Towards a mathematical theory of intentional systems

Nathaniel Virgo

Martin Biehl
Manuel Baltieri
Matteo Capucci
Simon McGregor
timorl

University of Hertfordshire UH
ELSI Earth-Life Science Institute
Tokyo Institute of Technology

Funded by the John Templeton Foundation

Big motivating questions

What makes an agent, an agent?

Must every agent have a model?

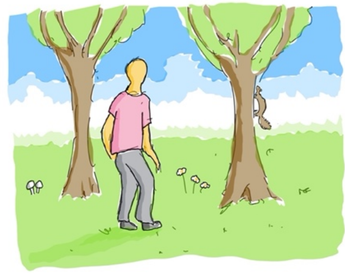
How does mind relate to the physical world?

In the end what I'll show you is mathematically very simple, but I hope I can get you interested in the questions.

Feel free to interrupt!

Start by thinking about "coupled dynamical systems"/ "open systems"

You probably know several ways to formalise this idea.



neither system can know the other, except by how it behaves at the interface.

At the level of physics it seems reasonable to think the world is made up of coupled systems

Yet somehow it feels like agency etc. are "something more than that".

Philosophers & cognitive scientists have many different ideas about why that is.

I pick on a few that will be relevant:

Conant & Ashby

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

and its relatives

Embodied cognition / dynamical approach:

Cognition happens largely in the coupling between the organism and its environment.

To a large extent cognition <u>can</u> be explained in terms of coupled dynamical systems.

"The world is its own best model"

The Bayesian brain

Things inside the brain <u>parametrise</u> <u>Bayesian</u> <u>priors</u> — we'll see in detail what that means later

Constitutive enactivism

Our biological nature is important!

We're not <u>just</u> dynamical systems because we can die. The interface doesn't just exist, our biology has to create it.

Meaning arises from our need to stay alive.

Dennett & the intentional stance.

I'll unpack these in more detail

The "good regulator theorem"

Famous paper from early control theory,
back when it was called cybernetics

Conant & Ashby (1970)
'Every good regulator of a system must be a model of that system'

A flawed paper – it doesn't really show what its title claims

But what they were trying to do is interesting – can we improve on it?

↳ the main thing I'll present is an attempt to do that.

We know we often <u>can</u> solve control problems (like POMDPs) by model-building,

but is that the only way?

If we try to build a model some other way (e.g. by training a neural network), will it secretly have a model anyway?

What about "the world is its own best model"?

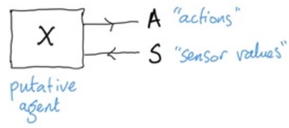
Do embodied agents provide counterexamples?

Dennett's intentional stance

(rough characterisation)

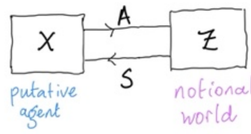
When we say something is an agent, we mean we can predict its behaviour by treating it <u>as if</u> it is an agent.

Suppose we have a system



To "treat it as if it is an agent" we must supply extra data:

- a <u>notional world</u> that the agent can couple to,



- a <u>goal</u> that the agent is supposed to be trying to achieve
  (e.g. utility function, desired state $z \in Z$, etc.)

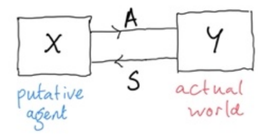
Then we ask,

"What are the optimal actions an agent <u>should</u> take, to achieve this goal?"

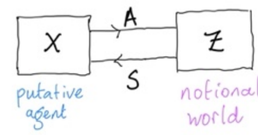
If this yields good predictions of the agent's <u>actual</u> behaviour, then we've successfully applied the intentional stance.
(Even if the predictions weren't perfect.)

This sounds tautological, but for some systems / goals / notional worlds this will be much more successful than others. The patterns in X's behaviour that make it possible are a genuine property of X.

(**aside:** I see this as similar to the ("Real Patterns") MaxEnt approach to statistical mechanics (E.T. Jaynes) — we assume optimality "by default", and when it fails we learned something about the system. Biologists use flux-balance analysis in a similar way.)



