The Role of the Free Energy Principle in AI Safety: Markov Blankets and Beyond

Manuel Baltieri

Araya Inc.

TAIS - 6th April 2024

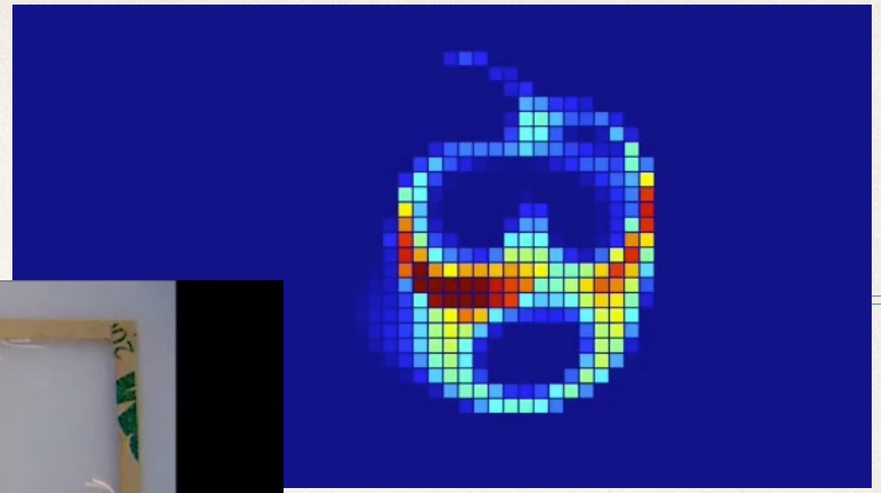
ARAYA

Outline

- ALife & AI Safety
- Boundaries: The free energy principle vs. active inference
- Beyond the FEP (w/ Martin Biehl, Matteo Capucci, Nathaniel Virgo)

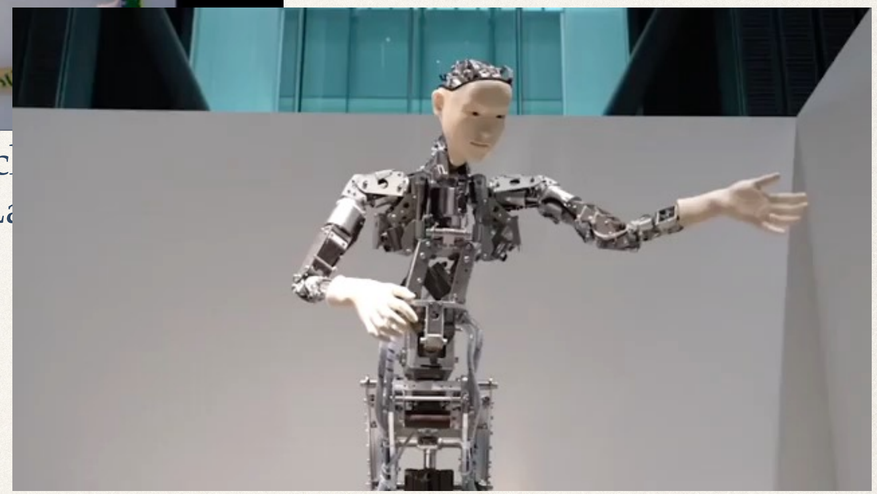
ALife in brief

- “Life as it could be”
- Wet, soft and hard ALife
- Study of (generalised forms of):
  - Adaptation
  - Autonomy
  - Self-organisation
  - Evolution
  - …



e version) + Project with code

Čejková J. et al. (2014), Dynamics of c
in salt concentration gradients, La



Yoshida, T. et al. (2023). From Text to Motion: Grounding GPT-4 in a Humanoid Robot" Alter3". *arXiv:2312.06571*.

ALife and AI Safety

Ideas where ALife and AI Safety seem to naturally meet:

- **Boundaries**
- (Minimal) Agency
- Open-ended evolution
- Meta-learning (Learning to learn), adaptivity (ability to adapt) and evolvability (ability to evolve)
- …

Photo by Shoeib Abolhassani on Unsplash

Boundaries

Theories of agency in ALife

- What an agent is
- Defining entities, some are agents
- Boundaries as part of the definition
- Example frameworks: Dynamical systems theory, information theory
- Inspiration: biology, ALife itself

Agent foundations in AI Safety / Alignment

- W
- A
- B
- B
- t
- Inspiration: economics, computer science

‹ «Boundaries» Sequence ›

«Boundaries», Part 1: a key missing concept from utility theory

158
Ω 59

by **Andrew_Critch** 9 min read 27th Jul 2022 32 comments

Boundaries / Membranes [technical] | Game Theory | Group Rationality | World Optimization | Rationality | Curated

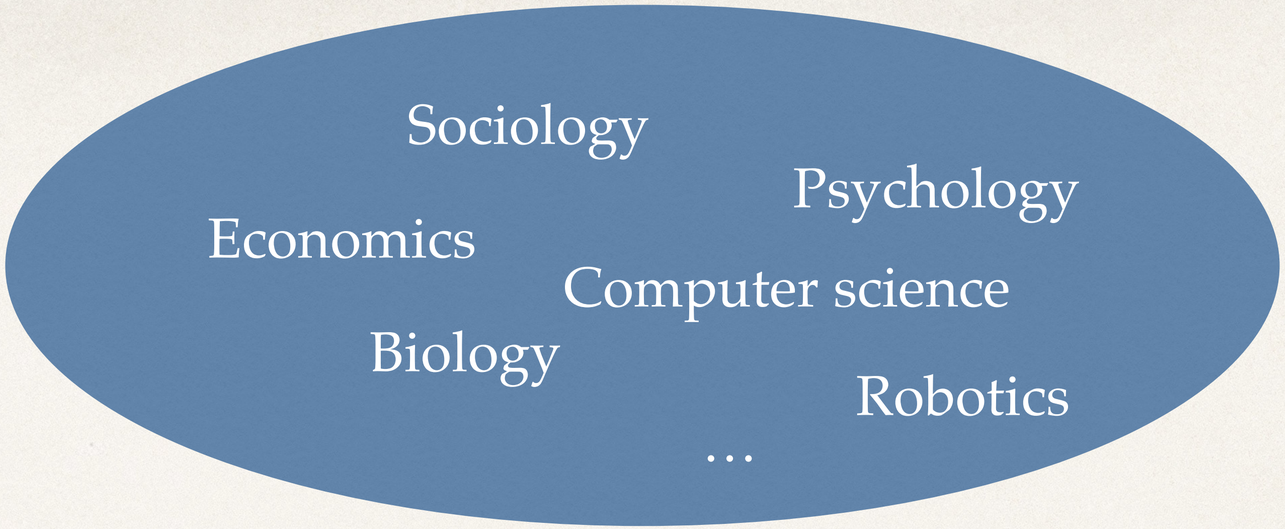
Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This post has been recorded as part of the LessWrong Curated Podcast, and can be listened to on Spotify, Apple Podcasts, and Libsyn.

