mapped to the same state $\mu_\mathsf{s}(\mathsf{upd}_\mathsf{X}(x)) \in M$ (i.e. to the fibre over $\mu_\mathsf{s}(\mathsf{upd}_\mathsf{X}(x))$.) In this sense, these states are indistinguishable from the perspective of M.

This approximation is derived from the map on states μ_s defined as a surjective function, see Eq. (47). If the map on states was *bijective*, then $\overline{\mathsf{upd}_\mathsf{X}} = \mathsf{upd}_\mathsf{X}$ and beliefs would be (trivially) concentrated on a single hidden state.

Our result also shows a rather simplistic form of Bayesian filtering where observations are essentially ignored. This is to be expected due to our definition of model in Definition II.9. As mentioned in Remark II.10, models of autonomous systems are also autonomous systems, and so M must be autonomous. We thus only have an update function $\mathsf{upd}_\mathsf{M} : M \to M$

that doesn't take any inputs to build a Bayesian filtering interpretation.

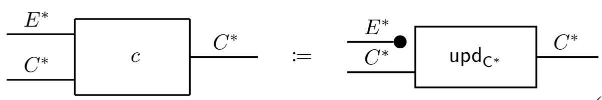
Nonetheless, even while ignoring observations, the Bayesian filtering interpretation is still consistent. This is possible since, due to the approximation of $\mathsf{upd}_{\mathsf{X}}$, for any prior given by $\mu_{\mathsf{s}}^{-1} : M \nrightarrow X$, the observations do not affect the posterior.

There is another way in which Bayesian filtering interpretations induced by models of autonomous systems are special. Their beliefs are necessarily disjoint. This is a direct consequence of the interpretation map being given by a right inverse μ_{s}^{-1} of the surjective map on states μ_{s}.

C. Applications to the Internal Model Principle

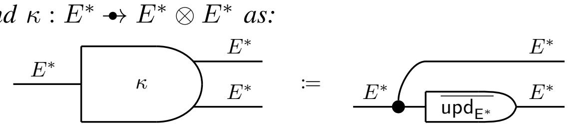
Recall that, in the IMP setting, the error feedback structure of Assumption 3 let us define the attracting controller C^* which is an autonomous system. Theorem II.13 then shows that the attracting controller models the attracting environment E^*. Theorem IV.4 tells us that we then also have the following Bayesian filtering interpretation.

Example IV.5. *Define* $c : E^* \otimes C^* \to C^*$ *as*



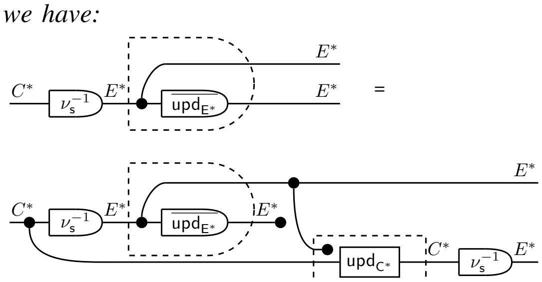
(54)

and $\kappa : E^* \nrightarrow E^* \otimes E^*$ *as:*



(55)

Then κ *is the hidden Markov model, and* $\nu_{\mathsf{s}}^{-1} : C^* \nrightarrow E^*$ *the interpretation map of a Bayesian filtering interpretation of* c*, i.e. we have:*



(56)

This example explicitly shows the precise sense in which one can understand a controller as modelling its environment from a Bayesian perspective: under the IMP assumptions [25]–[28], a model in the control theoretic sense described by Definition II.9 admits a Bayesian filtering interpretation with reasoner and (hidden Markov) model, as described by Definition III.10 following work by [33], [49], given above.

It also provides a clearer understanding of the statement that the environment can be seen as a "convenient fiction by which the designer may specify (or analyst describe) precisely the class of tracking and disturbance rejection tasks which the controlled system is to accomplish with zero (asymptotic) error" [25], since the environment now just appears in Eq. (55) as a special case of a full-fledged *epistemic* (see [50]) Bayesian model κ for a reasoner, i.e. while it could in principle capture properties of the physical world where the controller is instantiated, it effectively only needs to obey the consistency equation expressed by a Bayesian filtering interpretation.

V. Conclusions and future work

The idea of "internal models" has appeared in a number of different research fields, including control theory, biology, artificial intelligence and cognitive science [1], [3], [6], [9], [12], [14], [24]. These notions of internal models appeal, at least on the surface, to a common intuition of a system modelling another system in order to achieve a goal, e.g. a reinforcement learning agent forming a model of its world to maximise the sum of expected rewards, or a cognitive system performing Bayesian inference on the hidden states generating its observations with a model of the environment to perform a certain task. It is however unclear in the literature whether this goes beyond a simple analogy, and whether different notions of internal models can be captured by a common mathematical theory.

