When does a system model another system?

Manuel Baltieri, 6th March 2025, 数理科学と神経科学の異分野連携研究会2025

ARAYA

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

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].

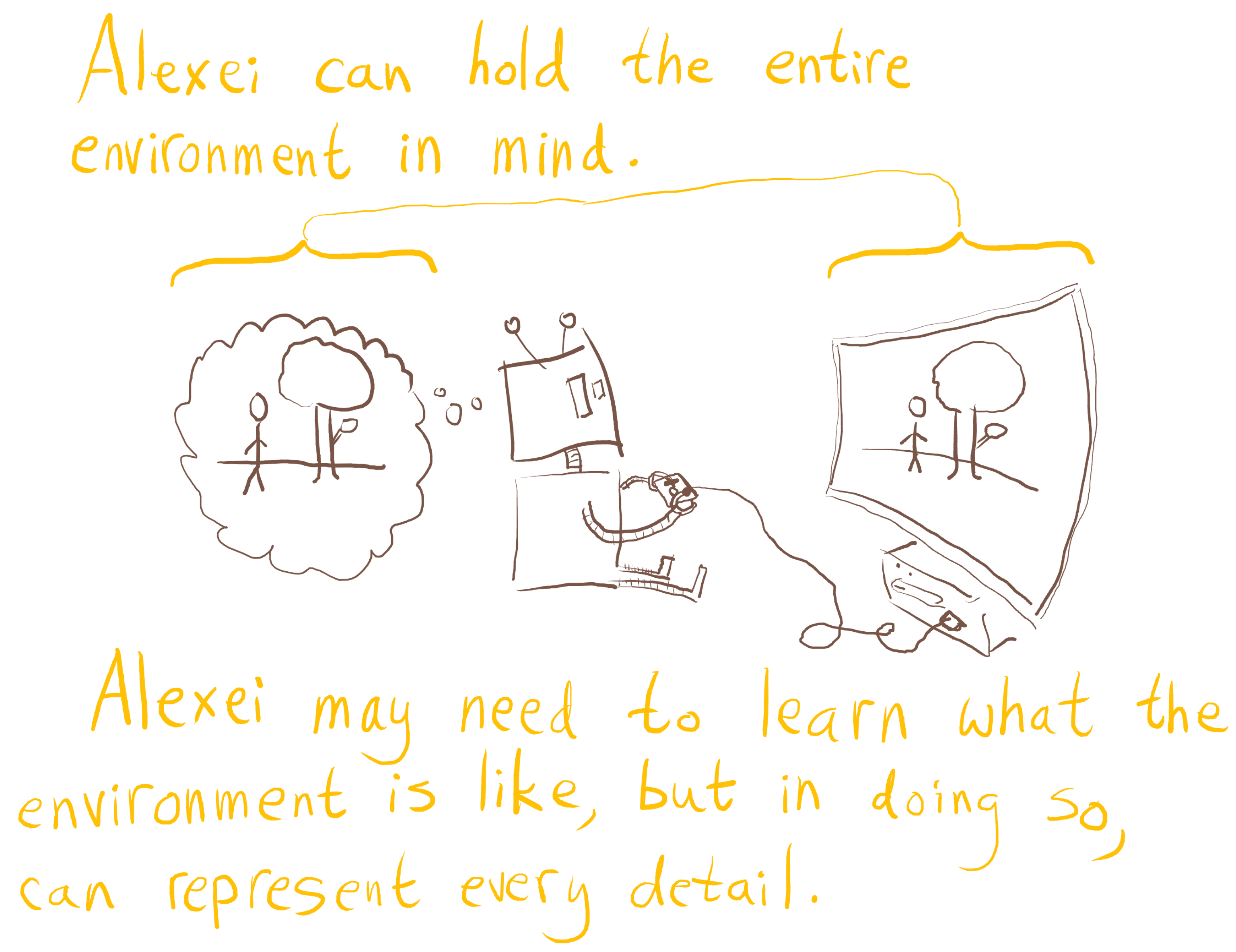
While the label "internal model," or just "model" is used across different disciplines, it is unclear whether it always refers to the same underlying formal concept. If cognitive scientists propose internal models for the study of cognition, are they referring to the same kind of mathematical objects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

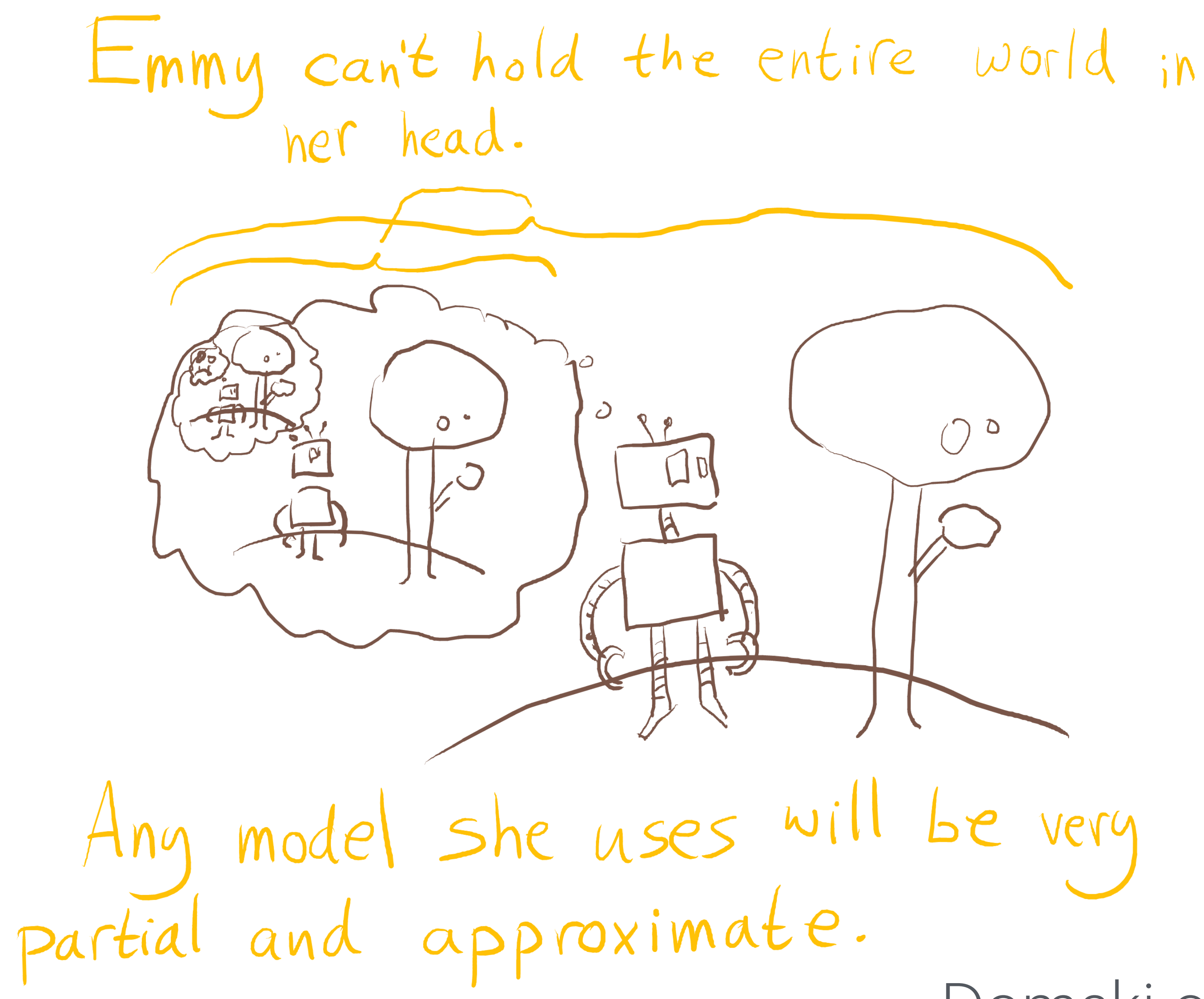
To do so, we structure this work in two main parts. In the first part (Section II), we present the IMP developed by [25]–

Contents

Different flavours of models

- **Internal models in control**
  - Systems, and
  - Models
- **Process theories from categories**
  - Probabilities, possibilities, and
  - Bayes theorem and conjugate priors
- **Take home message: Internal model principle implies a Bayesian filtering interpretation, the converse is not true**





Demski et al. (2018)

The problem

Representations? World models? Internal models?

- “Evidence for neural representations in area XYZ"
- “Perception updates internal representations of the external world”
- “Biological organisms use internal models to navigate dynamic environments”
- “Brains build predictive world models to anticipate future events”
- ...

I have no idea what any of these things mean, mathematically

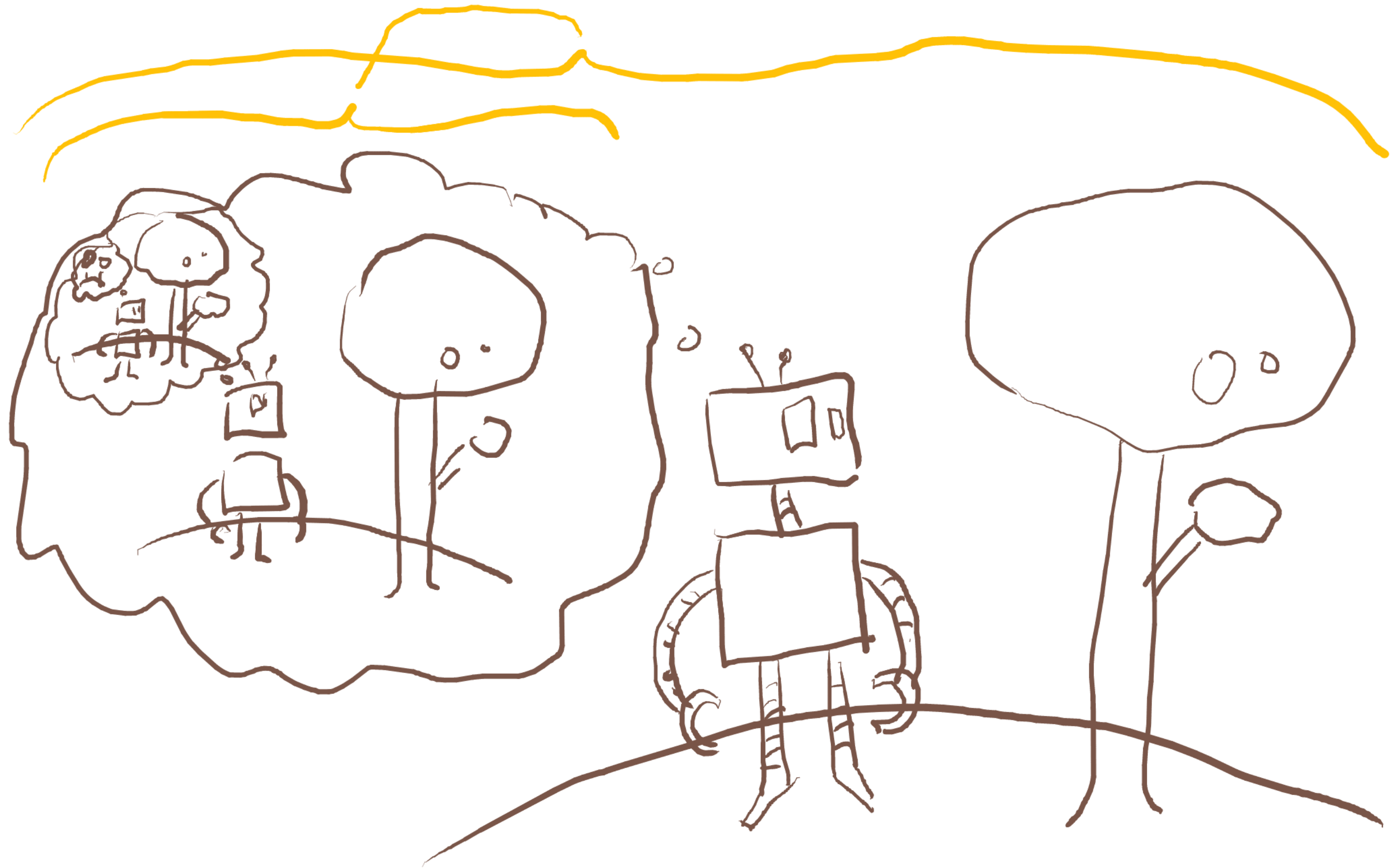
The intuition

“Brains model the environment”

Alexei can hold the entire environment in mind.

Alexei may need to learn what the environment is like, but in doing so, can represent every detail.

Emmy can't hold the entire world in her head.



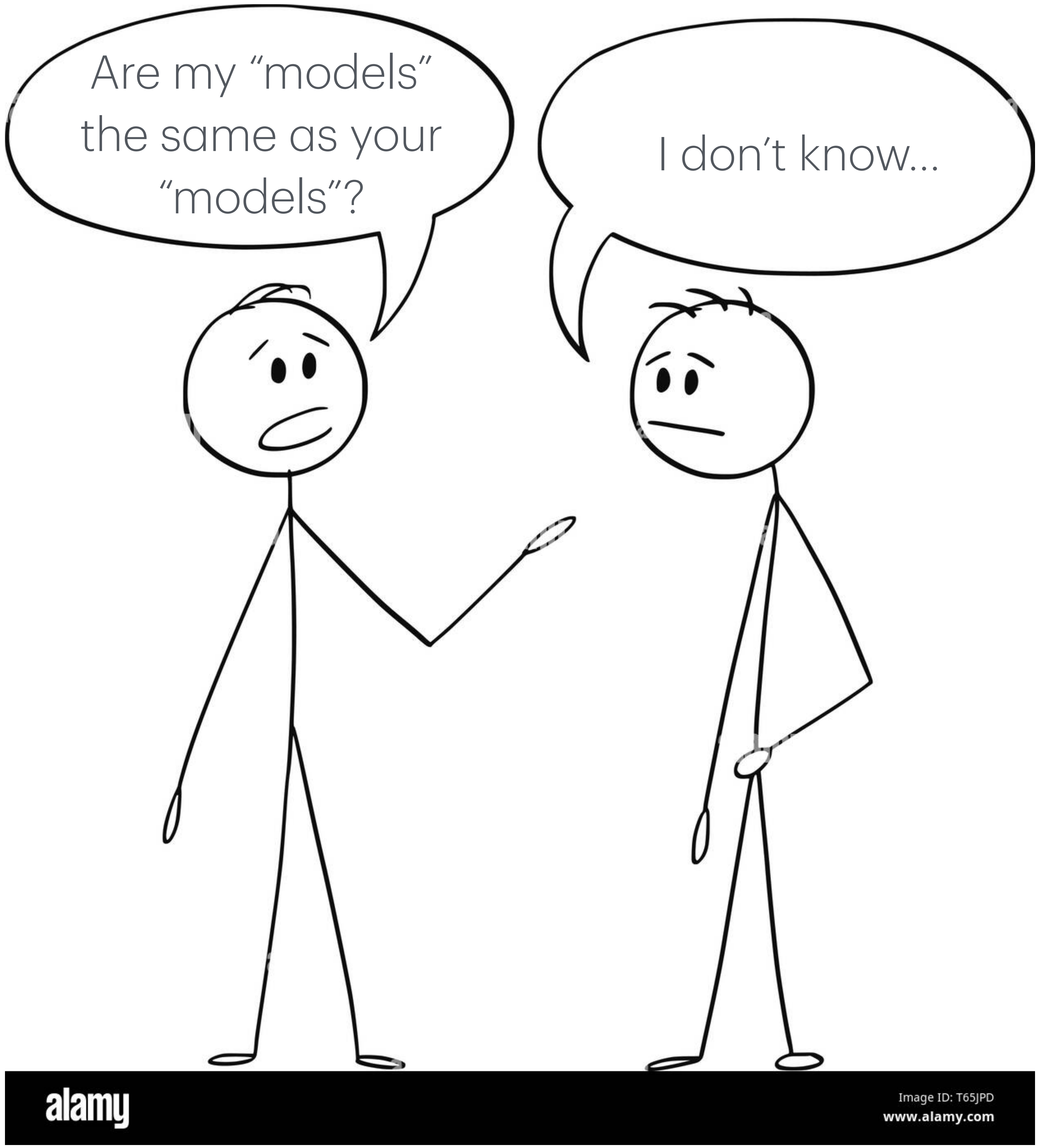
Any model she uses will be very partial and approximate.

Demski et al. (2018)

The literature

Looking in different areas

- Machine/reinforcement learning ("world models")
- Biology ("internal models")
- Cognitive science/philosophy of mind ("internal models", "Bayesian models")
- Neuroscience ("internal models")



The literature

Different definitions

INT. J. SYSTEMS SCI., 1970, VOL. 1, NO. 2, 89–97

Every good regulator of a system must be a model of that system†

ROGER C. CONANT

Department of Information Engineering, University of Illinois, Box 4348, Chicago, Illinois, 60680, U.S.A.

and W. ROSS ASHBY

Biological Computers Laboratory, University of Illinois, Urbana, Illinois 61801, U.S.A.‡

[Received 3 June 1970]

Automatica, Vol. 12, pp. 457–465. Pergamon Press, 1976. Printed in Great Britain

The Internal Model Principle of Control Theory*

B. A. FRANCIS† and W. M. WONHAM‡

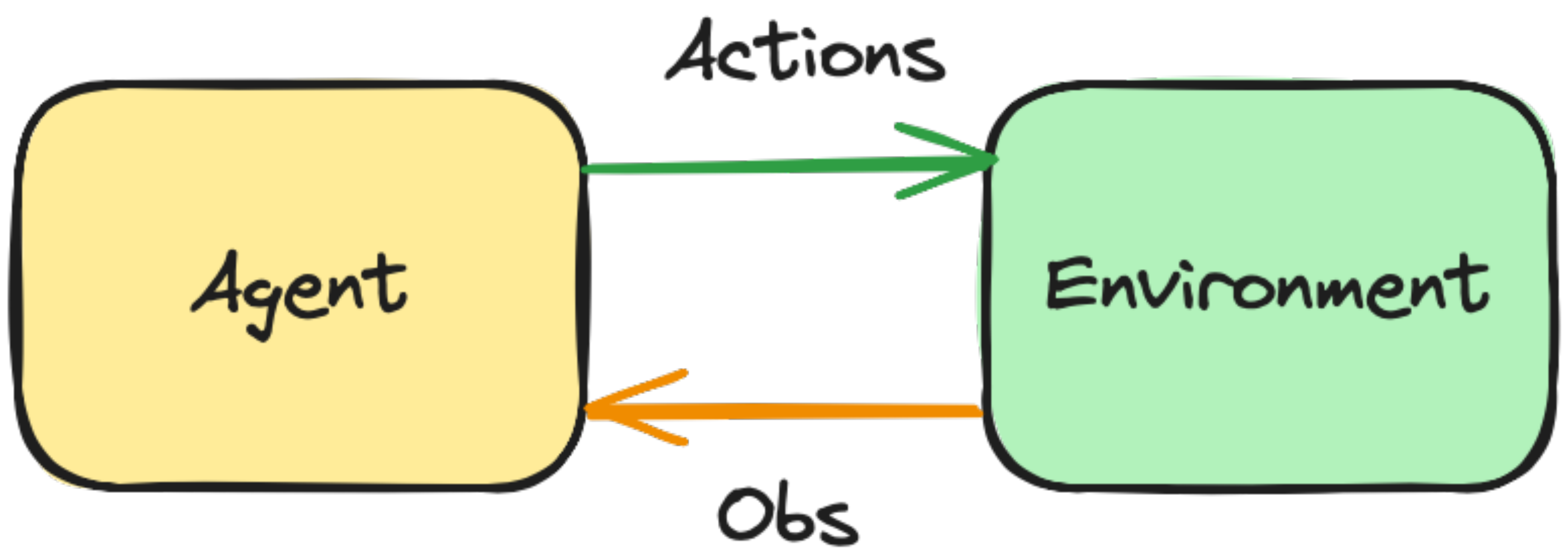
In multivariable servomechanisms designed to accommodate parameter uncertainty, the controller must have special qualitative structural features which may be derived for linear and weakly nonlinear systems.