≠



in general

The plan

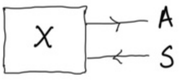
1. simple model of coupled systems
2. beliefs & Bayes
3. a "good regulator theorem" for embodied agents

+ optional extras

Coupled systems

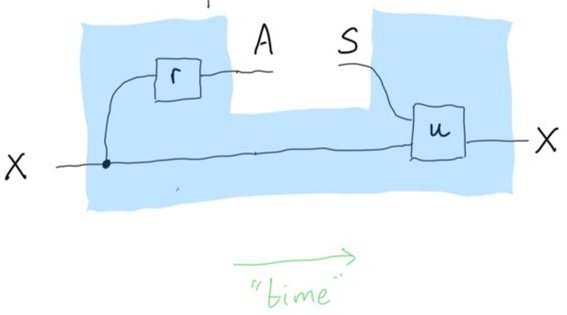
Moore Machines
(the simplest version)

A Moore machine consists of



- an interface, which is
  - a set A of possible actions (outputs)
  - a set S of possible sensor values (inputs)
- a state space X
- a readout map $r: X \to A$
- an update map $u: X \times S \to X$

A useful picture:



Many variations of this concept

- optics
- generalised lenses ($C^{op} \to Cat$)
- polynomial functors
- ⋮

Categorical Systems Theory

(David Jaz Myers, Matteo Capucci, Sophie Libkind, ...)

gives a very nice way to reason about Moore machine like things and how to couple them.

We won't need most of it for this talk, though!

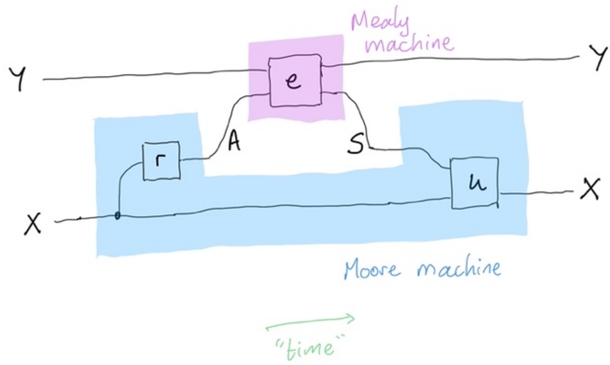
We'll need the concept of a Mealy machine:

A Mealy machine with interface (S, A) consists of

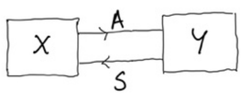
- a state space Y
- a map $e : A \times Y \to S \times Y$

We use Mealy machines to model environments, so
actions A are inputs
sensor values S are outputs.

Useful picture (of how Moore & Mealy machines compose)

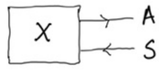
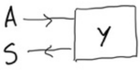


a.k.a.



Unnecessary but interesting detail:

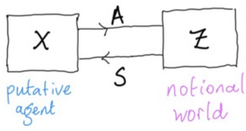
In this diagram,

-  is an object in a monoidal double category of interfaces (**Lens** in this case)
-  is an element of a kind of copresheaf on this double category (lax monoidal, lax double functor $\mathbb{I}\mathbf{nterfaces}^{\mathsf{T}} \to \mathbb{S}\mathbf{pan}$)
-  is an element of a corresponding kind of presheaf.

Beliefs and belief updating

Neither Dennett nor Conant & Ashby say much about this, but I think it's crucial.

Given

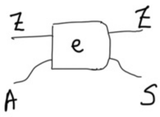


what could it mean for X to "have beliefs about Z"?

Start by asking:
what would it mean for *us* to have beliefs about Z?

To make it more fun (& appeal to Bayesian intuitions), let's instantiate things in a Markov category.

So now



is a stochastic map

At each time step

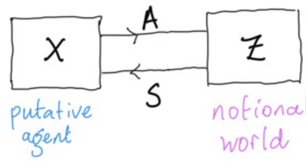
- we get to choose an action $a \in A$
  (for now don't worry about how this choice is made)
- we recieve a sensor value $s \in S$
- but we don't get to observe the state Z directly.

Still, the sensor values give us some partial information about Z. How can we keep track of it?

Bayesian answer:
Maintain a prior over Z and update to a posterior at each time step,

$$\underbrace{p_1(z'|s,a)}_{\text{posterior}} = \frac{\overbrace{\sum_z}^{} \overbrace{p_0(z)}^{\text{prior}} \overbrace{p_e(z',s|a)}^{\text{likelihood}}}{\underbrace{p_0(s)}_{\text{normalisation} = \sum_{z,s} p(z)p_e(z',s|a)}}$$

This is a version of Bayesian filtering
(≈ the prior is over the current state, but the posterior is over the next one.)