In this work we provided a first investigation on different notions of "models": one that is used in the control theoretic context of the internal model principle [3]–[5], and one from work on Bayesian interpretations [33], [49]. We formally connected the two by showing that the notion of model for autonomous systems in internal model principle as formulated in [26]–[28] can be seen as a special case of the more general Bayesian filtering interpretations proposed by [33], [49].

More specifically, Theorem IV.4 tells us that the definition of a model used in the IMP literature, i.e. the special case of our Definition II.9 for autonomous systems [25]–[28], [40], induces a Bayesian filtering interpretation.

Moreover, it also tells us that the IMP definition of a model is too restrictive from a Bayesian perspective: Bayesian filtering interpretations are a more general formalisation of what it means for a system to model another one. The Bayesian filtering interpretations induced by the control-theoretic models are such that (1) the observations *need not* be taken into account in order to update beliefs about the hidden states, and (2) those beliefs are always disjoint.

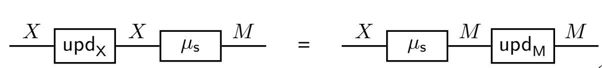
However, the purpose of Bayesian filtering is to use observations to update beliefs. An explicit example of a system with a non-trivial Bayesian filtering interpretation, solving a partially observable Markov decision problem, can be found in [49].

One direction for future work is to generalise Theorem IV.4 beyond autonomous systems, to general open systems. A more challenging task may be to generalise the IMP itself in particular in a way that doesn't require Assumption 3 or Assumption 4.

Appendix

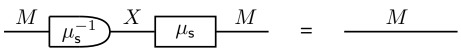
A. Proof of Proposition IV.2

Proof. For this proof, we will use string diagrams as presented in Section III to denote the arrows of the Markov category $\mathbf{Rel}^+$ since all the maps we need are morphisms in that category. First, we translate Definition II.9 for autonomous systems (see Remark II.10) into string diagrams, obtaining:



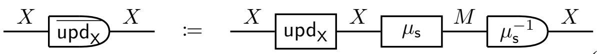
(57)

Then we note that in $\mathbf{Rel}^+$, we have that $\mu_s^{-1} : M \nrightarrow X$ is a right inverse of the surjective map $\mu_s : X \to M$, i.e.:



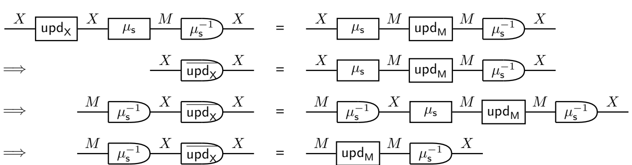
(58)

Furthermore, we also have the definition of the updates induced by the Hepburn–Wonham IMP, $\overline{\mathsf{upd}_X}$, see Eq. (49):



(59)

We then prove the following, starting from Eq. (57), we can (i) post-compose both sides with μ_s^{-1}, (ii) apply Eq. (59) on the left hand side, (iii) pre-compose both sides with μ_s^{-1}, and then (iv) use the surjectivity of μ_s (Eq. (58)) to get:

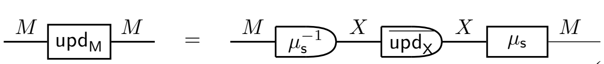


(60)

which gives us the string diagram version of Eq. (48). □

B. Proof of Theorem IV.4

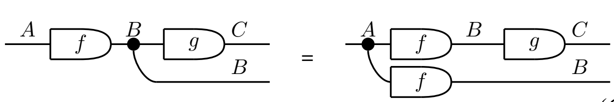
Proof. For this proof, we will use string diagrams for the Markov category $\mathbf{Rel}^+$. We start with the following equation, derived from post-composing both sides of Eq. (60) with μ_s^{-1} and subsequently using the surjectivity property of μ_s, see Eq. (58),



(61)