Summary—The classical regulator problem is posed in the context of linear, time-invariant, finite-dimensional systems with deterministic disturbance and reference signals. Control action is generated by a compensator which is required to provide closed loop stability and output regulation in the face of small variations in certain system parameters. It is shown,

Second, it is to regulate a variable z which is given function of the plant output c and tl reference signal r; typically z may be tl tracking error $r - c$. A plant-compensator comt nation with these two properties is termed

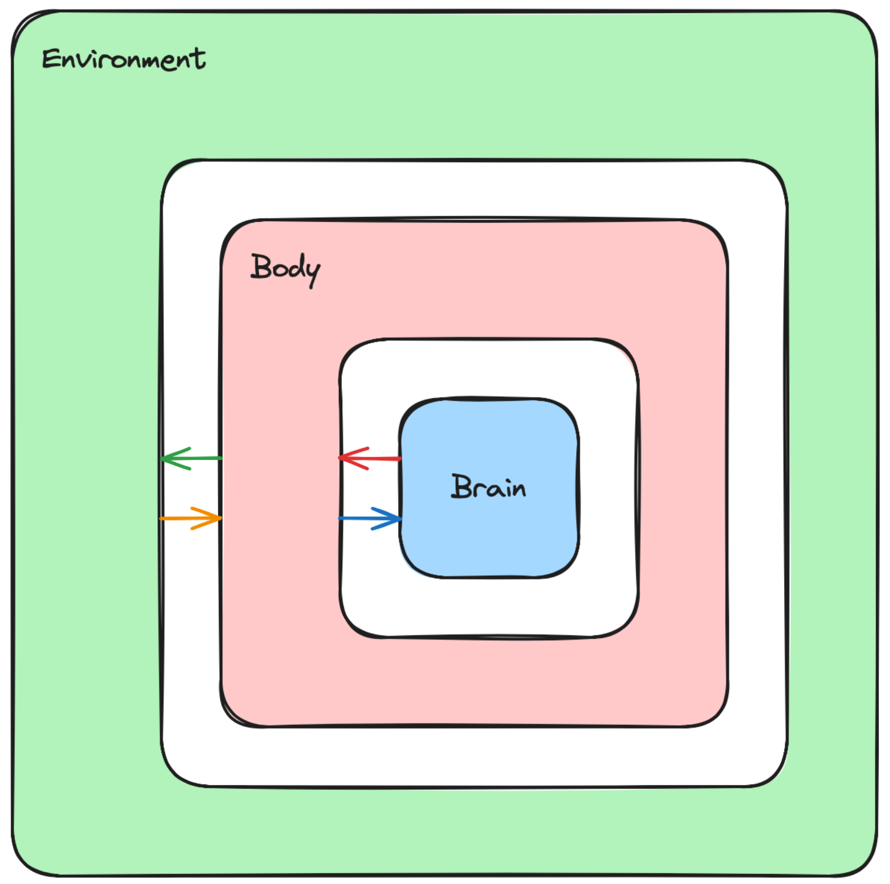
Background

Agents and environments

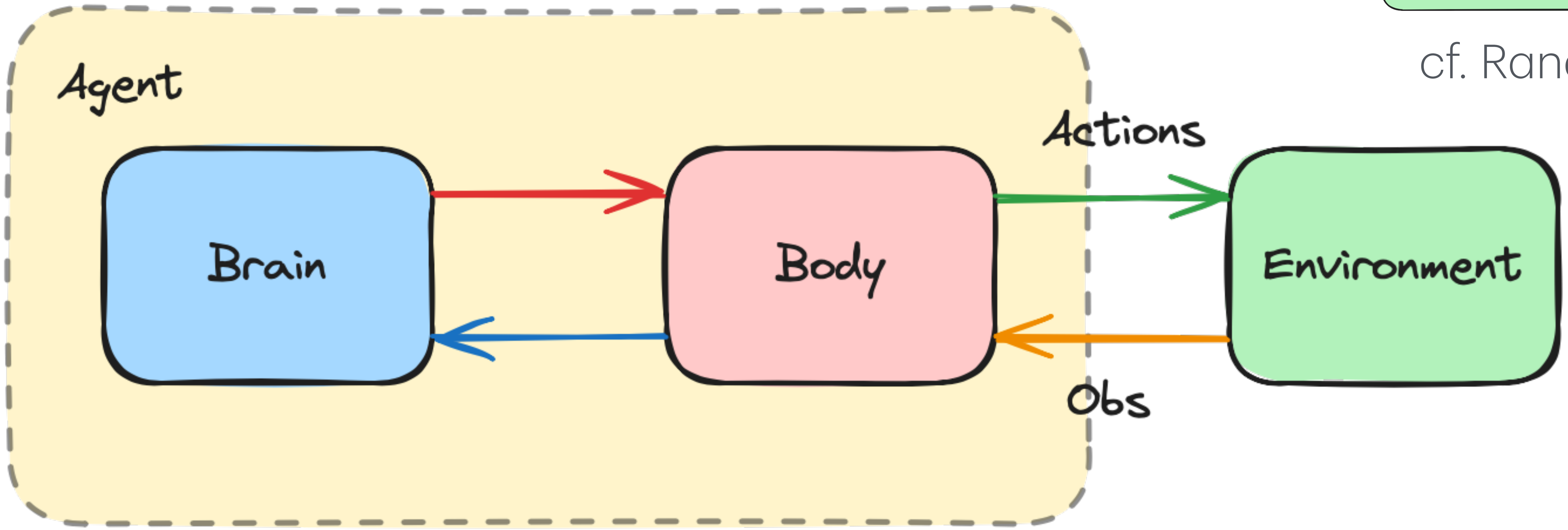


Unpacking that a little

Factorising the agent

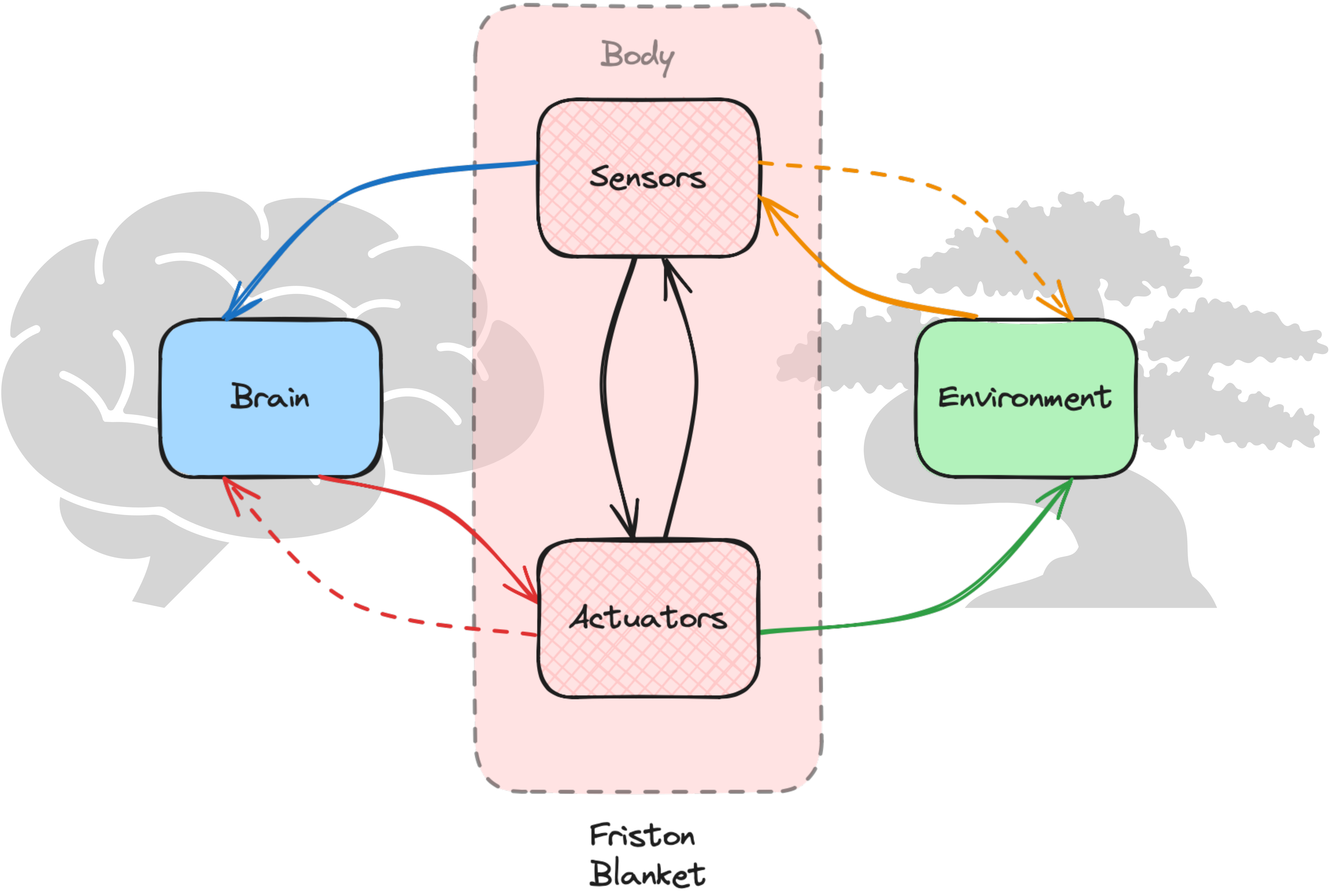


cf. Randy Beer



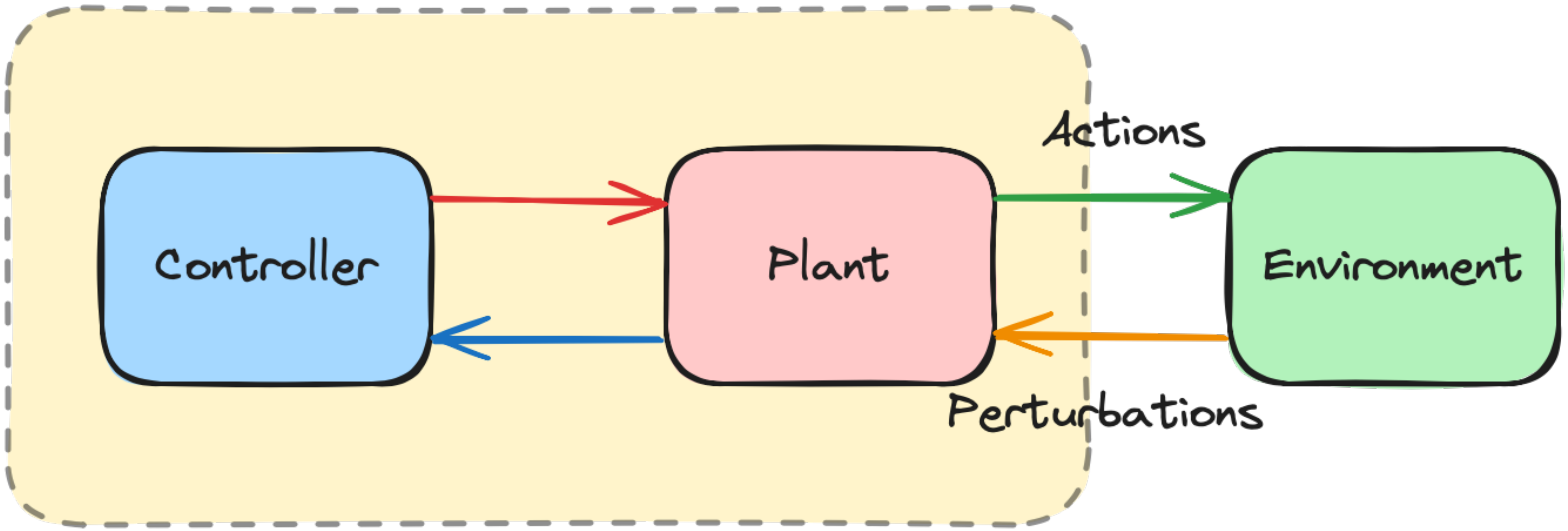
Meanwhile, the FEP

Friston blankets, boundary factored into sensors and actuators



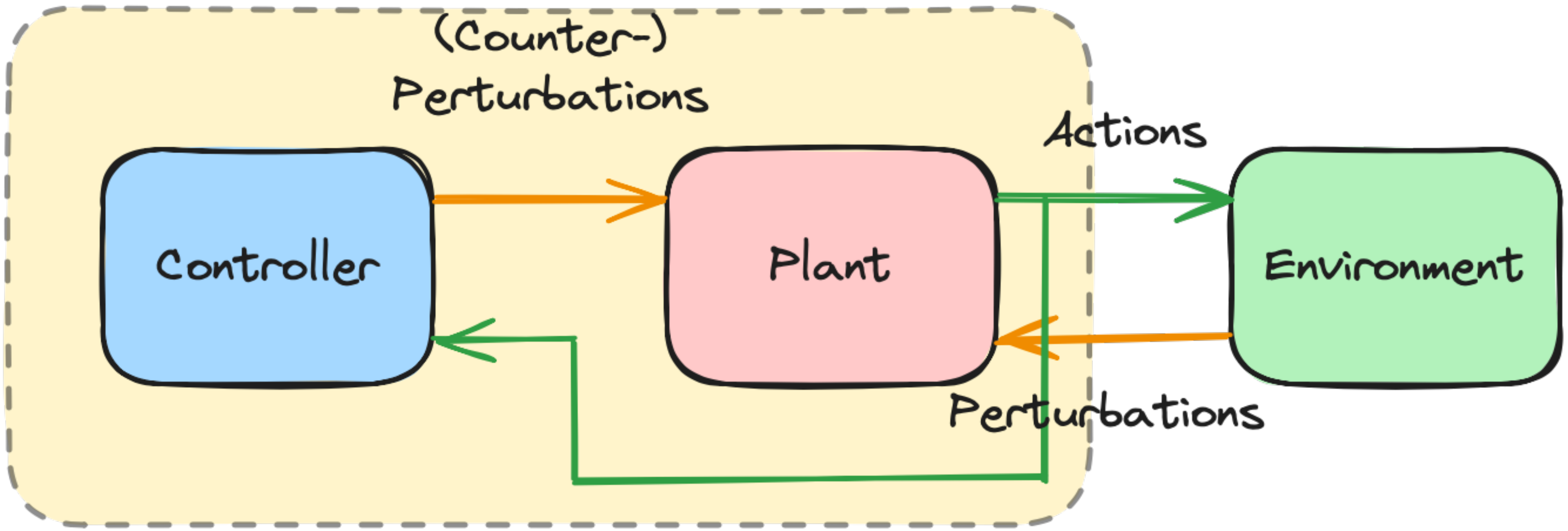
Meanwhile, in control theory

Control-plant-environment factorisation



Internal model principle (IMP)

A model of homeostasis (implying a model?)



Controller models environment because it
“knows” how to counteract perturbations

“When does a system model another system?”

1. What do we mean by “system”?

Systems (fully observable)

Some definitions

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.

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:

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

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

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

Systems	Inputs	Outputs
Open	Yes	Yes
Autonomous	No	Yes
Closed	No	No

Factorisation of systems

Our setup: full system and components

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*

$$\mathsf{upd}_\mathsf{P} : P \times E \times C \to P, \tag{9}$$

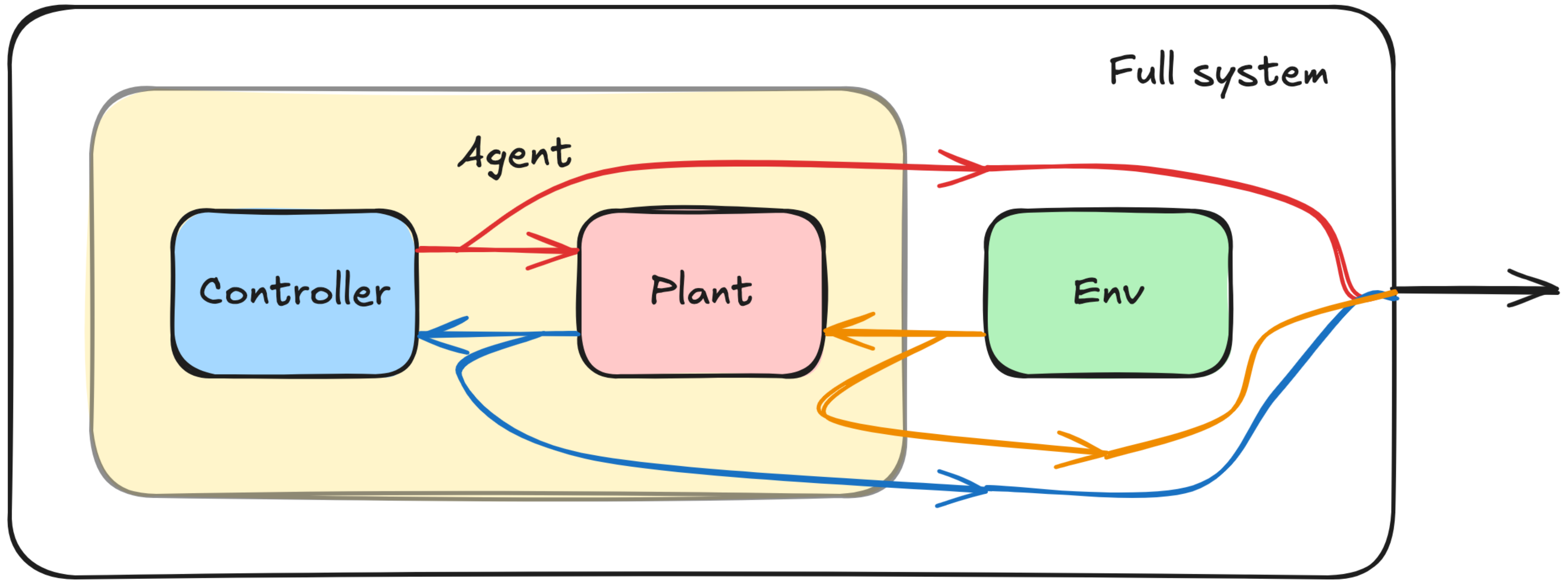
3) *the controller* $\mathsf{C} : \mathbf{Sys}\binom{P}{C}$ *is a system*

$$\mathsf{upd}_\mathsf{C} : C \times P \to C. \tag{10}$$

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.

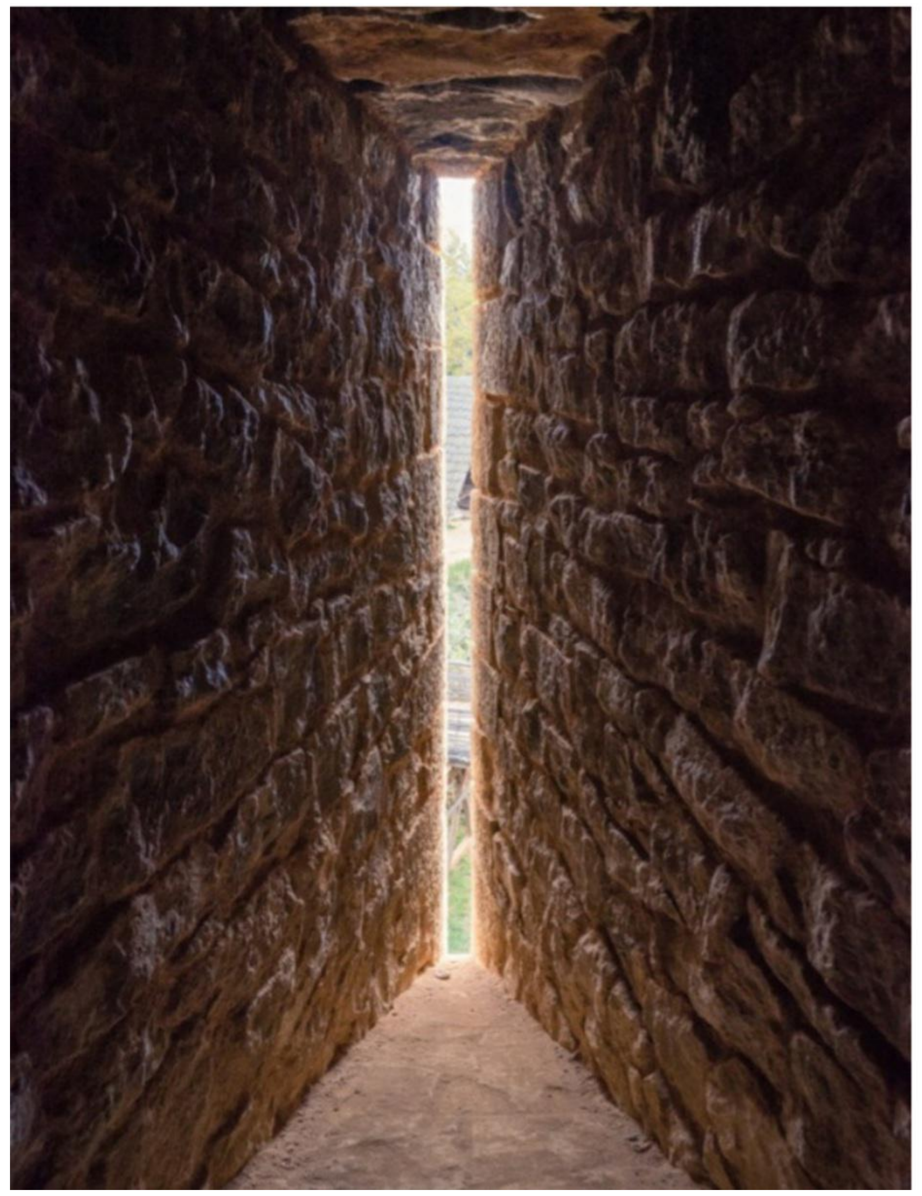


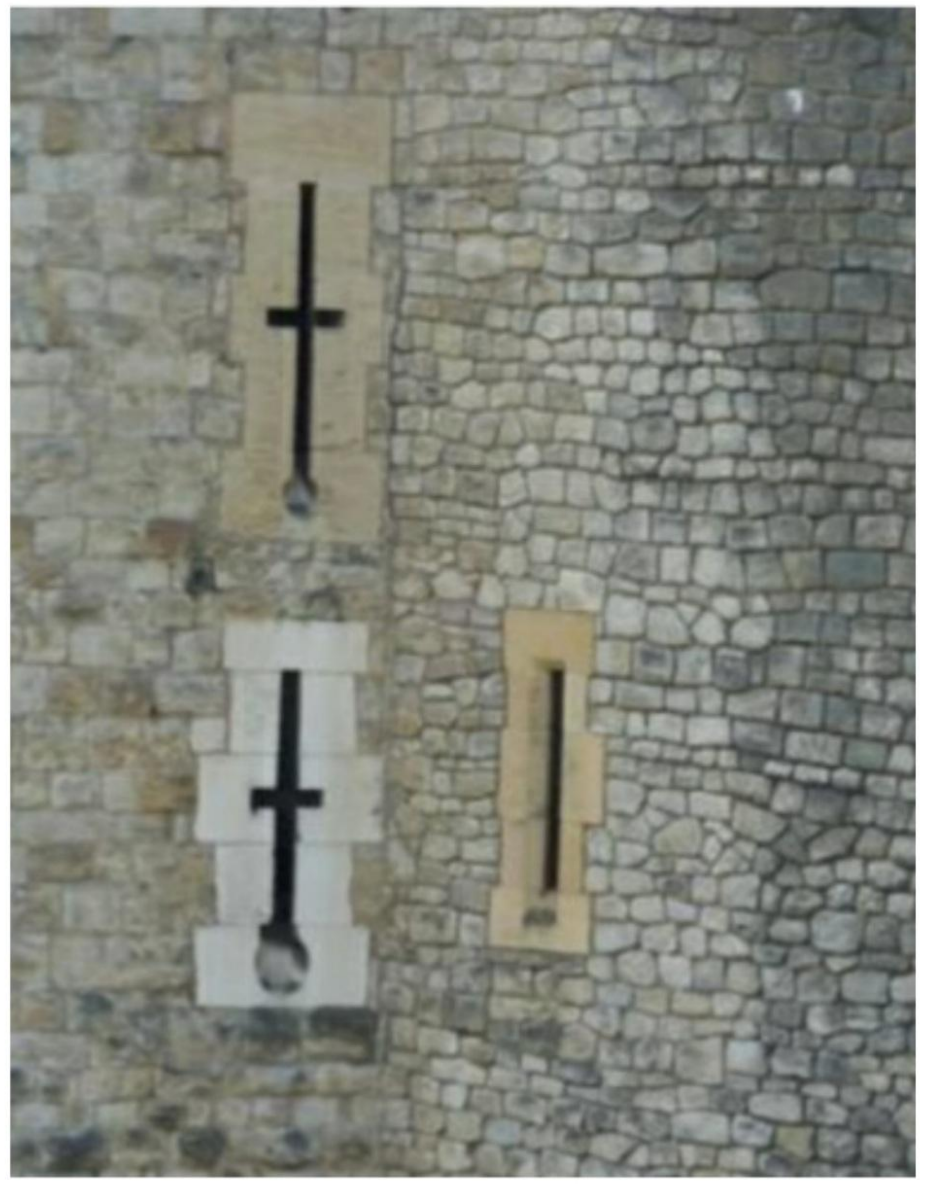
“When does a system model another system?”

