

A structural view of predictive coding

A mathematical perspective

Draft

PREDICTIVE PROCESSING FROM A COALGEBRAIC PERSPECTIVE *

Manuel Baltieri

Araya Inc.
Tokyo
{manuel_baltieri}@araya.org

Filippo Torresan

University of Sussex
Brighton
{f.torresan}@sussex.ac.uk

Tomoya Nakai

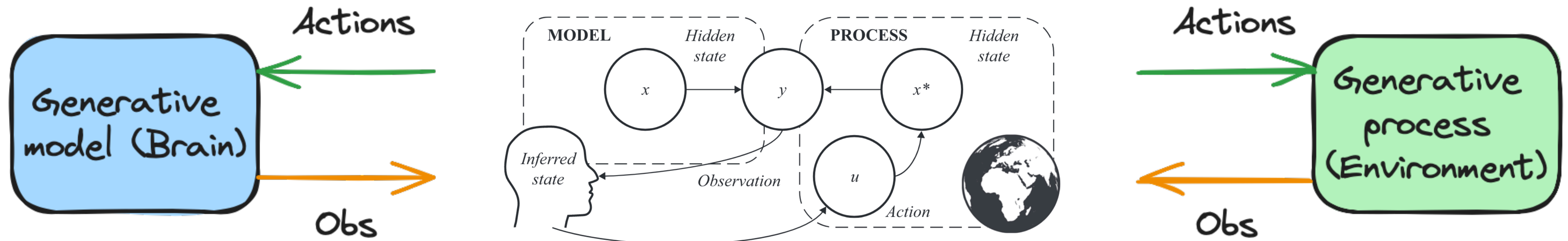
Araya Inc.
Tokyo
{nakai_tomoya}@araya.org

ABSTRACT

Predictive processing posits the brain as a prediction-generating machine. According to this theory, the brain continuously generates and updates generative models of the external world, seen as a generative process producing incoming sensory information that needs to be predicted. On this

Predictive coding under the FEP

Some intuitive ideas



Parr et al. 2023

1. Prediction error minimisation: generative model produces observations consistent generative process
2. Generative model need not be a mirror of the generative process (structure vs. behaviour)

Duality of structure and behaviour

Algebras and coalgebras

Structure vs. Observation

Posted by Emily Riehl

guest post by [Stelios Tsampas](#) and [Amin Karamlou](#)



Today we'll be talking about the theory of universal algebra, and its less well-known counterpart of universal coalgebra. We'll try to convince you that these two frameworks provide us with suitable tools for studying a fundamental duality that arises between *structure* and *behaviour*. Rather than jumping straight into the mathematical details we'll start with a few motivating examples that arise in the setting of functional programming. We'll talk more about the mathematics at play behind the scenes in the second half of this post.

With that our whirlwind tour comes to a close. We've seen how universal algebra gives us tools for exploring the *structure* of things, while universal coalgebra allows us to explore their *behaviour*. Together they gave us a way to rigorously analyse the duality between structure and behaviour. Earlier in the article we made the rather bold claim that this duality transcends the examples we've seen here and goes up all the way to the foundations of thought. We'll end on a similarly dramatic note by giving you a philosophical question to ponder:

Is a "thing" best defined by its constituent parts (structure) or by its observable actions(behaviour).

Algebras (using constructors)

$$\begin{aligned}(3 + 1) * (4 - 2) \\ &= (4) * (2) \\ &= 8\end{aligned}$$

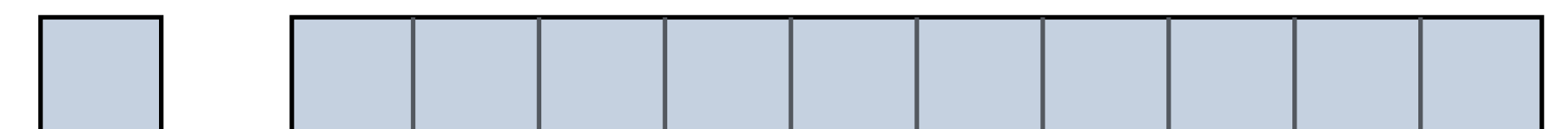
Coalgebras (using destructors)

Stream



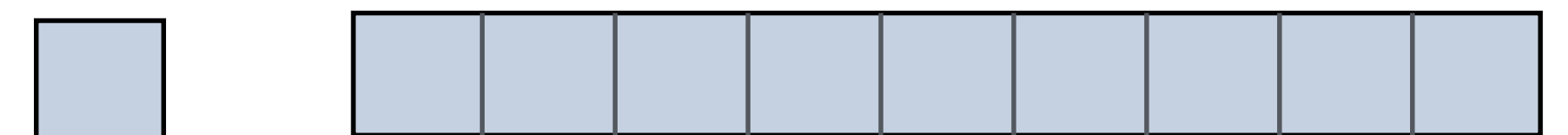
Head

Stream



Head

Stream



A 1/2 new perspective

Let's simplify things

- Core ideas behind the FEP same from “internal model principle” (old)
- Disentangle structure from algorithms and approximations (new)

4.5.1 A Generative Model for Predictive Coding

To motivate the form of generative model used for continuous states, we start with the following pair of equations:

$$\begin{aligned}\dot{x} &= f(x, v) + \omega_x \\ y &= g(x, v) + \omega_y\end{aligned}\tag{4.15}$$

The first of these expresses the evolution of a hidden state over time, according to a deterministic function ($f(x, v)$) and stochastic fluctuations (ω). The second equation expresses the way in which data are generated from the hidden state. In each case, the fluctuations are assumed normally distributed, giving the following probability densities for the dynamics and likelihood:

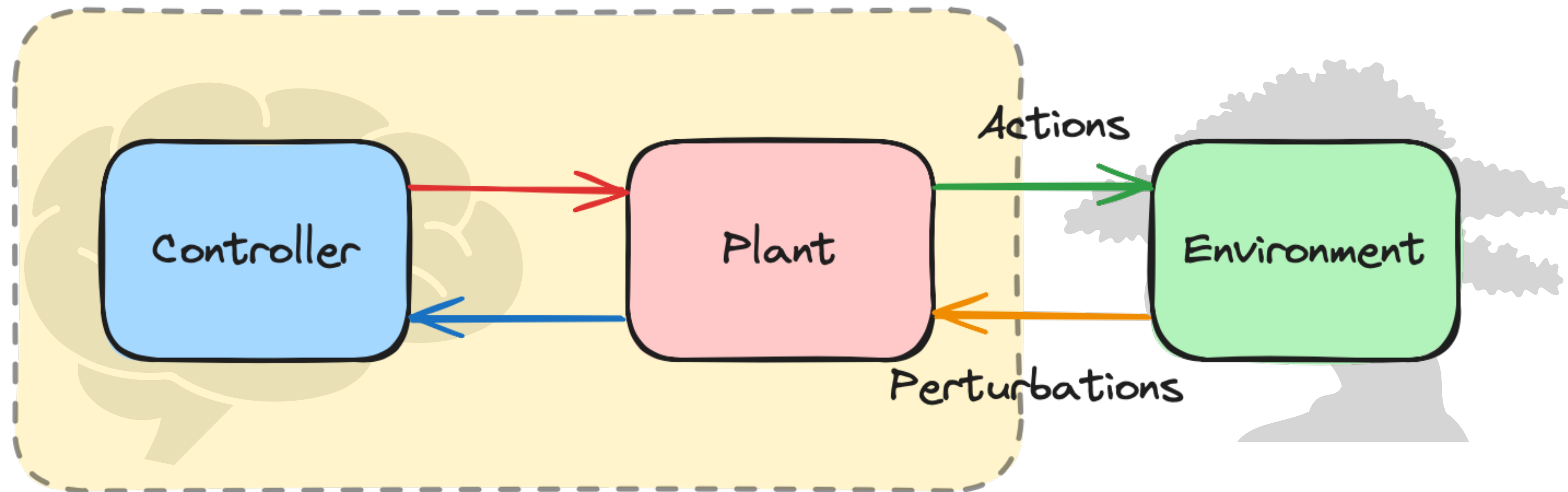
$$\begin{aligned}p(\dot{x}|x, v) &= \mathcal{N}(f(x, v), \Pi_x) \\ p(y|x, v) &= \mathcal{N}(g(x, v), \Pi_y)\end{aligned}\tag{4.16}$$

$$\left. \begin{aligned}D\tilde{x} &= \tilde{f}(\tilde{x}, \tilde{v}) + \tilde{\omega}_x \\ \tilde{y} &= \tilde{g}(\tilde{x}, \tilde{v}) + \tilde{\omega}_y\end{aligned} \right\} \Rightarrow \begin{aligned}p(\tilde{x}|\tilde{v}) &= \mathcal{N}(D \cdot \tilde{f}, \tilde{\Pi}_x) \\ p(\tilde{y}|\tilde{x}, \tilde{v}) &= \mathcal{N}(\tilde{g}, \tilde{\Pi}_y)\end{aligned}\tag{4.18}$$

$$\begin{aligned}F[\mu, y] &= -\ln p(\tilde{y}, \tilde{\mu}_x, \tilde{\mu}_v) \\ &= \frac{1}{2} \tilde{\epsilon} \cdot \tilde{\Pi} \tilde{\epsilon} \\ &= \frac{1}{2} (\tilde{\epsilon}_y \cdot \tilde{\Pi}_y \tilde{\epsilon}_y + \tilde{\epsilon}_x \cdot \tilde{\Pi}_x \tilde{\epsilon}_x + \tilde{\epsilon}_v \cdot \tilde{\Pi}_v \tilde{\epsilon}_v) \\ \tilde{\epsilon} &= \begin{bmatrix} \tilde{\epsilon}_y \\ \tilde{\epsilon}_x \\ \tilde{\epsilon}_v \end{bmatrix} = \begin{bmatrix} \tilde{y} - \tilde{g}(\tilde{\mu}_x, \tilde{\mu}_v) \\ D\tilde{\mu}_x - \tilde{f}(\tilde{\mu}_x, \tilde{\mu}_v) \\ \tilde{\mu}_v - \tilde{\eta} \end{bmatrix} \\ \tilde{\Pi} &= \begin{bmatrix} \tilde{\Pi}_y & & \\ & \tilde{\Pi}_x & \\ & & \tilde{\Pi}_v \end{bmatrix}\end{aligned}\tag{4.19}$$

Meanwhile, in control theory

Control-plant-environment factorisation



Internal

A model of h

1

A Bayesian Interpretation of the Internal Model Principle

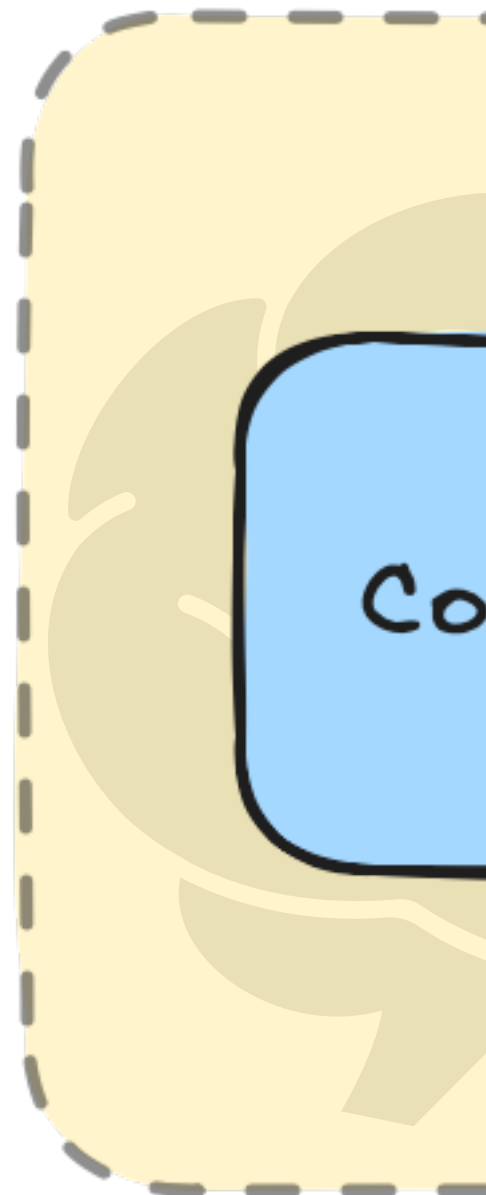
Manuel Baltieri, Araya Inc.[†]

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

arXiv:2503.00511v2 [math.OC] 19 Apr 2025



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].