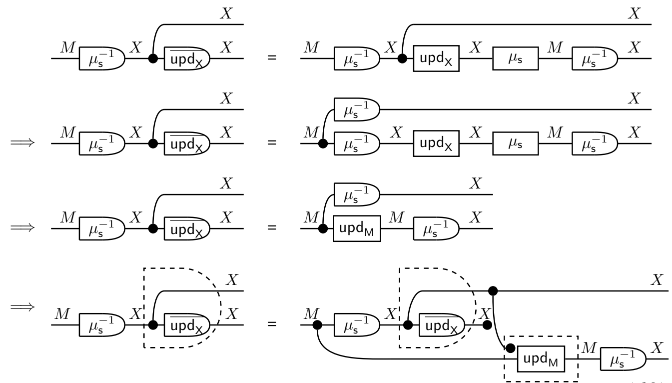
This tells us that the composite $\mu_s^{-1} \fatsemi \overline{\mathsf{upd}_X} \fatsemi \mu_s$ is deterministic, and thus allows us to apply the definition of *positivity* for Markov categories [35, Definition 11.22] (since every category with conditionals, thus including $\mathbf{Rel}^+$ [67], is positive). We briefly recall the definition of a positive Markov category.

Definition A.1. A Markov category $\mathbf{C}$ is positive if given two morphisms $f : A \to B$ and $g : B \to C$ such that their composite $f \fatsemi g$ is deterministic implies the following



(62)

It then follows that, starting from Eq. (59), we can (i) parallel compose both sides with id_X and pre-compose both sides with $\mu_s^{-1} \fatsemi \Delta_X$ (where Δ_X is the copy map for the system X) (ii) apply the definition of positivity to the deterministic composite in Eq. (61) on the right hand side, and (iii) use the surjectivity of μ_s, see Eq. (58), to obtain the following:



(63)

□

Acknowledgments

Matteo Capucci is an Independent Researcher funded by the Advanced Research+Innovation Agency (ARIA). Work by Martin Biehl on this paper was made possible through the support of Grants 62828 and 62229 from the John Templeton Foundation. Nathaniel Virgo's work on this paper was made possible through the support of grant 62229 from the John Templeton Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

References

[1] R. C. Conant and W. R. Ashby, "Every good regulator of a system must be a model of that system," *International journal of systems science*, vol. 1, no. 2, pp. 89–97, 1970.

[2] W. R. Ashby, "Requisite variety and its implications for the control of complex systems," *Cybernetica*, vol. 1, pp. 83–99, 1958.

[3] B. A. Francis and W. M. Wonham, "The internal model principle of control theory," *Automatica*, vol. 12, no. 5, pp. 457–465, 1976.

[4] E. D. Sontag, "Adaptation and regulation with signal detection implies internal model," *Systems & control letters*, vol. 50, no. 2, pp. 119–126, 2003.

[5] M. Bin, J. Huang, A. Isidori, L. Marconi, M. Mischiati, and E. D. Sontag, "Internal models in control, bioengineering, and neuroscience," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, pp. 55–79, 2022.

[6] T.-M. Yi, Y. Huang, M. I. Simon, and J. Doyle, "Robust perfect adaptation in bacterial chemotaxis through integral feedback control," *Proceedings of the National Academy of Sciences*, vol. 97, no. 9, pp. 4649–4653, 2000.

[7] C. Briat, A. Gupta, and M. Khammash, "Antithetic integral feedback ensures robust perfect adaptation in noisy biomolecular networks," *Cell systems*, vol. 2, no. 1, pp. 15–26, 2016.

[8] M. H. Khammash, "Perfect adaptation in biology," *Cell Systems*, vol. 12, no. 6, pp. 509–521, 2021.

[9] D. Ha and J. Schmidhuber, "World models," *arXiv preprint arXiv:1803.10122*, 2018.

[10] P. Wu, A. Escontrela, D. Hafner, P. Abbeel, and K. Goldberg, "Daydreamer: World models for physical robot learning," in *Conference on Robot Learning*. PMLR, 2023, pp. 2226–2240.

[11] T. Taniguchi, S. Murata, M. Suzuki, D. Ognibene, P. Lanillos, E. Ugur, L. Jamone, T. Nakamura, A. Ciria, B. Lara *et al.*, "World models and predictive coding for cognitive and developmental robotics: frontiers and challenges," *Advanced Robotics*, pp. 1–27, 2023.

[12] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Prentice Hall Press, 2009.

[13] S. Grossberg, "Competitive learning: From interactive activation to adaptive resonance," *Cognitive science*, vol. 11, no. 1, pp. 23–63, 1987.

[14] M. Kawato, “Internal models for motor control and trajectory planning,” *Current opinion in neurobiology*, vol. 9, no. 6, pp. 718–727, 1999.

[15] M. Ito, “Control of mental activities by internal models in the cerebellum,” *Nature Reviews Neuroscience*, vol. 9, no. 4, pp. 304–313, 2008.

[16] B. Baker, B. Lansdell, and K. P. Kording, “Three aspects of representation in neuroscience,” *Trends in cognitive sciences*, 2022.

