# A "Good Regulator Theorem" for Embodied Agents

Nathaniel Virgo[1,2], Martin Biehl[3], Manuel Baltieri[4], and Matteo Capucci[5]
[1]University of Hertfordshire, UK [2]Earth-Life Science Institute (ELSI), Institute of Science Tokyo, Japan
[3]Cross Labs, Cross Compass Ltd., Japan [4]Araya, Inc., Japan [5]Independent researcher

**Abstract**

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

## 1 Introduction

The work of W. Ross Ashby had a profound impact on the development of Artificial Life, and of cognitive science more broadly. In particular, Ashby (1960) led to the idea of the *viability boundary* of an organism, a notion of current interest in ALIFE (McShaffrey and Beer, 2023; Beer et al., 2024; Egbert and Pérez-Mercader, 2018) and a foundational concept in enactivism and the evolutionary robotics approach to cognition (e.g. Beer, 1995, 1997, 2004; Di Paolo, 2005).

In parallel, (Conant and Ashby, 1970) claimed in its title that "every good regulator of a system must be a model of that system", an idea that was taken up in control theory under the name 'internal model principle' (Francis and Wonham, 1976) and whose influence on cognitive science can still be felt today, for example, in some works under the banner of the free energy principle (FEP) such as (Friston, 2019; Da Costa et al., 2021), or in the information-theoretic definition of 'semantic information' (Kolchinsky and Wolpert, 2018).

Although the ideas in (Ashby, 1960) and (Conant and Ashby, 1970, hereafter C&A) are not dissimilar, the bodies of work descending from them appear to have reached disparate conclusions: Artificial Life has produced many examples of agents that apparently do not need an internal model in order to achieve their tasks (e.g. Braitenberg, 1986; Beer, 2003), apparently contradicting (C&A).

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

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

For Conant and Ashby, for a system $X$ to *be a model* of a system $Y$ there should be a " '-morphic' relation", i.e. a homomorphism of some kind from, $Y$ to $X$. This means roughly that the model $X$ behaves like a coarse-grained version of the thing it's modelling, $Y$, with an intuition along the lines that a controller must behave like the system it's controlling in order to predict it.

In contrast, our notion of model builds on a recent notion of *interpretation map* (Virgo et al., 2021; Biehl and Virgo, 2023. The idea is that it must be possible for an observer to attribute belief states (priors) to the states of $X$ that change consistently with the dynamics of model $Y$. In the terminology of Seth and Tsakiris (2018) this is a version of "having a model" (as opposed to "being a model") in that it is about states encoding priors, although see the comments on 'as-if' agency below. In the current work we consider "possibilistic" belief states, as in (Baltieri et al., 2025, section III), which are simpler than Bayesian priors, though closely analogous.

We show that this version of 'has a model' has its own "good regulator theorem" (theorem 3.4), which doesn't need

any assumptions beyond a fairly minimal definition of regulation. Although this definition of regulation is minimal, it is flexible enough to apply to embodied agents. The theorem shows that every good regulator has a model, but it allows that the model might be trivial, and this is how the apparent counterexamples are accounted for.

A related result is proved in (Baltieri et al., 2025) in the context of the Internal Model Principle (IMP). Here we take a more first-principles approach, which doesn't need the assumptions of the IMP and leads to a simpler yet stronger theorem. In particular, (Baltieri et al., 2025) requires that the controller can fully observe the plant, which limits the extent to which it can be applied to embodied agents. Nevertheless, this prior work can be seen as showing that "being a model" implies "having a model" in senses similar to ours.

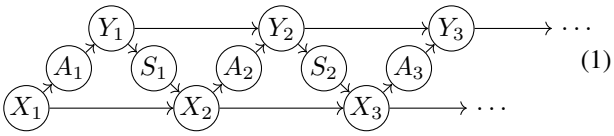
The study of interpretation maps builds on a research programme of 'as-if' agency (McGregor, 2016, 2017; McGregor et al., 2025a,b), which aims to formalise Dennett's ideas about the intentional stance (e.g. Dennett, 2006). The idea of a *stance* is central to our work: our notion of model depends fundamentally on choices that an external observer must make about how to interpret the system's internal state and behaviour. A model is part of a stance the observer can optionally take with respect to the system, rather than a property of the system alone. (In this respect our work could be characterised as "is a model" rather than "has a model" in Seth and Tsakiris's terms, but we will continue to use the "has a model" terminology.) At the same time, the choice is not arbitrary: the patterns in a system's dynamics that make an interpretation possible are a "real" property of that system (c/f Dennett, 1991).

Our work could be seen as laying theoretical groundwork for similar ideas in machine learning, for example in work focusing on extracting Bayesian models (Ortega et al., 2019) or world-models (Richens et al., 2025) from agents trained on multiple tasks. There are also has some similarities to recent work on emergence (Rosas et al., 2024), and we suspect the two approaches will turn out to support each other, since our models are probably best thought of as interpretations of the macroscopic "software" level rather than the micro. In the conclusion section we also discuss potential relations to the free energy principle (FEP).

Our notion of regulation includes both 'extrinsic' cases such as a thermostat regulating the temperature of a room and 'intrinsic' cases such as an organism regulating its own body temperature. Because of this, we can claim as an informal corollary that *every system that is a good regulator of itself, must have a model of its environment.* This is perhaps surprising, and suggests a possible connection to notions of sense-making in enactive cognitive science, where meaning is said to emerge from an organism's need to maintain itself as a physical metabolic system (Di Paolo, 2005; De Jaegher and Di Paolo, 2007). The term "model" might seem counter to enactivism, but our approach is closer to a dynamical systems approach than a traditionally computational one. Our models arise from the dynamics rather than being invoked to explain them, and in a biological context they might have more in common with the enactive notion of meaning than with a traditional computationalist notion of model.