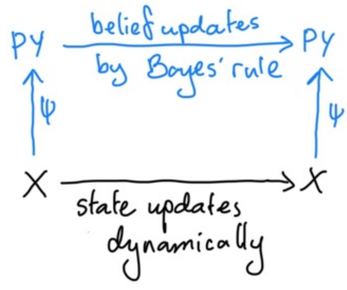
what could it mean for X to "have beliefs about Z"?

Our answer: it means we have specified a map $\psi : X \to PZ$

(PZ: probability distributions over Z)

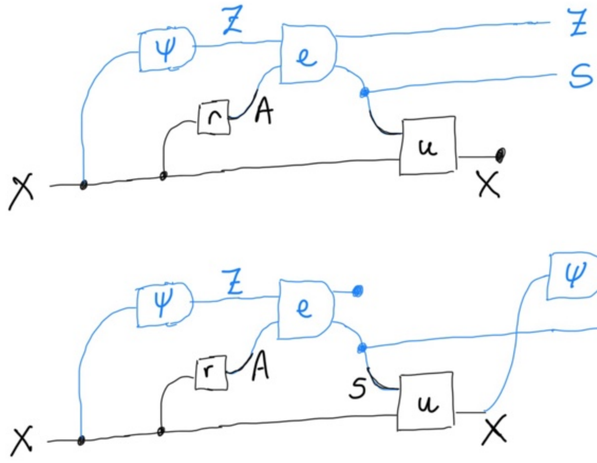
(equivalently just $X \to Z$ in a Markov category)

such that, informally,



commutes.

In string diagrams, one version of that looks like this:



I won't go through this in detail. The equation expresses that Bayes' rule is obeyed, as above.

We say the state X parametrises the prior / beliefs — this sort of thing is common in practical applications of Bayes.

For us, the important thing is that it provides a kind of semantic interpretation of the machine X. When the machine is in state $x \in X$ it has Bayesian beliefs $\Psi(x)$ about Y.

But as you might expect, a given machine may have many different semantic interpretations.

Possibilistic Bayes

Sometimes we don't care about assigning probabilities to states, only about which ones are possible and which aren't. (which ones are in the support.)

We can model this as a map $\Psi : X \longrightarrow \mathcal{P}(Y)$.

Those complicated string diagrams become

$$\underbrace{\Psi(u(x,s))}_{\text{belief at next time step}} = \underbrace{\{z' \in Z \mid \exists z \in Z.\ (z', s) = e(z, r(x))\}}_{\text{every state that could be entered, given the current beliefs \& known values of } s \text{ and } a.}$$

This is closer to a logical interpretation where states of X map to predicates on Y, rather than probability measures.

We can also allow forgetting

$$\Psi(u(x,s)) \supseteq \{z' \in Z \mid \exists z \in Z.\ (z', s) = e(z, r(x))\}$$

belief at next time step

every state that could be entered, given the current beliefs & known values of s and a.

The agent can forget constraints, so its posterior beliefs can be a superset of the logical ideal ones.

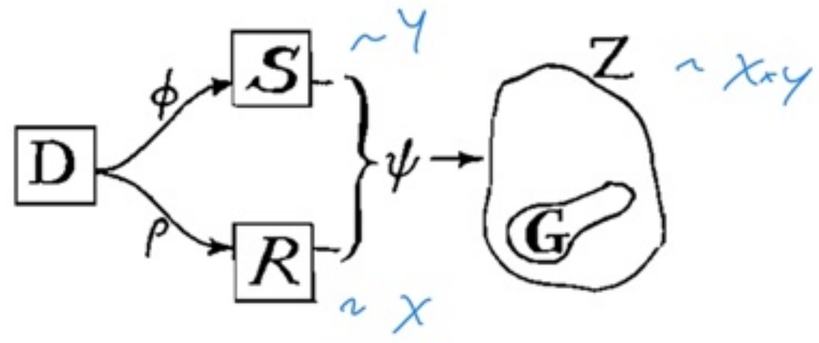
This is necessary to make the next step work.

Open questions ...