This is Part 1 of my «Boundaries» Sequence° on LessWrong.

Summary: «Boundaries» are a missing concept from the axioms of game theory and bargaining theory, which might help pin down certain features of multi-agent rationality (this post), and have broader implications for effective altruism discourse and x-risk (future posts).

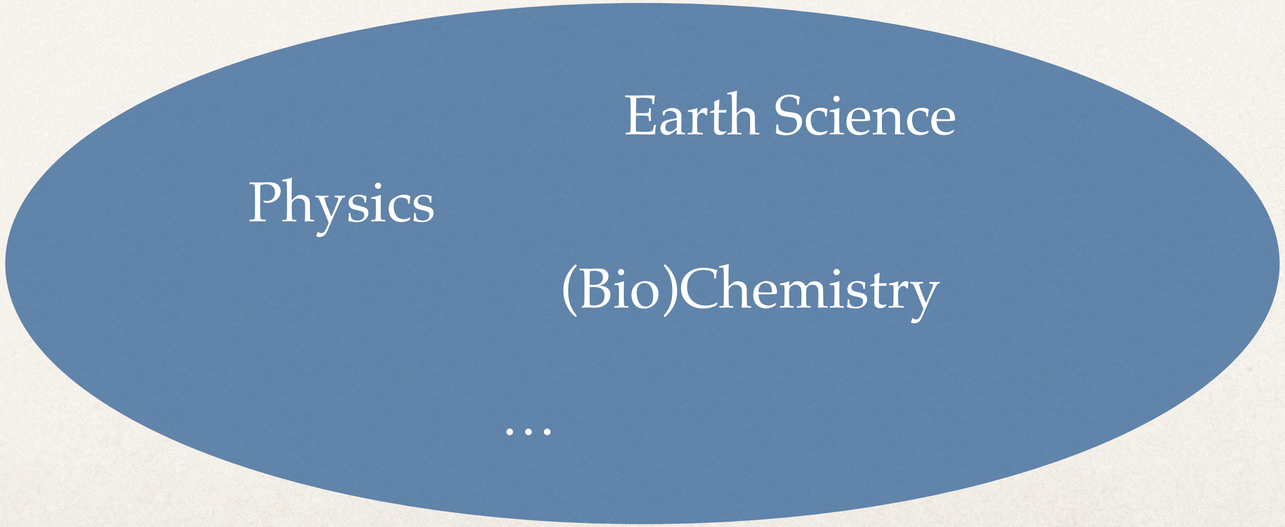
Agency obviously a thing,
no agreement
on what it is though