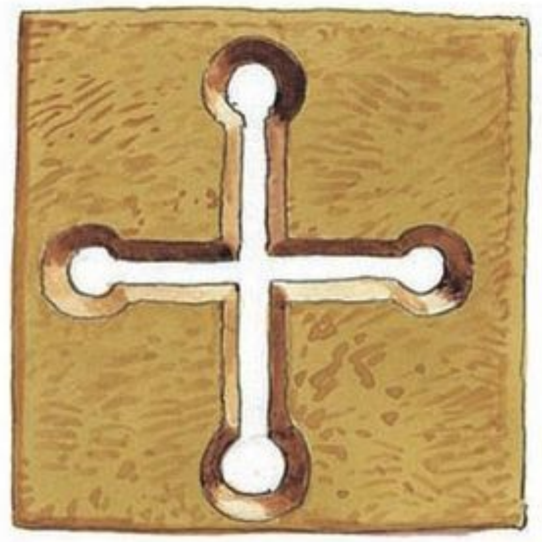
2. What do we mean by “model”?

Informally

What is this?







Two perspectives

An example

- Controller: the army **outside** the castle
- Environment: the army **inside** the castle

Model

Definition

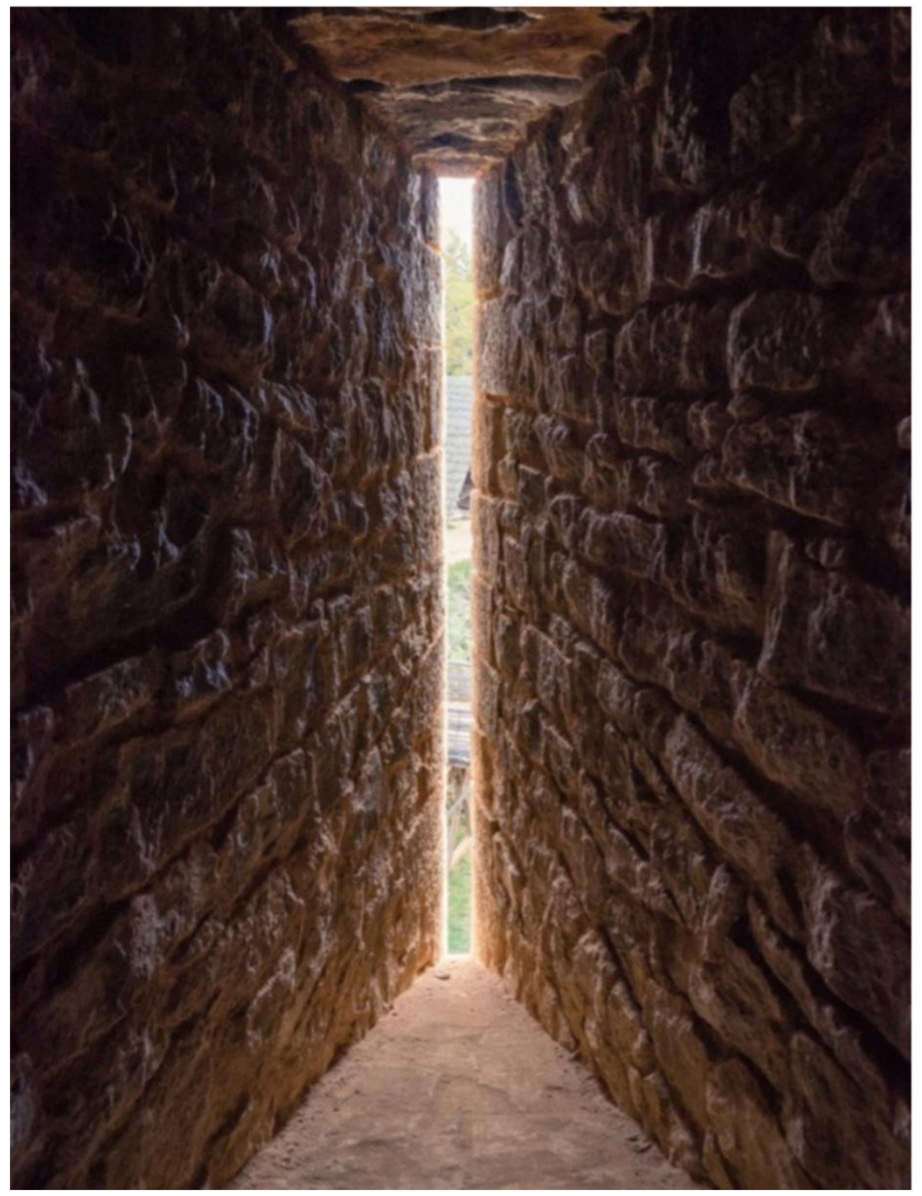
Definition II.9 (Model)**.** A *model* of a system $\mathsf{X} \in \mathbf{Sys}\binom{I}{X}$ is:

- 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 $x \in X$.



Generalising ideas such as:

- Coarse grainings
- Lumpability
- Variable aggregation
- State aggregation
- Model reduction/compression (PCA, SVD, t-SNE, UMAP, etc.)
- Dynamical consistency
- Macrostates
- ϵ-machines
- ...

“When does a system model another system?”

3. “When” does this happen?

Controllers modelling systems

Sufficient conditions for models of the full system and of the environment

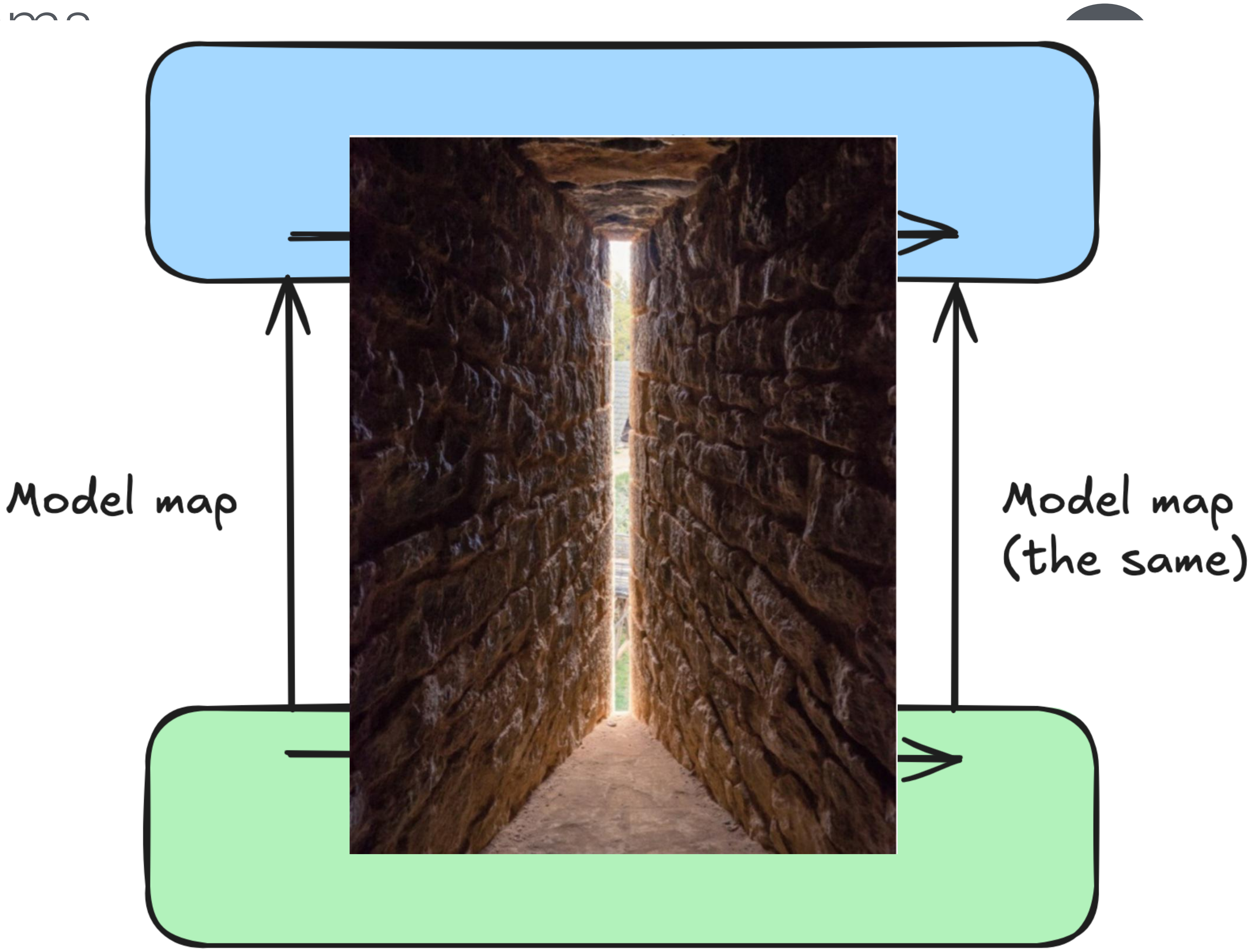
- Controllers solve proble…

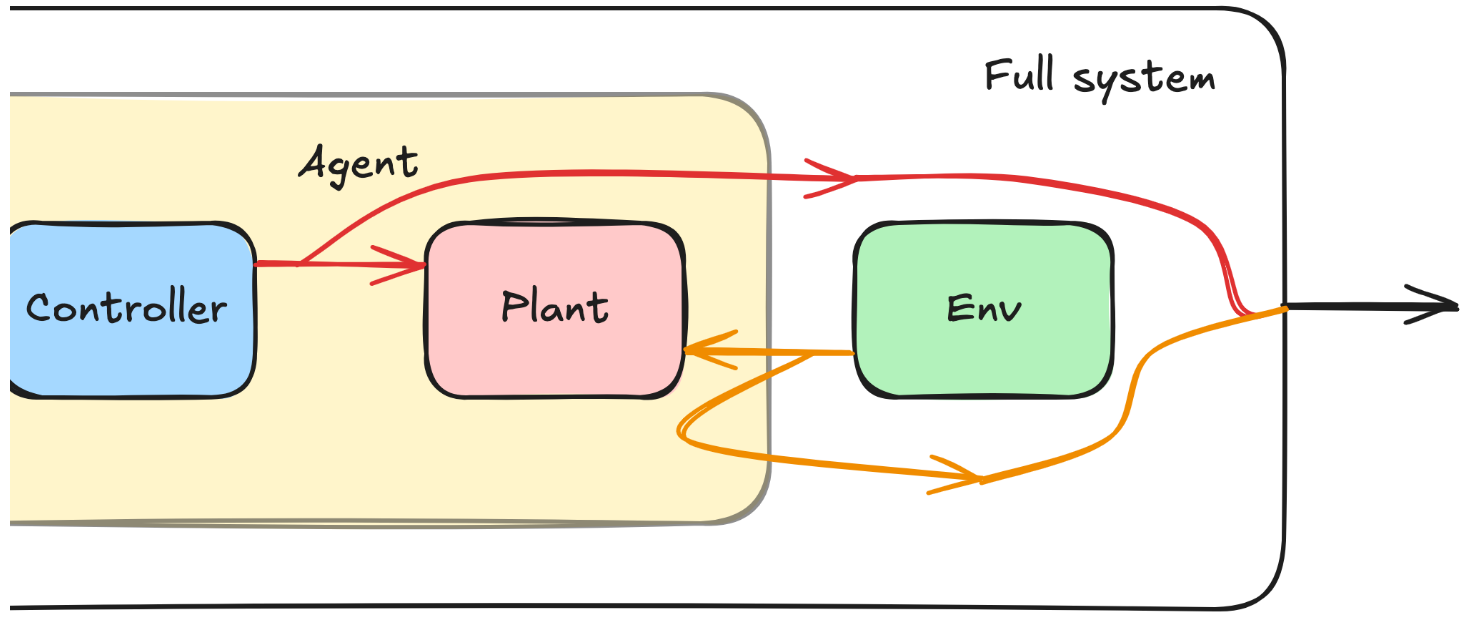
Assumption 2 (Regulation cond… *problem, meaning there exists a… nition II.6)* $\mathsf{S}^* \to \mathsf{S}$ *such that, o…*

- Controllers are autono… when solving problems

Assumption 3 (Error feedback… *autonomous dynamics* $\mathsf{upd}_{\mathsf{C}^*}$: … *system* C^* : $\mathbf{Sys}\binom{1}{C^*}$*, that we c… making* π_{C^*} *a full-fledged map o…*

$$\pi_{\mathsf{C}^*} : \mathsf{S}^* \to \ldots$$





…odel Principle (attracting environ-… …tem subject to Assumptions 1 to 4. …ting environment E^* *via the dashed*

$$\xrightarrow{\nu} \mathsf{C}^* \qquad \to \mathsf{S}^* \xrightarrow[\text{Assumption 3}]{\pi_{\mathsf{C}^*}} \mathsf{C}^* \tag{18}$$

- Kind of mysterious…

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* $\pi_E(s) = s_{\mathsf{E}}$.

"When does a system model another system?"

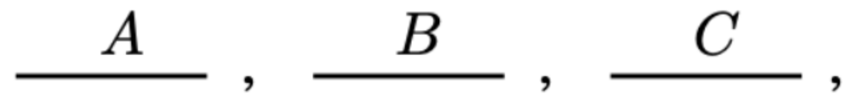
4. What does Bayes have to do with this?

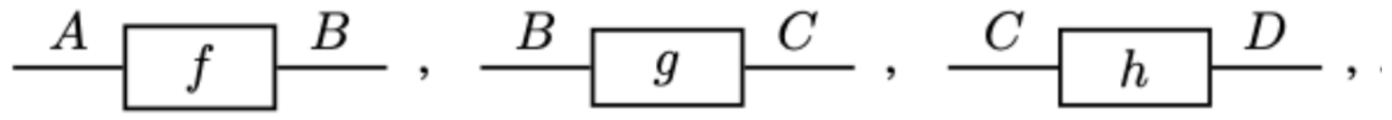
Categories

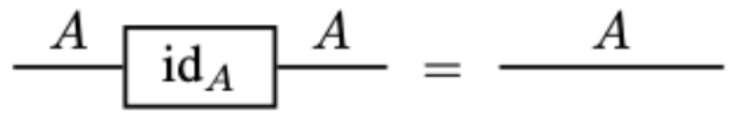
String diagrams

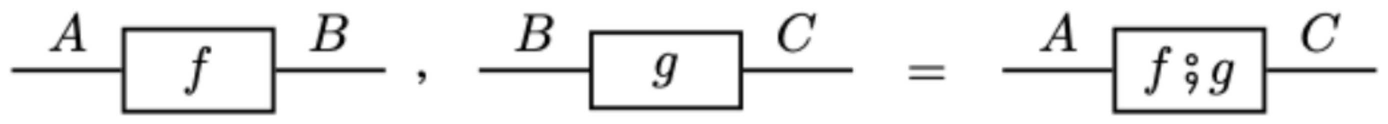
Definition III.1 (Category)**.** A category $\mathbf{C}$ consists of:

- 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 $\fatsemi$ on arrows called the *composition rule*, such that given morphisms $f : A \to B$ and $g : B \to C$, their composite $f \fatsemi g$ 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 $f \fatsemi (g \fatsemi h) = (f \fatsemi g) \fatsemi h$,
  - left and right *unit laws*: for every pair of objects A, B and morphism $f : A \to B$, we must have $\mathsf{id}_A \fatsemi f = f = f \fatsemi \mathsf{id}_B$.

- Objects 

- Morphisms 

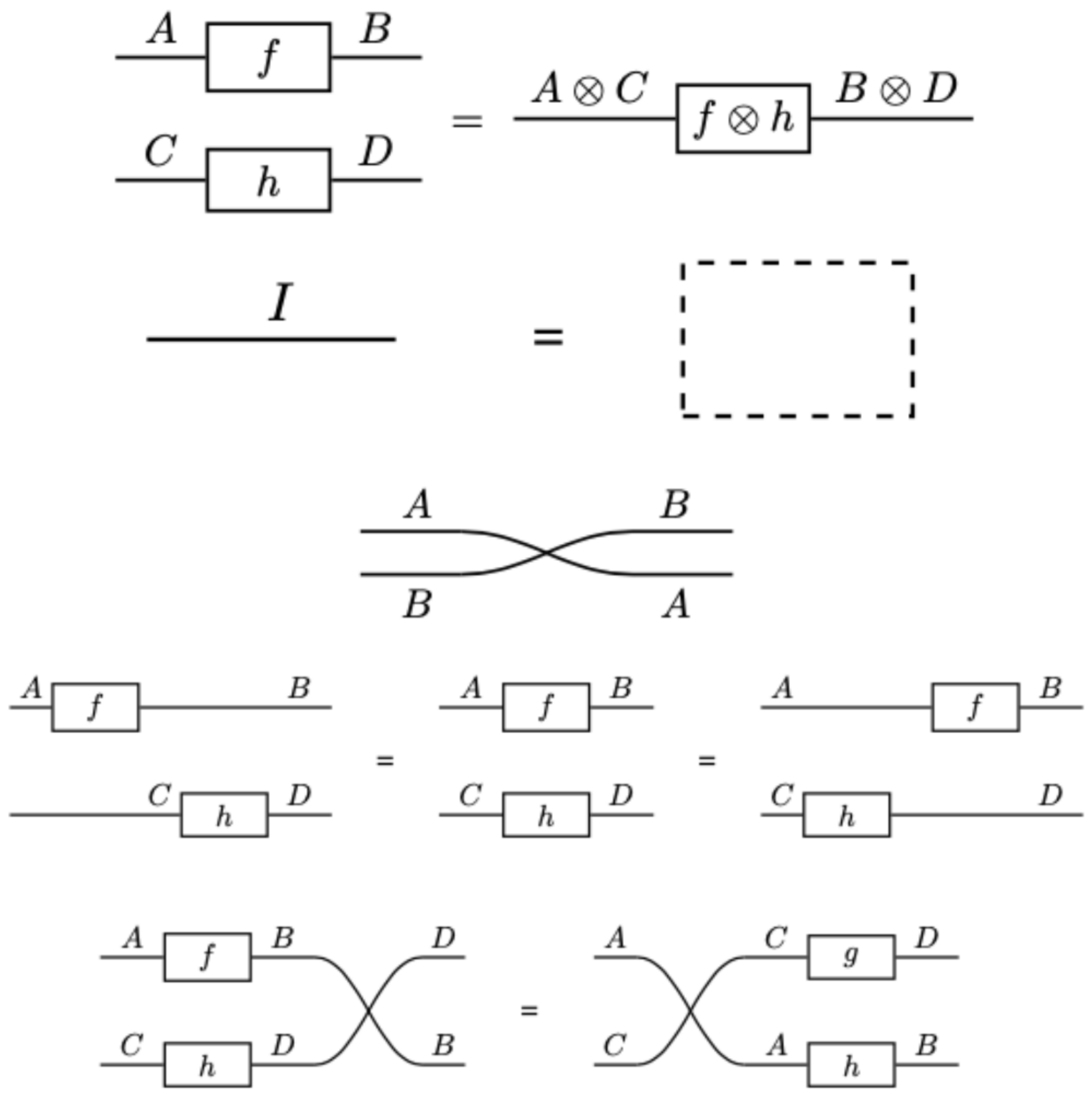
- Identity 

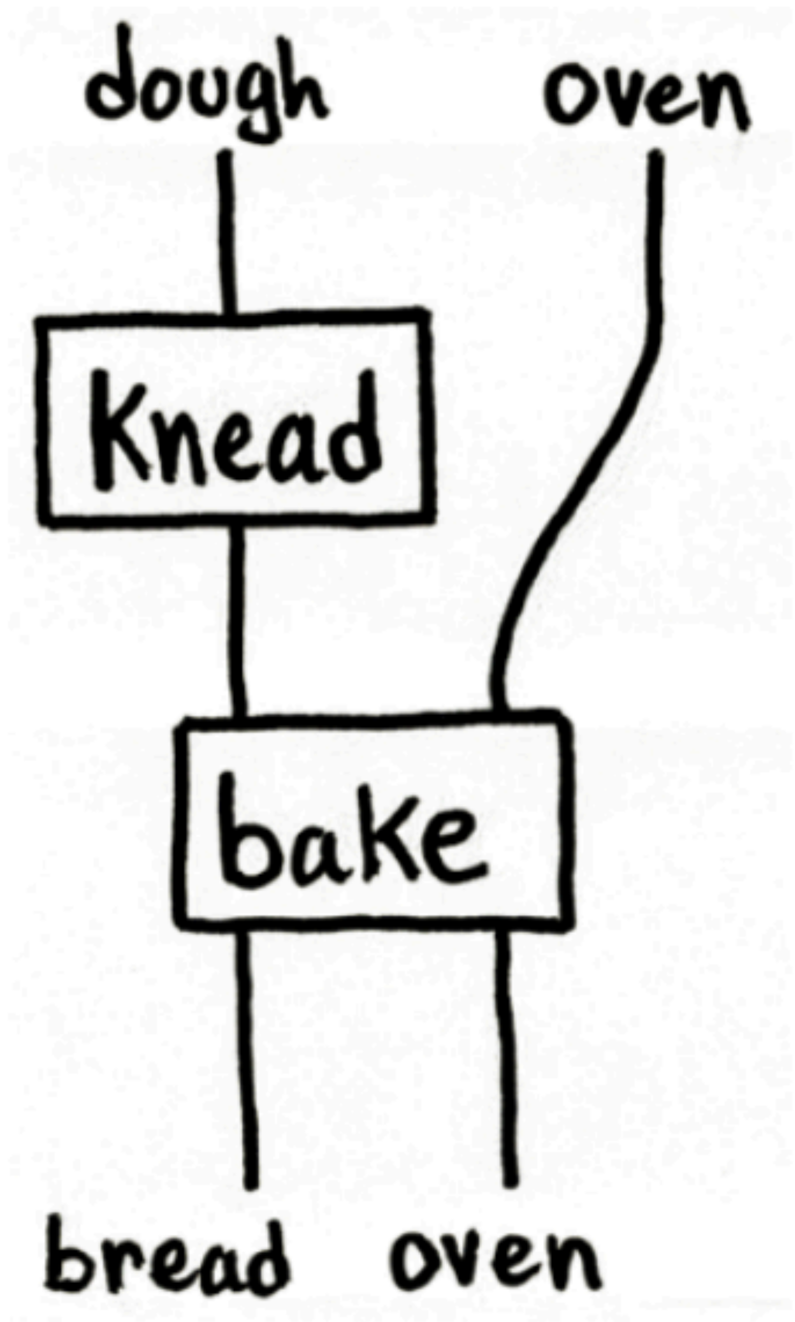
- Composition 

Process theories

Putting things in parallel in string diagrams

- Parallel composition
- Identity object
- Swap map
- Interchange law
- Naturality of swap