## 2 Coupled systems and regulation

A common framework for modelling embodied agents is the *sensorimotor loop* (Klyubin et al., 2004; Bertschinger et al., 2006; Zahedi et al., 2010). The precise details can differ, but one typically assumes, for simplicity, that time proceeds in discrete steps and that the states of the agent and its environment change over time according to a causal Bayesian network along the lines of the following:



(1)

Here, $X_1, X_2, \ldots$ are the agent's state at different times, $Y_1, Y_2, \ldots$ the state of its environment (which may be taken to include the agent's position within it), $A_1, A_2, \ldots$ the actions it takes on each time step and $S_1, S_2, \ldots$ the inputs it gets to its sensors.

We will formalise the sensorimotor loop in a way that's close to this in spirit but different in look-and-feel, based on the concept of a 'machine'. Although the language of machines might feel computational in nature, machines are really just a formalism for talking about coupled dynamical systems in discrete time. One should think of them as modelling an embodied agent coupled to its environment, in the same way that coupled continuous-time dynamical systems are used in (Beer, 1995, 1997) for example.

We restrict our attention to deterministic systems in discrete time, but the concept of machine can be extended to the stochastic case and the time-continuous case, among others—see (Myers, 2023) for a common framework. Our results are somewhat specialised to the discrete time non-stochastic case. Their extension to the stochastic case is not obvious, though it would be straightforward to generalise them to the possibilistic machines used in (Baltieri et al., 2025).

We define two different types of machine, for reasons that will become apparent shortly. We first fix sets $A$ and $S$, which are the set of actions the agent can take on each time step, and the set of states its sensors can take; together we call them the *interface*.

**Definition 2.1.** A *Moore machine* consists of a *state space*, which is a set $X$, together with a *readout function* $r : X \to A$ and an *update function* $u : X \times S \to X$.

We will use a Moore machine to model the agent part of the sensorimotor loop. The idea is that the map $r$ determines the action the agent takes, as a function of its state, while the map $u$ determines how the agent's internal state updates, as a

function of its previous state and its sensor input. Although we use the language of action selection and state updating, these should really just be thought of in dynamical terms; they are nothing but a description of how the system couples to its environment and changes over time. We will think of readout as happening *before* the update, so that on each time step the agent first takes an action and then receives a sensor input, at which point its state changes.

We model the environment with a different kind of machine, called a Mealy machine. We define it with its inputs and outputs swapped compared to a Moore machine, since it takes the agent's action as an input and produces a sensor value as an output. Its definition is slightly simpler than that of a Moore machine, in that it only has one function.

**Definition 2.2.** A *Mealy machine* consists of a *state space*, which is a set $Y$, together with an *evolution function* $e : Y \times A \to Y \times S$.

The basic idea of the sensorimotor loop is that when an agent is coupled to an environment, they form a (closed) dynamical system. Since we are dealing with discrete time and discrete state spaces, for us a "dynamical system" is just a set $W$ called the *state space*, together with a function $h : W \to W$ called the *update map*. The update map $h$ takes the state at the current time and returns the next state.

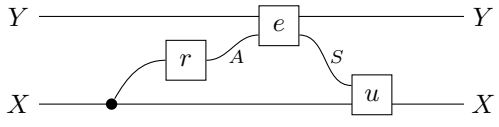
We can now define what it means to couple an agent to its environment:

**Definition 2.3.** Given a Moore machine $(X, r : X \to A, u : X \times S \to X)$ and a Mealy machine $(Y, e : Y \times S \to Y \times A)$, the corresponding *coupled system* is a dynamical system with state space $X \times Y$ and update map $h : X \times Y \to X \times Y$ given by

$$h(x, y) = (u(x, s)\,,\, y'), \tag{2}$$

where $y'$ and $s$ are defined by $(y', s) = e(y, r(x))$.

In the formal language of *string diagrams* (see e.g. Baez and Stay, 2010; Selinger, 2010; Coecke and Kissinger, 2017), Equation (2) can be represented as

$h\ =$ 

(3)

Although this is a formal diagram it can be read in a very intuitive way, from left to right: first the agent takes an action determined by the map $r$, which gets the initial value of $X$ as an input. Then the environment's state updates as a function of this action and the environment's previous state via the map $e$, which generates both a new value of $Y$ and a sensor value; and finally the agent's state updates via the map $u$, producing a new value of $X$. See (Biehl and Virgo, 2023; Baltieri et al., 2025) for more on string diagram approaches to the sensorimotor loop. The relationship between string diagrams and Bayesian networks is explored in (Fritz and Klingler, 2023; Fong, 2013; Jacobs et al., 2019).

The reason for defining two types of machine was so that we could couple them in this way, where an action takes place, the environment reacts and a sensor value is received all in the same time step. This is a modelling choice, however—there is nothing inherently "agent-like" about Moore machines or "environment-like" about Mealy machines. We could have made both systems Moore machines at a slight cost in generality, or we could have made the agent a Mealy machine and the environment a Moore machine, with only small changes to what follows.

## 2.1 Defining regulation

From now on we fix an agent, in the form of a Moore machine $(X, r : X \to A, u : X \times S \to X)$ and environment in the form of a Mealy machine $(Y, e : Y \times S \to Y \times A)$. Given this, we can ask what it means for the agent to perform regulation.

In a classical control theory setup, we would be thinking about the agent regulating the environment. In that case, we would have a subset $G_Y \subseteq Y$ of the environment's state space called the *good set*, and seek to find a controller (that is, a Moore machine) that, when coupled to the environment, keeps its state inside the good set. For example, a "good" thermostat might be one that keeps its environment's temperature between $19\,^\circ\mathrm{C}$ and $25\,^\circ\mathrm{C}$, and the good set would consist of all environment states whose temperature is in this range.

There is a lot that can be said about this classical setup, but our interest here is in embodied biological agents. Generally speaking, biological agents aren't concerned with regulating their environments but with regulating *themselves*, i.e. their own internal physiological variables (Ashby, 1960; Beer, 2004)[1]. One way to set this up would be to have the good set be a subset $G_X \subseteq X$ of the *agent*'s states, e.g. perhaps including a suitable range of body temperatures. However, it turns out that for our purposes it is not hard to set things up in a more general way that encompasses both the classical setup and the Ashbian one. This leads to the following simple, albeit slightly counterintuitive, definition:

**Definition 2.4.** A *regulation situation* consists of the agent $(X, r, u)$ and environment $(Y, e)$, together with a subset $G \subseteq X \times Y$, which we refer to as the *good set*.