Minimal agency

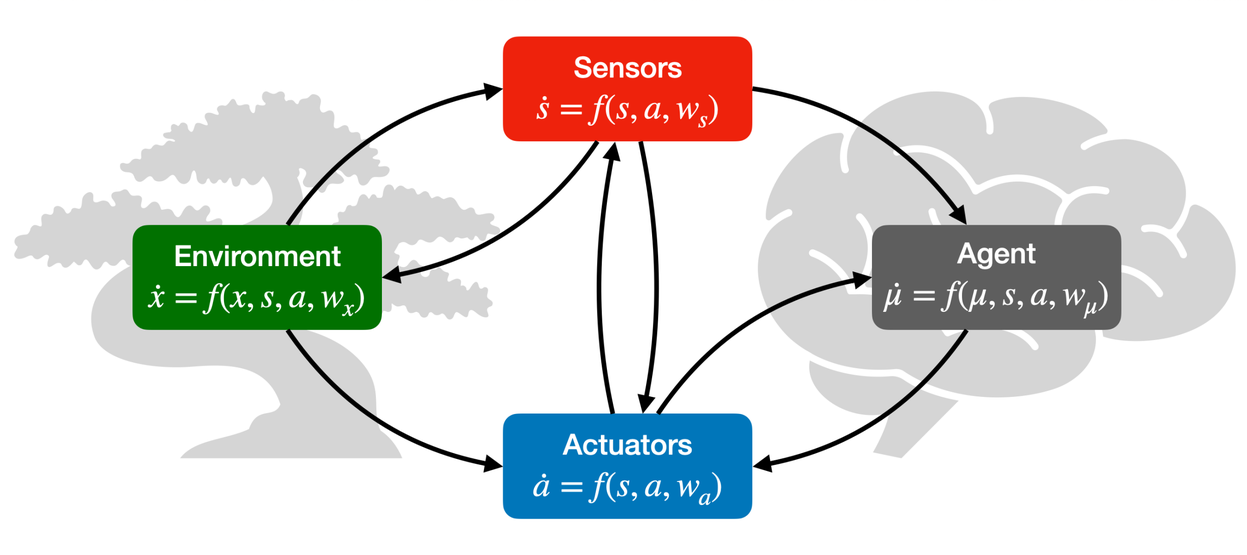
ALife

No agency / no need
for a different framework



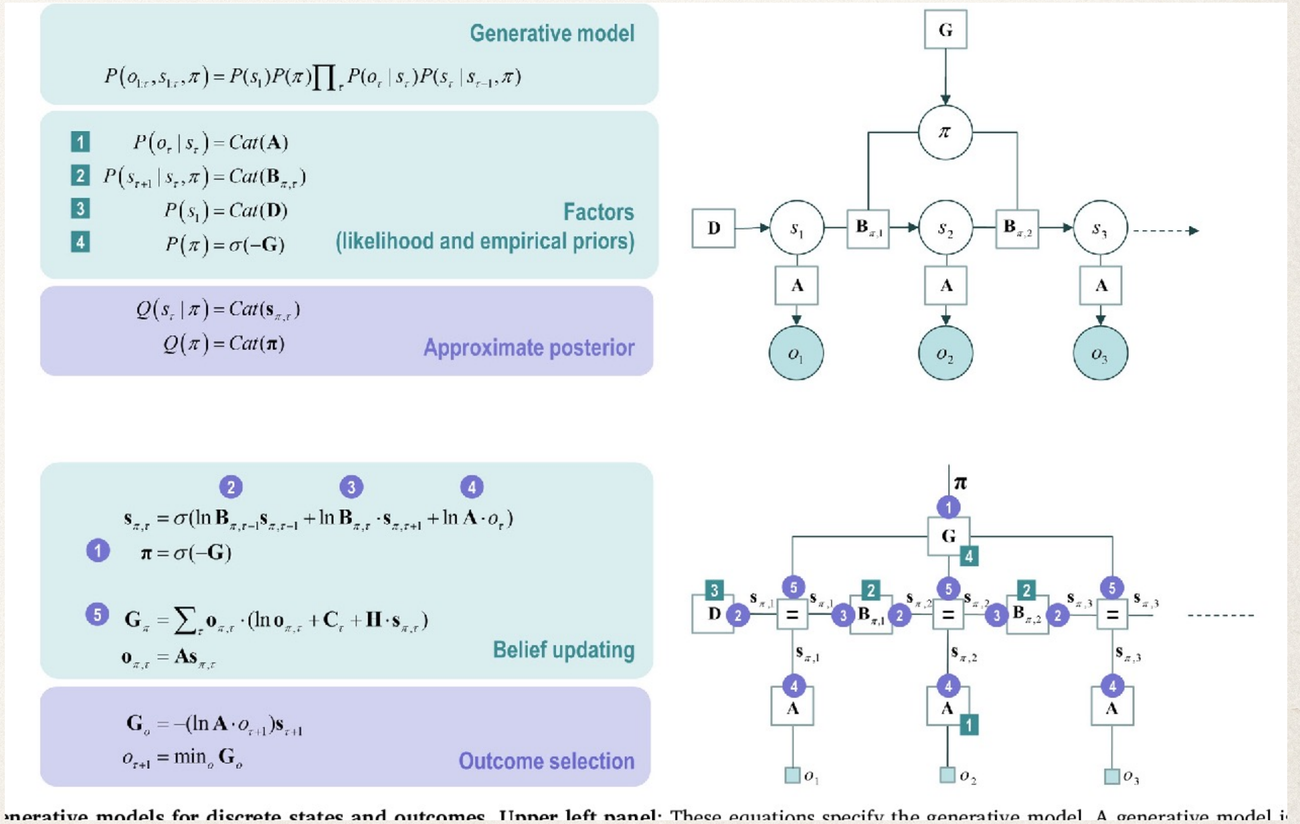
The free energy principle

- A foundational theory of agents, (living) systems, “things”
- A thing is a “thing” if and only if it (appears to) minimise(s) free energy
- Markov blankets as a “veil” that separates internal from external states



Active inference

- Assumes POMPDs/state-space models structure (~ RL setup)
- Requires specifying preferences
- Provides an alternative cost function (expected free energy)
- …ideally one that is derived from the FEP, but <u>it can stand without it</u>



nerative models for discrete states and outcomes. Upper left panel: These equations specify the generative model. A generative model i

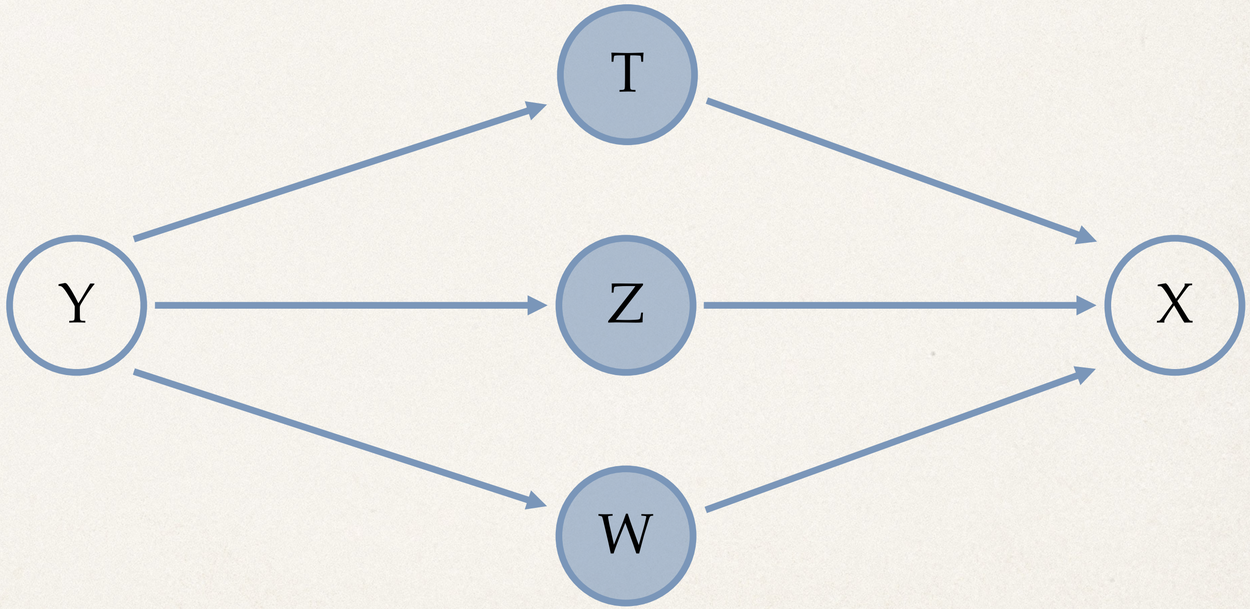
The FEP 1.01 - as of early 2021

The FEP targets:

1. systems which can be modelled as **random dynamical systems** with
2. a **unique steady-state distribution** (= weak mixing for recurrent but a-periodic Markov chains),
3. whose vector field can be **decomposed (via the Ao decomposition)**, uniquely and in a special way (= there's a number of equally valid alternatives), into orthogonal curl-free and divergence-free flows of a quasi-potential,
4. such that the set of random variables at steady-state (the stochastic process is effectively studied at steady-state) can be **partitioned into internal, external and blanket "states"** via an assumption (this is not an implication) of conditional independence between internal and external variables given the blanket (variables), based on a some **selection of** either **internal or external "states"** (the process is complementary),
5. under the additional assumption (a conjecture as seen in Friston et al. 2021, "Stochastic chaos and markov blankets") of "sparse coupling" that allows mapping of steady-state independencies to independencies on dynamical components, i.e., orthogonal curl-free and divergence-free flows,
6. and with a conditional synchronisation map assumed to connect the most likely internal and external states (see Aguilera et al. 2021 for possible issues) to try and ensure that internal variables *model* in some non-trivial sense external ones,
7. such systems can be said to contain a partition of internal states that appear to perform inference on a partition of external states via a gradient descent on variational free energy (*"Approximate Bayesian inference lemma"*).