[17] I. Harvey, “Untimed and misrepresented: Connectionism and the computer metaphor,” University of Sussex, School of Cognitive and Computing Sciences, Tech. Rep., 1992.

[18] R. D. Beer, “A dynamical systems perspective on agent-environment interaction,” *Artificial Intelligence*, vol. 72, no. 1-2, pp. 173–215, 1995.

[19] T. Van Gelder, “What might cognition be, if not computation?” *The Journal of Philosophy*, vol. 92, no. 7, pp. 345–381, 1995.

[20] I. Harvey, “Misrepresentations,” in *Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems*. MIT Press, 2008.

[21] D. C. Knill and W. Richards, *Perception as Bayesian inference*. Cambridge University Press, 1996.

[22] D. McNamee and D. M. Wolpert, “Internal models in biological control,” *Annual review of control, robotics, and autonomous systems*, vol. 2, pp. 339–364, 2019.

[23] E. Todorov, “Optimal control theory,” *Bayesian brain: probabilistic approaches to neural coding*, pp. 269–298, 2006.

[24] A. K. Seth, “The Cybernetic Bayesian Brain,” in *Open MIND*, W. Wiese and T. K. Metzinger, Eds. Frankfurt am Main, Germany: MIND Group, 2014, pp. 9–24.

[25] W. M. Wonham, “Towards an abstract internal model principle,” *IEEE Transactions on Systems, Man, and Cybernetics*, no. 11, pp. 735–740, 1976.

[26] J. S. A. Hepburn and W. M. Wonham, “The internal model principle of regulator theory on differentiable manifolds,” *IFAC Proceedings Volumes*, vol. 14, no. 2, pp. 319–323, 1981.

[27] ——, “Structurally stable nonlinear regulation with step inputs,” *Mathematical systems theory*, vol. 17, no. 1, pp. 319–333, 1984.

[28] ——, “Error feedback and internal models on differentiable manifolds,” *IEEE Transactions on Automatic Control*, vol. 29, no. 5, pp. 397–403, 1984.

[29] J. Huang, A. Isidori, L. Marconi, M. Mischiati, E. D. Sontag, and W. M. Wonham, “Internal models in control, biology and neuroscience,” in *2018 IEEE Conference on Decision and Control (CDC)*. IEEE, 2018, pp. 5370–5390.

[30] D. J. Myers, “Double categories of open dynamical systems,” *Electronic Proceedings in Theoretical Computer Science*, vol. 333, pp. 154–167, 2021.

[31] ——, “Categorical systems theory,” http://davidjaz.com/Papers/DynamicalBook.pdf, 2023.

[32] M. Capucci, B. Gavranović, J. Hedges, and E. F. Rischel, “Towards foundations of categorical cybernetics,” *Electronic Proceedings in Theoretical Computer Science*, no. 372, pp. 235–248, 2022.

[33] N. Virgo, M. Biehl, and S. McGregor, “Interpreting Dynamical Systems as Bayesian Reasoners,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2021, pp. 726–762.

[34] K. Cho and B. Jacobs, “Disintegration and bayesian inversion via string diagrams,” *Mathematical Structures in Computer Science*, vol. 29, no. 7, pp. 938–971, 2019.

[35] T. Fritz, “A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics,” *Advances in Mathematics*, vol. 370, p. 107239, 2020.

[36] B. Coecke, T. Fritz, and R. W. Spekkens, “A mathematical theory of resources,” *Information and Computation*, vol. 250, pp. 59–86, 2016.

[37] B. Coecke and A. Kissinger, *Picturing Quantum Processes: A First Course in Quantum Theory and Diagrammatic Reasoning*. Cambridge University Press, 2017.

[38] A. Isidori and C. I. Byrnes, “Output regulation of nonlinear systems,” *IEEE transactions on Automatic Control*, vol. 35, no. 2, pp. 131–140, 1990.

[39] W. M. Wonham, *Linear multivariable control*. Springer, 1974, vol. 101.

[40] W. M. Wonham and K. Cai, *Supervisory control of discrete-event systems*. Springer, 2019.

[41] A. Isidori, A. Krener, C. Gori-Giorgi, and S. Monaco, “Nonlinear decoupling via feedback: a differential geometric approach,” *IEEE transactions on automatic control*, vol. 26, no. 2, pp. 331–345, 1981.

[42] I. Thobani, “A triviality worry for the internal model principle,” *Synthese*, vol. 204, no. 1, p. 36, 2024.

[43] W. G. Noid, J.-W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, and H. C. Andersen, “The multiscale coarse-graining method. i. a rigorous bridge between atomistic and coarse-grained models,” *The Journal of chemical physics*, vol. 128, no. 24, 2008.