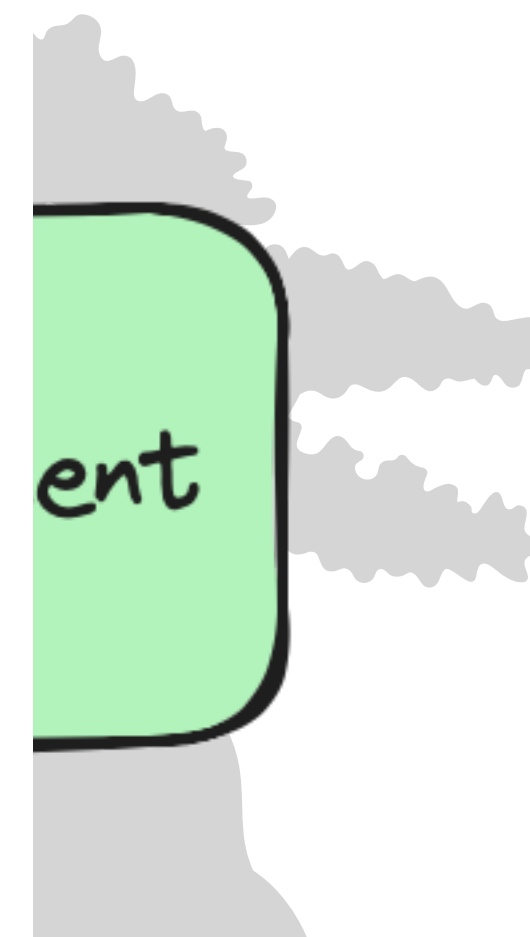
organisms at all scales, including microorganisms such as bacteria [6]–[8].

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

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

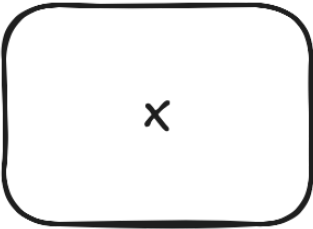
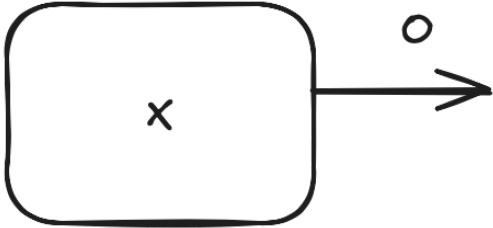
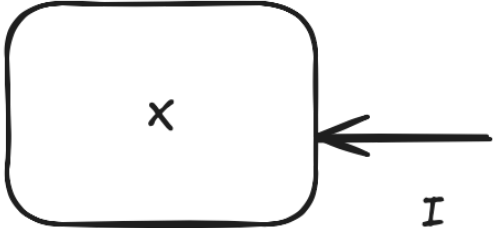
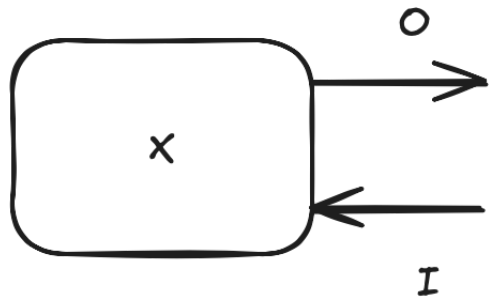
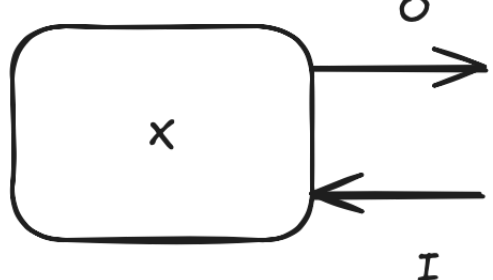
While the label “internal model,” or just “model” is used across different disciplines, it is unclear whether it always refers to the same underlying formal concept. If cognitive scientists propose internal models for the study of cognition, are they referring to the same kind of mathematical objects 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



Abstracting things

Coalgebras as a language for dynamical systems

| | The “standard” way | The coalgebraic way | Graphically (informal) |
|--|---|--|---|
| A (closed) dynamical system | $(X, \alpha : X \rightarrow X)$ | $(X, f : X \rightarrow X)$ |  |
| A dynamical system with outputs | $(X, O, \alpha : X \rightarrow X, \gamma : X \rightarrow O)$ | $(X, f_{\text{Out}} : X \rightarrow X \times O)$ |  |
| A dynamical system with inputs | $(X, I, \beta : X \times I \rightarrow X)$ | $(X, f_{\text{In}} : X \rightarrow X^I)$ |  |
| A dynamical system with inputs&outputs | $(X, I, O, \beta : X \times I \rightarrow X, \gamma : X \rightarrow O)$ | $(X, f_{\text{Moore}} : X \rightarrow X^I \times O)$ |  |
| A probabilistic system with inputs&outputs | $(X, I, O, \beta_P : X \times I \rightarrow P(X), \gamma_P : X \rightarrow P(O))$ | $(X, f_{\text{PrMoore}} : X \rightarrow P(X)^I \times P(O))$ |  |