Here we define the good set as a subset of the states of the coupled system, rather than of the agent or the environment alone. This encompasses both the classical setup and the Ashbian one, because given a subset $G_Y \subseteq Y$ of environment states we can take $G$ to be the set of pairs $(x, y)$ with $x \in X, y \in G_Y$, whereas given a subset $G_X \subseteq X$ of agent states we can take $G$ to consist of pairs $(x, y)$ with $x \in G_X, y \in Y$.

[1] From the perspective of autopoiesis they regulate more than this: they maintain their autopoietic organisation (Maturana and Varela, 1980, p. 79). Describing this mathematically is a challenge beyond the scope of this paper, however.

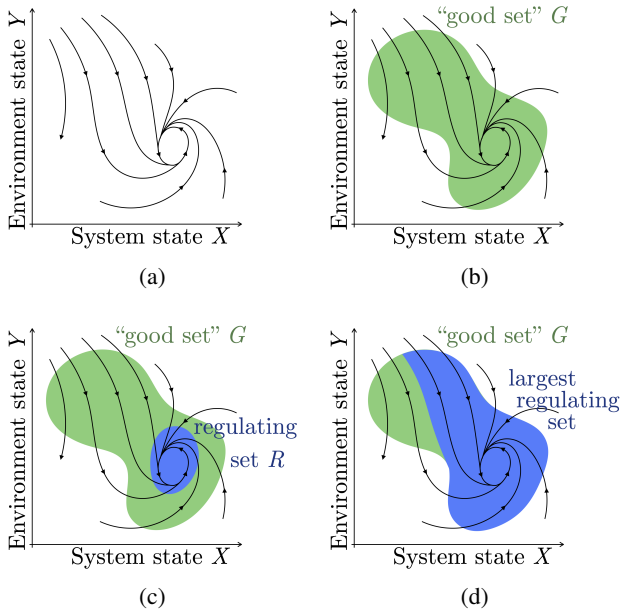



Figure 1: Schematics drawn in continuous time (cf. Beer, 1995), although the definitions in the main text are discrete. (a) A pair of coupled systems together form a dynamical system, some of whose variables belong to each system; in this schematic we assume one variable each, resulting in a two dimensional phase space. (b) The same phase space, equipped with a "good set" $G$. Not every trajectory that starts within the good set stays within it indefinitely, but some do. (c) A *regulating set* is a nonempty subset of $R$ that is forward-closed. The existence of a regulating $R$ set witnesses that there are some trajectories that stay inside $G$ indefinitely. (d) There may be many regulating sets; here we show the largest one.

We refer to definition 2.4 as a regulation *situation* rather than a regulation *problem* since we are not trying to find a controller that regulates a given environment; instead we follow C&A's approach of taking both the controller and the environment as given, along with the good set.

Intuitively, we want to say that the agent performs regulation if it keeps the state of the coupled system inside the good set. However, in general, whether this happens or not is a function not just of the dynamics but also of the initial state of the coupled system, i.e. the combined initial state of *both* the agent *and* its environment. This is illustrated in fig. 1, where trajectories stay inside $G$ indefinitely if and only if the initial state is in the region indicated in fig. 1d.

In order to take the state of the coupled system into account, we use the notion of *forward-closed set*. For our simple notion of dynamical system consisting only of a set $W$ and a function $h : W \to W$, a *forward-closed set* is a subset $V \subseteq W$ such that for each $v \in V$ we have $h(v) \in V$. In other words, to say $V$ is forward-closed means that if the state is in $V$ on the current time step then it will also be in $V$ on the next time step, and hence it will be inside $V$ in all future times as well. An example of a forward-closed set is the basin of attraction of a fixed point, although not all forward-closed sets are of this form. With this notion in hand we can define what it means to perform regulation in our framework.

**Definition 2.5.** Given a regulation situation as in definition 2.4, we say that the agent is a *good regulator* if we can exhibit a set $R \subseteq X \times Y$ that is (1) non-empty, (2) forward-closed, and (3) contained in $G$, i.e. $R \subseteq G$. We call such a set $R$ a *regulating set*.

If the initial state is in $R$ then the system will remain in $G$ indefinitely, so the existence of $R$ implies that regulation is possible, given an appropriate initial state.

**Some philosophical context.** The reader might have noticed a strange asymmetry in definition 2.5, and this seems a good point to address this, along with other philosophical issues that have arisen by this point. The asymmetry is that we say the *agent* is a good regulator, even though definitions 2.4 and 2.5 are otherwise symmetric between agent and environment (aside from the modelling choice about Moore *vs.* Mealy machines). This is because we are really describing a *stance* one can take toward the agent, of a similar flavour to Dennett's intentional stance.

We think of the agent and its environment as observed physical things (although one or both of them could equally be a model in the mind of an observer as we discuss below). The two machines and the coupled system specify the dynamics of these systems, which is something physical and in principle empirically testable. But the good set says something about what *should* happen, and as such we don't expect it to be any kind of physical property of the system itself. Instead, we think of it as something imposed on the system by the observer: the observer has decided to treat the Moore machine *as if* it is an agent whose goal is described by the good set $G$. This might be because the agent is a robot and the observer its designer, in which case $G$ is part of a specification of its desired behaviour, or because the system is an organism and the observer a biologist, in which case the good set could be thought of as a fiction standing in for some knowledge or a hypothesis about its evolutionary history.

We will explore the consequences of this, from the perspective of an observer who is committed to taking such an 'as-if' stance, treating this particular system as an agent with this particular good set $G$ and this particular regulating set $R$. We will discuss the agent's 'norms' and 'beliefs', but it is important to keep in mind that these belong to the observer as much as to the agent itself; they are attributed to the agent by the observer, and a different observer might attribute different norms and beliefs to the same system.

