The free energy principle and the internal model principle

A guide for the study of agents?

Manuel Baltieri

04/10/2023

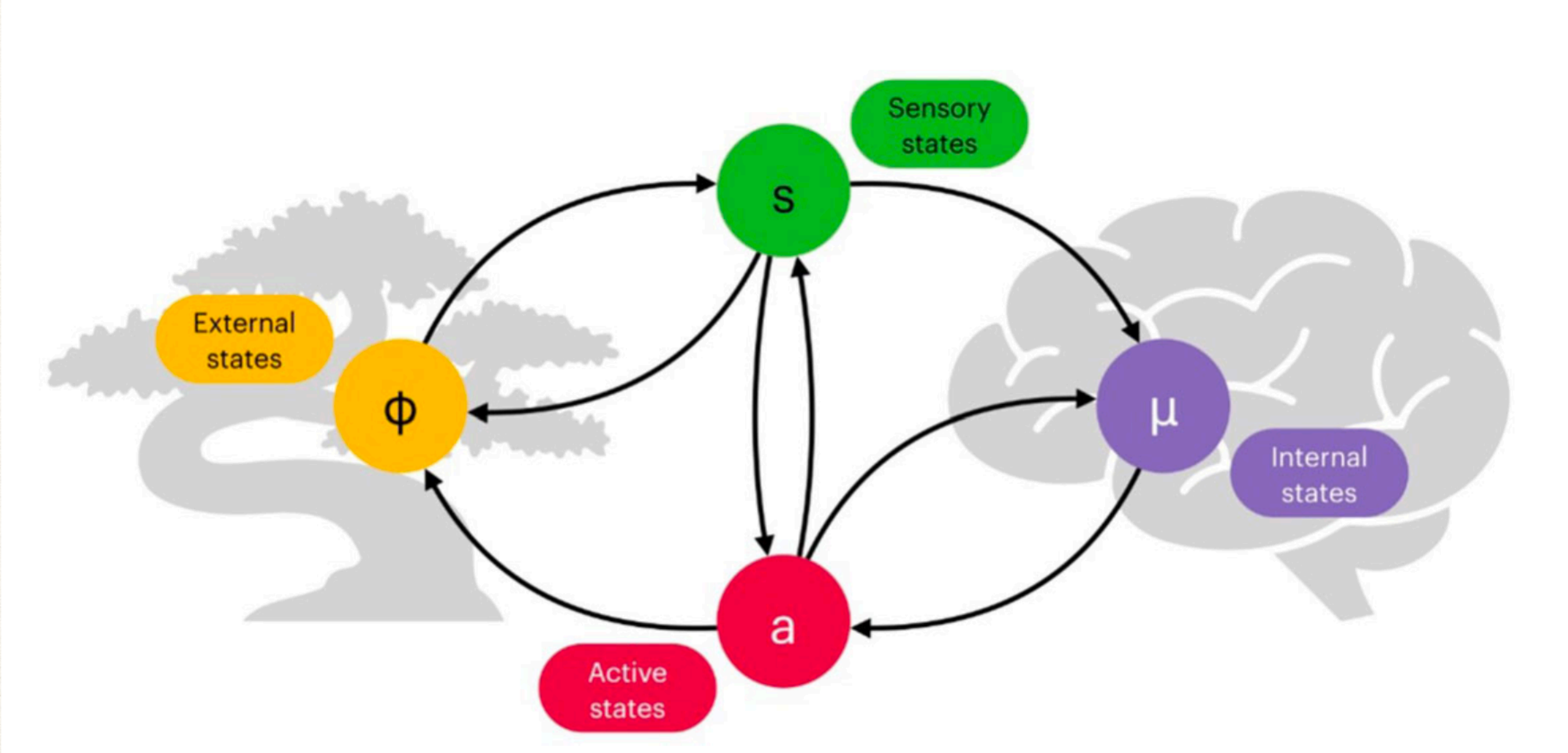
ARAYA

Outline

- The free energy principle vs. active inference
- Agency and alignment
- The internal model principle
- Viability theory