What is a Markov blanket?

Markov blanket ~ the set of random variables (e.g., T, W, Z) that render a set of random variables (e.g. Y) conditionally independent of a set of random variables, (e.g. X)



Pearl vs Friston blankets - claims

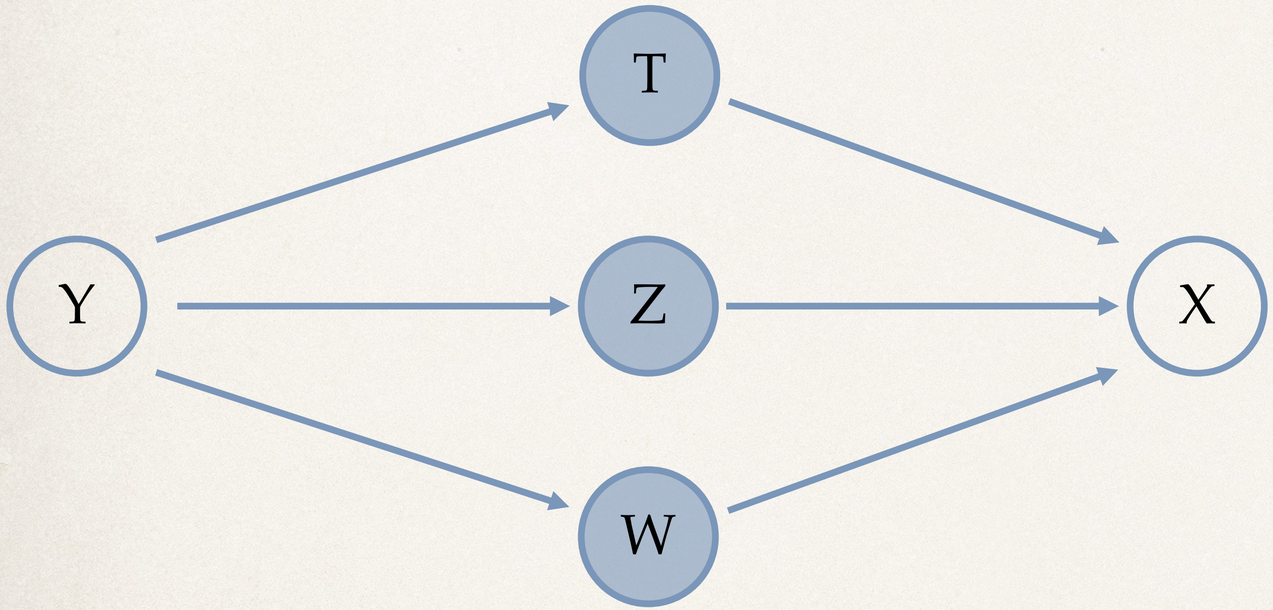
Pearl blankets (classical Markov blankets)

- Random variables (no time)
- (Usually) Epistemic
- Systems of interest are assumed
- Inference algorithms applied by a scientist after selecting a blanket for a modelled "thing"
- (Roughly, not exhaustively) active inference

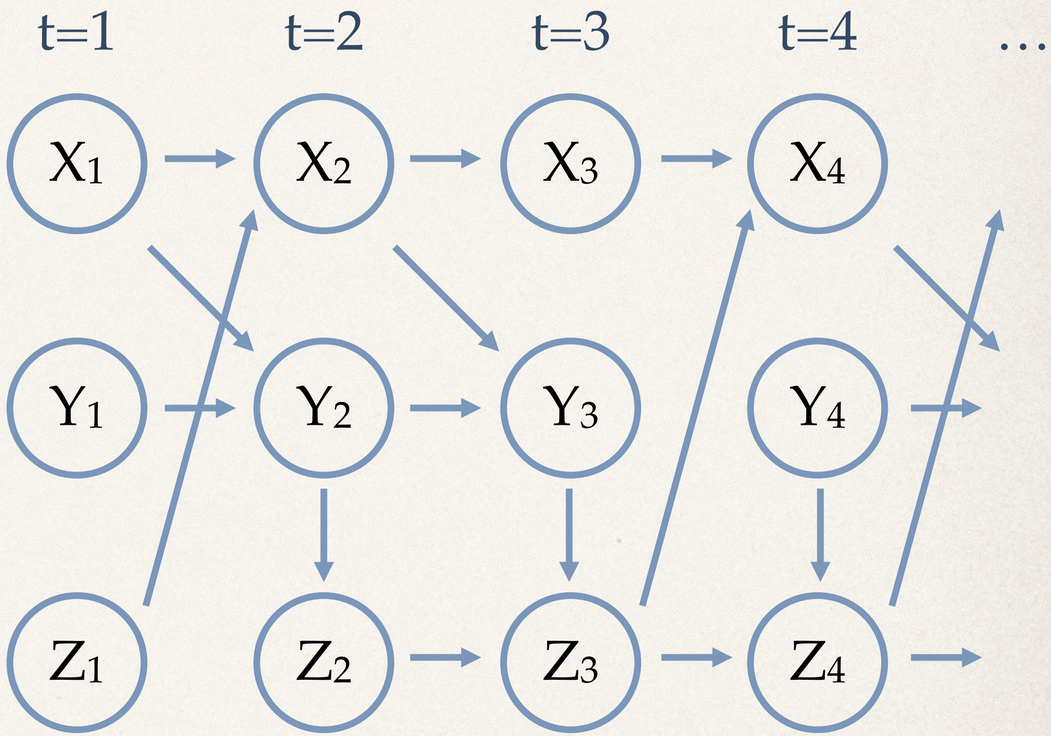
Friston blankets (Markov blankets in the FEP)

- (Stationary) Stochastic processes (time)
- (Usually) Metaphysical
- A foundational theory of "things"
- Inference emerging as the interaction between things/agents and their environments (no scientist)
- (Roughly, not exhaustively) FEP

From random variables to stochastic processes



vs.



What should be conditionally independent of what given what?
What is an "individual" over time?