A final philosophical point concerns the role of the regulating set $R$. In general there might be many possible regulating sets, i.e. many non-empty forward-closed subsets of $G$, but in order to apply our theorem, one must pick one of them. One can always take the largest regulating set (given by the union of all regulating sets, assuming at least one regulating set exists), but there may be reasons to consider another regu-

lating set besides this one. For example, proving that some particular set $R$ is a regulating set might be much easier than identifying the largest regulating set. Different choices of $R$ will lead to different belief interpretations in lemma 3.2.

## 3 Models and interpretations

While Conant and Ashby's notion of model is about a map from the system being modelled to the system doing the modelling, our theorem invokes a different kind of model, one that's closely related to so-called *Bayesian filtering interpretations*, though not probabilistic in nature.

To set the stage we give some brief background on Bayesian interpretations. The core idea is in some ways an old one, with roots going back to (Kalman, 1960), but interpretations were introduced as an explicit concept in 'as-if' agency in (Virgo et al., 2021), inspired by a category-theoretic approach to conjugate priors (Jacobs, 2020).

To build a Bayesian filtering interpretation, we start with a dynamical system with inputs (which might be an embodied agent but doesn't have to be). To *interpret* it as performing Bayesian inference about some hidden variable, we regard its internal states as parametrising a prior over the hidden variable. This means we have a map from the state space of the system to probability distributions over the hidden variable, which we think of as telling us "what the system's prior is", as a function of its state. This must be such that whenever the system's state updates in response to an input, the new prior assigned to the system's new state must equal the posterior that Bayes' theorem mandates. In other words, the following somewhat informal diagram should commute, where $\Delta(Y)$ is the space of priors over some hidden variable $Y$, the agent's state space is $X$, and $\psi : X \to \Delta(Y)$ is the interpretation map.

$$\begin{array}{ccc} \Delta(Y) & \xrightarrow{\text{update by Bayes conditioned on } s \in S} & \Delta(Y) \\ \uparrow{\scriptstyle \psi} & & \uparrow{\scriptstyle \psi} \\ X & \xrightarrow{\text{dynamical transition on receiving input } s \in S} & X \end{array} \tag{4}$$

Making this formal requires specifying some other data, including the agent's *model*, i.e. how the agent believes $S$ and $Y$ are correlated, and how it believes $Y$ changes over time, if $Y$ is not assumed to be constant. Once the details are filled in, eq. (4) becomes Equation 5 in (Virgo et al., 2021), the so-called *consistency equation for Bayesian filtering interpretations*. Our notion of interpretation below has this same general form but differs in its details from this previous work.

Regardless of how the details are filled in, there can be (and generally are) many interpretations of a given system. It doesn't make sense to ask which interpretation is the correct one, or to the extent that it does, it isn't a question about the system but about something external to it, such as the intention of its designer. To interpret a system as performing inference requires choosing an interpretation map, and this choice belongs to the observer and not to the system itself.

In (Biehl and Virgo, 2023) this idea was extended to the case of an agent that takes actions on the basis of its Bayesian beliefs, which we will also be concerned with below. In (Baltieri et al., 2025) the idea is instantiated in the realm of *possibilistic* reasoning, where the agent only cares what can or can't happen, as opposed to *probabilistic*, where the possibilities are assigned probabilities. This is also what we will do, although also in a slightly different way.

Let us now set out the specific notion of interpretation that's relevant for our main theorem, theorem 3.4. We want to consider an agent as having beliefs about its environment, which in our framework means a Moore machine having beliefs about a Mealy machine.

As before, we start with a Moore machine given by $(X, r : X \to A,\, u : X \times S \to X)$ and a Mealy machine which we will write as $(Z, f : Z \times A \to Z \times S)$. This Mealy machine is not necessarily the same as the environment $(Y, e)$ mentioned previously. Instead, we will think of it as the thing the agent *has beliefs about*. (Meaning: the thing about which some observer interprets it as having beliefs.) The picture to have in mind is that the Moore machine might be coupled to an environment $(Y, e)$, but independently of that it might 'believe' it is coupled to a different environment $(Z, f)$. As discussed in (Virgo et al., 2021), it might be that $(Y, e) = (Z, f)$, but it need not be, since we want to allow the possibility that the agent is not in its natural environment, or otherwise has an incorrect model of how the environment behaves. All the machines have the same interface, $(S, A)$.

We sometimes refer to the Mealy machine $(Z, f)$ as the agent's model (as attributed to it by an observer). This sense of 'model' is very different from Conant and Ashby's one. We sometimes talk about the agent *having* a model of $(Z, f)$, although to be precise we should say "the observer interprets the agent as having a model," since the model really belongs as much to the observer as it does to the system itself.

As discussed, to talk about the agent having beliefs about the environment we should define an interpretation map, which should obey certain conditions. In the present work, this will be a function $\psi : X \to \mathcal{P}(Z)$, where $\mathcal{P}$ means power set. The idea is that when the agent is in state $x \in X$, we interpret it as believing that the environment is in one of the states in $\psi(x)$, which is a subset of $Z$. If the set $\psi(x)$ is a singleton $\{z\}$, the agent is said to be certain that the environment is in state $z$. If $\psi(x) = Z$ then we say the agent is completely uncertain about the environment's state; it could be in any state at all. For anything in between the agent has some partial certainty—it is interpreted as knowing that the state is within some set but not precisely what the state is. It is also possible that $\psi(x) = \emptyset$, in which case we say the agent has 'absurd' beliefs. This is similar to believing that $0 = 1$; it signifies an inconsistency in the agent's beliefs, since there is no state of the environment that satisfies them.

What conditions should we require the interpretation map $\psi : X \to \mathcal{P}(Z)$ to have in order to be valid? To determine this, let us put ourselves in the mind of someone who knows that the state of the environment is within some set $B \subseteq Z$. Suppose that this person then takes some action $a \in A$ (for the moment we will not worry about how this action was selected) and then receives some sensor value $s \in S$. Knowing the values of $a$ and $s$, this person can then reason as follows:

"I know that initially the environment was in some state $z \in B$ and that I took action $a$, and that means the inputs to the function $f$ must have been $(z, a)$ for some $z \in B$. So, before observing the sensor value $s \in S$, I knew that the new state and the sensor value together had to be in the set $\{\, (z', s) \in Z \times S \mid \exists\, z \in B : (z', s) = f(z, a) \,\}$. But in fact I know the value of $s$, and so I can deduce that the new state of the environment must be within the set

$$\mathsf{update}(B, a, s) \coloneqq \{\, z' \in Z \mid \exists\, z \in B : (z', s) = f(z, a) \,\}, \quad (5)$$

where $s$ is the known sensor value."