Good Regulators

Back to Conant and Ashby:

Fig. 1

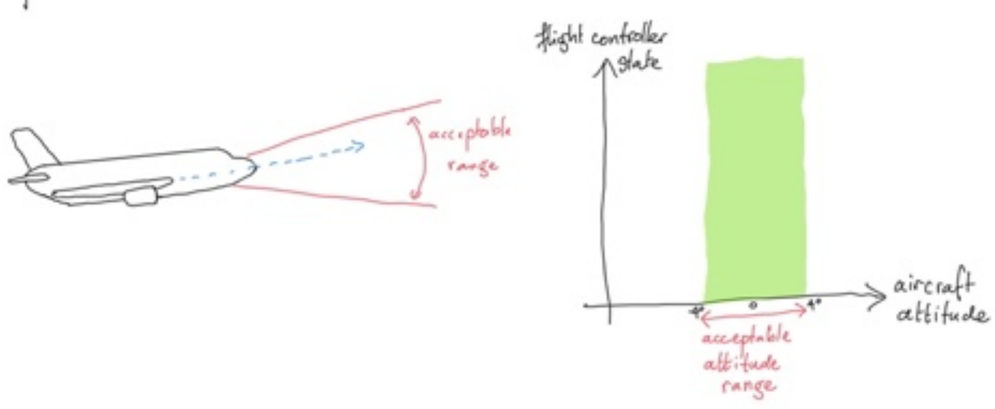


Their definition of good regulator involves a "good set" In our case this will be a subset $G \subseteq X \times Y$ of "good states."

If we pretend the state spaces are continuous, we can draw this as

We'll say the controller is a <u>good regulator</u> if it's <u>possible</u> for the combined system to stay in the good states indefinitely. (I'll unpack that shortly.)

In control theory, the good set usually arises from a set of desired states of the <u>plant</u>



In biology, the "controller" is an organism, and the good set usually arises from a set of "viable" states — states of the controller, not the plant.

W. Ross Ashby - Design for a Brain

Randall Beer — Autopoiesis and Cognition in the Game of Life (and many other works)

connection to enactivism.

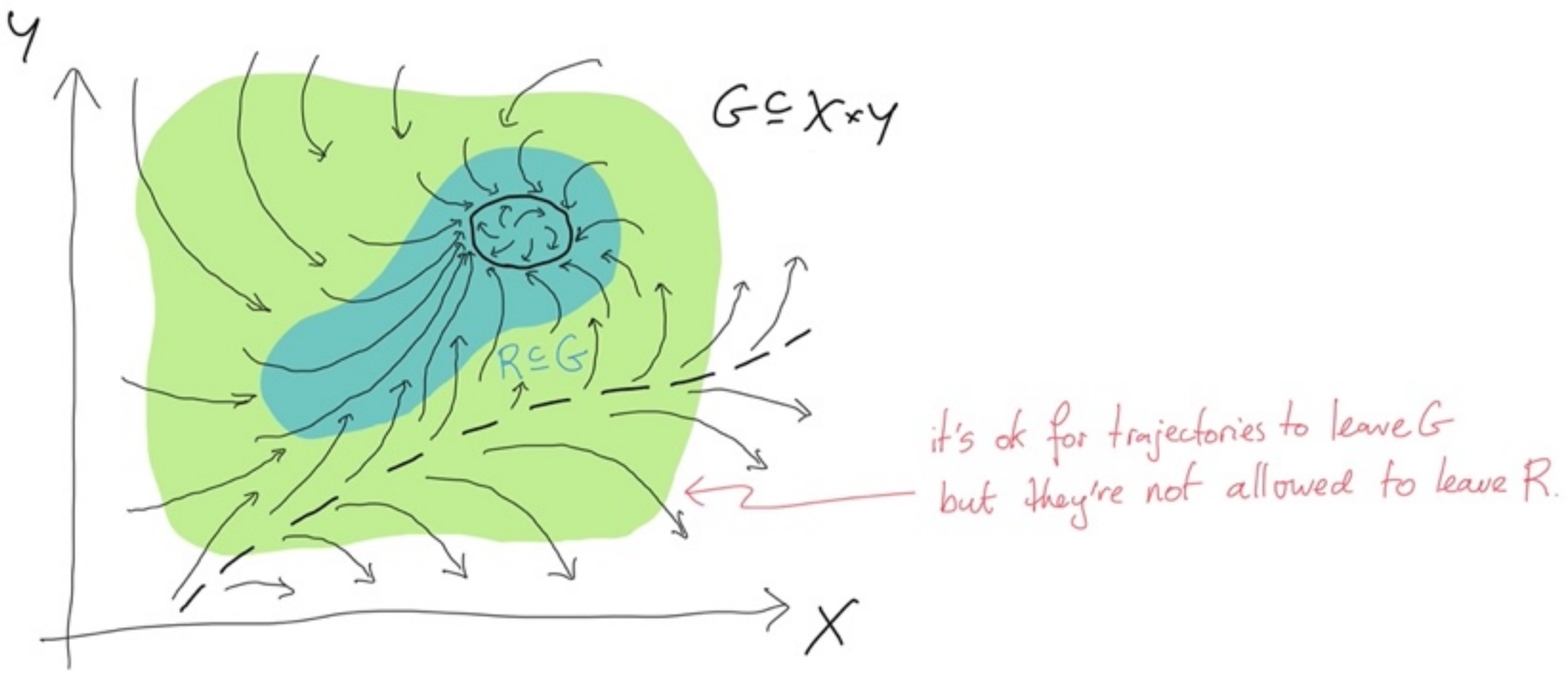
I said it has to be possible for the joint system to stay in the good states.