See Biehl et al., 2021; Aguilera et al., 2021; Virgo et al. 2022 (or ask me at the end)

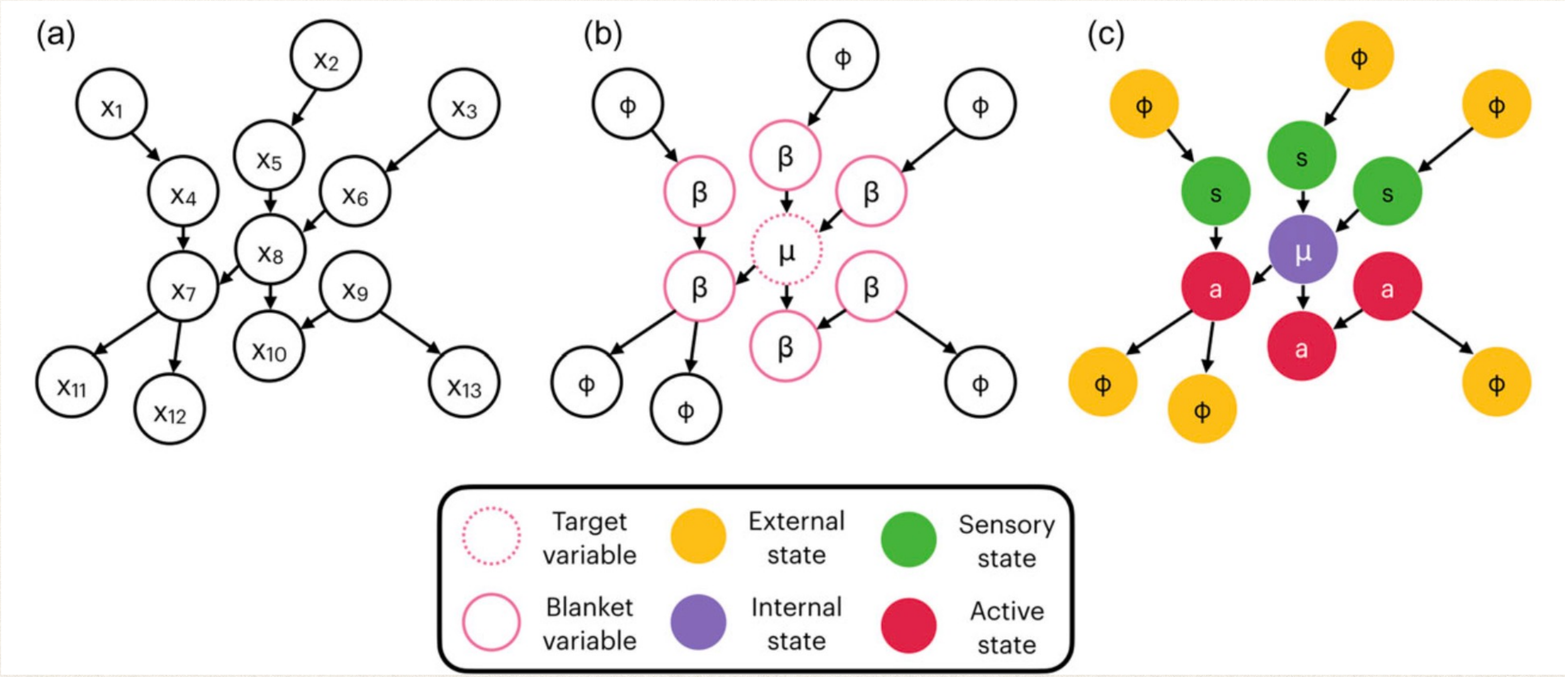
	Pearl blanket	Friston blanket	Other blankets
Markov blankets as conditional independent for random variables (no time involved)	O		
Markov blankets within a Markov chain (the present shields future from past, see Pearl et al., 1989)	O	X (after Biehl et al., 2021)	
Markov blankets within a steady-state distribution (Friston, 2013, "Life as we know it")	O	O?	
Markov blankets within a stochastic process with off-block-diagonal solenoidal couplings and extra constraints (Biehl et al., 2021)	required on steady-state distribution	X (after Biehl et al., 2021)	
Markov blankets within a stochastic process from conjectured lack of off-block-diagonal solenoidal couplings (Friston et al., after 2021)	required on steady-state distribution	O?	
Asymptotic approximation to a weak-coupling equilibrium (Friston et al., 2021, "Parcels and particles: Markov blankets in the brain)	required on steady-state distribution	O?	
Causal blanket (Rosas et al., 2020)			O
History-dependent blanket (Virgo et al., 2022)			O
Standard definitions of conditional independence for stochastic processes (see our reply for a few references)			O
«Boundaries», Part 3a: Defining boundaries as directed Markov blankets			O

The elephant in the room

Draw a Bayesian network (if it helps)

Assume a-priori a set of variables of interest (target variables)

Apply a sensorimotor interpretation



A foundational theory of agency…?

More open questions: what are active states?