Equation (5) expresses a kind of belief updating, somewhat akin to Bayesian filtering, but in a possibilistic rather than probabilistic framework. $\mathsf{update}(B, a, s)$ is the posterior beliefs about the next environment state, given prior beliefs $B$ about the current environment state and data about the action taken and sensor value received, $a$ and $s$.

Before turning this into a formal definition of consistency for an interpretation map, we will allow our imaginary person one extra step in their reasoning: they can decide to forget information. Although the ideal posterior is given by $\mathsf{update}(B, a, s)$, we will allow the person to adopt any posterior $C$ as long as $\mathsf{update}(B, a, s) \subseteq C$. The set $C$ can contain less information than $\mathsf{update}(B, a, s)$, in the sense that it puts less constraint on what the environment's state might be. This 'forgetting' step might be a pragmatic choice if the more specific information is not needed any more to achieve a task that the person is trying to do.

Returning to our goal of attributing beliefs to a Moore machine, the idea is that as the agent takes actions and receives sensor values, the image of $\psi$ should change over time in the same way that a person's beliefs would, if they were reasoning in the way described above. This leads to the following definition:

**Definition 3.1.** Given the agent $(X, r, u)$, a model in the form of a Mealy machine $(Z, f)$ and a function $\psi : X \to \mathcal{P}(Z)$, we say $\psi$ is a *consistent belief map* if for every $x \in X$ and every $s \in S$ we have $\mathsf{update}(\psi(x), r(x), s) \subseteq \psi(u(x, s))$. That is,

$$\{\, z' \in Z \mid \exists\, z \in \psi(x) : (z', s) = f(z, r(x)) \,\} \subseteq \psi(u(x, s)). \quad (6)$$

If $\psi$ is a consistent belief map then we say that $(X, r, u)$ *can be interpreted as having model* $(Z, f)$, since the beliefs $\psi$ assigns to $(X, r, u)$ update in a way consistent with $(Z, f)$. The idea is that $\psi(x)$ represents the beliefs that the observer attributes to the agent when it is in state $x$. For these beliefs to update consistently over time, the beliefs attributed to $x'$ must behave like the posterior beliefs of the person in the informal argument above, i.e. they must be a superset of $\mathsf{update}(B, a, s)$, where $B = \psi(x)$ is the prior beliefs, $a = r(x)$ is the action taken and $s$ is the sensor value received.[2]

We now show that there is a correspondence between belief interpretations (that is, consistent belief maps) and forward-closed sets. Given the coupled system from definition 2.3 (which involves the 'true' environment $(Y, e)$) and a forward-closed set $R$, we can define a map $\psi : X \to \mathcal{P}(Y)$ by

$$\psi(x) \coloneqq \{\, y \in Y \mid (x, y) \in R \,\}. \quad (7)$$

In the following lemma, we show that eq. (7) defines a consistent belief map, where the model $(Z, f)$ is the same as the true environment $(Y, e)$. This is the main technical result needed to prove our "good regulator theorem" for embodied agents (the upcoming theorem 3.4).

**Lemma 3.2.** *Consider an agent $(X, r, u)$ and the environment $(Y, e)$ together with a subset $R \subseteq X \times Y$. We have that $R$ is forward-closed if and only if $\psi : X \to \mathcal{P}(Y)$ as defined by eq.* (7) *is a consistent belief map as per definition 3.1, with model $(Z, f) = (Y, e)$.*

*Proof.* We express the proof as a chain of inferences.

$$\begin{aligned}
& R \text{ is forward-closed} \\
\iff & \text{for every } (x, y) \in R \text{ we have } (u(x, s), y') \in R, \\
& \quad \text{where } (y', s) = e(y, r(x)) \\
\iff & \text{for every } x \in X, y \in \psi(x) \text{ we have}[3] \\
& \quad y' \in \psi(u(x, s)), \text{ where } (y', s) = e(y, r(x)) \\
& \quad \text{(by eq. (7))} \\
\iff & \text{for every } x \in X, y \in \psi(x), s \in S \text{ we have} \\
& \quad \{\, y' \in Y \mid (y', s) = e(y, r(x)) \,\} \subseteq \psi(u(x, s)) \\
\iff & \text{for every } x \in X, s \in S \text{ we have} \\
& \quad \bigcup_{y \in \psi(x)} \{y' \in Y \mid (y', s) = e(y, r(x))\} \subseteq \psi(u(x, s)),
\end{aligned}$$

where the left-hand side of the last inequality is equal to

$$\{y' \in Y \mid \exists y \in \psi(x) : (y', s) = e(y, r(x))\} = \mathsf{update}(\psi(x), r(x), s). \qquad \square$$

[2] It is possible for the left-hand side of eq. (6) to be the empty set, which means that $s$ is a "subjectively impossible" sensor value, as considered in (Virgo et al., 2021). That is, it can't occur, according to the agent's prior beliefs and model. In this case eq. (6) is always satisfied, so it imposes no constraint on the posterior beliefs. If $B$ is non-empty then there is always at least one $s \in S$ such that $\mathsf{update}(B, a, s)$ is non-empty, i.e. there is always at least one subjectively possible sensor value.

[3] Note when $\psi(x) = \emptyset$, universal quantification over it is always 'vacuously' true.

Lemma 3.2 establishes a correspondence between forward-closed sets and consistent belief maps. However, we don't *just* want to interpret the agent as having beliefs about its environment, we also want to interpret it as taking actions that are consistent with its beliefs. To do that we need to reason about what the agent *wants* (or rather what the observer chooses to interpret it as wanting), and for that reason we introduce a second function, $\phi : X \to \mathcal{P}(Z)$, called the *normative map*. Unlike the belief map $\psi$, the map $\phi$ doesn't have to obey any consistency condition.