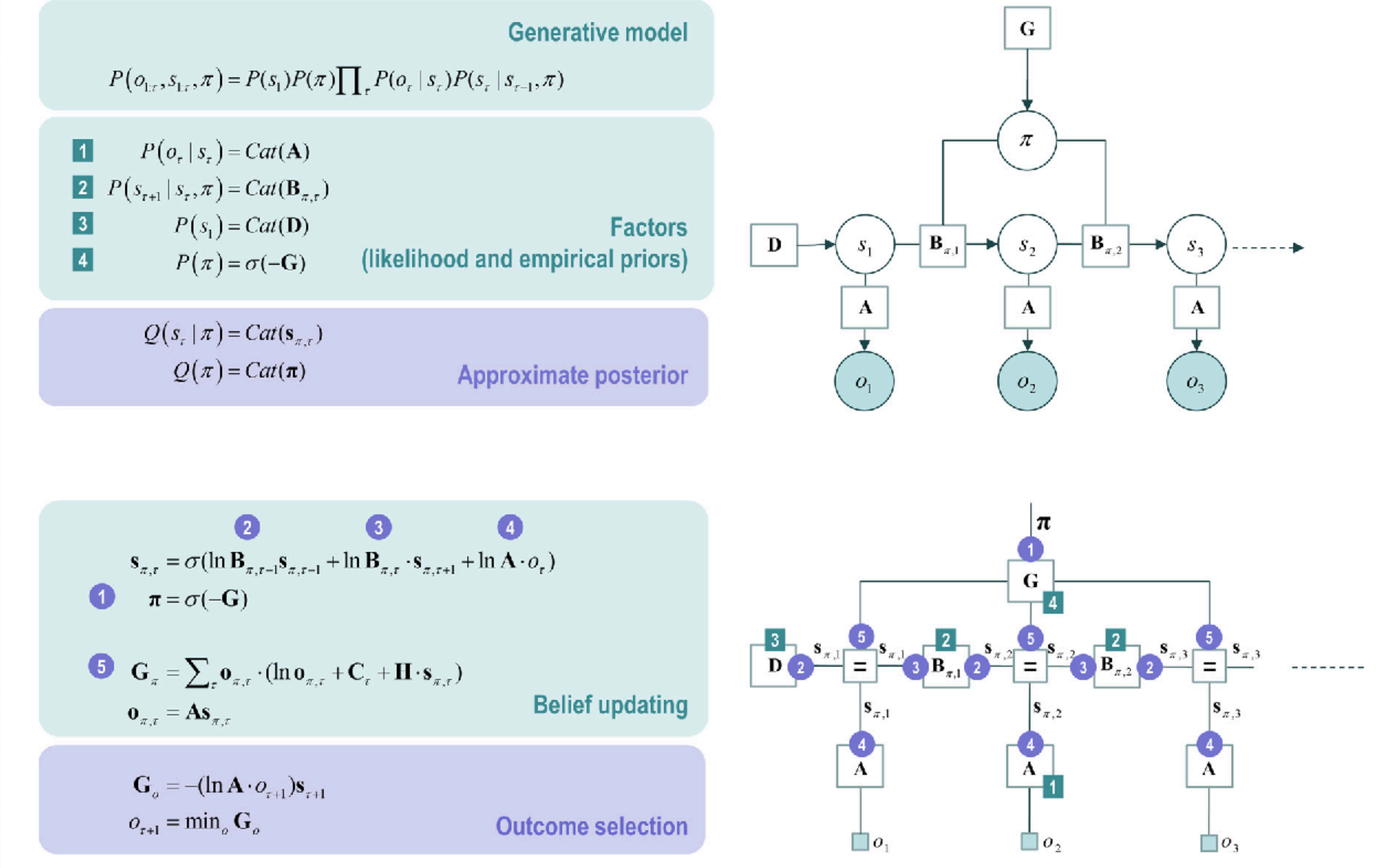
The free energy principle

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



Active inference

- Assumes POMPDs/state-space models structure (~ RL setup)
- 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:

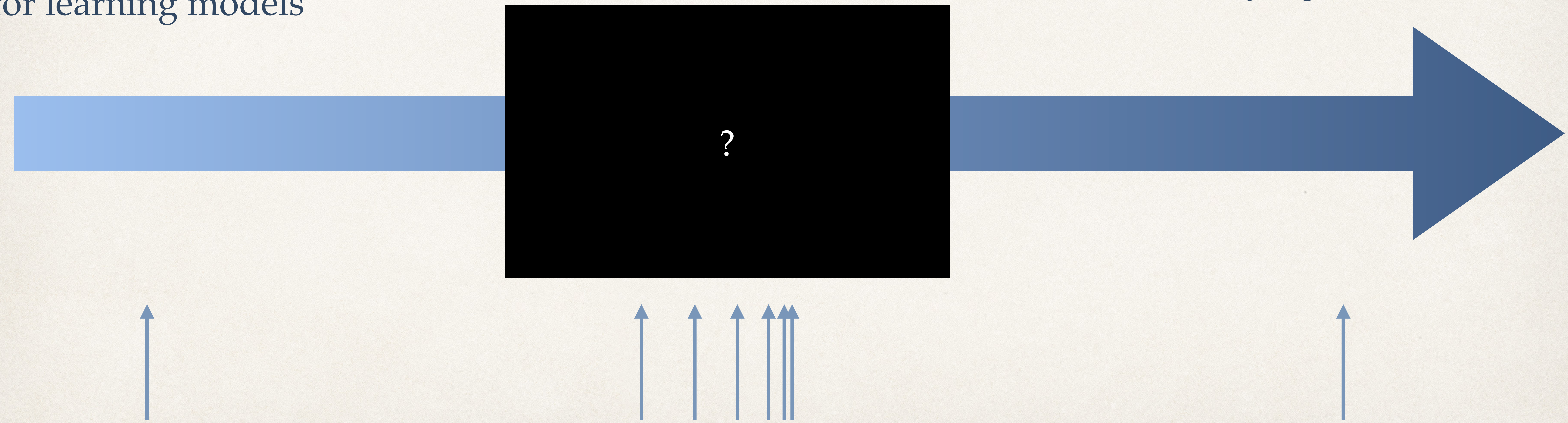
1. systems which can be modelled as **random dynamical systems** with
2. a **unique steady-state distribution** (= weak mixing for recurrent but a-periodic Markov chains),
3. whose vector field can be **decomposed (via the Helmholtz-Hodge(+ Ao?) decomposition)**, uniquely and in a special way (= there's 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"*).

AI Alignment

(Intersections, bifurcations, dead ends for)
Black-box models, agency, human feedback,
reward hacking, goal emergence, …