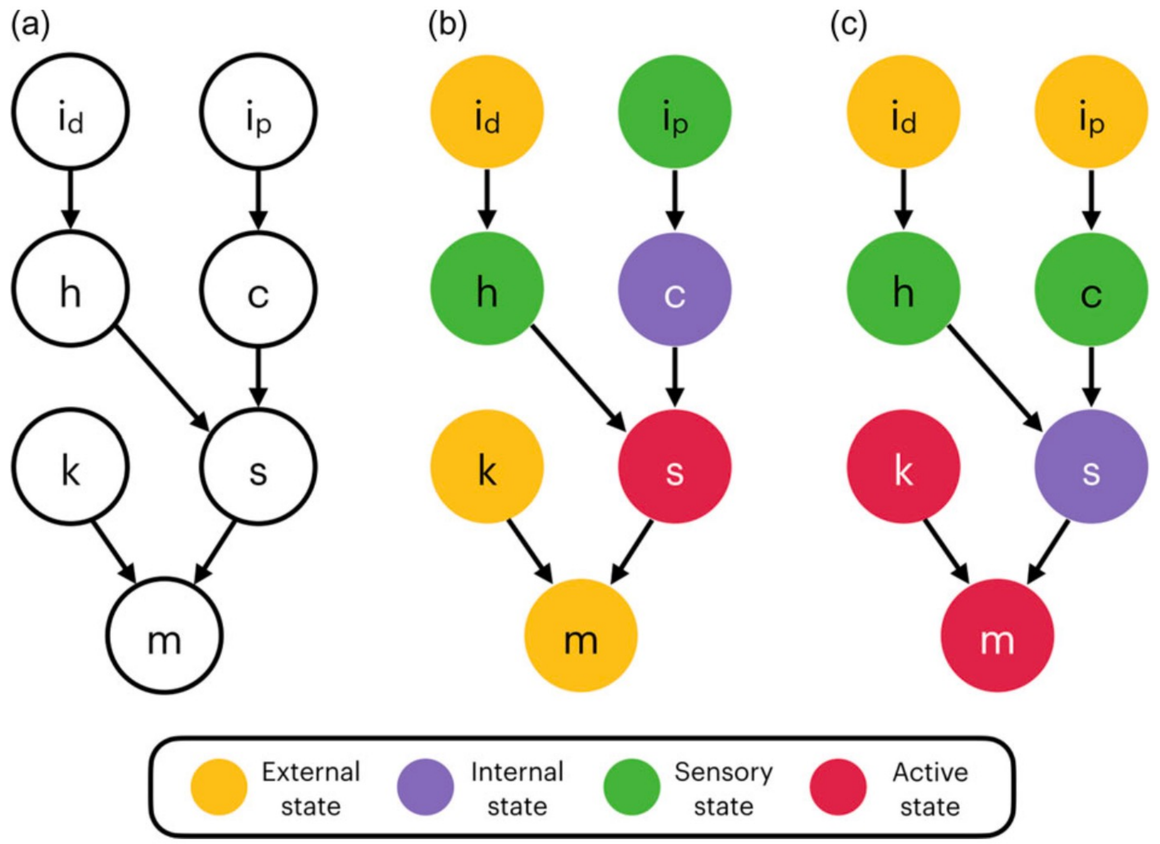
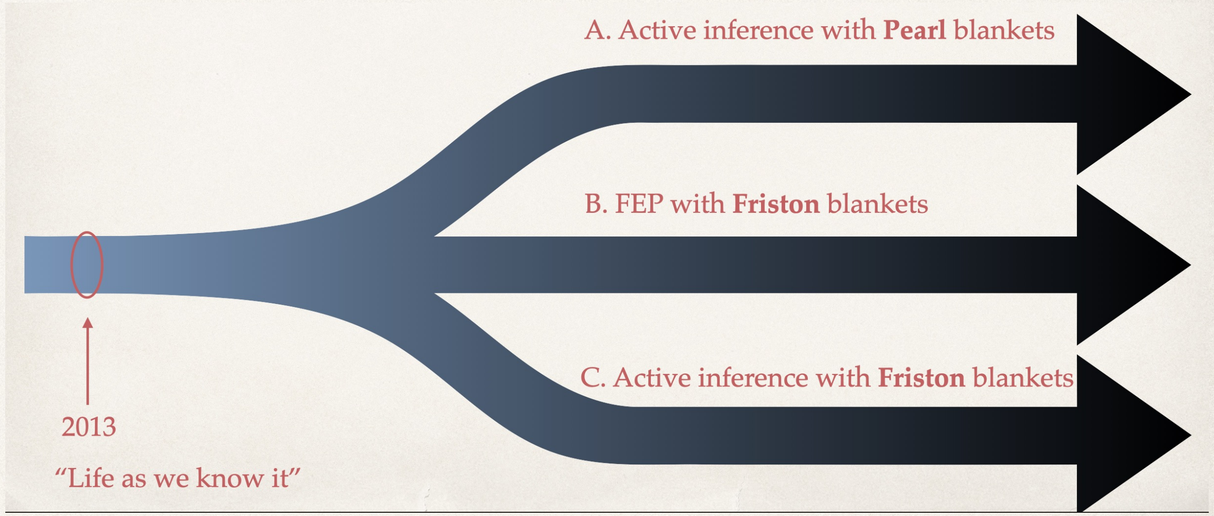


Figure 7. Conditions leading up to the knee-jerk reflex. On the left, a Bayesian network where i_d and i_p denote the motor intentions of the doctor and the patient respectively. Node *s* denotes the spinal neurons that are directly responsible for causing the kicking movement *m*. Node *h* indicates a medical intervention with a hammer, while *c* stands for a motor command sent to *s* from the central nervous system. Finally, node *k* stands for a third way of moving the patient's leg, for example, by someone else kicking it to move it mechanically. The middle (b) and the right figures (c) with the coloured-in nodes show two different ways of partitioning the same network using a "naive" Friston blanket with different choices of internal states, *c* and *s* respectively.

Active inference without FEP in AI Safety?

Agency in ALife : Agent Foundations :: FEP : Active inference

- Can do Agent Foundations of course, but what about boundaries, def. of agent, etc.
- Active inference as an alternative to standard utility theory formulations
- … it just has to be tested



The big(ger) picture (beyond boundaries)

What an agent is

- Minimal agency (ALife)
- Examples:

  FEP (maybe not)

  next talk (Martin Biehl)

  work w/ Martin, Nathaniel, Matteo

  boundaries

  …

What an agent does

- Agent foundations (AI Safety)
- Examples:

  active inference

  reinforcement learning

  decision theory

  game theory

  …

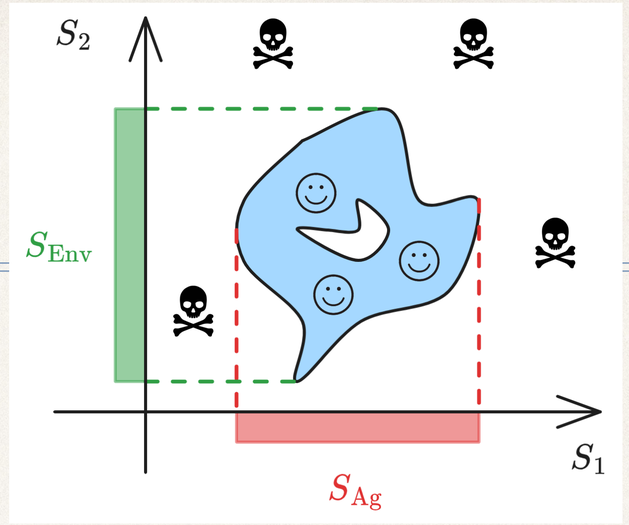
Adaptive fit as a bisimulation

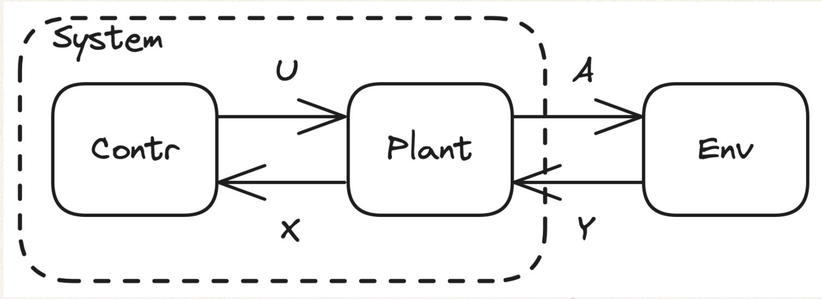
Goal: provide a mathematical formulation to describe "good couplings" between an agent and its environment over time