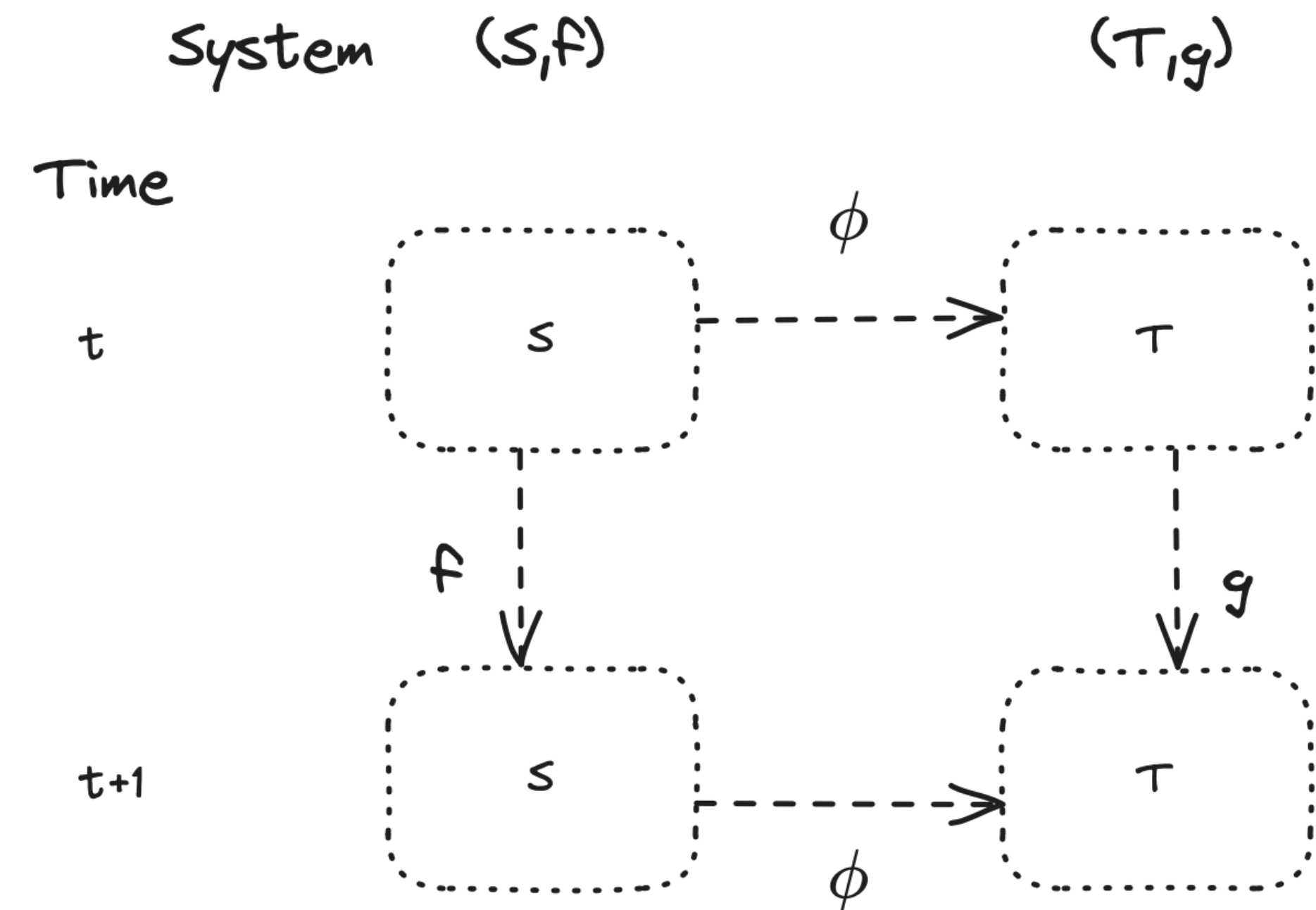
Maps between (closed) systems

Coalgebra (homo)morphisms by example

Take two closed systems, (S, f) and (T, g) . A map between these systems is a function ϕ such that the following diagram commutes

$$\begin{array}{ccc} S & \xrightarrow{\phi} & T \\ f \downarrow & & \downarrow g \\ S & \xrightarrow{\phi} & T \end{array}$$

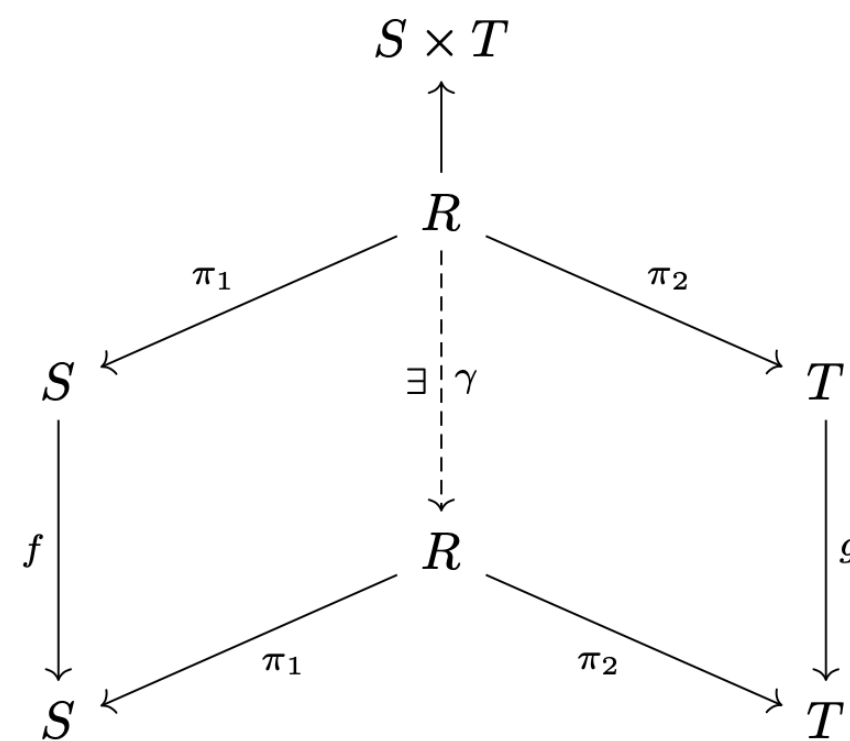
or in other words, if $g(\phi(S)) = \phi(f(S))$.