[44] H. A. Simon and A. Ando, “Aggregation of variables in dynamic systems,” *Econometrica: journal of the Econometric Society*, pp. 111–138, 1961.

[45] Z. Ren and B. H. Krogh, “State aggregation in markov decision processes,” in *Proceedings of the 41st IEEE Conference on Decision and Control, 2002.*, vol. 4. IEEE, 2002, pp. 3819–3824.

[46] J. G. Kemeny, J. L. Snell *et al.*, *Finite markov chains*. van Nostrand Princeton, NJ, 1969, vol. 26.

[47] B. Moore, “Principal component analysis in linear systems: Controllability, observability, and model reduction,” *IEEE transactions on automatic control*, vol. 26, no. 1, pp. 17–32, 1981.

[48] P. E. Caines and Y.-J. Wei, “The hierarchical lattices of a finite machine,” *Systems & Control Letters*, vol. 25, no. 4, pp. 257–263, 1995.

[49] M. Biehl and N. Virgo, “Interpreting systems as solving pomdps: a step towards a formal understanding of agency,” in *International Workshop on Active Inference*. Springer, 2022, pp. 16–31.

[50] N. Virgo, “Unifilar Machines and the Adjoint Structure of Bayesian Models,” *Electronic Proceedings in Theoretical Computer Science*, vol. 397, p. 299–317, 2023.

[51] D. Horsman, S. Stepney, R. C. Wagner, and V. Kendon, “When does a physical system compute?” *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 470, 2014.

[52] F. Borceux, *Handbook of Categorical Algebra: Volume 2: Categories and Structures*, ser. Encyclopedia of Mathematics and Its Applications. Cambridge University Press, 1994, vol. 2.

[53] J. C. Baez and M. Stay, “Physics, topology, logic and computation: a Rosetta Stone,” in *New structures for physics*. Springer, 2010, pp. 95–172.

[54] P. Selinger, “A survey of graphical languages for monoidal categories,” in *New structures for physics*. Springer, 2010, pp. 289–355.

[55] P. Perrone, “Markov categories and entropy,” *IEEE Transactions on Information Theory*, 2023.

[56] D. M. Stein and S. Staton, “Probabilistic programming with exact conditions,” *J. ACM*, 2023. [Online]. Available: https://doi.org/10.1145/3632170

[57] C. I. Byrnes and A. Isidori, “Limit sets, zero dynamics, and internal models in the problem of nonlinear output regulation,” *IEEE Transactions on Automatic Control*, vol. 48, no. 10, pp. 1712–1723, 2003.

[58] J. B. Rawlings, D. Q. Mayne, and M. Diehl, *Model predictive control: theory, computation, and design*. Nob Hill Publishing Madison, WI, 2017, vol. 2.

[59] J.-P. Aubin, A. M. Bayen, and P. Saint-Pierre, *Viability Theory: New Directions*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[60] A. Tanwani, B. Brogliato, and C. Prieur, “Well-posedness and output regulation for implicit time-varying evolution variational inequalities,” *SIAM Journal on Control and Optimization*, vol. 56, no. 2, pp. 751–781, 2018.

[61] B. Brogliato and A. Tanwani, “Dynamical systems coupled with monotone set-valued operators: Formalisms, applications, well-posedness, and stability,” *Siam Review*, vol. 62, no. 1, pp. 3–129, 2020.

[62] T. Denœux, Z. Younes, and F. Abdallah, “Representing uncertainty on set-valued variables using belief functions,” *Artificial Intelligence*, vol. 174, no. 7-8, pp. 479–499, 2010.

[63] M. O. Rabin and D. Scott, “Finite automata and their decision problems,” *IBM journal of research and development*, vol. 3, no. 2, pp. 114–125, 1959.

[64] B. Jacobs, “A channel-based perspective on conjugate priors,” *Mathematical Structures in Computer Science*, vol. 30, no. 1, pp. 44–61, 2020.

[65] A. Caticha, “Entropy, information, and the updating of probabilities,” *Entropy*, vol. 23, no. 7, p. 895, 2021.

[66] T. Fritz, A. Klingler, D. McNeely, A. Shah-Mohammed, and Y. Wang, “Hidden Markov Models and the Bayes Filter in Categorical Probability,” *arXiv preprint arXiv:2401.14669*, 2024.

[67] T. Fritz and A. Klingler, “The d-separation criterion in categorical probability,” *Journal of Machine Learning Research*, vol. 24, no. 46, pp. 1–49, 2023.