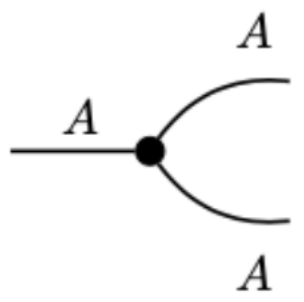


Boisseau et al. (2022)

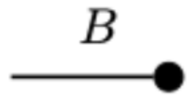
Markov categories

Process theories for non-deterministic processes

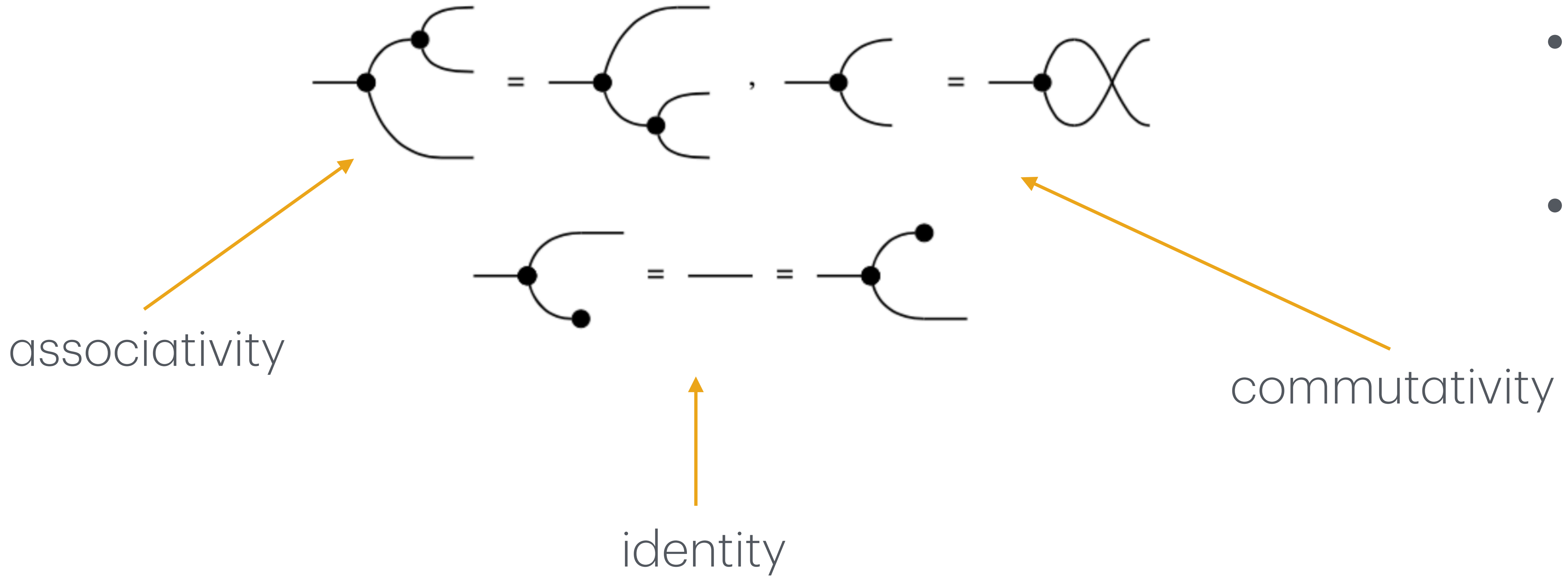
- Copy



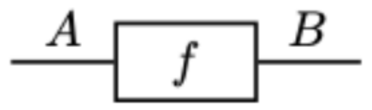
- Delete



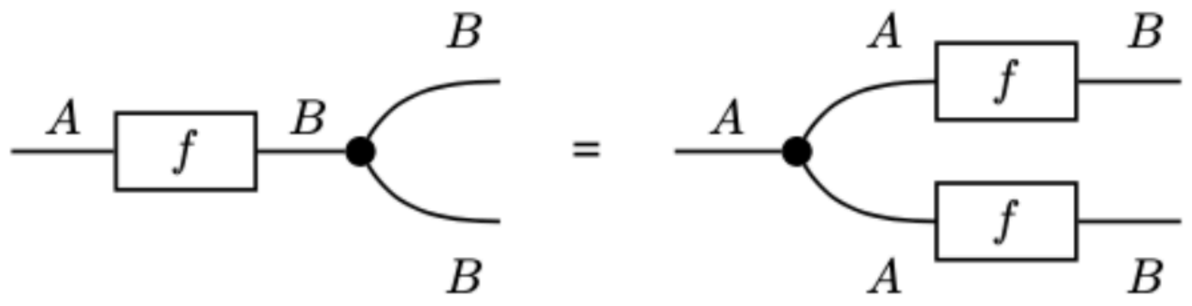
subject to the following



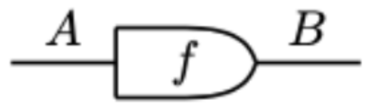
- Deterministic morphism



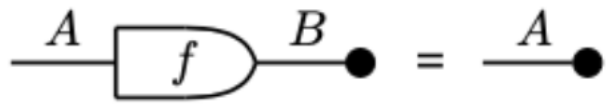
such that



- Non-deterministic morphisms



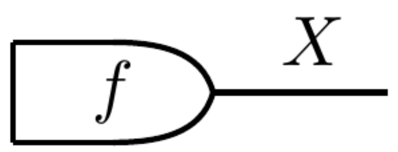
- Normalised probabilities



Markov categories

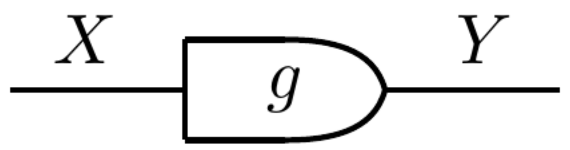
By example, with probabilities

- Probability distribution



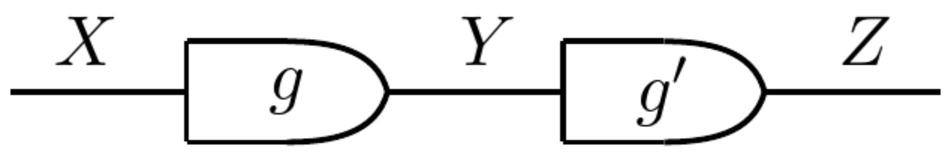
$f(x)$ or $P(x)$

- Conditional probability



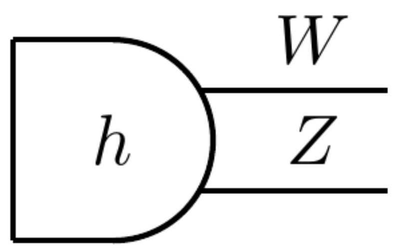
$g(y \mid x)$ or $P(y \mid x)$

- Chapman-Kolmogorov



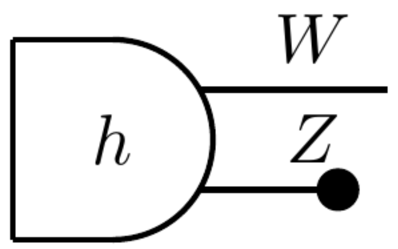
$$\sum_{y \in Y} g'(z \mid y) g(y \mid x) = g''(z \mid x)$$

- Joint probability



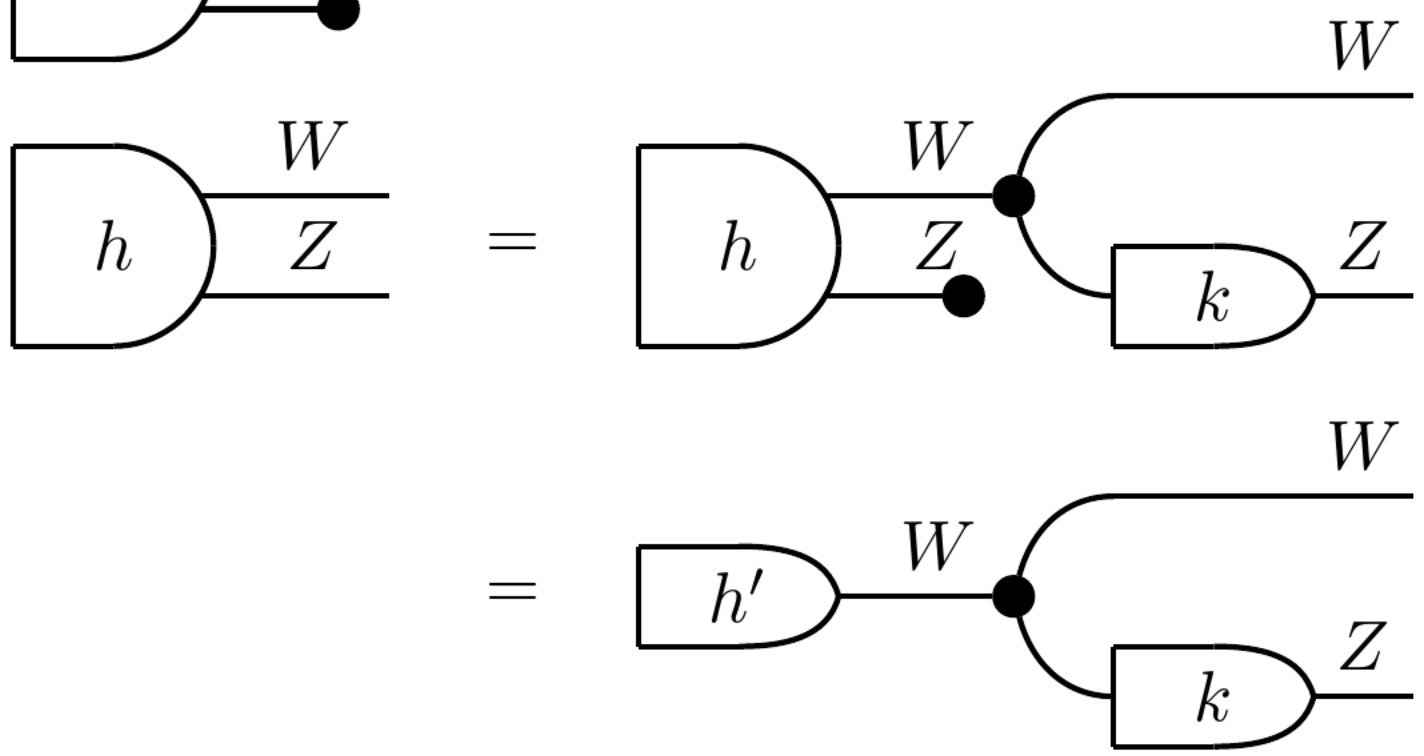
$h(w, z)$

- Marginalisation



$$\sum_{z \in Z} h(w, z) = h'(w)$$

- Chain rule

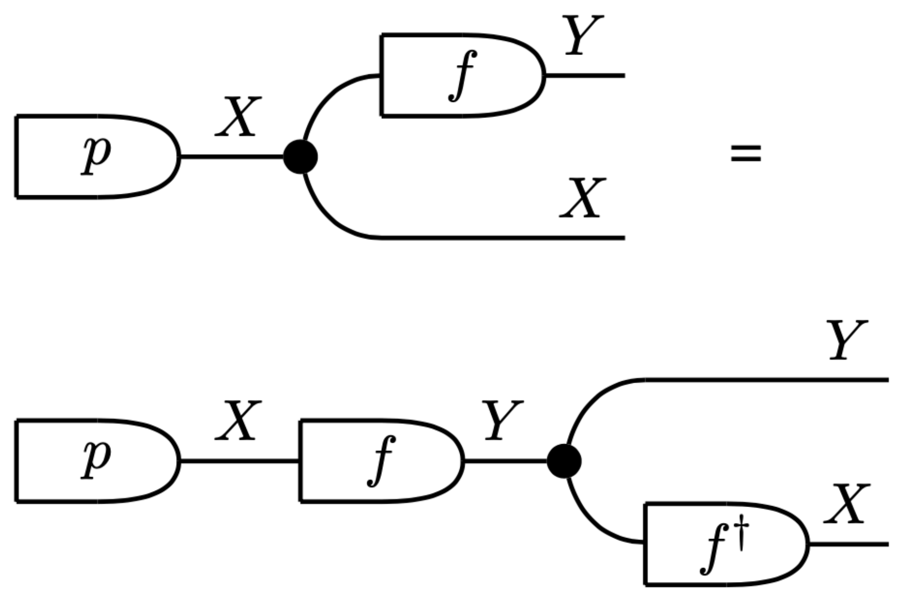


$h(w, z) = h''(z \mid w)\ h'(w)$

Bayesian inference

In string diagrams

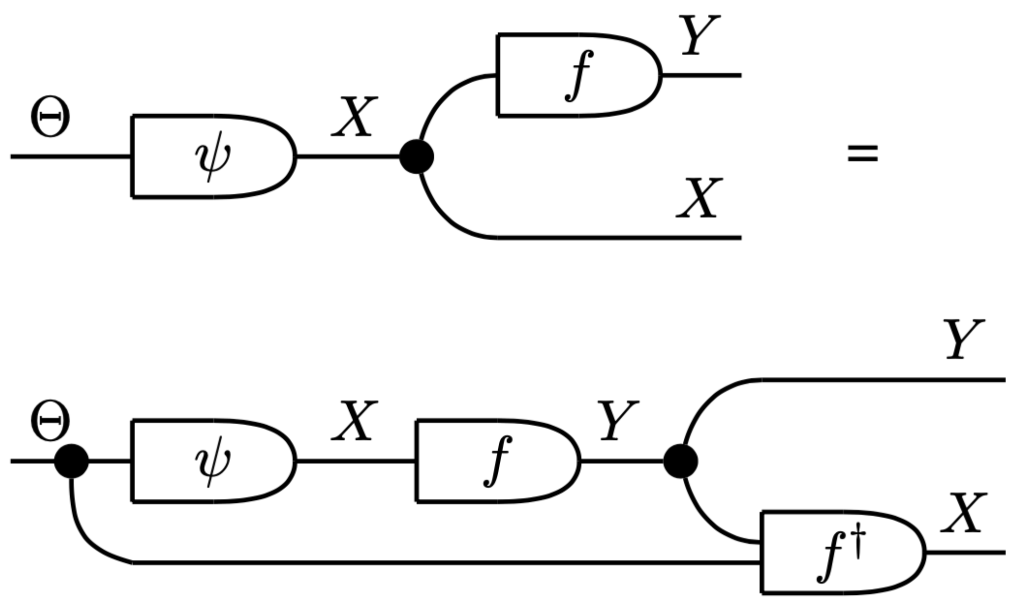
The map $f^\dagger$ is a Bayesian inversion (think, a *posterior*) of f if



$$f(y\,|\,x)\ p(x) = f^\dagger(x\,|\,y) \sum_{x \in X} f(y\,|\,x)\ p(x)$$

$$\implies f^\dagger(x\,|\,y) = \frac{f(y\,|\,x)\ p(x)}{\sum_{x \in X} f(y\,|\,x)\ p(x)}$$

With (hyper)parameters, $f^\dagger$ is a Bayesian inversion of f if



$$f(y\,|\,x)\ \psi(x;\theta) = f^\dagger(x\,|\,y;\theta) \sum_{x \in X} f(y\,|\,x)\ \psi(x;\theta)$$

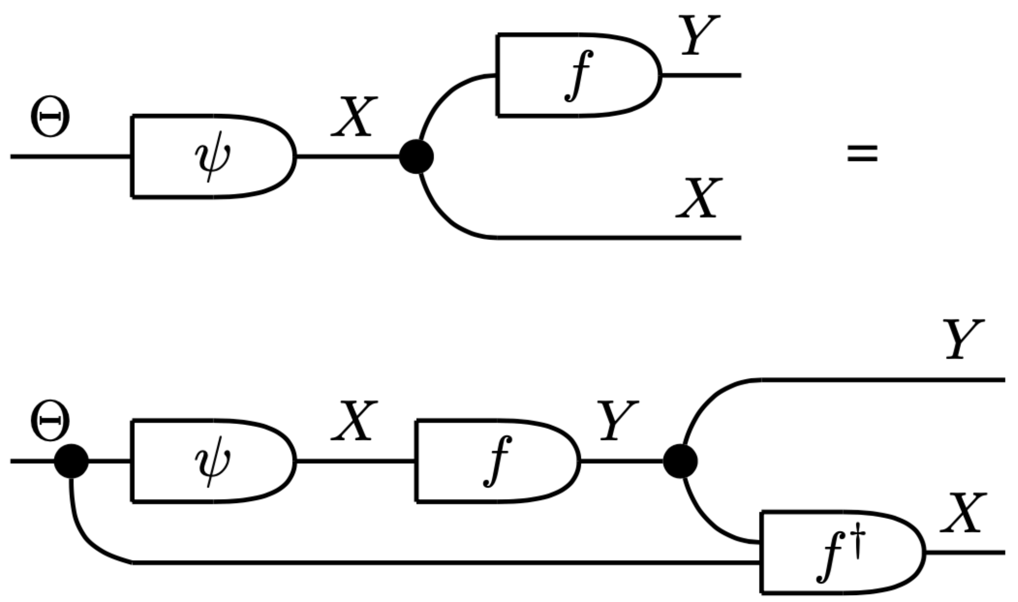
$$\implies f^\dagger(x\,|\,y;\theta) = \frac{f(y\,|\,x)\ \psi(x;\theta)}{\sum_{x \in X} f(y\,|\,x)\ \psi(x;\theta)}$$

Conjugate priors in categories

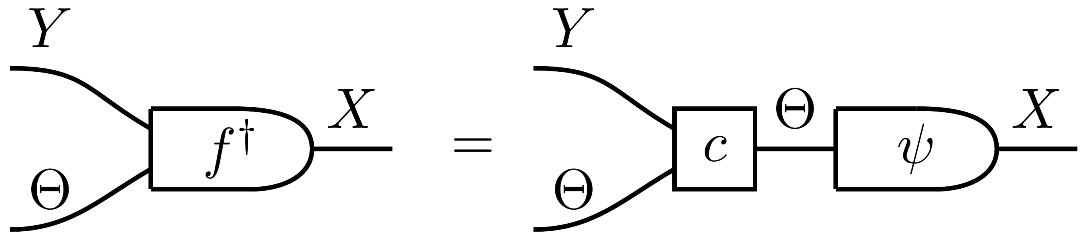
In string diagrams

What are conjugate priors?

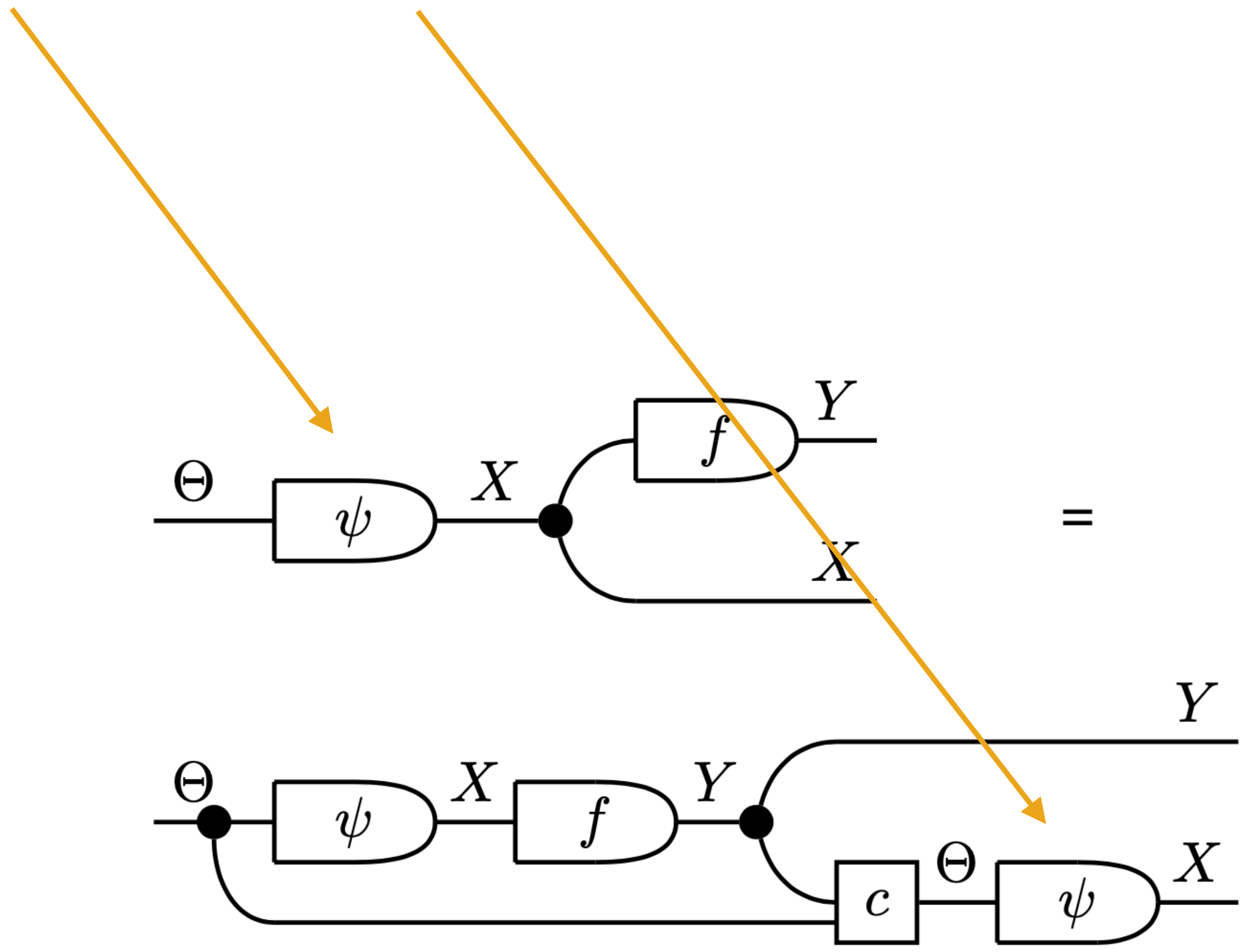
- take parametrised Bayes



- impose the following



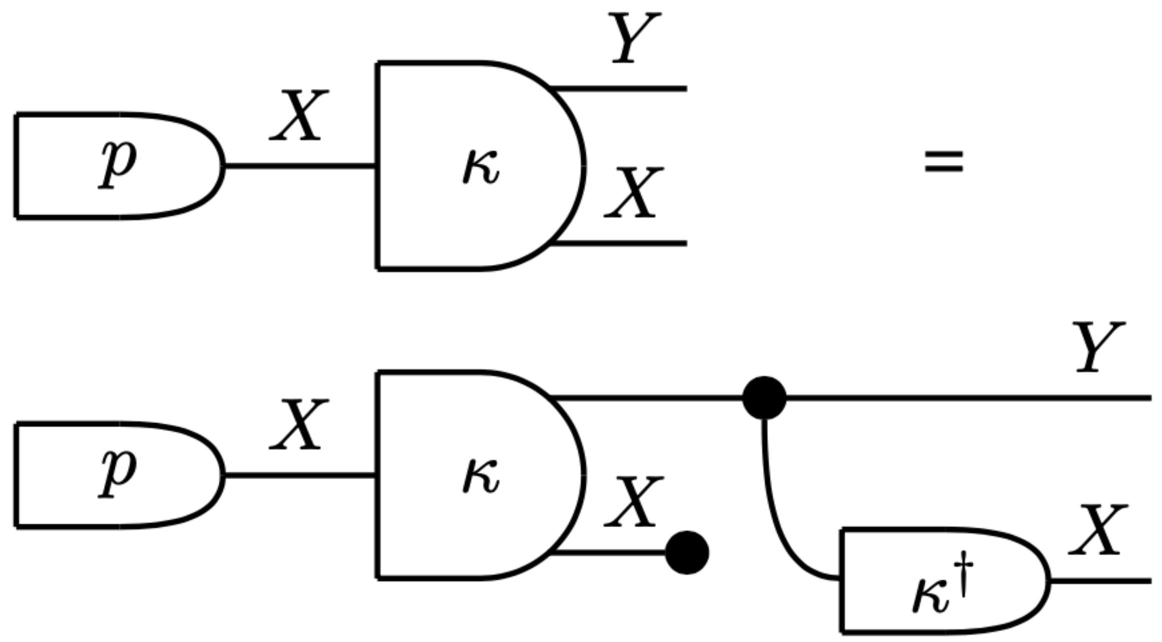
"Prior and posterior are of the same family"