Biased data,
algorithms, etc.
for learning models

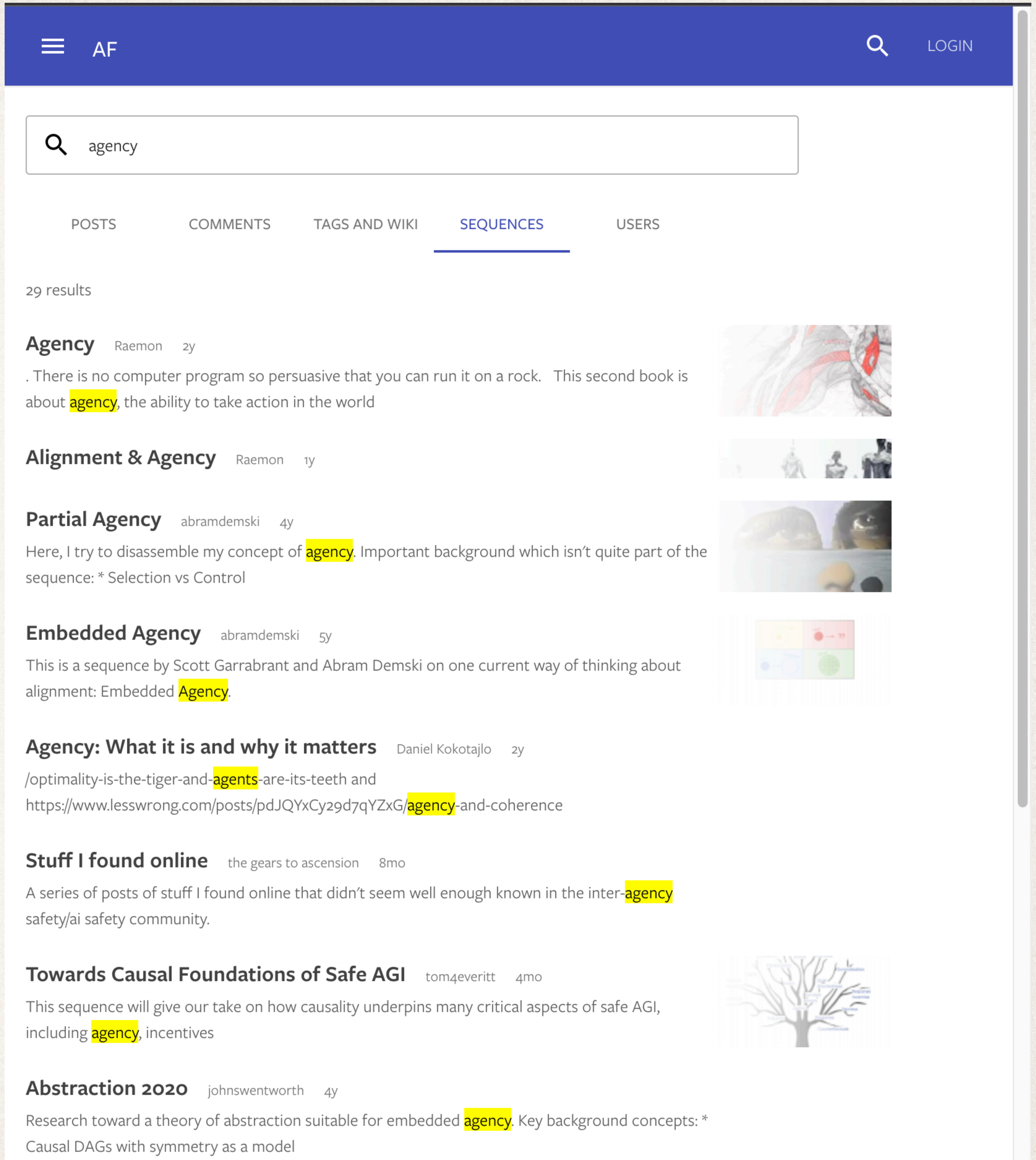
Super-human AIs
trying to kill us

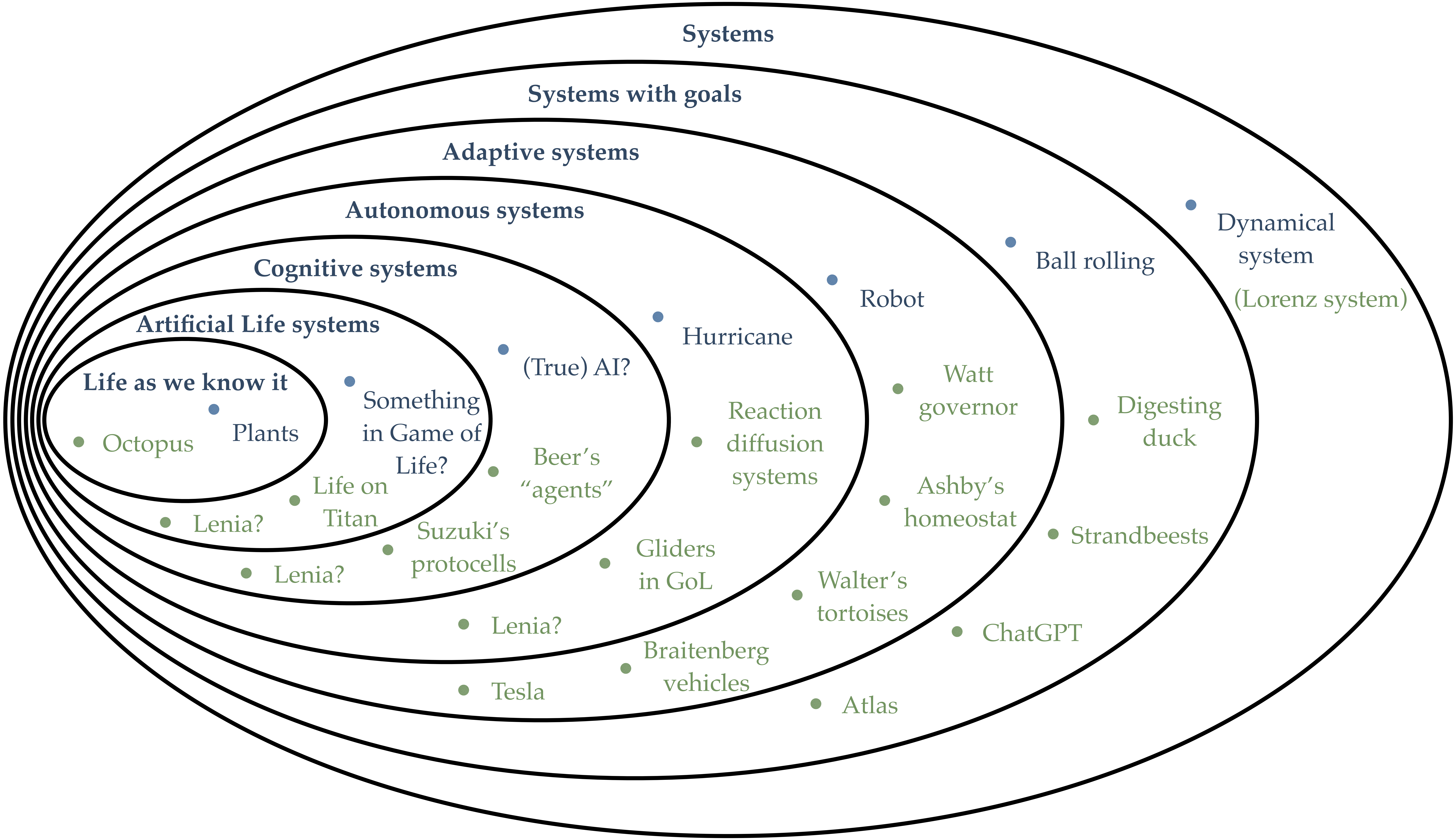


Alignment and agency

Agents: goal-directed autonomous systems that interact with, but are fundamentally distinct from, their environments

- Keywords for alignment research: goals and autonomy
  - Systems with misaligned goals are often not great
  - Autonomous systems with misaligned goals ~~are~~ can be scary
- My interest here: agency (not necessarily having to do with human-centric notions of agency)