The idea is that if the agent is in state $x \in X$ then its goals are satisfied as long as the environment is in one of the states in $\phi(x) \subseteq Z$. If this ever fails to be the case then the agent has failed at its task. The normative map will end up relating to $G$ in the same way that the belief map relates to $R$.

It is possible for $\phi(x)$ to be the empty set, which means that the agent never wants to enter state $x$, regardless of the environment's state. It is also possible that $\phi(x) = Z$, in which case the agent's goal is satisfied when the agent is in state $x$, regardless of the state of the environment.

The following definition summarises this "subjective" notion of regulation, in which the agent should satisfy its own subjective goals (given by $\phi$) with respect to its own beliefs about its assumed environment (given by $\psi$). This is subjective in that it's regulation "from the agent's point of view," as attributed to it by an observer.

**Definition 3.3.** Given the agent $(X, r, u)$, a model $(Z, f)$, a consistent belief map $\psi : X \to \mathcal{P}(Z)$ and a normative map $\phi : X \to \mathcal{P}(Z)$, we say the agent is a *subjective good regulator* if $(i)$ we have $\psi(x) \subseteq \phi(x)$ for all $x \in X$, and $(ii)$ there exists $x_0 \in X$ such that $\psi(x_0)$ is non-empty.

Thus, firstly, there should be some state $x_0$ in which the agent should have consistent beliefs (i.e. $\psi(x_0)$ is non-empty) and it should believe that its goal is currently satisfied when it's in that state, i.e. $\psi(x_0) \subseteq \phi(x_0)$. If this were not the case the agent would already have failed in its task, and hence would not be a good regulator from its own point of view. We can assume the agent starts in state $x_0$.

In addition to this, requiring that $\psi$ be a consistent belief interpretation implies that given the action $a = r(x)$ that the agent takes and any sensor value $s \in S$ that can occur according to the agent's model, its posterior beliefs will also be such that its goal is satisfied. We can interpret this as saying that the agent will always take an action such that, from its subjective point of view (as attributed by the observer) its goal will still be satisfied in the next time step—and hence also in all future time steps.

There may be states $x$ where the agent's beliefs are inconsistent (i.e. such that $\psi(x) = \emptyset$). This doesn't cause problems because such states can never be entered, as long as the sensor values the agent receives are subjectively possible ones (i.e. possible according to its beliefs). For states with inconsistent beliefs, the condition $\psi(x) \subseteq \phi(x)$ is satisfied automatically.

### 3.1 Every good regulator has a model

As mentioned previously, the interpretation maps $\psi : X \to \mathcal{P}(Z)$ and $\phi : X \to \mathcal{P}(Z)$ are really just the regulating set $R \subseteq X \times Y$ and the good set $G \subseteq X \times Y$ in disguise. Given $G$ and $R$, we can set $(Z, f) = (Y, e)$ and define $\psi(x) = \{\, y \in Y \mid (x, y) \in R \,\}$ as in eq. (7), along with

$$\phi(x) \coloneqq \{\, y \in Y \mid (x, y) \in G \,\}. \tag{8}$$

We claim that with $\phi$ and $\psi$ defined this way, our two definitions of regulation are equivalent, i.e. that a Moore machine is a good regulator of a system if and only if it is a subjective good regulator of that system. If this is so then we have a theorem along the lines of "every good regulator has a model," since being a subjective good regulator involves having beliefs about the environment, which are updated in a similar way to a Bayesian model.

This claim is the subject of our main theorem, whose proof is by now rather straightforward:

**Theorem 3.4.** *Consider the agent $(X, r, u)$ and the environment $(Y, e)$, together with sets $G \subseteq X \times Y$ and $R \subseteq X \times Y$. Then the agent is a good regulator in the sense of definition 2.5, with good set $G$ and regulating set $R$, if and only if it is a subjective good regulator in the sense of definition 3.3, with model $(Z, f) = (Y, e)$, belief map $\psi$ given by eq.* (7) *and normative map $\phi$ given by eq.* (8)*.*

*Proof.* We only need to note that the three components of the two definitions correspond to each other:

1. $R$ is forward-closed if and only if $\psi$ is a consistent belief interpretation. (This was shown in lemma 3.2.)
2. $R$ is non-empty if and only if there exists $x \in X$ such that $\psi(x)$ is non-empty, since $R = \{(x, y) \mid x \in X, y \in \psi(x)\}$.
3. $R \subseteq G$ if and only if for every $x \in X$ we have $\psi(x) \subseteq \phi(x)$. This follows from the definitions of $\psi$ and $\phi$. □

Our theorem only says that the agent *can* be interpreted such that the model $(Z, f)$ is identical to the 'true' environment $(Y, e)$, not that it must be. It might be that a different interpretation with a simpler model would make more pragmatic sense, as in the example in the following section.

In fact the theorem doesn't require $(Y, e)$ to be the true environment either. It only says that if the agent can perform regulation when placed in environment $(Y, e)$, then it admits an interpretation whose model is $(Y, e)$. There is nothing to stop $(Y, e)$ from itself being a counterfactual that the observer is considering, rather than the true environment in which the agent is situated.

### 3.2 . . . but some models are trivial

Conant and Ashby concluded from their theorem that because every good regulator must have a model, there is no point in trying to design regulators in any way other than first

building a model. Can we draw a similar conclusion from our work? We think not, and to see why we propose the following informal examples.

Consider a doorstop, whose job is to hold doors open when humans desire it. Despite having no state and no dynamics, the doorstop 'achieves its goal' by merely existing, and thus, intuitively, it doesn't seem to make sense to interpret it as an agent with a model. Nevertheless our theorem says it can be so interpreted; how can we make sense of this?

More formally, consider the case where the state space $X$ of the agent and the spaces of sensor values and actions $S$ and $A$ have only one element each, so that effectively the agent has no state besides the state of existing. The environment $(Y, e)$ can still have nontrivial dynamics, and the good set $G$ and regulating set $R$ can be arbitrarily complicated.