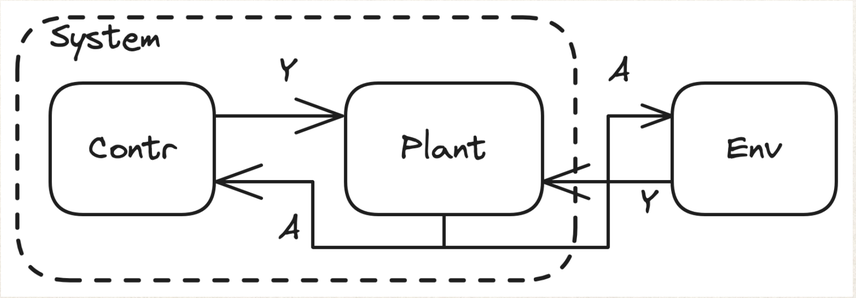
Use a version of the internal model principle from control theory

Define homeostatic regulation for $X = A, U = Y$

Contr/Env now have the same input-output type







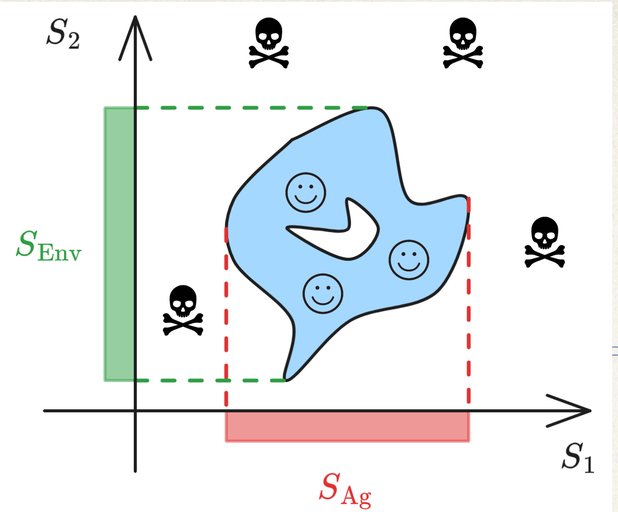
w/ Martin Biehl, Matteo Capucci, Nathaniel Virgo

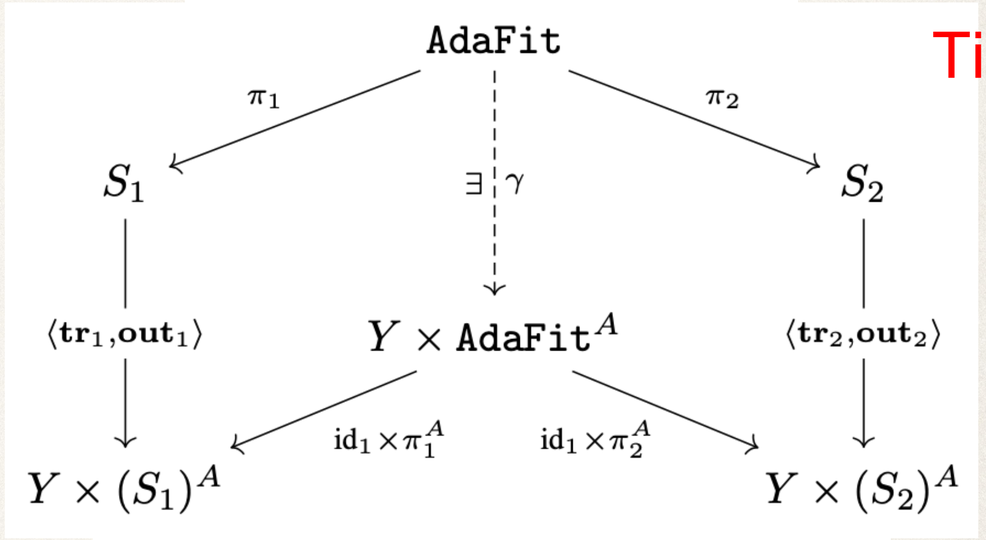
Adaptive fit as a bisimulation (2)

Describe a relation on state-spaces (top figure, blue area), $\_ \subseteq S_1 \times S_2$

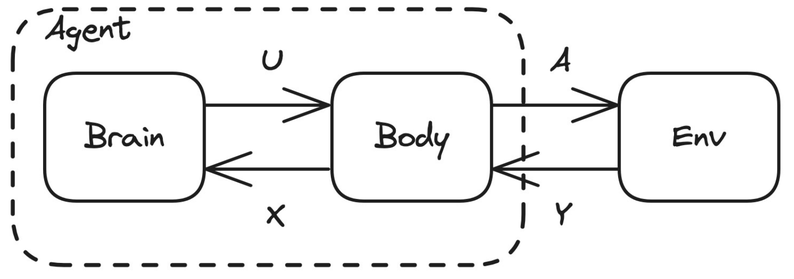
Make it structure preserving (consistent over time), and obtain the $AdaFit$ bisimulation

$AdaFit$ (roughly) selects pairs of coupled dynamical systems where an "agent" can be said to be in homeostatic equilibrium with its "environment"





Time



w/ Martin Biehl, Matteo Capucci, Nathaniel Virgo

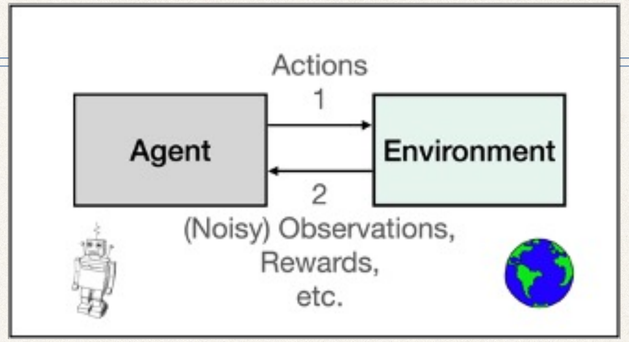
A Bayesian interpretation of the Internal Model Principle