Models that change over time

Bayesian filtering and conjugate priors for filtering

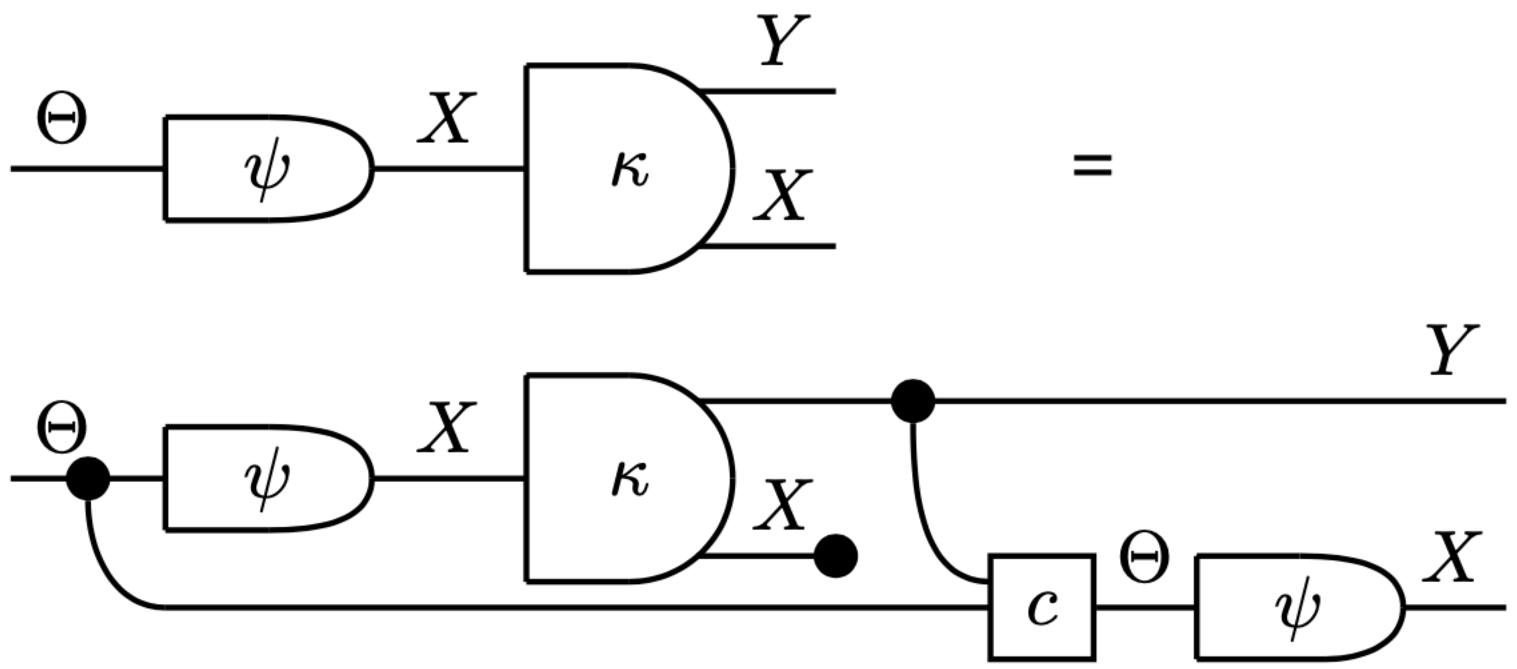
The map $\kappa^\dagger$ is a Bayesian filtering inversion of κ if



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

Conjugate priors for Bayesian filtering

—> There exists a map c such that



Bayesian interpretations

Special case

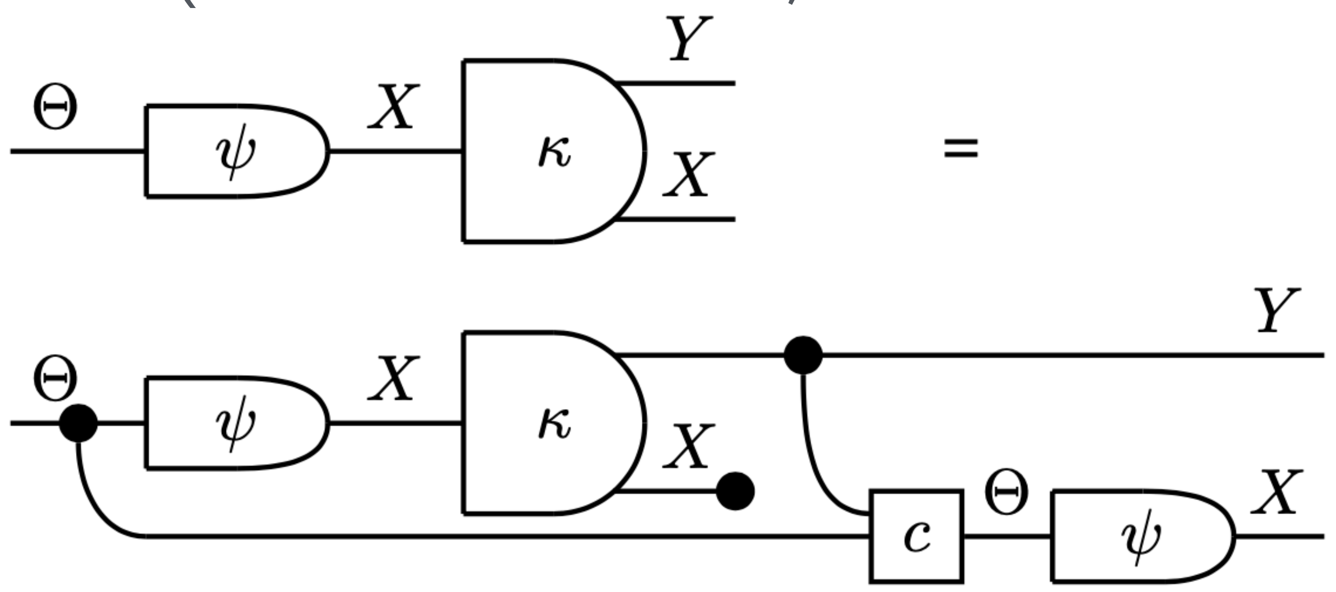
Turn definition of conjugate priors for Bayesian filtering (and inference as a special case) around:

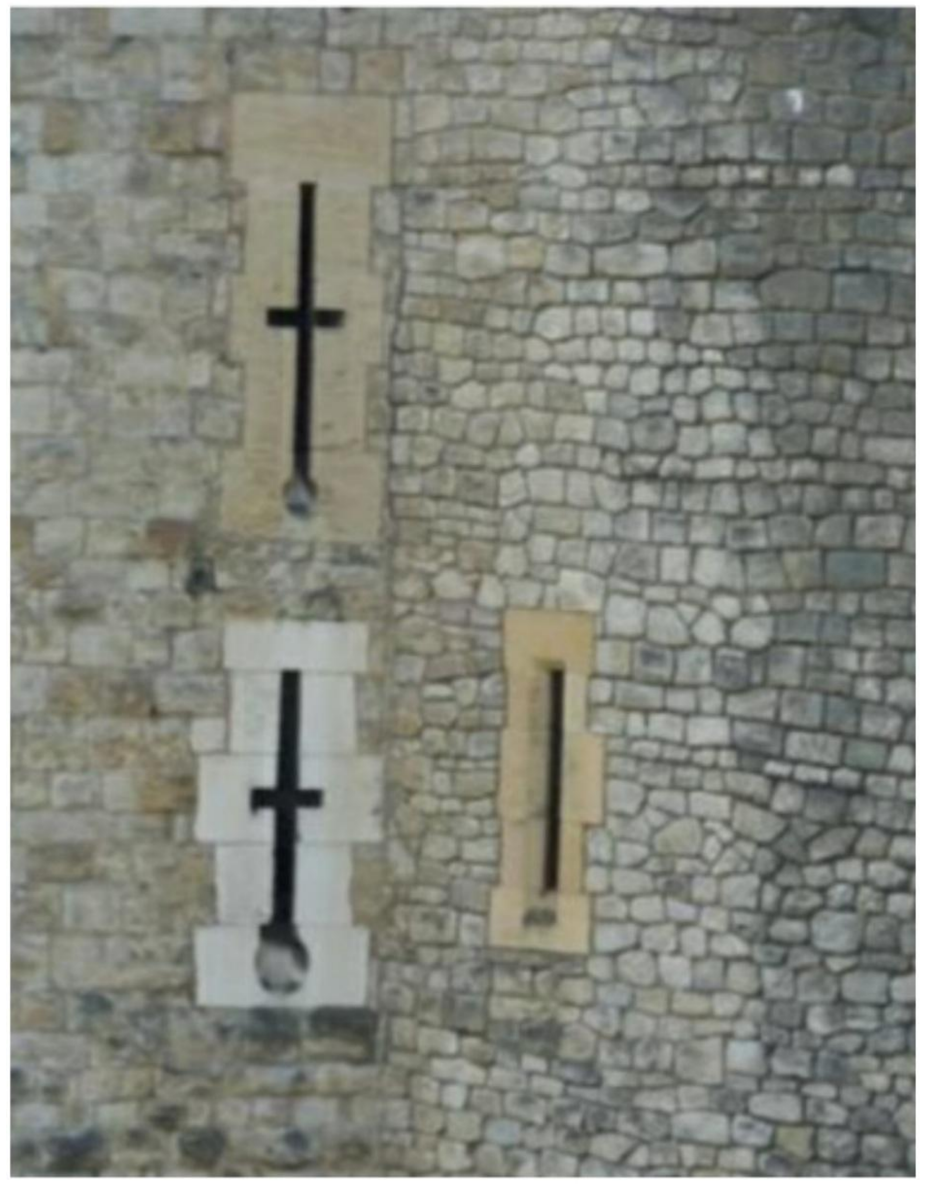
- assume a map c (controller, brain, maybe agent, etc.)

and find

- interpretation (or belief) map ψ, and
- Bayesian model κ (environment, whole world, etc.)

such that...



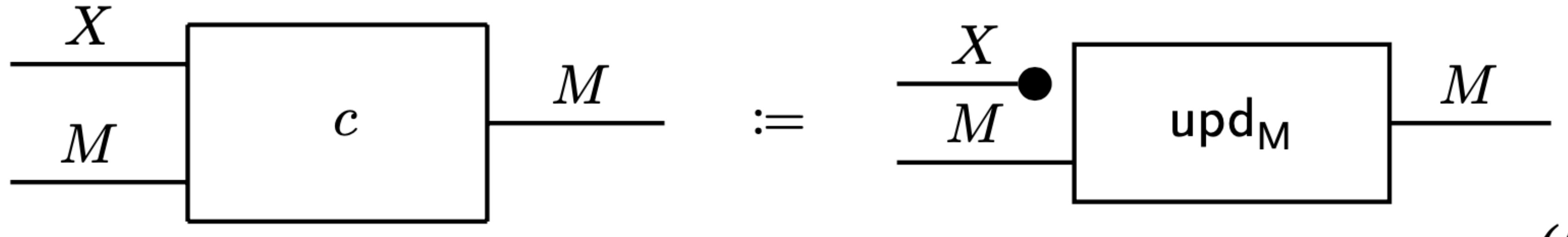


The theorem

Main result

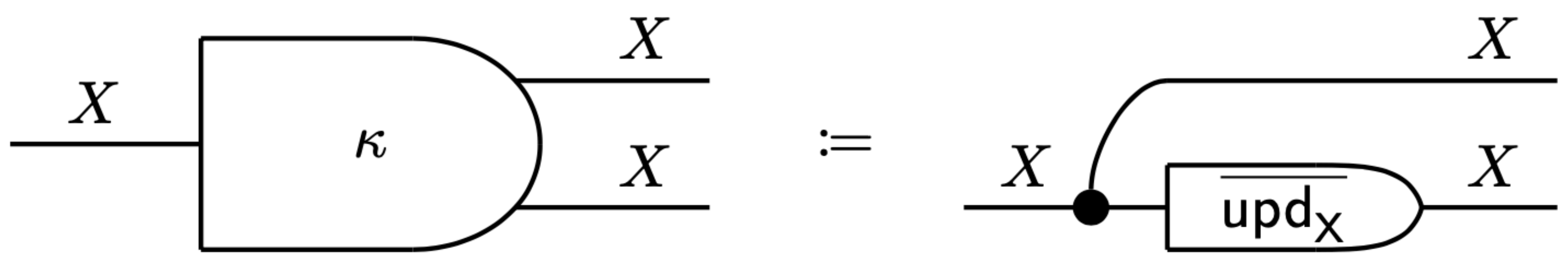
Informally: for every "model",
we have a Bayesian filtering interpretation
(actually, more than one, but at least this one).

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)

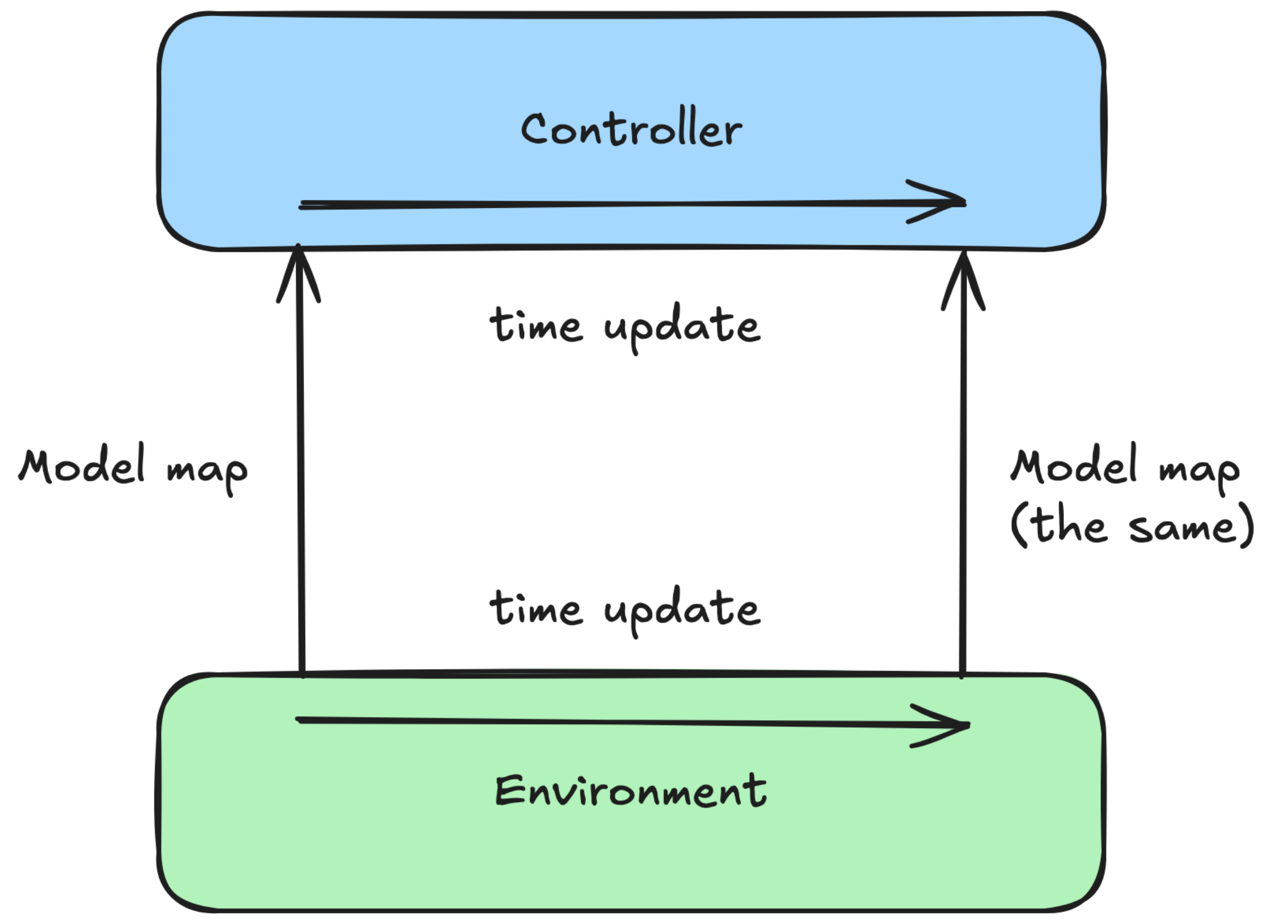
and $\kappa : X \nrightarrow X \otimes X$ *as:*

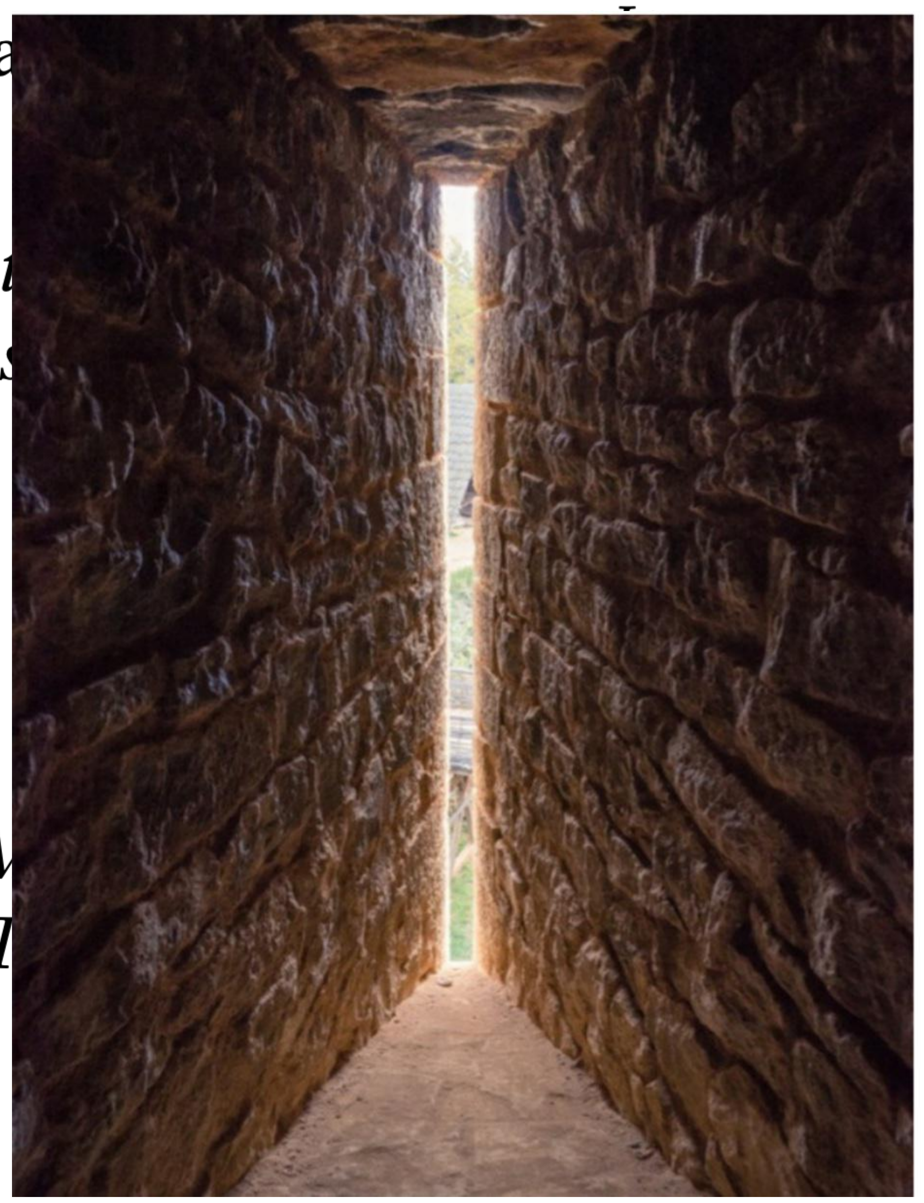


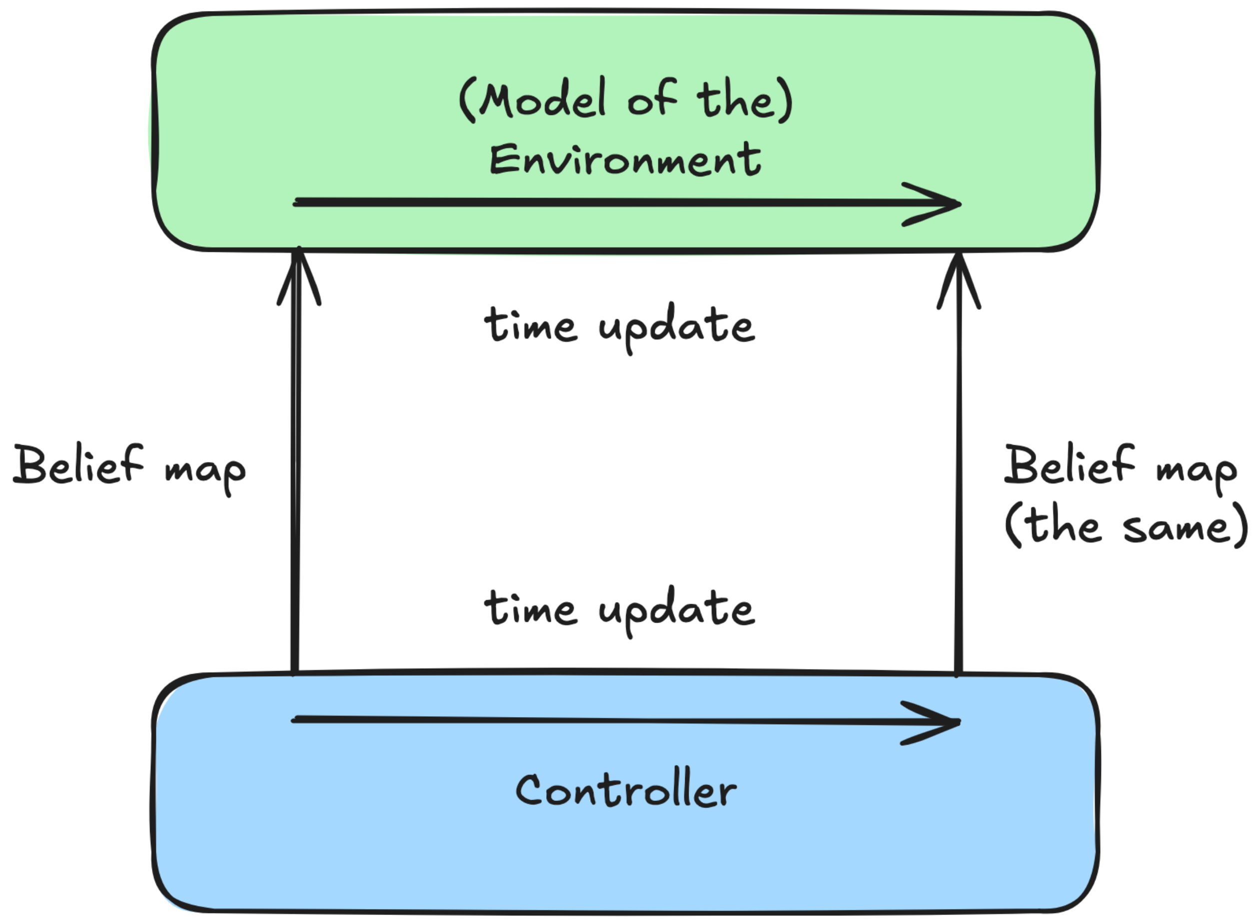
(52)