Because $X$ has only one element, say $X = \{*\}$, we have $X \times Y \cong Y$. For this reason we can think of $G$ and $R$ as if they are subsets of $Y$ instead of $X \times Y$. It's then not hard to see that $\phi(*)$ must essentially just be $G$ and similarly $\psi(*)$ is $R$. Because there is only one $* \in X$ the beliefs $\psi(*)$ don't change over time. This is because the consistent updating that our theorem attributes to the agent is of the form "I believe the system's state is in $R$ (hence also in $G$) and I know that if it's in $R$ now then whatever state it's in on the next time step will also be in $R$, so it's consistent to also adopt $R$ as my posterior belief."

This is a trivial interpretation (cf. the trivial models of Baltieri et al., 2025) in that although the model $(Z, f)$ might be complicated and $\phi(*)$ and $\psi(*)$ could be complicated subsets of it, they don't depend in a non-trivial way on the agent's state. Because of this, although our theorem says the doorstop can be interpreted this way, it isn't necessarily a good idea to take this interpretation seriously.

This will not always be the case. If we were to model a detective, who searches for clues, speaks to witnesses etc., eliminating possibilities until they can identify the culprit, then we would expect the model our theorem attributes to genuinely reflect the detective's knowledge about the case, updating in a nontrivial way according to what they find out. The nontrivial interpretation arises because the detective is highly *intertwined* with their environment, solving their task in a way that relies on complex feedbacks that involve both the detective's internal state and that of their environment.

Many systems of interest will lie between these two extremes, becoming intertwined with their environment to some extent while also relying on its dynamics (including the dynamics of their embodiment within it); such examples are common in ALIFE (e.g. Braitenberg, 1986; Beer, 2003).

We believe there is much more that can be said about intertwinedness and non-triviality. For now we simply conclude that *any* theorem with a statement along the lines of "every good regulator has a model" must somehow account for all these examples, and ours does this by allowing such systems to have models that can be more or less trivial.

# 4 Conclusions and Discussion

We have proved a theorem whose statement can be read as "every good regulator must have a model," or more precisely, every good regulator can be interpreted as having a model. This notion of admitting an interpretation is more sophisticated than Conant and Ashby's notion of "being a model" in that it involves the concept of belief updating. Using this different concept of model allows our theorem to apply to *all* good regulators, without the additional assumptions that Conant and Ashby and their successors have needed to make.

Our theory also has some resemblance to some versions of the free energy principle (FEP). In particular the versions described in (Friston, 2019; Da Costa et al., 2021) involve a "synchronisation map" that has some similarity to our $\psi$, while the overall argument has some qualitative resemblance to our lemma 3.2: the synchronisation map arises from a stationary state in a somewhat similar fashion to the way our $\psi$ arises from a forward-closed set. Indeed, the current work arose in part from an effort to remove some of the approximations and technical assumptions behind (Friston, 2019) in order to understand what it's really saying. We don't know whether our result relates to the FEP in a precise way.

We have emphasised that in order to interpret a system as doing regulation (whether of the objective or subjective kind), an observer must make a series of choices. These are, roughly, $(i)$ choosing which system to treat as an agent and where to draw its boundary; $(ii)$ choosing what the system is to be interpreted as trying to do; and $(iii)$ choosing *how* it is to be interpreted as achieving its task. None of these seem to be inherent properties of a system, so both the observer and the system seem to play unavoidable roles.

One might, however, try to eliminate the observer after all. One way to do this is to study *every* possible division of a system into agent and environment, *every* possible goal and *every* valid way to achieve them. The toolkit of category theory would be well suited to this approach. Another approach would be to evaluate such choices numerically, giving metrics by which one could be considered better than another.

Alternatively, one might seek biological arguments to regard some choices as more fundamental than others. This approach is taken by autopoietic theory and enactivism, which tend to emphasise the importance of metabolic self-maintenance over other goals that might be attributed to a system. Perhaps this can be formalised, with the good set consisting of all those states in which the system still exists.

We allowed the possibility that the agent can't fully observe its environment; this is the situation that any embodied agent finds itself in. As a consequence, the model our theorem attributes to the agent includes its uncertainty about the environment's unknown state. Our result hints that the notion of model can be rehabilitated even within a purely dynamics-based, behavioural approach: models and dynamical coupling appear to be two sides of the same coin.

## 5 Acknowledgements

The authors thank Simon McGregor and Inman Harvey for productive discussions that influenced this work.

This paper was made possible through the support of Grants 62229 (supporting Virgo and Biehl) and 62828 (supporting Biehl) 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.

Manuel Baltieri is supported by JST, Moonshot R&D Grant Number JPMJMS2012.

Matteo Capucci is an Independent Researcher funded by the Advanced Research + Innovation Agency (ARIA).

## References

Ashby, W. R. (1960). *Design for a brain*. Wiley New York.

Baez, J. (2016). The internal model principle. Blog post on 'Azimuth' `https://johncarlosbaez.wordpress.com/2016/01/27/the-good-regulator-theorem/`. Accessed: 2025-03-26.

Baez, J. C. and Stay, M. (2010). Physics, topology, logic and computation: a Rosetta Stone. In *New structures for physics*, pages 95–172. Springer.

Baltieri, M., Biehl, M., Capucci, M., and Virgo, N. (2025). A Bayesian interpretation of the internal model principle. arXiv preprint 2503.00511.

Beer, R. D. (1995). A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence*, 72(1-2):173–215.

Beer, R. D. (1997). The dynamics of adaptive behavior: A research program. *Robotics and Autonomous Systems*, 20(2-4):257–289.

Beer, R. D. (2003). The dynamics of active categorical perception in an evolved model agent. *Adaptive Behavior*, 11(4):209–243.

Beer, R. D. (2004). Autopoiesis and Cognition in the Game of Life. *Artificial Life*, 10(3):309–326.

Beer, R. D., McShaffrey, C., and Gaul, T. M. (2024). Deriving the intrinsic viability constraint of an emergent individual from first principles. In *The 2024 Conference on Artificial Life*, ALIFE 2024. MIT Press.