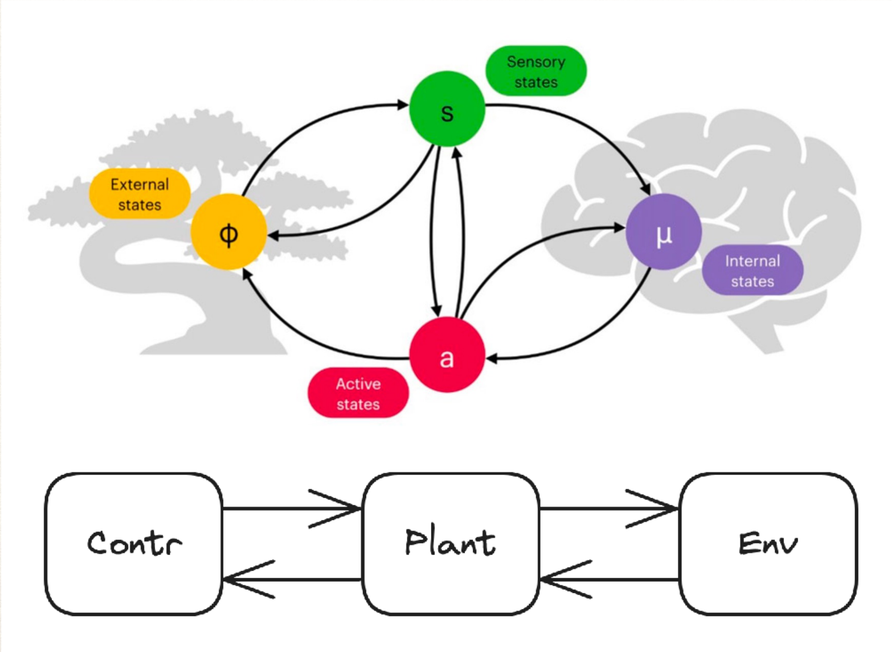
Goal: Understand different notions of systems modelling other systems (in the future, without assuming systems of interest (??))

Agents are thought to model their environment to act in a purposeful way and achieve their goals

Bayesian frameworks (used in cognitive science, ~ FEP) tell us that "models ~ Bayesian inference on hidden properties of the environment"

Internal model principle (used in biology) tell us that "models ~ consistency with env. dynamics"





w/ Martin Biehl, Matteo Capucci, Nathaniel Virgo

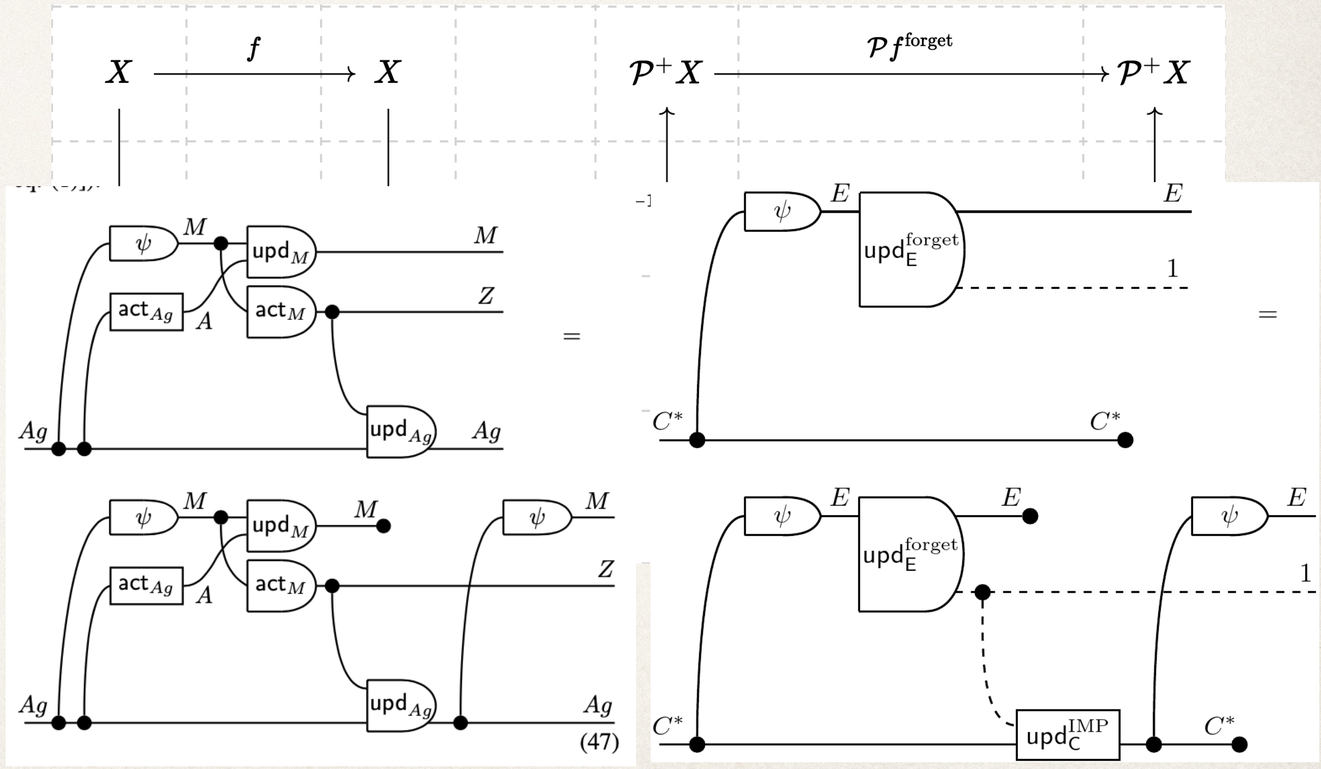
A Bayesian interpretation of the Internal Model Principle (2)

Apply category theoretic notions of Bayesian inversion to Markov categories (here, Rel^+)

+ assumptions → we obtain a version of the IMP from control theory as a forgetful coarse-grained Bayesian model

IMP ~ "systems modelling other systems to control them"

One of the central results from control theory can be vastly generalised



w/ Martin Biehl, Matteo Capucci, Nathaniel Virgo

Summary

Promising
Skeptical
To do

- More ALife and AI safety cross-pollination! —> e.g. boundaries, next open-ended evolution, etc
- FEP for AI safety —> unclear status
- Active inference and AI safety —> maybe, show it working
- Categorical approaches generalising cybernetics/control theory —> general (ALife~like) agency
    - Preference plasticity & corrigibility
    - Mesa-optimisers
    - AI boxing / containment
    - …