Then κ *is the hidden Markov model, and* $\mu_s^{-1} : M \nrightarrow X$ *the* …*esian filtering interpretation of* c,

Definition II.9 (Model). A *model* of a…







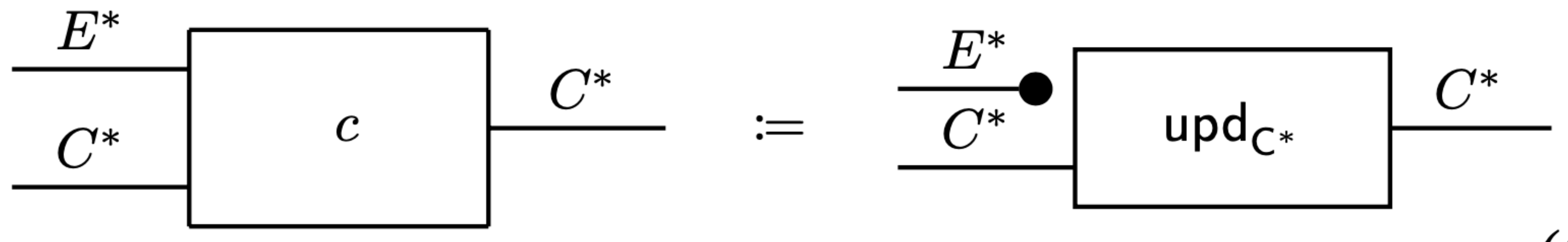
(45).

Proof. See Appendix B. □

Bayesian filtering for controllers

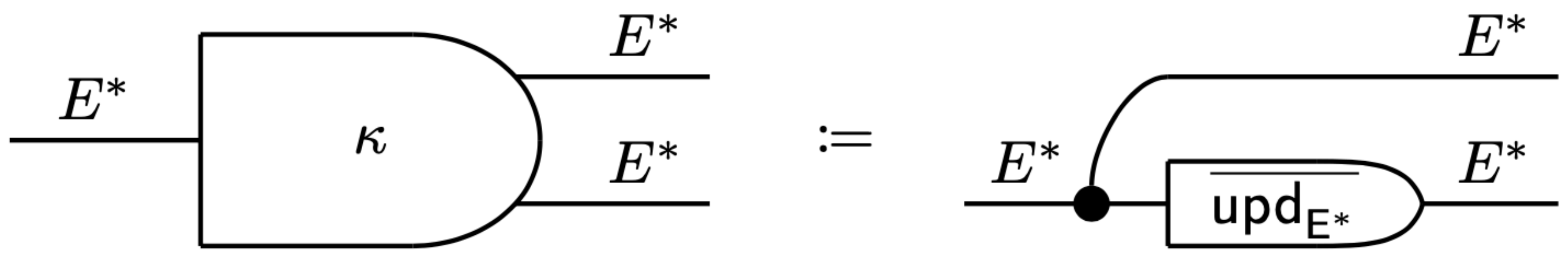
Controllers model environments in a Bayesian sense

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



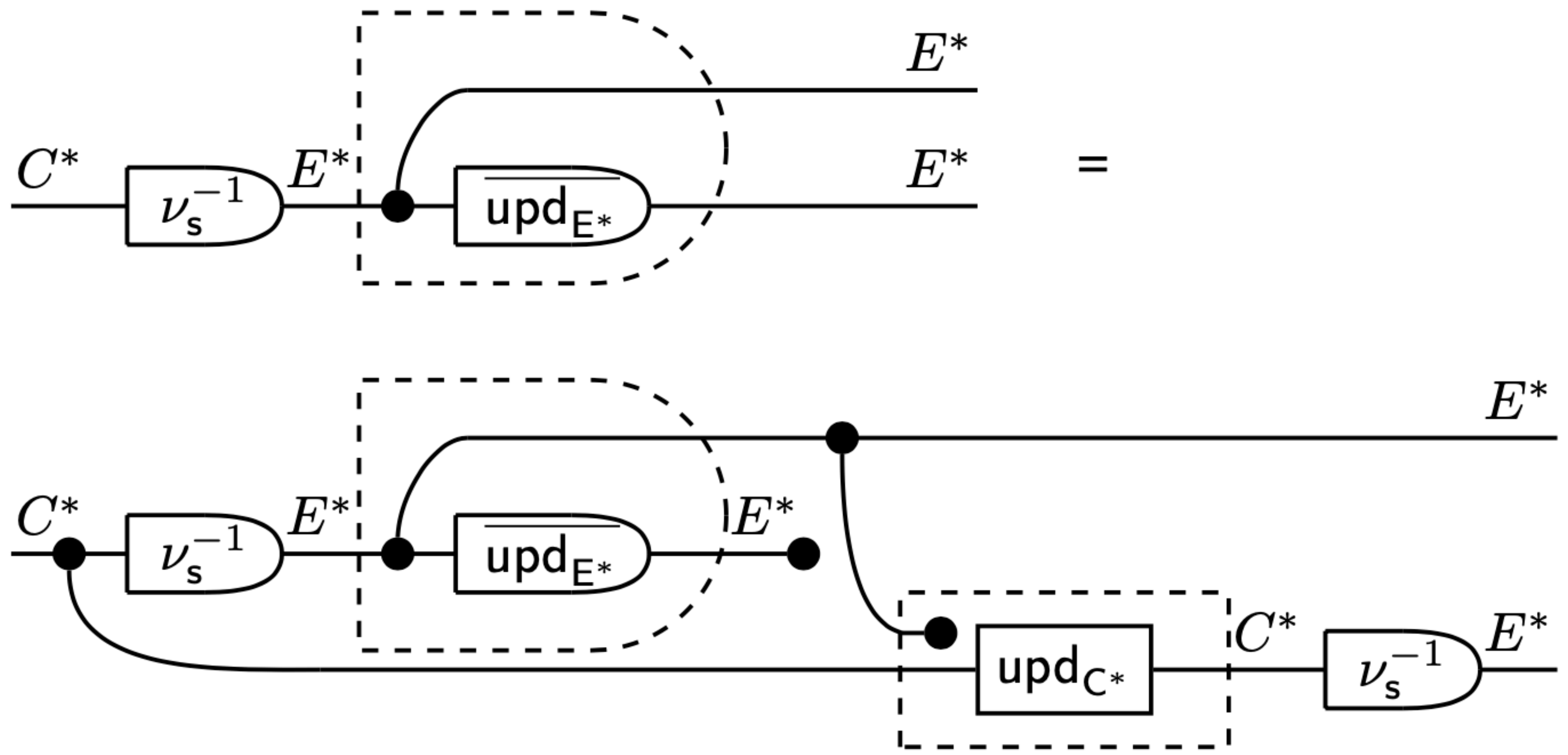
(54)

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

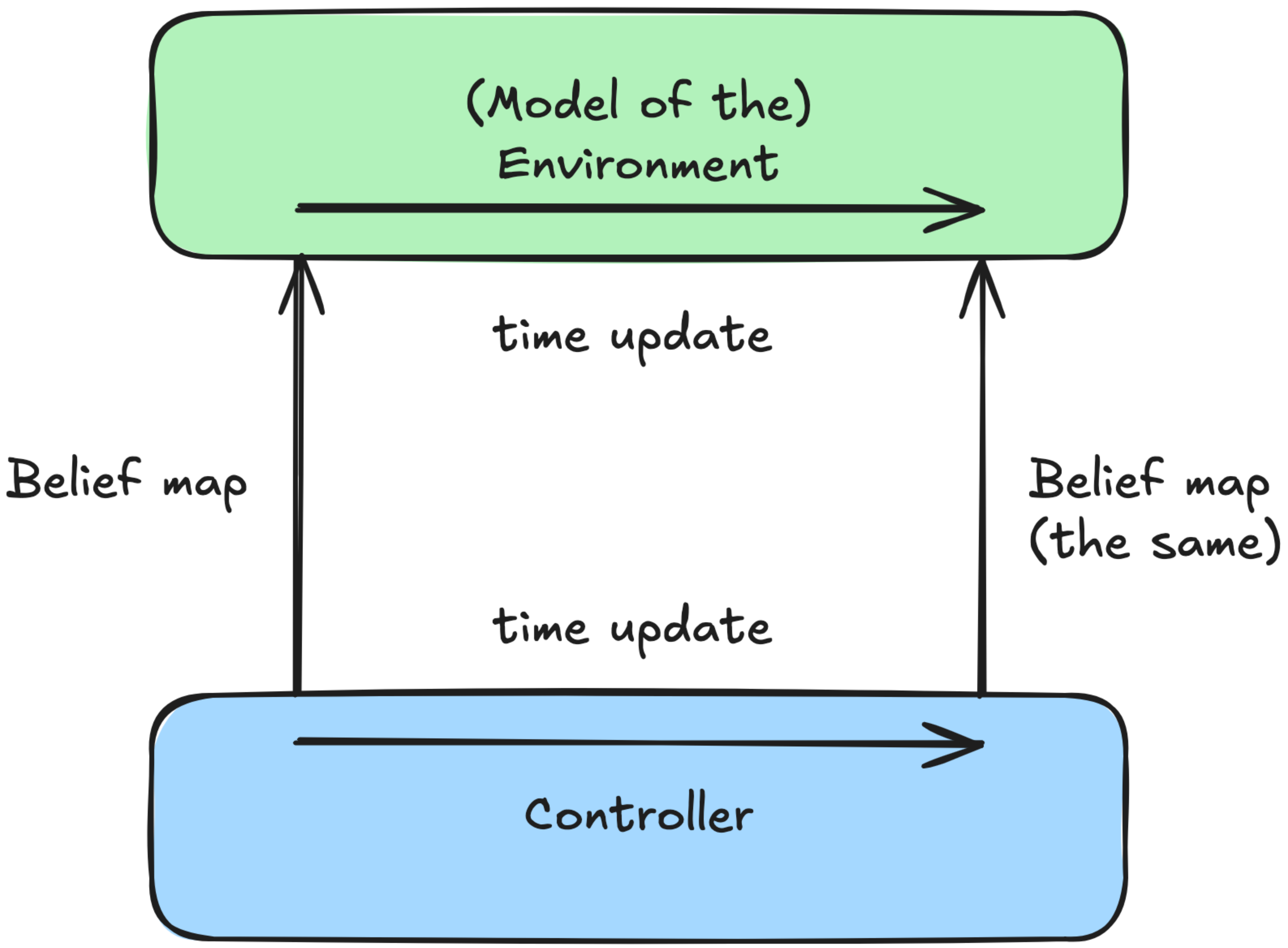


(55)

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



(56)

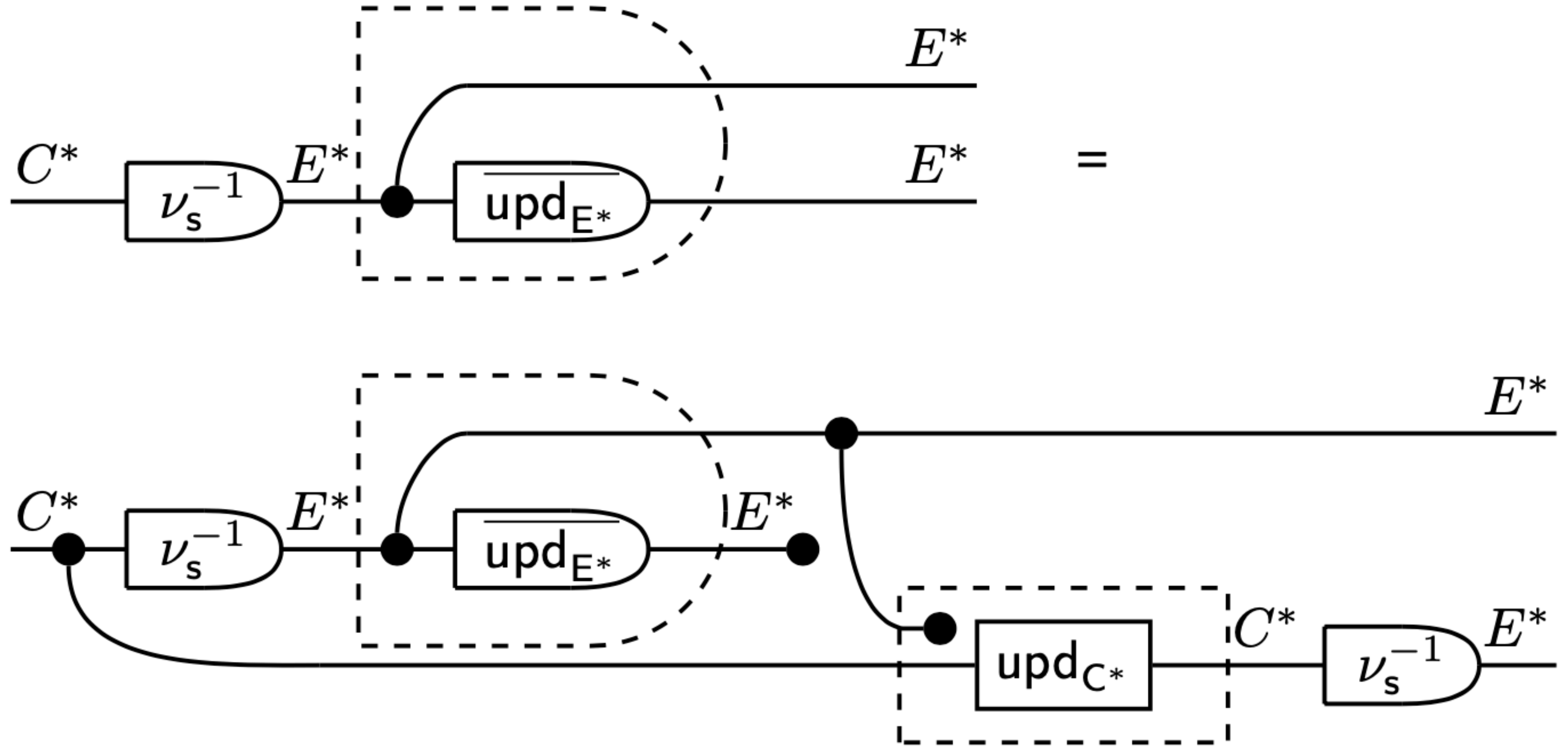


A special interpretation

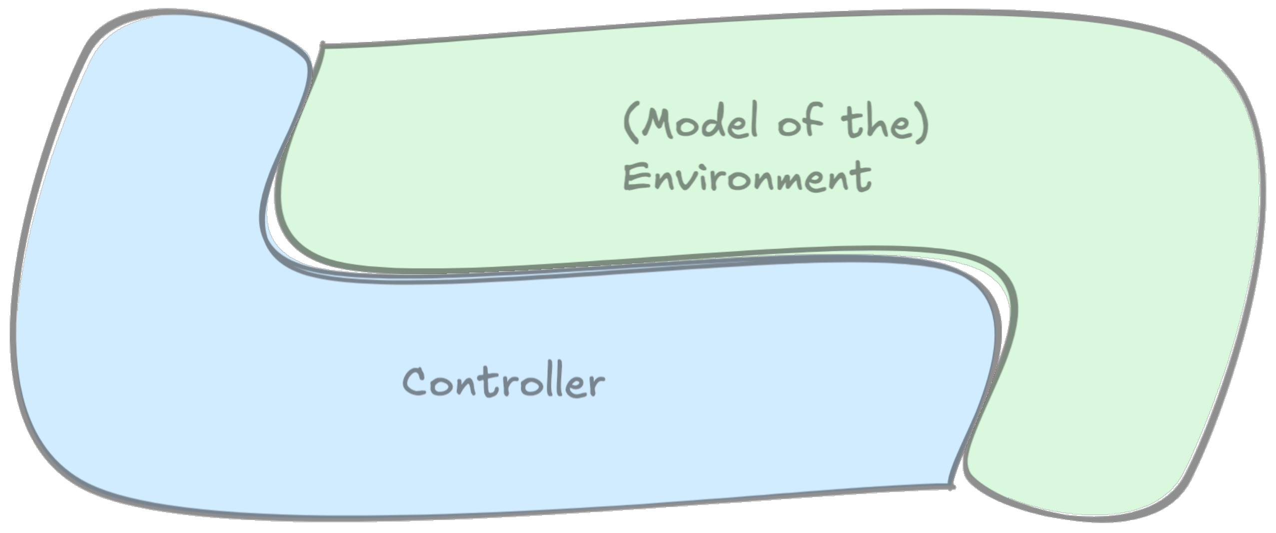
Control theoretic models are "trivial" from a Bayesian perspective

This interpretation is however:

- possibilistic (not probabilistic)
- the Bayesian model κ is an approximation (it groups and updates *indistinguishable* states of the env.)
- "trivial" since observations are ignored
- one where controller updates are deterministic



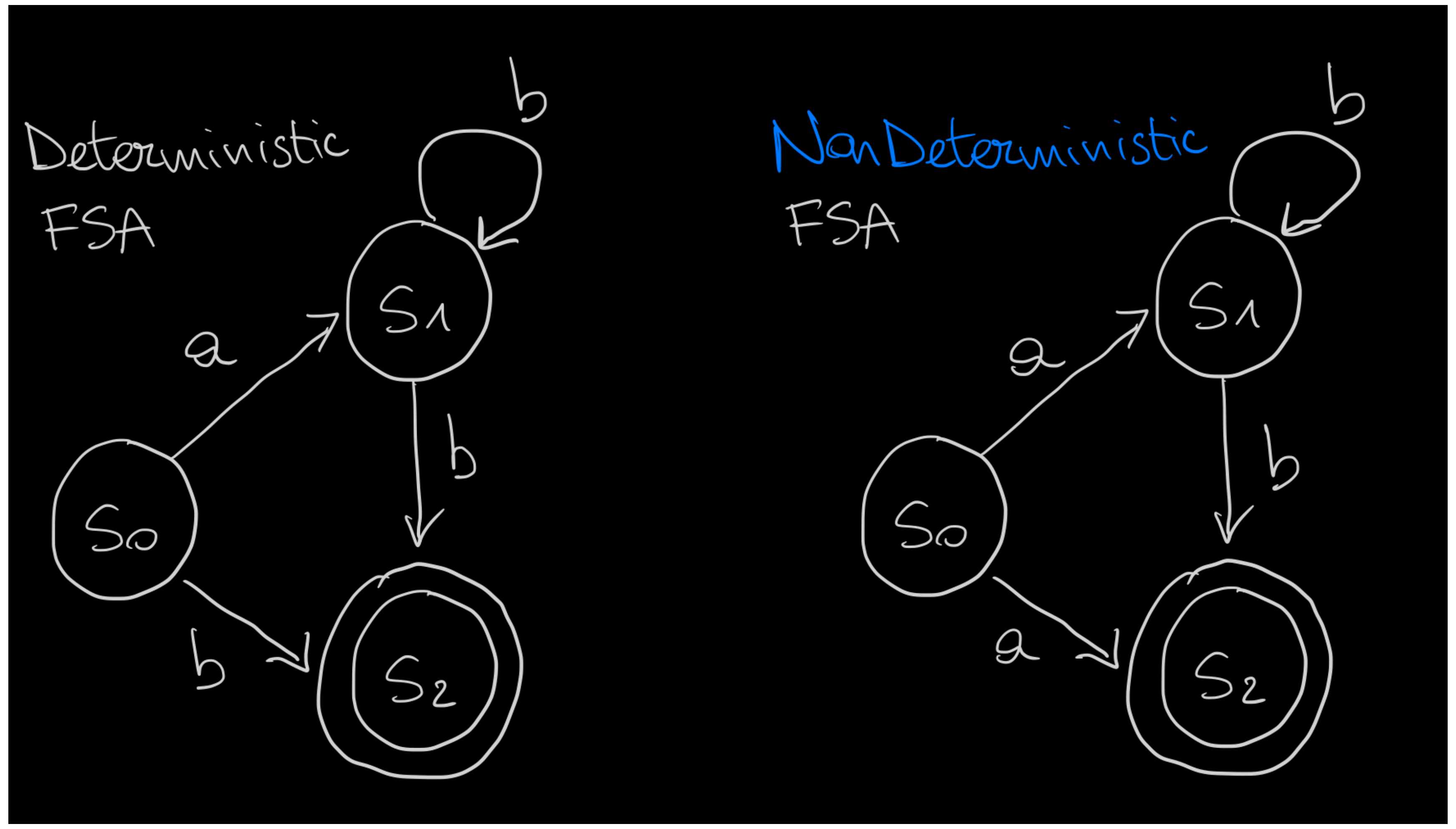
(56)