Behavioural equivalence

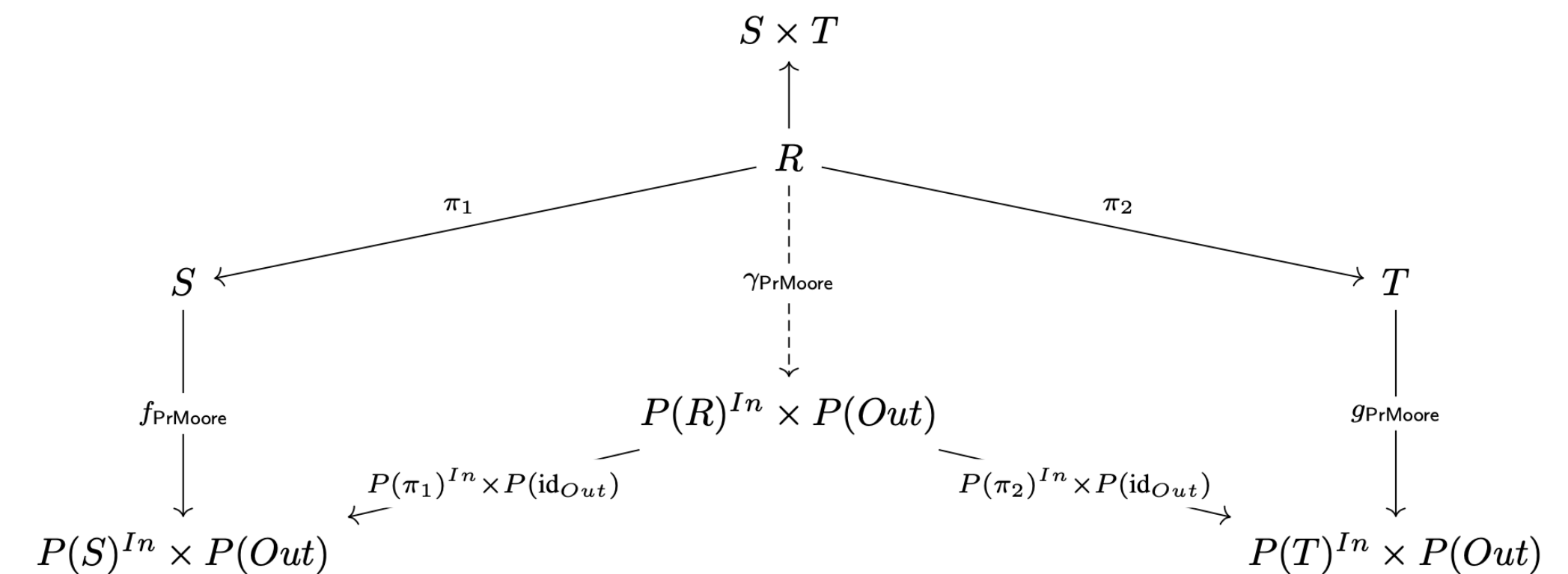
Bisimulations, congruences on *behaviour*, by example

Take two closed systems, (S, f) and (T, g) . A bisimulation between these systems is a relation R such that the following diagram commutes



or in other words, if $g(\pi_2(R)) = \pi_2(\gamma(R))$ and $f(\pi_1(R)) = \pi_1(\gamma(R))$.