Systems
Systems with goals
Adaptive systems
Autonomous systems
Cognitive systems
Artificial Life systems
Life as we know it
Plants
Octopus
Something in Game of Life?
Life on Titan
Lenia?
(True) AI?
Beer's "agents"
Suzuki's protocells
Lenia?
Hurricane
Reaction diffusion systems
Gliders in GoL
Lenia?
Robot
Watt governor
Ashby's homeostat
Walter's tortoises
Braitenberg vehicles
Tesla
Ball rolling
Digesting duck
Strandbeests
ChatGPT
Atlas
Dynamical system
(Lorenz system)

Alignment and FEP/active inference

Active-inference-style

- Assume agency
- Example problems:
  - Can goals differ from pre-assigned ones? Probably, see e.g., https://arxiv.org/abs/1710.11029 (funnily enough, related to FEP)
  - Alignment of inference/learning algorithms (see paper above)
  - Interactions with other agents (humans or other kinds)
  - …

FEP-style

- Define agency
- Example problems:
  - Agent/non-agent distinction? (In the AI Alignment community)
  - Theories of agency
  - Can non-agents *become* agents *over time*?
  - Can non-agentic *parts compose* to become agents?
  - How do agents develop their own goals?
  - …
  - + everything on the left

FEP-style alignment

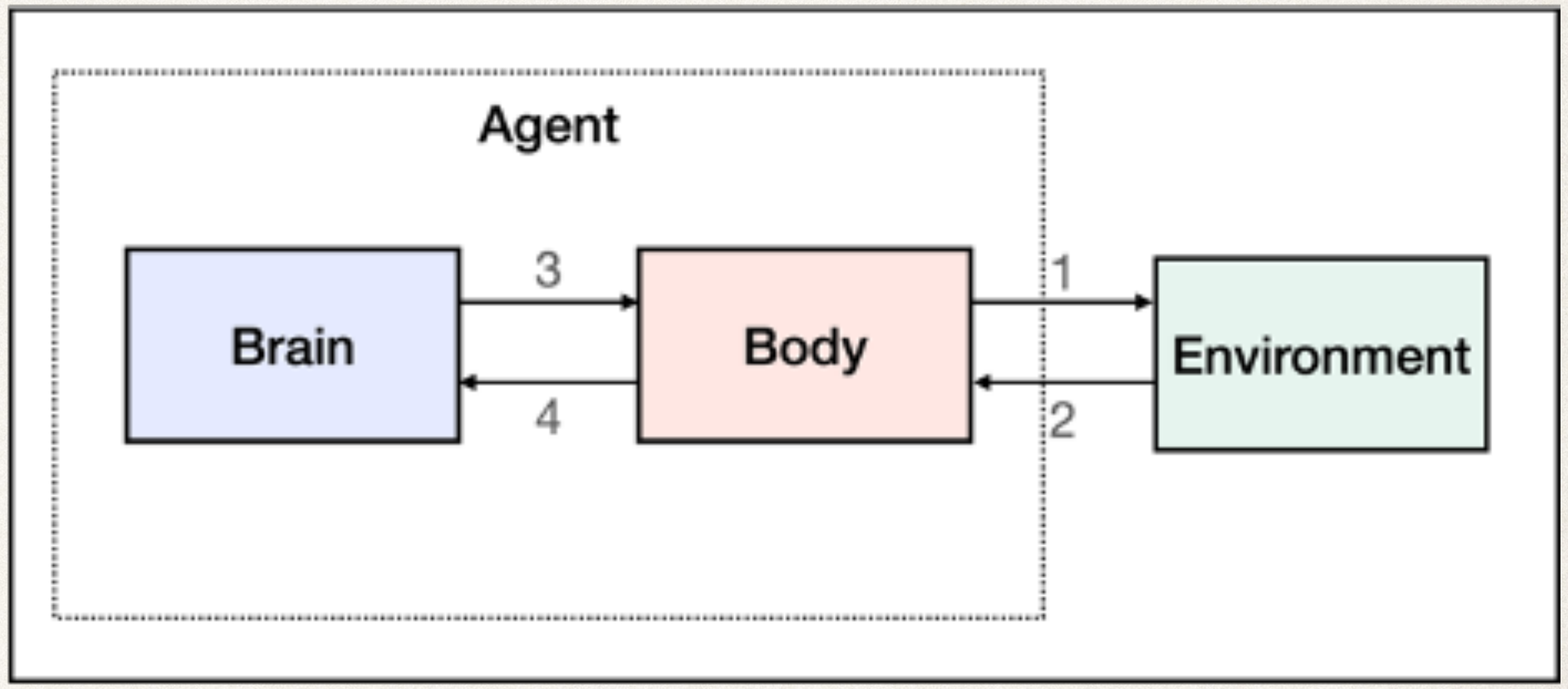
- Tl;dr: agents perform inference (~ model?) their environment
- Inspirations:
  - Cybernetics (good regulator “theorem”, law of requisite variety)
  - Control theory (internal model principle)
  - …