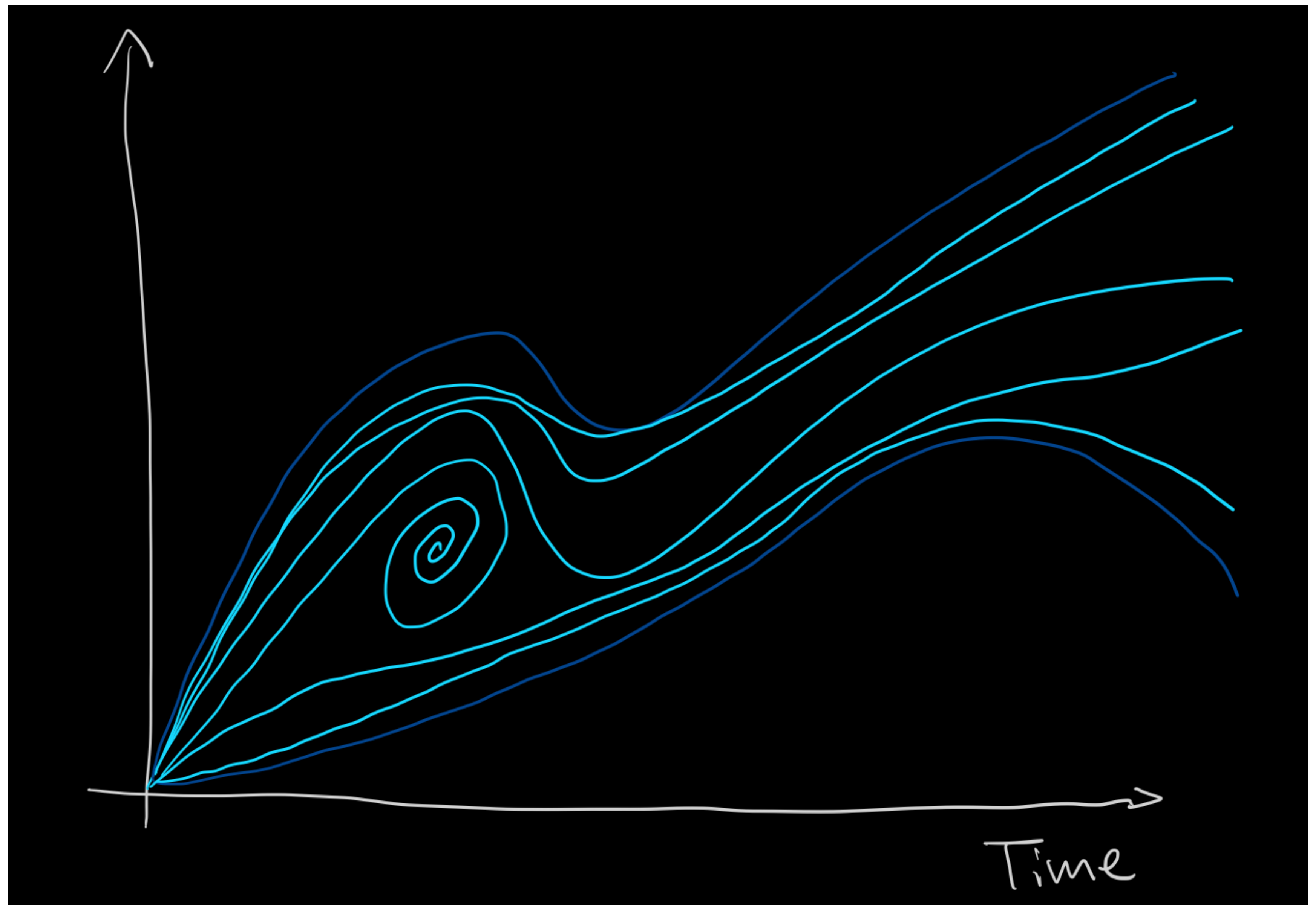


Possibilistic uncertainty

Beliefs without probabilities

- Non-deterministic automata (computer science)
- constructor theory (physics)
- viability theory (dynamical systems)

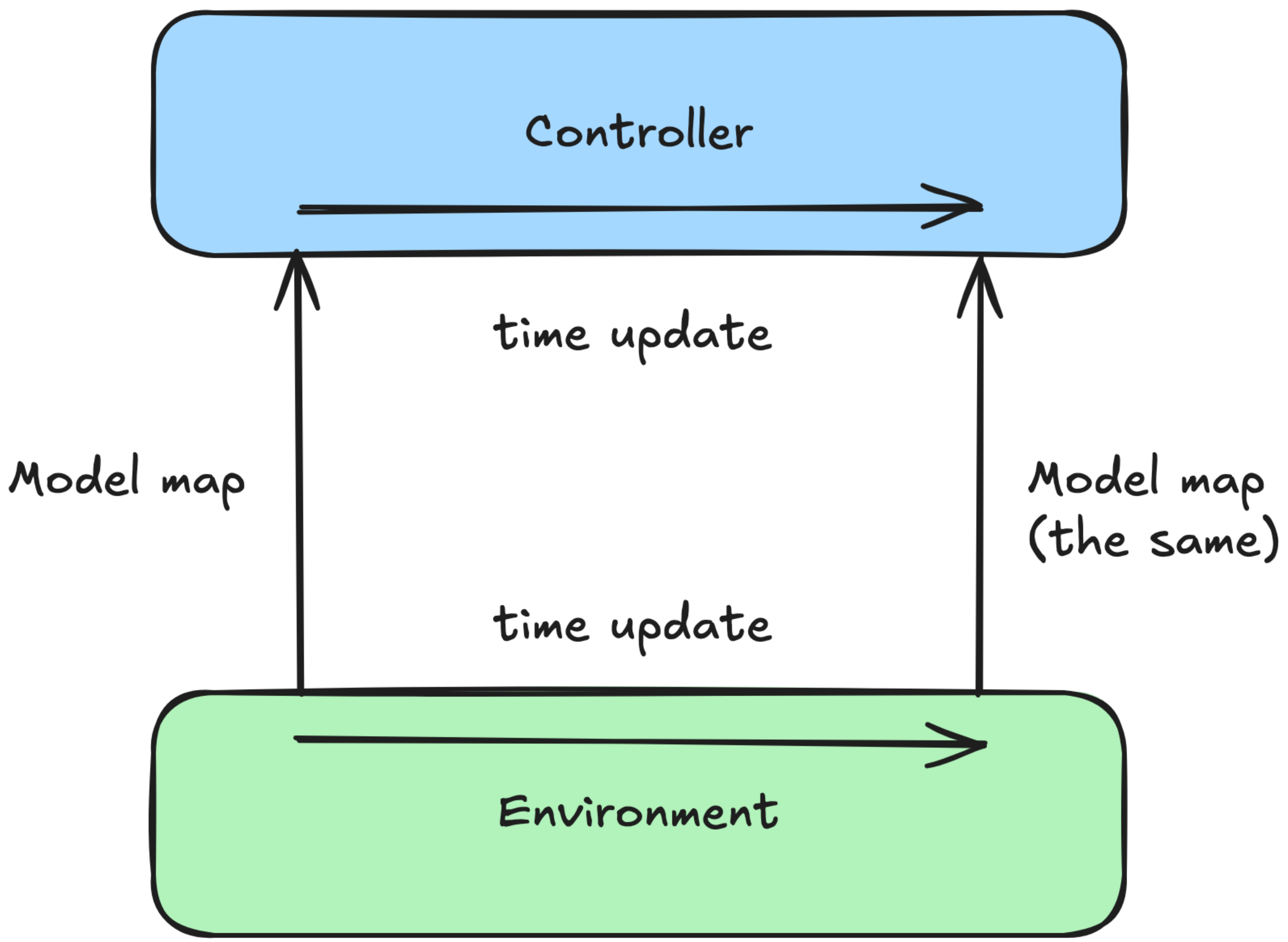




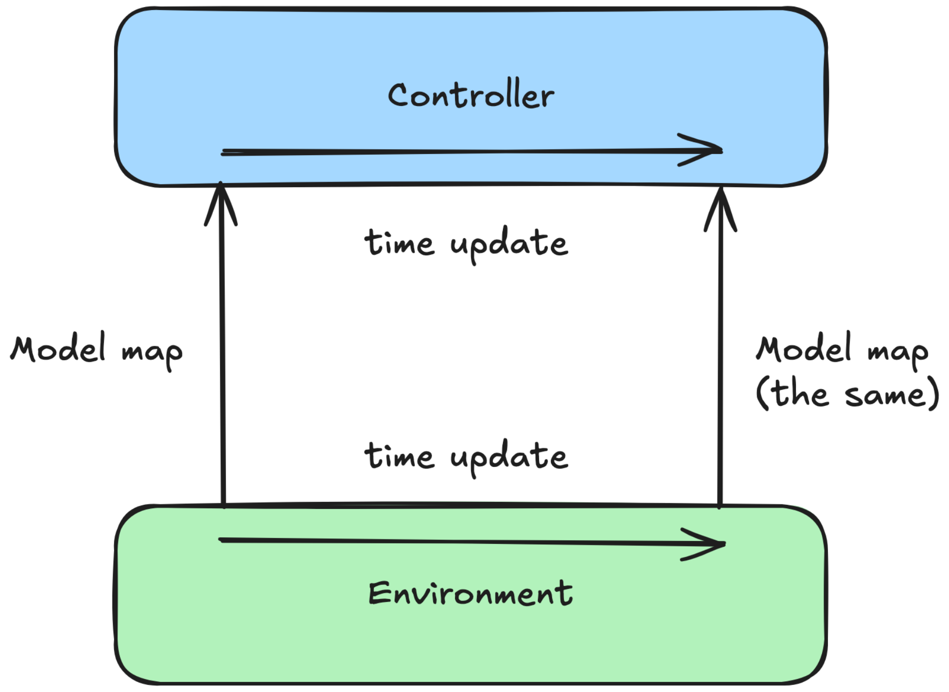
Two perspectives

An example

- Controller: the army **outside** the castle
- Environment: the army **inside** the castle
- Task for the controller: survive arrows from army inside castle


Controller
time update
Model map
Model map
(the same)
time update
Environment

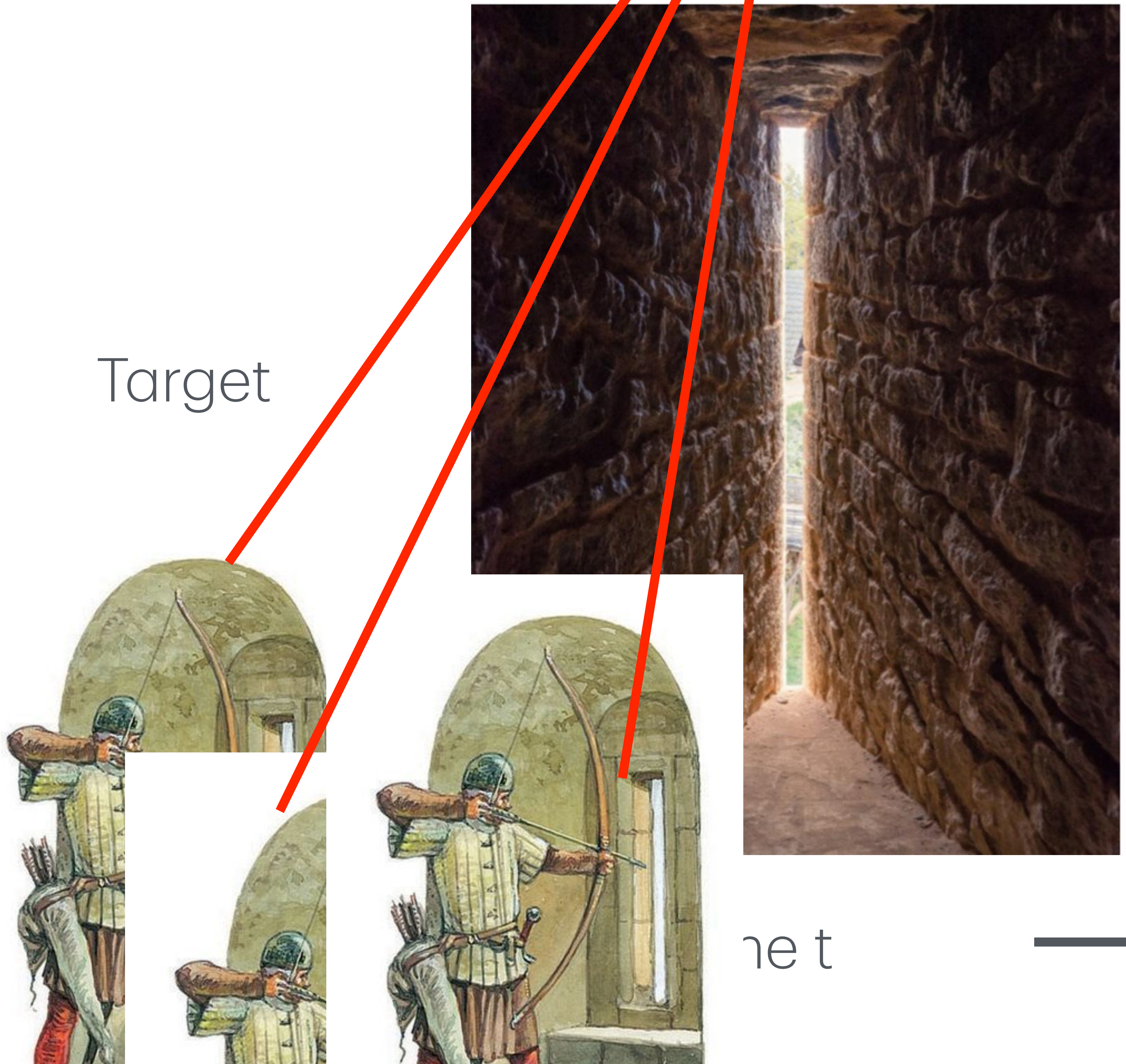
Controller
time update
Model map
Model map
(the same)
time update
Environment
Q: What should we have here?
A: The same target, in their new
position, targeted b he same 3
archers
Target
Model map
Model map
ne t
Time t + 1
Move place/window



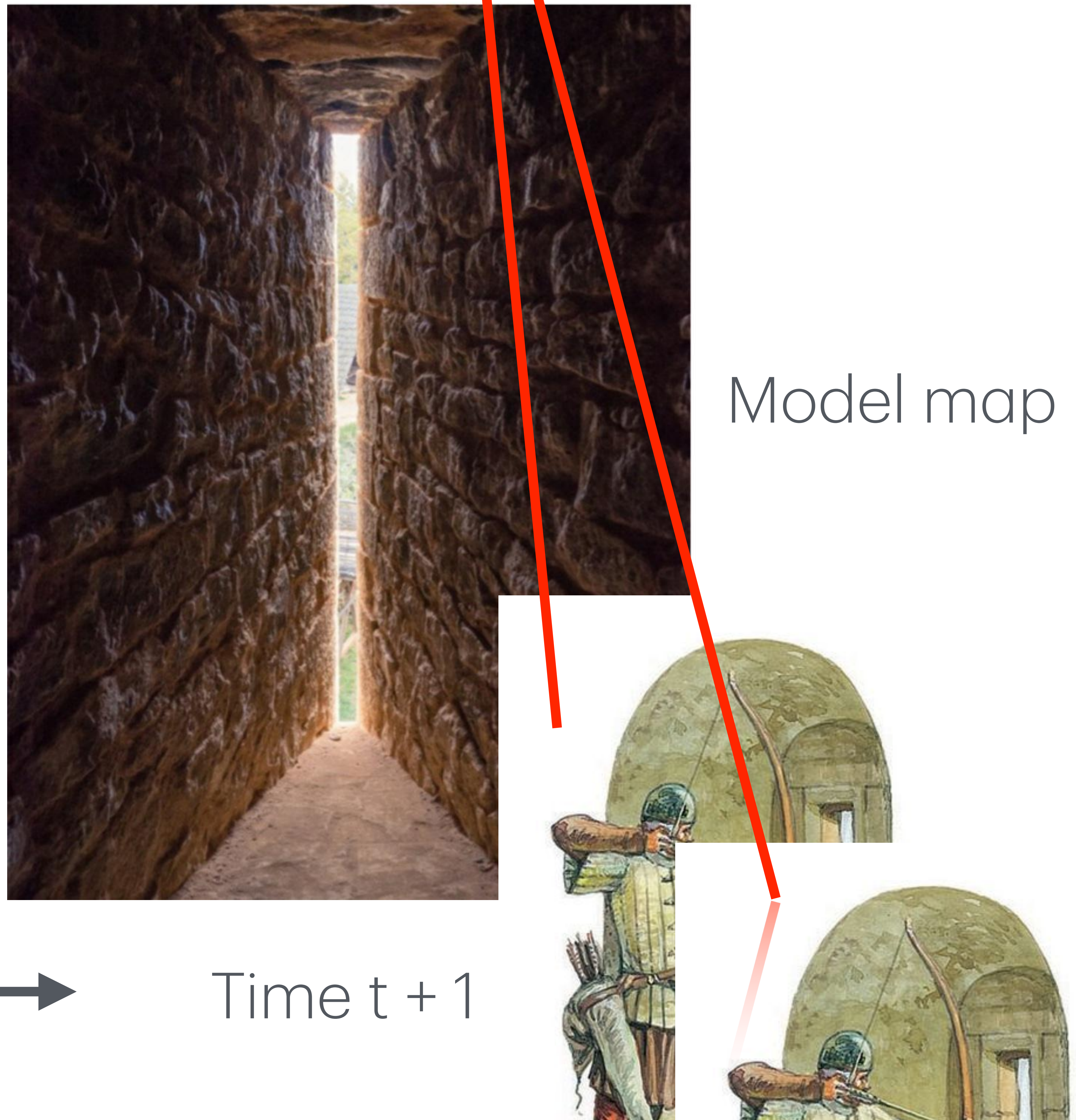
An archer gets hit, can't reach the window