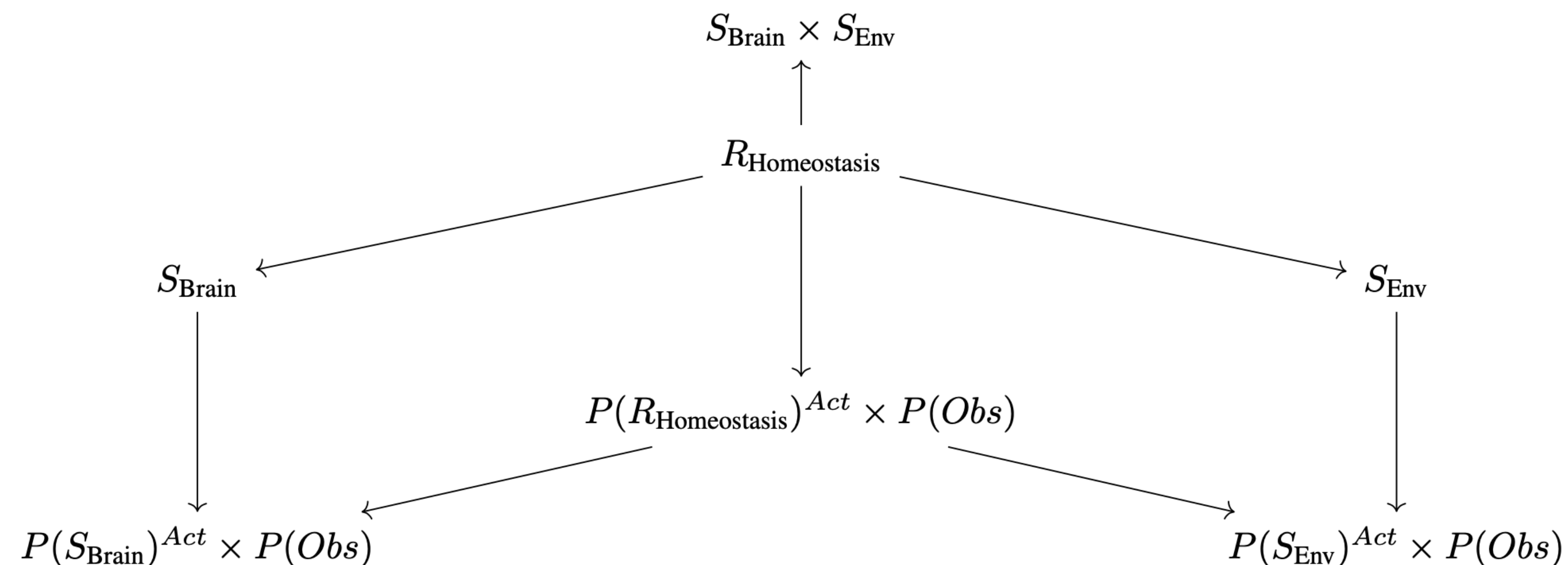
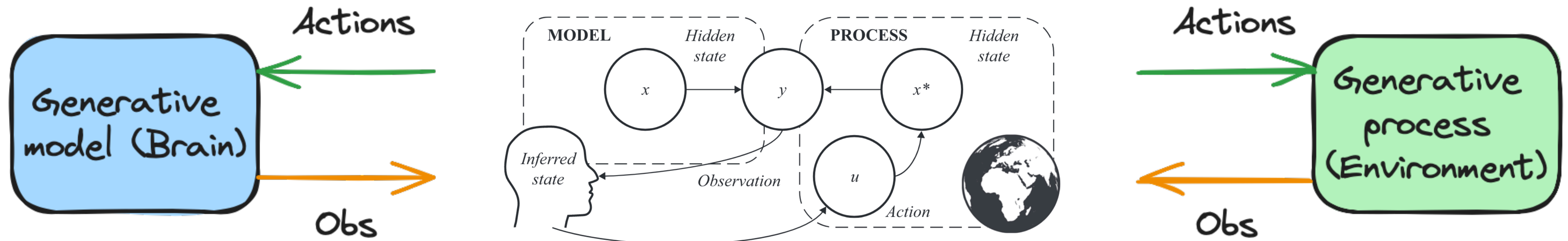
Take two open systems, (S, f_{PrMoore}) and (T, g_{PrMoore}) . A bisimulation between these systems is a relation R such that the following diagram commutes



Maps of systems vs. bisimulations?

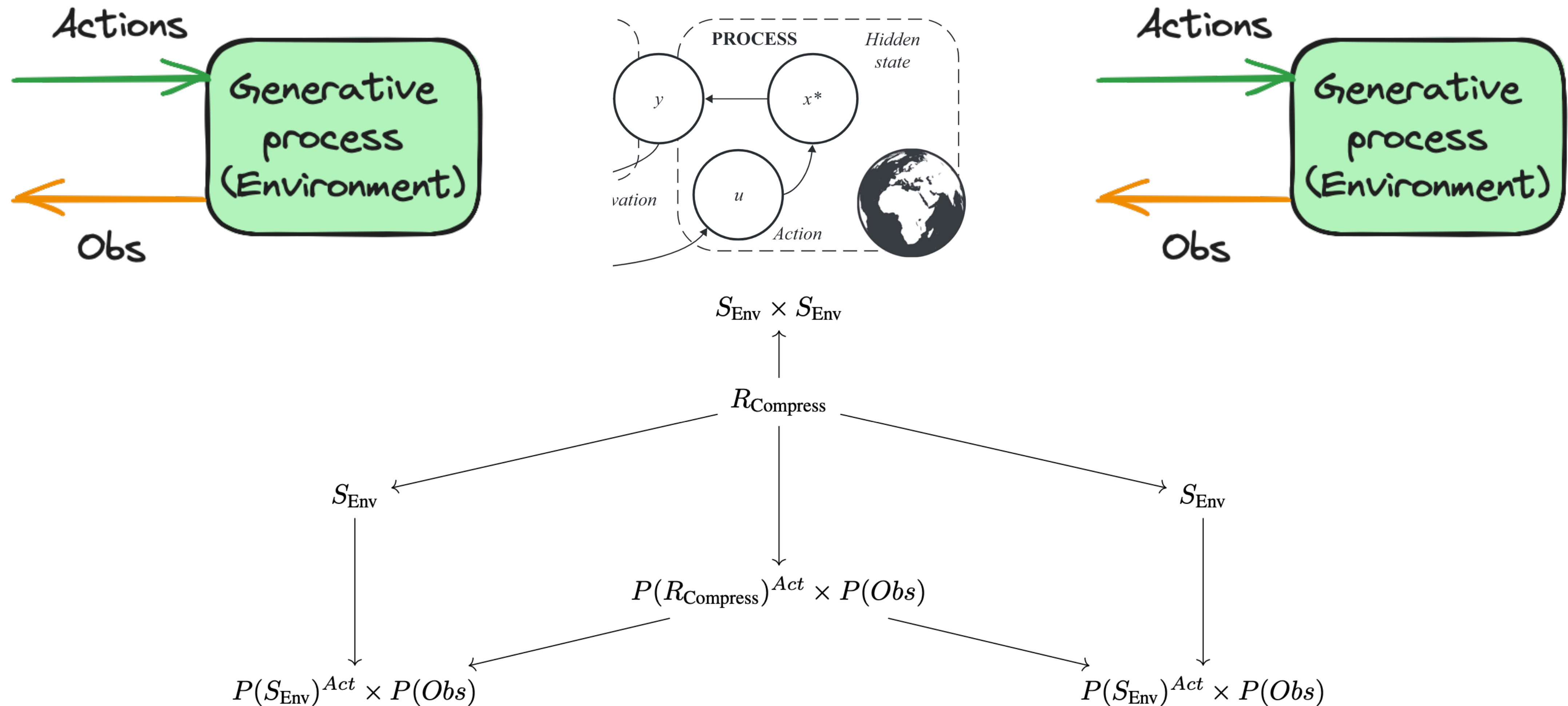
1. The relation between agent and environment