This means there exist trajectories that start in G and stay in G forever.

But we don't require all trajectories that start in G to stay in G.

i.e. G doesn't have to be forward invariant.

Instead there must exist a non-empty $R \subseteq G$ that is forward invariant.



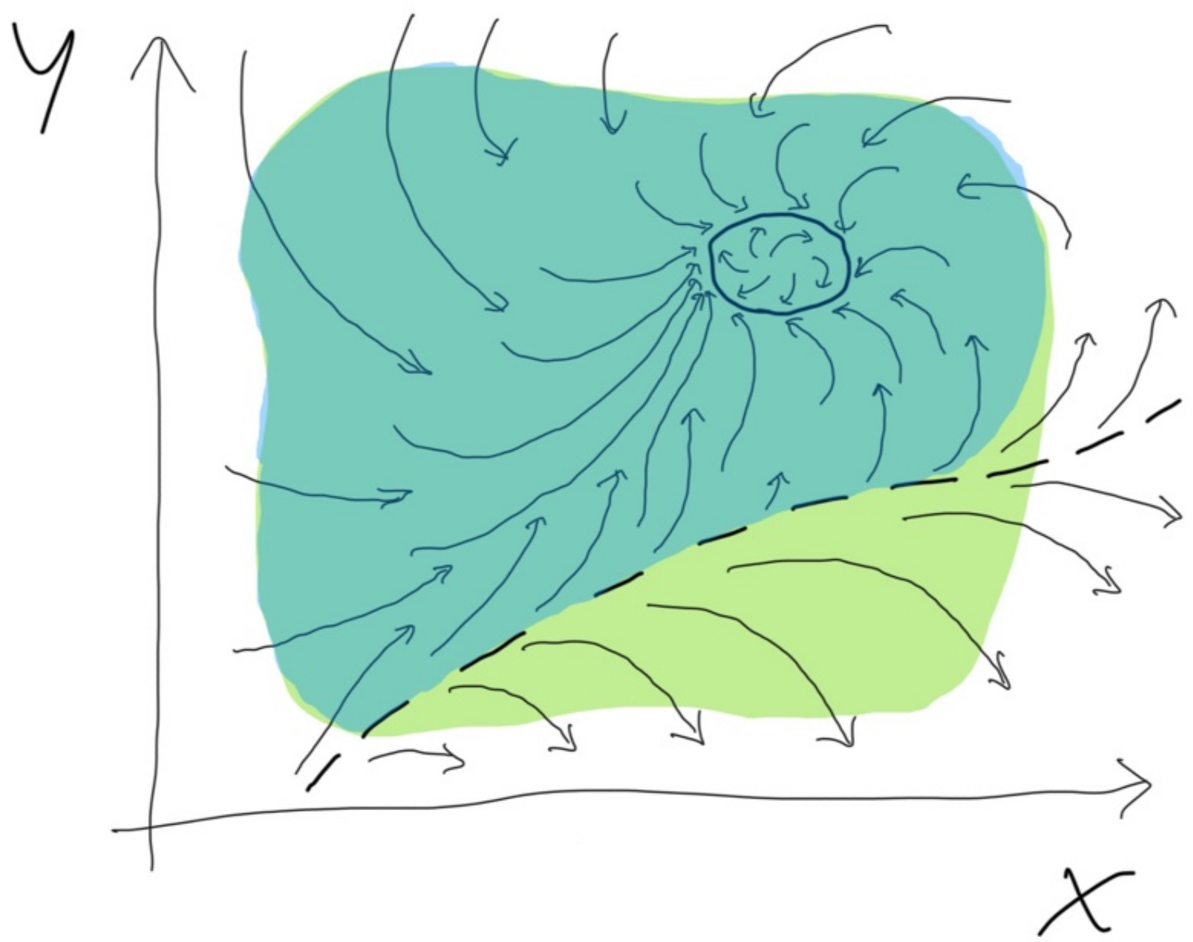
We call $R \subseteq X \times Y$ the "regulation relation," though it might have other names, e.g. "controlled invariant subspace."