Not a model

Target



Model map

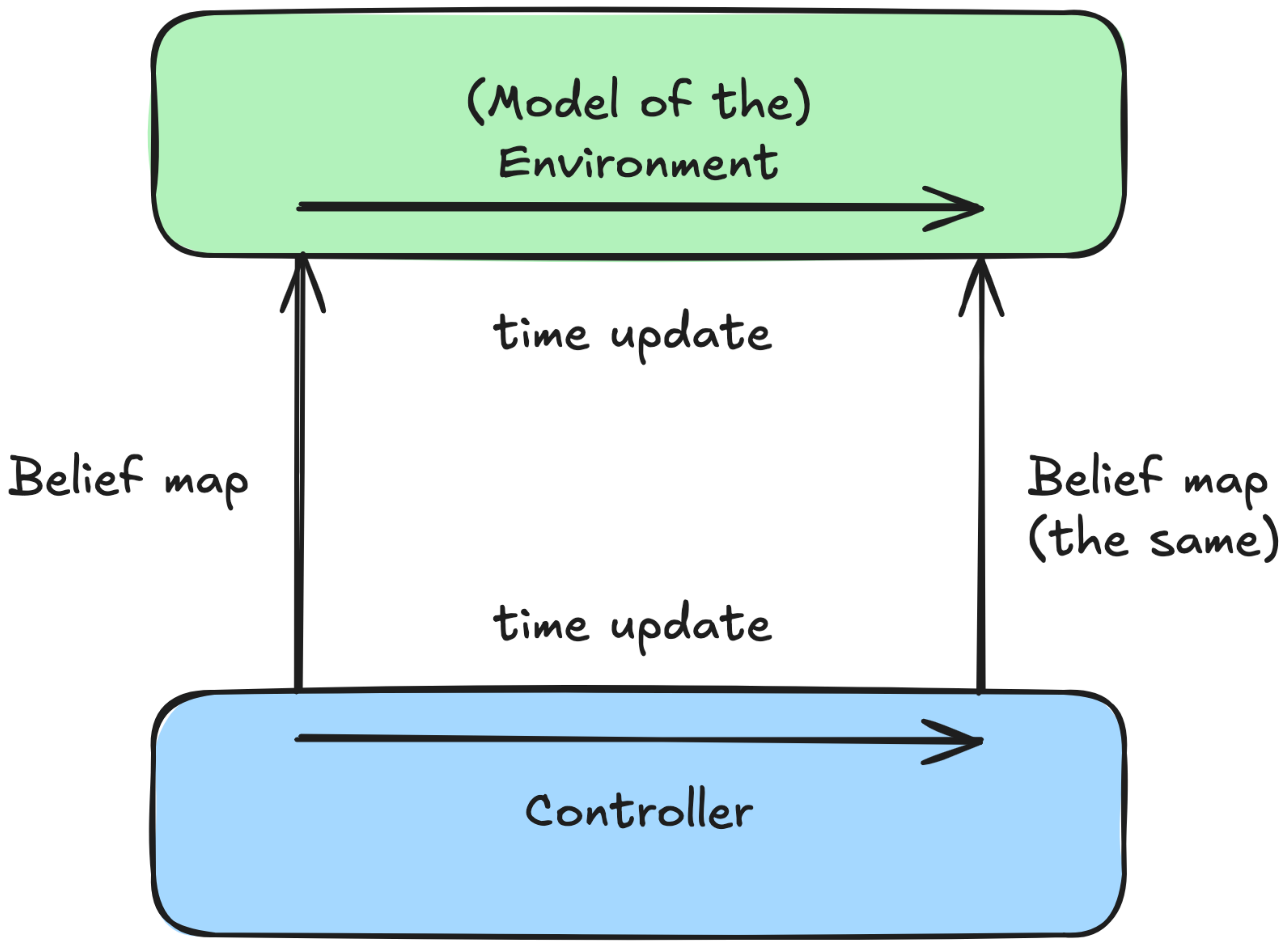


Model map

ne t

Move place/window

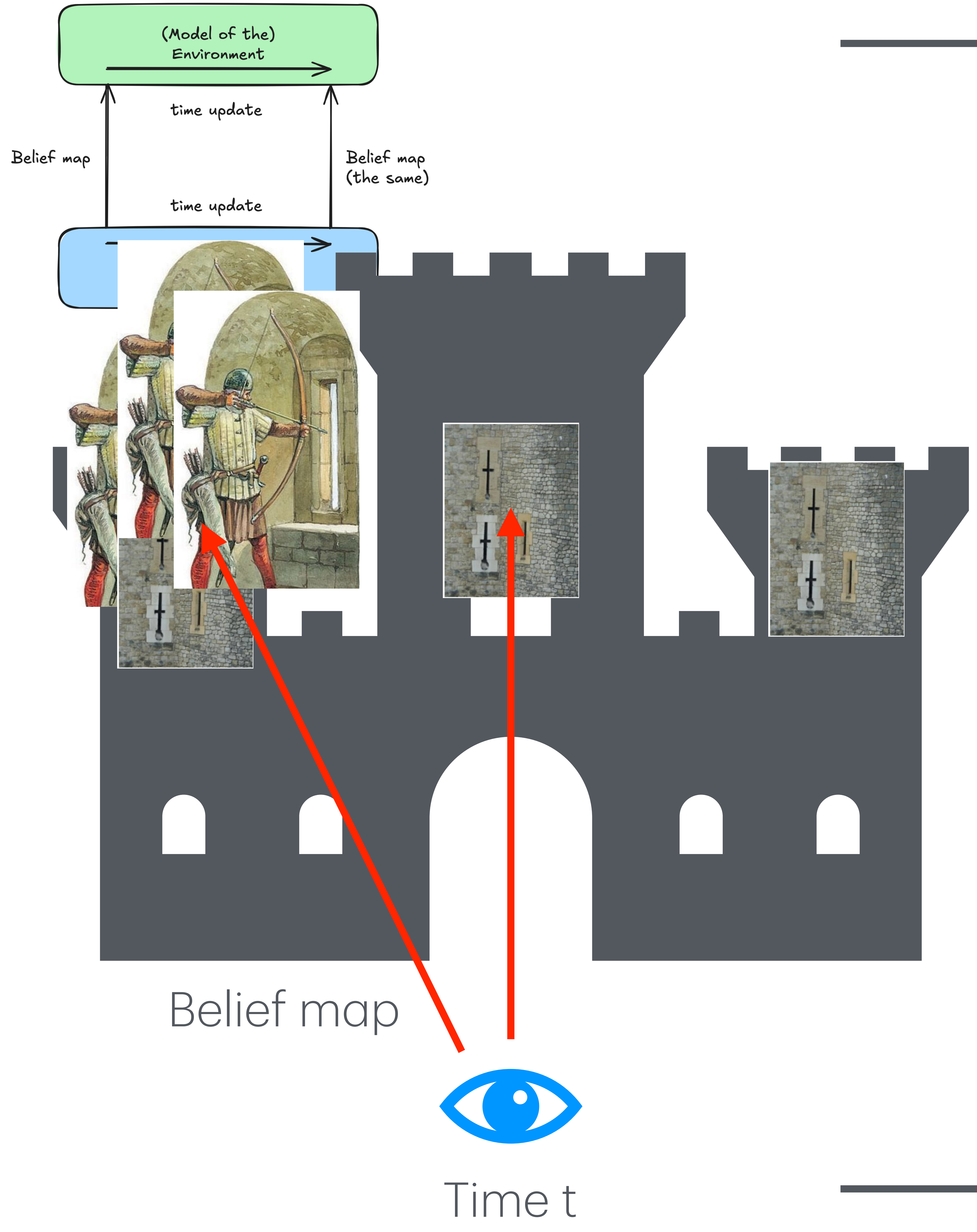
Time t + 1

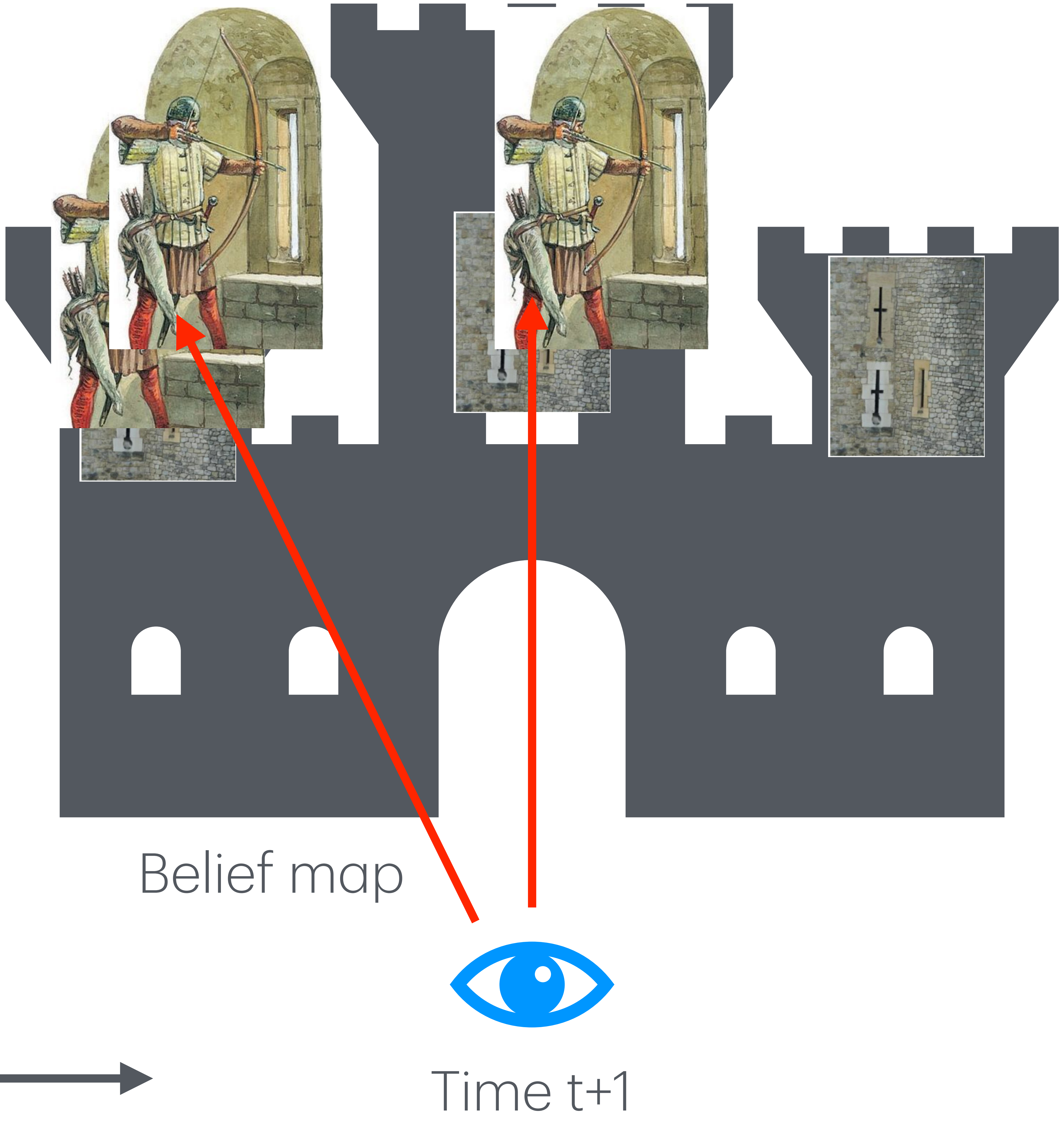


(Model of the)
Environment
time update
Belief map
Belief map
(the same)
time update
Controller

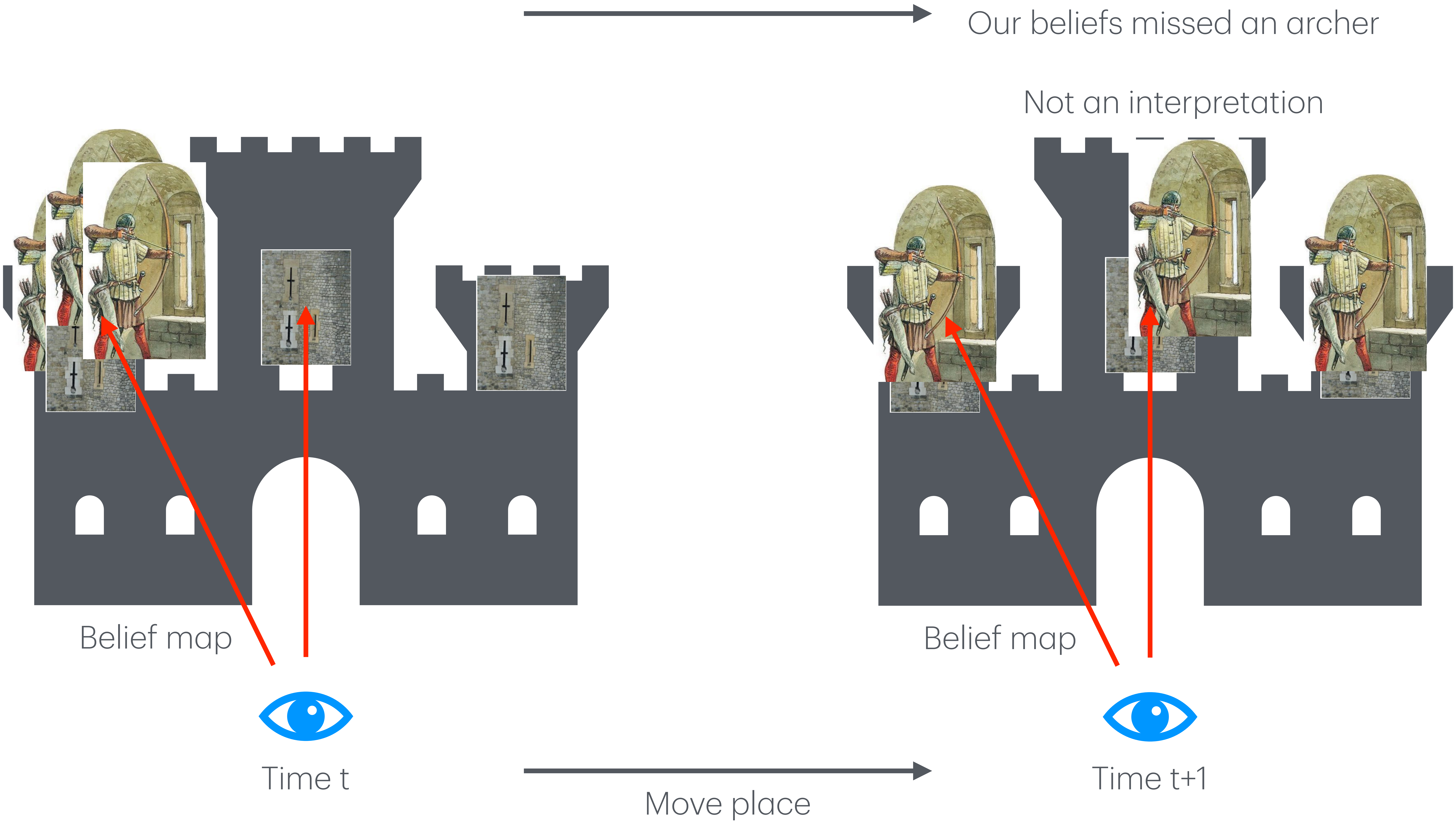
Q: What should we have here?

A: Position of archers with same target (us), consistent with previous beliefs





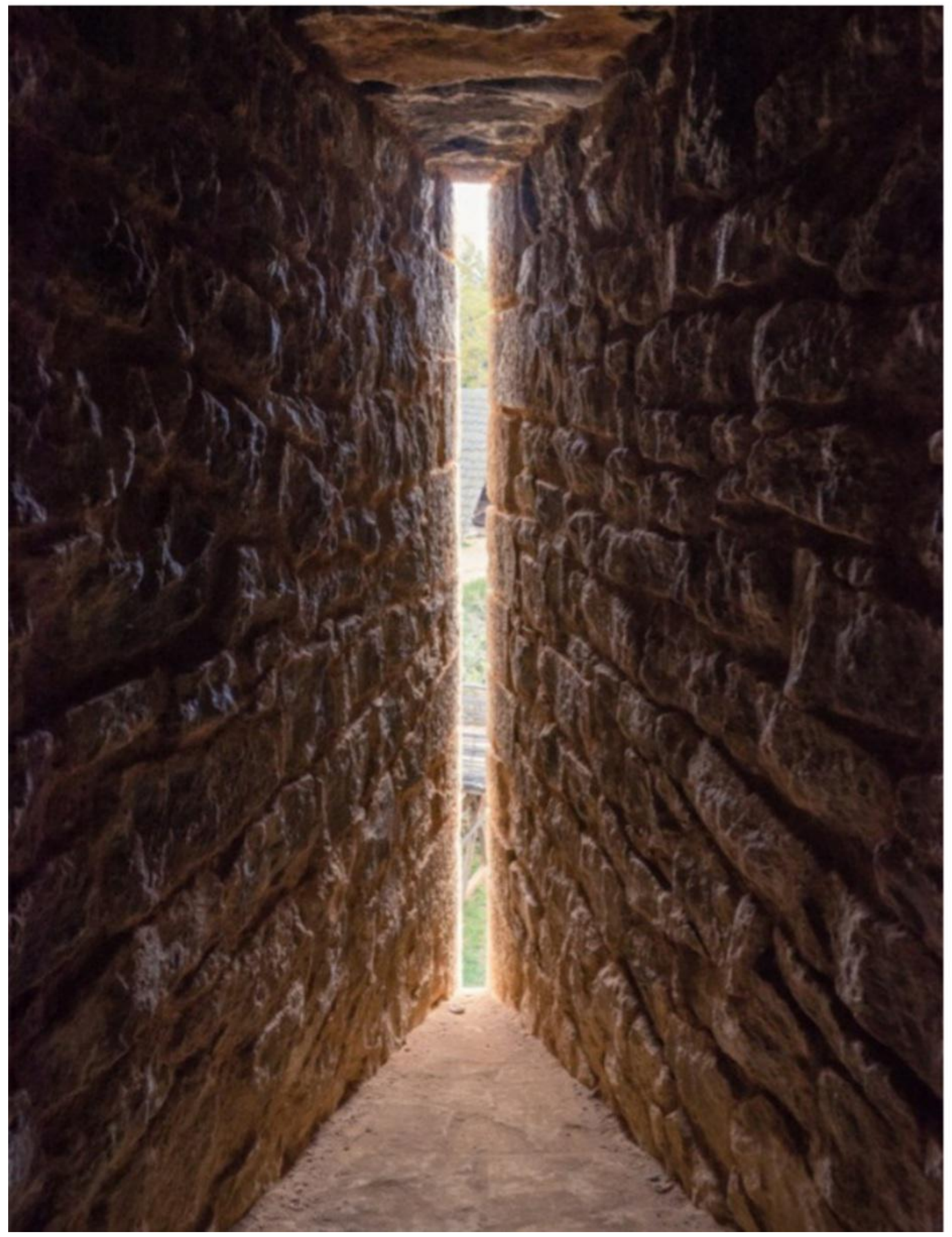
Move place



Our beliefs missed an archer
Not an interpretation
Belief map
Belief map
Time t
Time t+1
Move place

Model map/belief map

They decide how narrow the slit is (modulo probabilities)



Implications

And applications

IMP describes consistency between systems

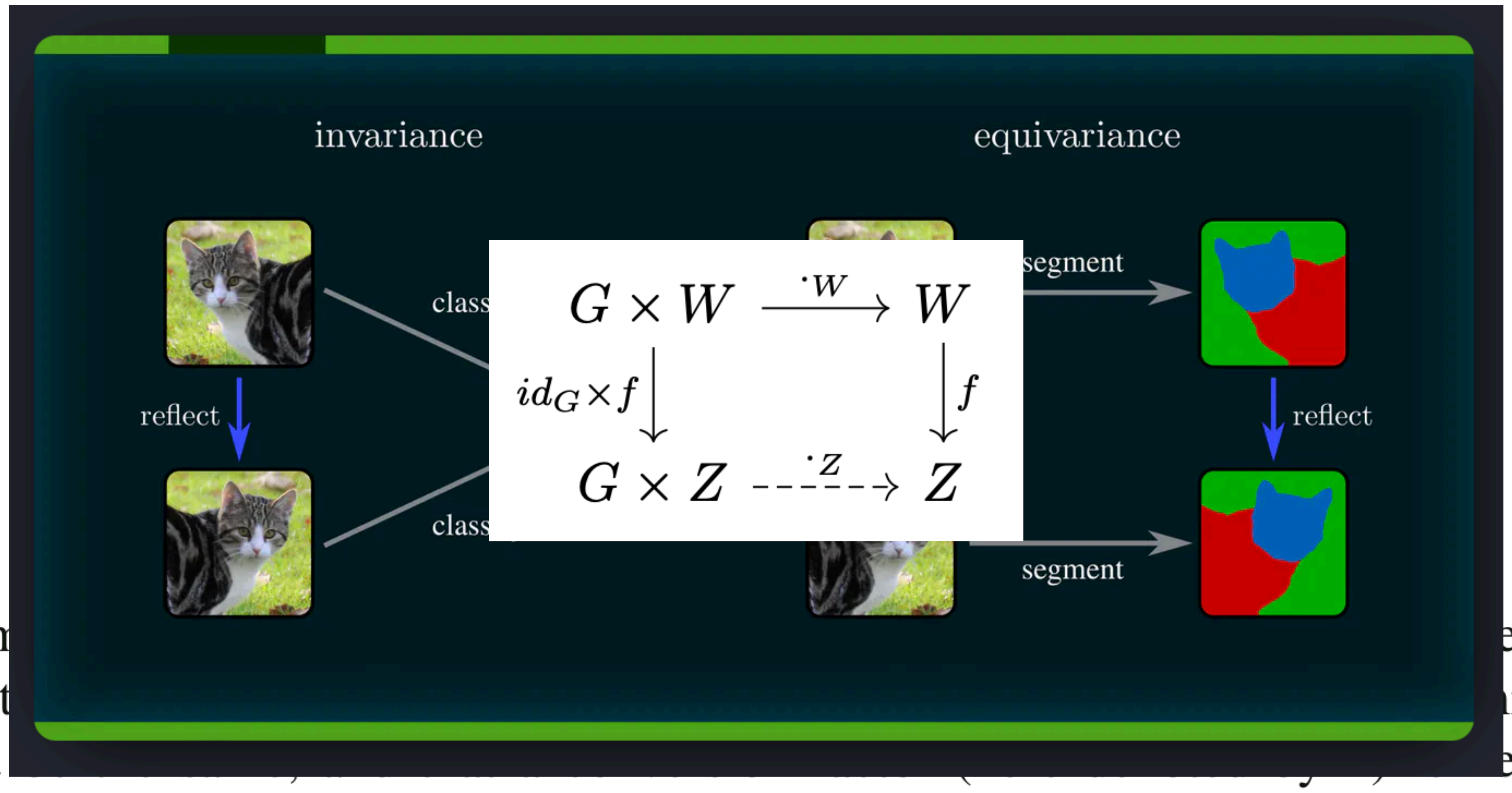
- brain-environment
- agent-environment
- controller-environment
- …

A pre-requisite for any good notion of model?

This appears in:

- theoretical biology (Rosen)

Finally, we must now introduce some dynamical considerations, to capture the idea that M is a *predictive* model. To do this, we must recall some properties of temporal encodings of dynamics, as they were described in Sect. 4.5 above. Let us suppose that $T_t : S_1 \to S_1$ is an abstract dynamics on S_1. If M is to be a dynamical model of this abstract dynamics, then there must exist a dynamics $\bar{T}_{x(t)} : M \to M$ such that the diagram



(6.3)

con … e instant t; it … ics will not … ecessary between them.

Results

In summary

1. Definition of “model” generalising coarse grainings and the likes, compatible with physics/control theory definitions

2. Proved that every “model” implies a Bayesian filtering interpretation (the reverse is not true because…)

3. This interpretation is very special

Martin Biehl

Matteo Capucci

Nathaniel Virgo

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

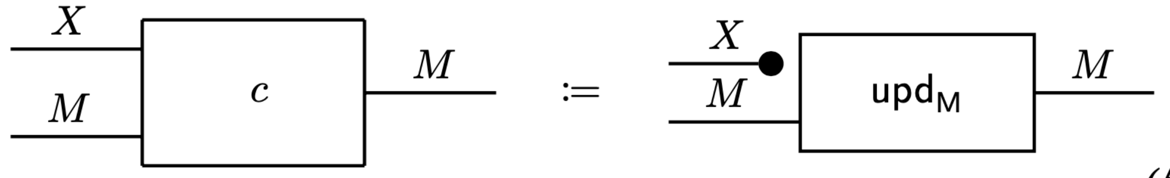
- a system $\mathsf{M} \in \mathbf{Sys}\binom{J}{M}$ (the *archetype*), and
- a map of systems (the *model per se*)

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

such that

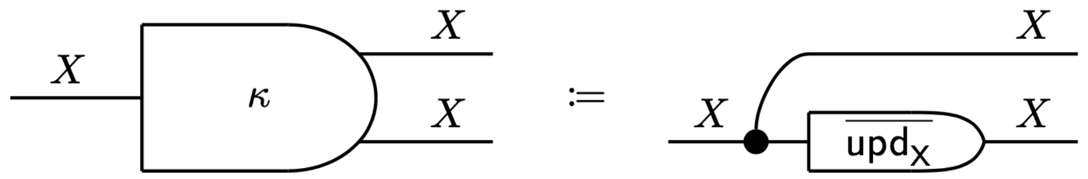
1) its part on states $\mu_s : X \to M$ is surjective, and
2) its part on inputs $\mu_i(x, -) : I \to J$ is surjective for each $x \in X$.

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*



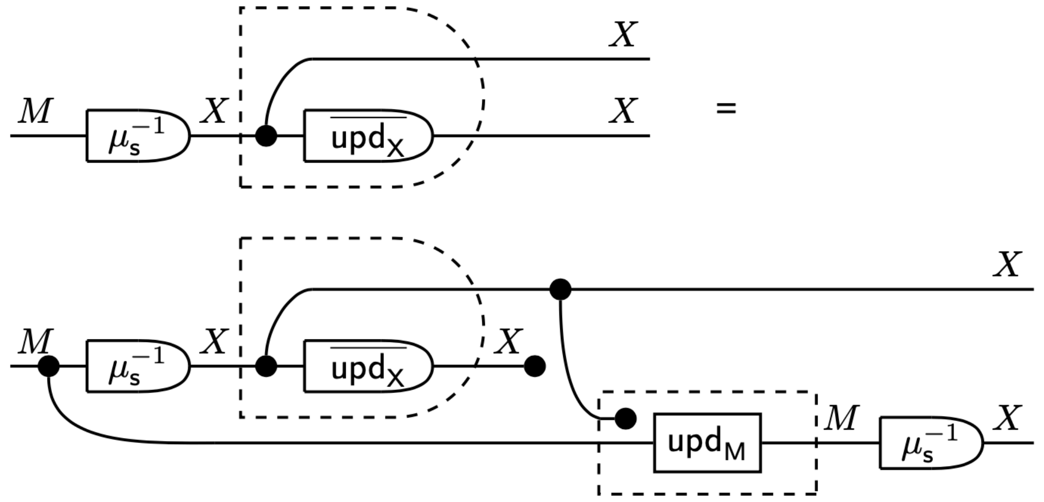
(51)

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



(52)

Then κ *is the hidden Markov model, and* $\mu_s^{-1} : M \nrightarrow X$ *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. □