Prediction error minimisation = Agent-environment attunement



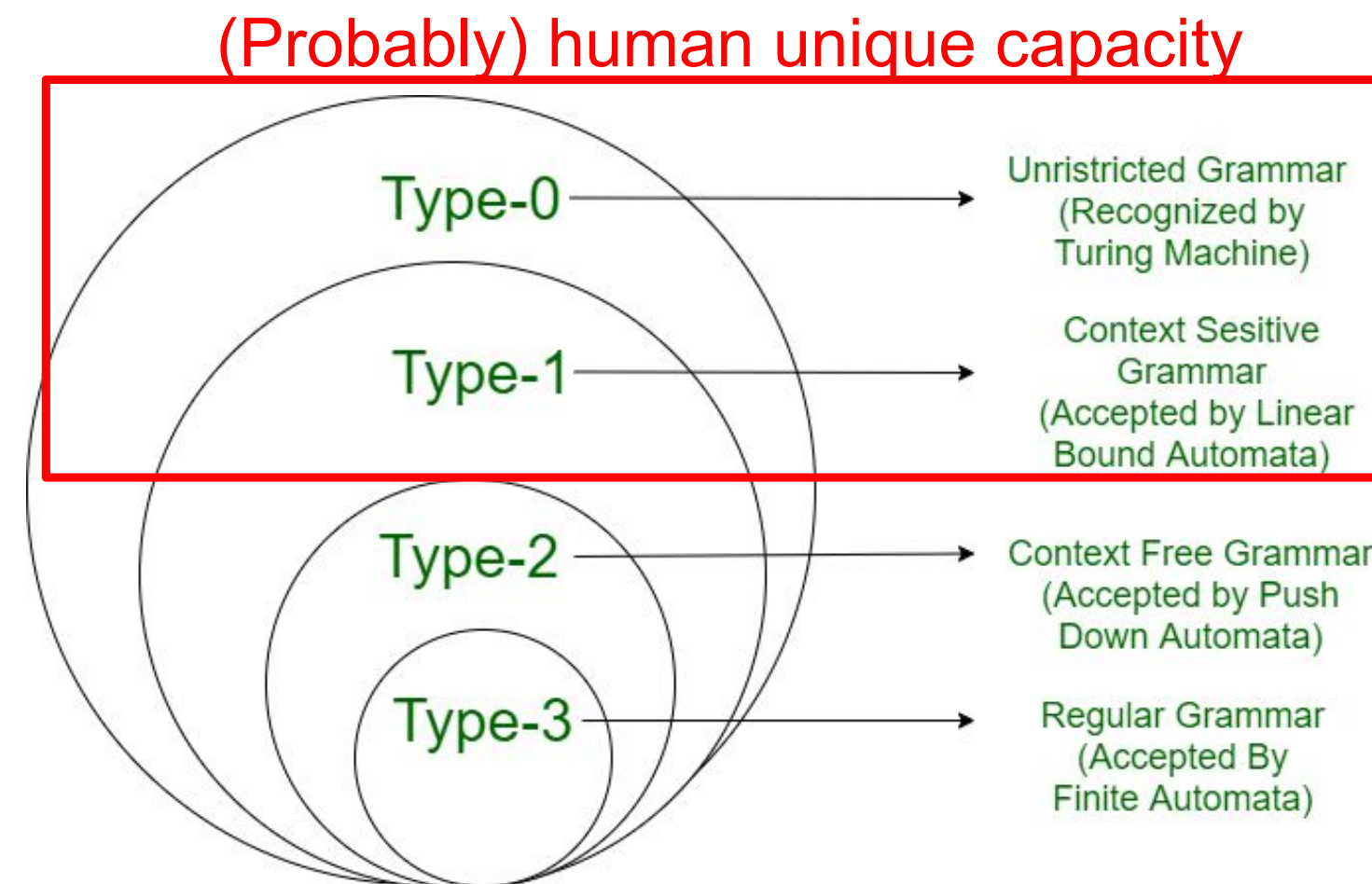
Compressing environments' models

Bisimulation equivalences of environments for a particular goal



Automata theory by changing functors

Background: classical theory of symbol sequences



<https://www.geeksforgeeks.org/chomsky-hierarchy-in-theory-of-computation/>

Examples:

Regular grammar

$\{a^n b^m\}$: aabbbbb

Context free grammar

$\{a^n b^n\}$: aaabbb

Context sensitive grammar

$\{a^n b^n c^n\}$: aaabbbccc

But... Seek for more computational approaches => **Surprisal**

Towards a Coalgebraic Chomsky Hierarchy (Extended Abstract)*

Sergey Goncharov¹, Stefan Milius¹, and Alexandra Silva²

¹ Lehrstuhl für Theoretische Informatik, Friedrich-Alexander-Universität Erlangen-Nürnberg

² Radboud University Nijmegen and Centrum Wiskunde & Informatica, Amsterdam

Abstract. The Chomsky hierarchy plays a prominent role in the foundations of theoretical computer science relating classes of formal languages of primary importance. In this paper we use recent developments on coalgebraic and monad-based semantics to obtain a generic notion of a \mathbb{T} -*automaton*, where \mathbb{T} is a monad, which allows the uniform study of various notions of machines (e.g. finite state machines, multi-stack machines, Turing machines, weighted automata). We use the *generalized powerset construction* to define a generic (trace) semantics for \mathbb{T} -automata, and we show by numerous examples that it correctly instantiates for some known classes of machines/languages captured by the Chomsky hierarchy. Moreover, our approach provides new generic techniques for studying expressivity power of various machine-based models.