Internal model principle

- Like GR"T", but for dynamical systems, and actually does what it says (under several assumptions)

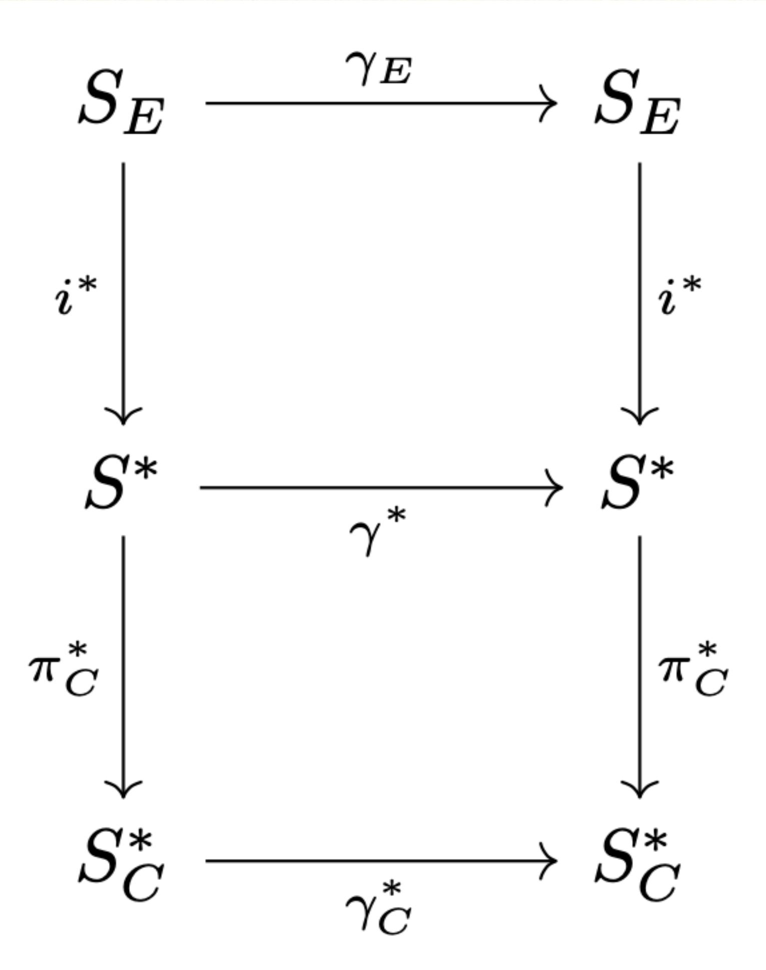
- ~~ to control a system (plant/body + environment) a controller/brain contains a model of (parts of) the environment when at equilibrium/the goal/control is achieved