Bertschinger, N., Olbrich, E., Ay, N., and Jost, J. (2006). Information and closure in systems theory. In *Proceedings of the 7th German Workshop on Artificial Life*, pages 9–19.

Biehl, M. and Virgo, N. (2023). Interpreting systems as solving POMDPs: A step towards a formal understanding of agency. In Buckley, C. L. e. a., editor, *Active Inference. IWAI 2022. Communications in Computer and Information Science*, pages 16–31. Springer.

Braitenberg, V. (1986). *Vehicles: Experiments in synthetic psychology*. MIT press.

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

Conant, R. C. and Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International journal of systems science*, 1(2):89–97.

Da Costa, L., Friston, K., Heins, C., and Pavliotis, G. A. (2021). Bayesian mechanics for stationary processes. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 477(2256).

De Jaegher, H. and Di Paolo, E. (2007). Participatory sense-making: An enactive approach to social cognition. *Phenomenology and the Cognitive Sciences*, 6(4):485–507.

Dennett, D. (2006). Intentional systems theory. In McLaughlin, B., Beckermann, A., and Walter, S., editors, *The Oxford Handbook of Philosophy of Mind*. Oxford University Press.

Dennett, D. C. (1991). Real patterns. *Journal of Philosophy*, 88(1):27–51.

Di Paolo, E. A. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences*, 4(4):429–452.

Egbert, M. D. and Pérez-Mercader, J. (2018). Methods for measuring viability and evaluating viability indicators. *Artificial life*, 24(02):106–118.

Fong, B. (2013). Causal theories: A categorical perspective on Bayesian networks. arXiv preprint 1301.6201.

Francis, B. A. and Wonham, W. M. (1976). The internal model principle of control theory. *Automatica*, 12(5):457–465.

Friston, K. J. (2019). A free energy principle for a particular physics. *arXiv preprint arXiv:1906.10184*.

Fritz, T. and Klingler, A. (2023). The d-separation criterion in categorical probability. *Journal of Machine Learning Research*, 24(46):1–49.

Jacobs, B. (2020). A channel-based perspective on conjugate priors. *Mathematical Structures in Computer Science*, 30(1):44–61.

Jacobs, B., Kissinger, A., and Zanasi, F. (2019). Causal inference by string diagram surgery. In Bojańczyk, M. and Simpson, A., editors, *Foundations of Software Science and Computation Structures*, pages 313–329, Cham. Springer International Publishing.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45.

Klyubin, A., Polani, D., and Nehaniv, C. (2004). Organization of the information flow in the perception-action loop of evolved agents. In *2004 NASA/DoD Conference on Evolvable Hardware, 2004. Proceedings*, pages 177–180.

Kolchinsky, A. and Wolpert, D. H. (2018). Semantic information, autonomous agency and non-equilibrium statistical physics. *Interface Focus*, 8(6):20180041.

Maturana, H. R. and Varela, F. J. (1980). *Autopoiesis and cognition: The realization of the living*. Springer Science & Business Media.

McGregor, S. (2016). A more basic version of agency? As if! In Tuci, E., Giagkos, A., Wilson, M., and Hallam, J., editors, *From Animals to Animats 14*, pages 183–194, Cham. Springer International Publishing.

McGregor, S. (2017). The Bayesian stance: Equations for 'as-if' sensorimotor agency. *Adaptive Behavior*, 25(2):72–82.

McGregor, S., timorl, and Virgo, N. (2025a). Formalising the intentional stance 1: attributing goals and beliefs to stochastic processes. arXiv preprint 2405.16490.

McGregor, S., timorl, and Virgo, N. (2025b). Formalising the intentional stance 2: a coinductive approach. arXiv preprint 2501.09173.

McShaffrey, C. and Beer, R. D. (2023). Decomposing viability space. In *Artificial Life Conference Proceedings 35*, page 51. MIT Press.

Myers, D. J. (2023). Categorical systems theory. Online book draft `http://davidjaz.com/Papers/DynamicalBook.pdf`. Accessed: 2025-05-07.

Ortega, P. A., Wang, J. X., Rowland, M., Genewein, T., Kurth-Nelson, Z., Pascanu, R., Heess, N., Veness, J., Pritzel, A., Sprechmann, P., Jayakumar, S. M., McGrath, T., Miller, K., Azar, M., Osband, I., Rabinowitz, N., György, A., Chiappa, S., Osindero, S., Teh, Y. W., van Hasselt, H., de Freitas, N., Botvinick, M., and Legg, S. (2019). Meta-learning of sequential strategies. arXiv preprint 1905.03030.

Richens, J., Abel, D., Bellot, A., and Everitt, T. (2025). General agents need world models. arXiv preprint 2506.01622.

Rosas, F. E., Geiger, B. C., Luppi, A. I., Seth, A. K., Polani, D., Gastpar, M., and Mediano, P. A. M. (2024). Software in the natural world: A computational approach to hierarchical emergence. arXiv preprint 2402.09090.

Selinger, P. (2010). A survey of graphical languages for monoidal categories. In *New structures for physics*, pages 289–355. Springer.

Seth, A. K. and Tsakiris, M. (2018). Being a beast machine: The somatic basis of selfhood. *Trends in cognitive sciences*.

Virgo, N., Biehl, M., and McGregor, S. (2021). Interpreting dynamical systems as Bayesian reasoners. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 726–762. Springer.

Wentworth, J. S. (2021). Fixing the good regulator theorem. Post on 'AI Alignment Forum' `https://www.alignmentforum.org/posts/Dx9LoqsEh3gHNJMDk/fixing-the-good-regulator-theorem`, also posted on the 'LessWrong' blog at `https://www.lesswrong.com/posts/Dx9LoqsEh3gHNJMDk/fixing-the-good-regulator-theorem`. Accessed: 2025-03-26.

Zahedi, K., Ay, N., and Der, R. (2010). Higher coordination with less control—a result of information maximization in the sensorimotor loop. *Adaptive Behavior*, 18(3-4):338–355.