1 Introduction

In recent decades much interest has been drawn to studying generic abstraction devices not only formally generalizing various computation models and tools, but also identifying core principles and reasoning patterns behind them. An example of this kind is given by the notion of *computational monad* [21], which made an impact both on the theory of programming (as an organization tool for denotational semantics [17, 23]) and on the

Summary

国立研究開発法人科学技術振興機構 ムーンショット型研究開発事業



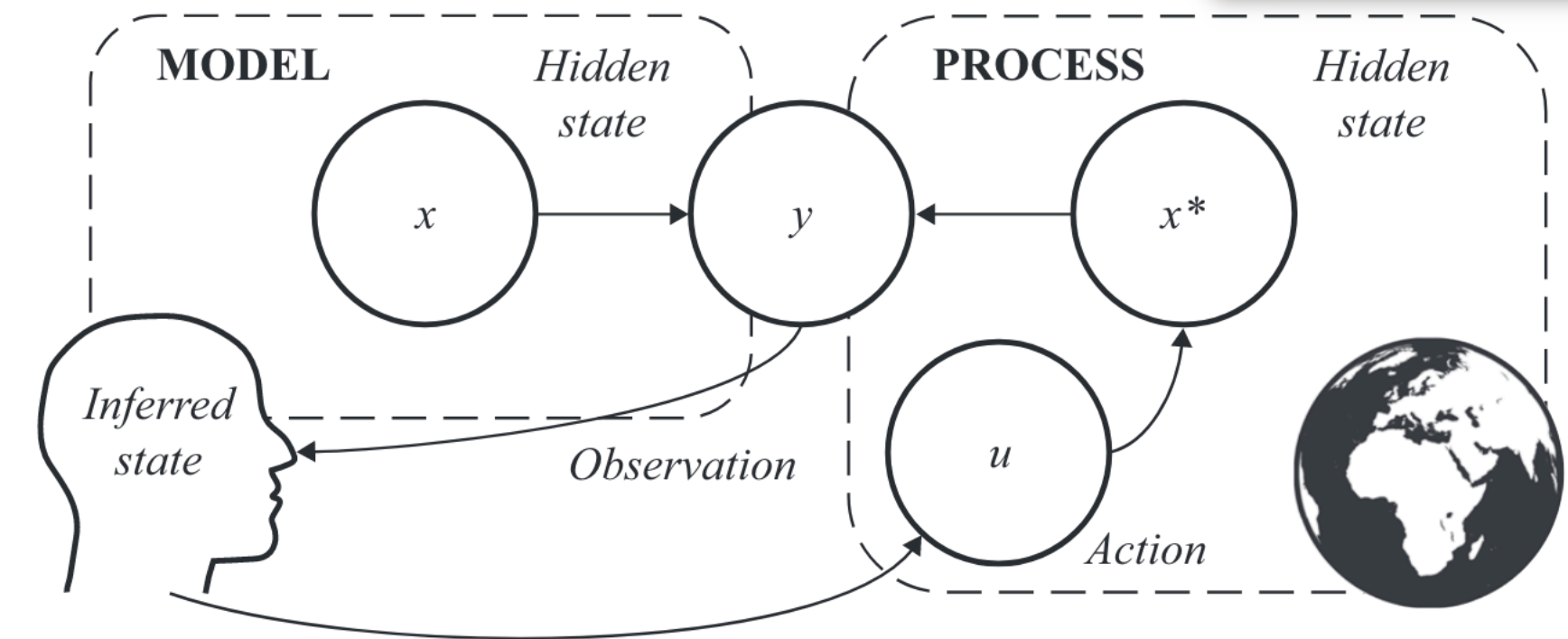
創発的研究支援事業
Fusion Oriented REsearch for disruptive Science and Technology

Using coalgebras to formalise a general treatment of predictive coding under the FEP:

- Prediction error as a bisimulation (equivalence on behaviours)
- Generative model as coarse-grained version of generative process (bisimulation equivalence) or as belief MPD of generative process/POMDP

Advantages:

- We can change functor (discrete probability, power set, tangent, etc.) and obtain automata, continuous-time systems, etc.



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

Fernando E. Rosas, Alexander Boyd, Manuel Baltieri

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

Summary

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

$P(S_{\text{Env}})^{A_t}$

3)

2. “Action-oriented” models

A way to look at compressed models

“World models” meaning “models the environment” is a pretty flashy but bad name

Surely they can’t be about the entire universe dynamics, so what are they talking about?

Action oriented models seem more reasonable (but not formal):

between brain, body, and world. Neural representations, this work has suggested, are not action-neutral mirrors of the world. Instead they are in some deep sense ‘action-oriented’ (Clark 1997, Engel et al. 2013). They are geared to promoting successful, fast, fluent actions and engagements for a creature with specific needs and bodily form. Such representations will be **as minimal as possible, neither encoding nor processing information in costly ways when simpler routines, combined with world-exploiting actions, can do the job.**

Clark 2015

Proposed formalisation: bisimulation equivalences.

These build (dynamical) compressions of environments, with various possible criteria, for instance:

- compression for all possible actions of all possible agents
- compression for all possible actions of a single agent
- compression for all possible actions of a single agent, given the same reward
- compression for the actions of a policy chosen by an agent, given the same reward
- ...

Maps between (open) systems

Coalgebra (homo)morphisms by example

Take two probabilistic dynamical systems, (S, f_{PrMoore}) and (T, g_{PrMoore}) . A map between these systems is a function ϕ such that the following diagram commutes

$$\begin{array}{ccc} S & \xrightarrow{\phi} & T \\ \downarrow f_{\text{PrMoore}} & & \downarrow g_{\text{PrMoore}} \\ P(O) \times P(S)^I & \xrightarrow{P(\text{id}_O) \times P(\phi)^I} & P(O) \times P(T)^I \end{array}$$

or in other words, if $g_{\text{PrMoore}}(\phi(S)) = \phi(f_{\text{PrMoore}}(S))$.

(Same thing as before, but requiring that (T, g_{PrMoore}) 's inputs and outputs are equal to (S, f_{PrMoore}) 's at each time step whenever there's a map between their states that commutes with the systems' dynamics.)