It is a subset of $X \times Y$ that is

- non-empty
- forward-invariant
- "good", i.e. contained in G

If we can exhibit such a set we say the controller is a good regulator.

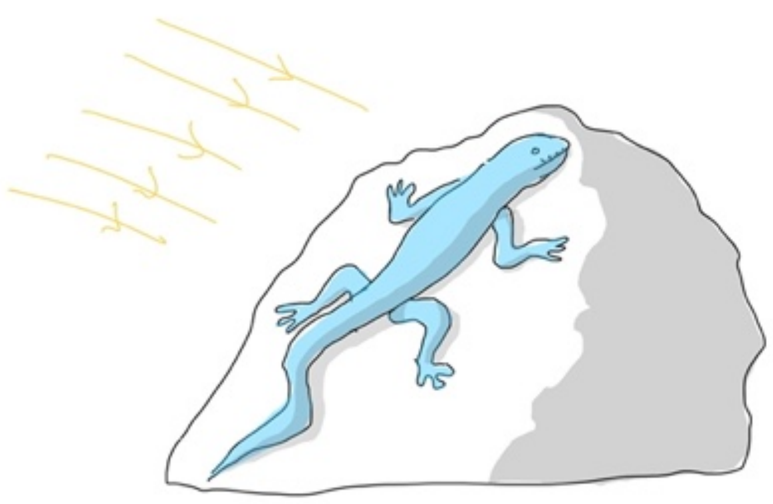
This set R is what we need for our "good regulator theorem."

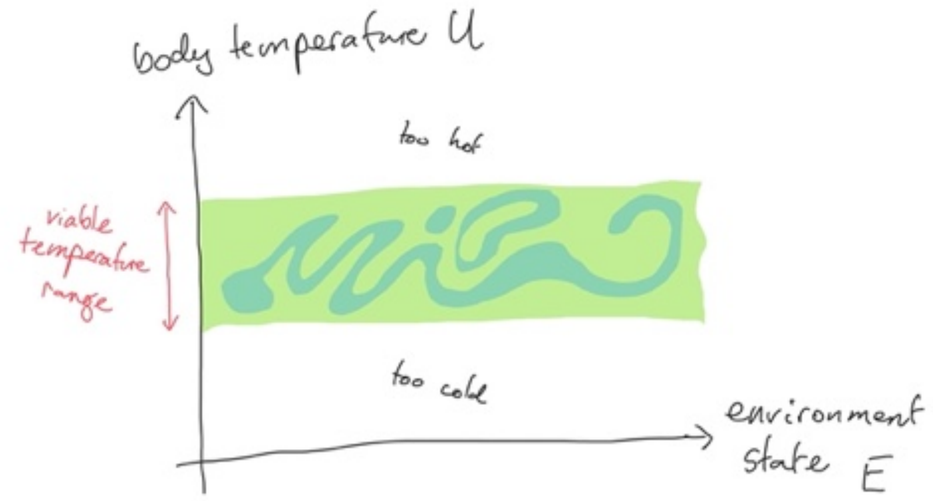


We can always let R be the largest subset of G, in which case we can define it coinductively.

Link to Bellman's principle

Even if the viability region is simple, R might be complex.





The "complexity" arises when the agent and environment need to be synchronised.

If agent and environment states are not correlated in the right way — if the system is outside R — then regulation will eventually fail.

"the future constrains the present"

Intuitively, this must be related to why the intentional stance is so effective for some systems. (Open questions.)

R is a relation

$$R \subseteq X \times Y$$

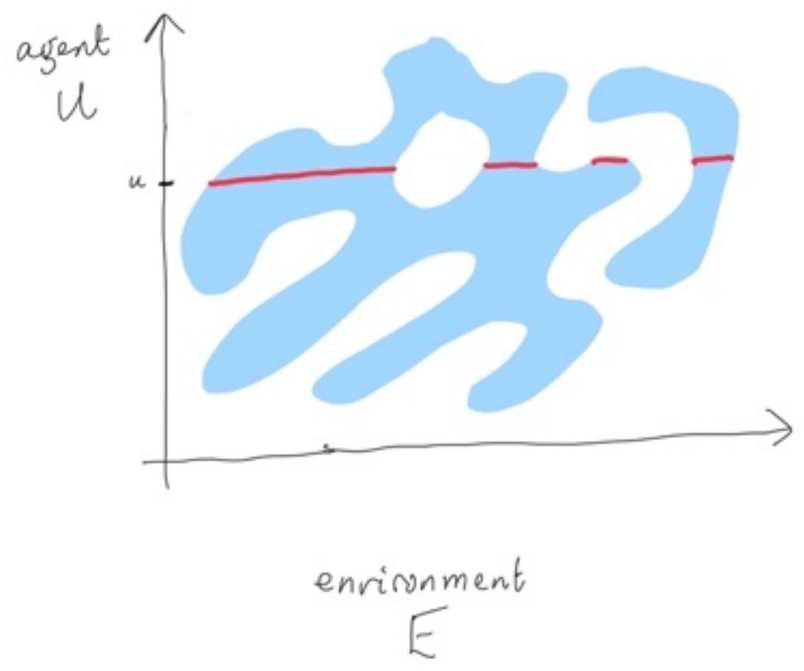
regulating set $\subseteq$ agent states $\times$ environment states

For each agent state u there is a set of environment states that it's related to. We call it

$$\Psi(x) = \{ y \in Y \mid (x, y) \in R \}$$

This is the interpretation map.

The idea is that if we know the agent is in state x, then the environment better be in one of the states $\Psi(x)$, otherwise the agent is in trouble.



Lemma

for any forward-closed set R, $\Psi(x)$ is a possibilistic Bayesian interpretation with forgetting,

$$\Psi(u(x,s)) \supseteq \{ z' \in Z \mid \exists z \in Z.\ (z', s) = e(z, r(x)) \}$$

This implies:

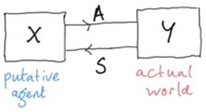
If we assume the agent thinks the environment is in one of the states $\Psi(x)$, then it will always "take actions" consistent with "trying" to stay inside the viable region.

(This can be made into a theorem.)

"every good regulator must have a model."

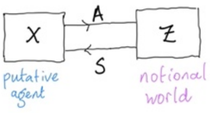
"every good regulator must have a model."

Specifically: if the agent is viable for some

$G \subseteq$ 

Then we can define from this a model

$\Psi : X \to \mathcal{P}(Z)$



where $Z = Y$

but models are not unique.

Counterexamples?

The "every good regulator..." claim always seemed odd to me...

Braitenberg Vehicle

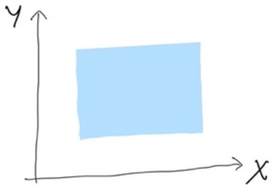
an open-loop controller (one with no sensors) can balance an inverted pendulum by vibrating the pivot up and down

+ many examples in biology & robotics where an appropriately designed body means less work for a controller
e.g. your legs.

Extreme counterexample:

Any "optimality implies model" result must account for these.

In our case, it's because the function $\Psi : X \to \mathcal{P}(Z)$ might not depend much on its argument.



What does the doorstop "know"?

→ Some tasks can be achieved without models, or with models that don't resemble the true environment, or are much simpler than it.

For these tasks, model building might not be helpful.

Conclusions:

- regarding Dennett:

  Formalising Dennet forces us to consider things Dennett doesn't say much about, such as the exact nature of beliefs, or multiple interpretability.

- regarding Conant & Ashby:

  If we take seriously that models are an "as-if" notion, then it's true that every good regulator has a model, in our sense.

  But this doesn't tell us much about simple regulators.
  (We suspect it will tell us a lot about systems that solve "representation-hungry tasks")

Open questions...

how to capture the MaxEnt-like aspect

how to model "hyperintensional" aspects of belief? (I know that $x = 123714$, but I don't know the value of x^2)

where to draw the system boundary?