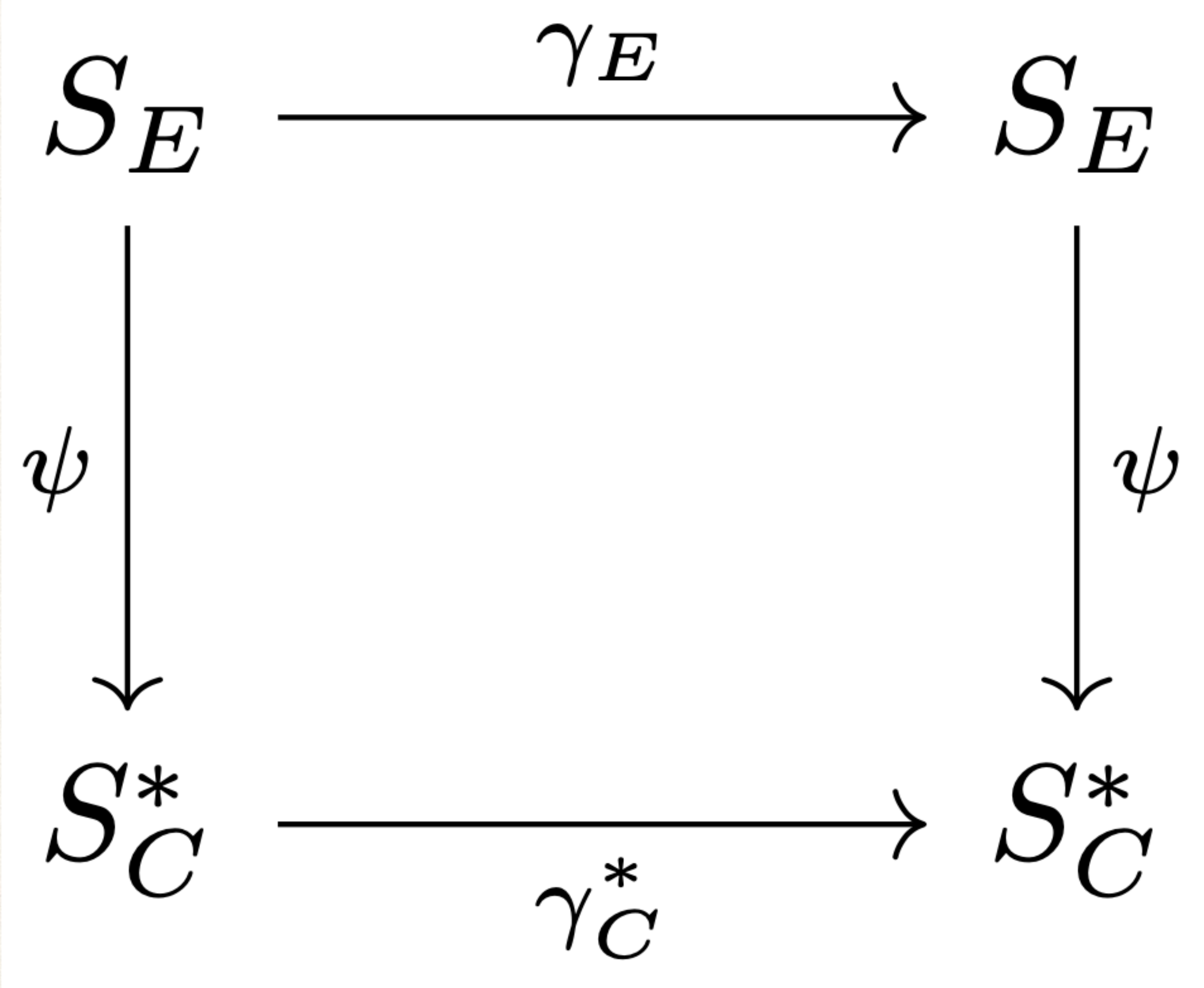
- Alignment
  - AI systems that achieve goals do so by modelling their environment (don't take it for granted!)
  - Systems scientists/control engineers regularly deal with control of black boxes (alignment vs control?)
  - Behavioural approaches to control (~ look at control in terms of relations between systems/how the behave)



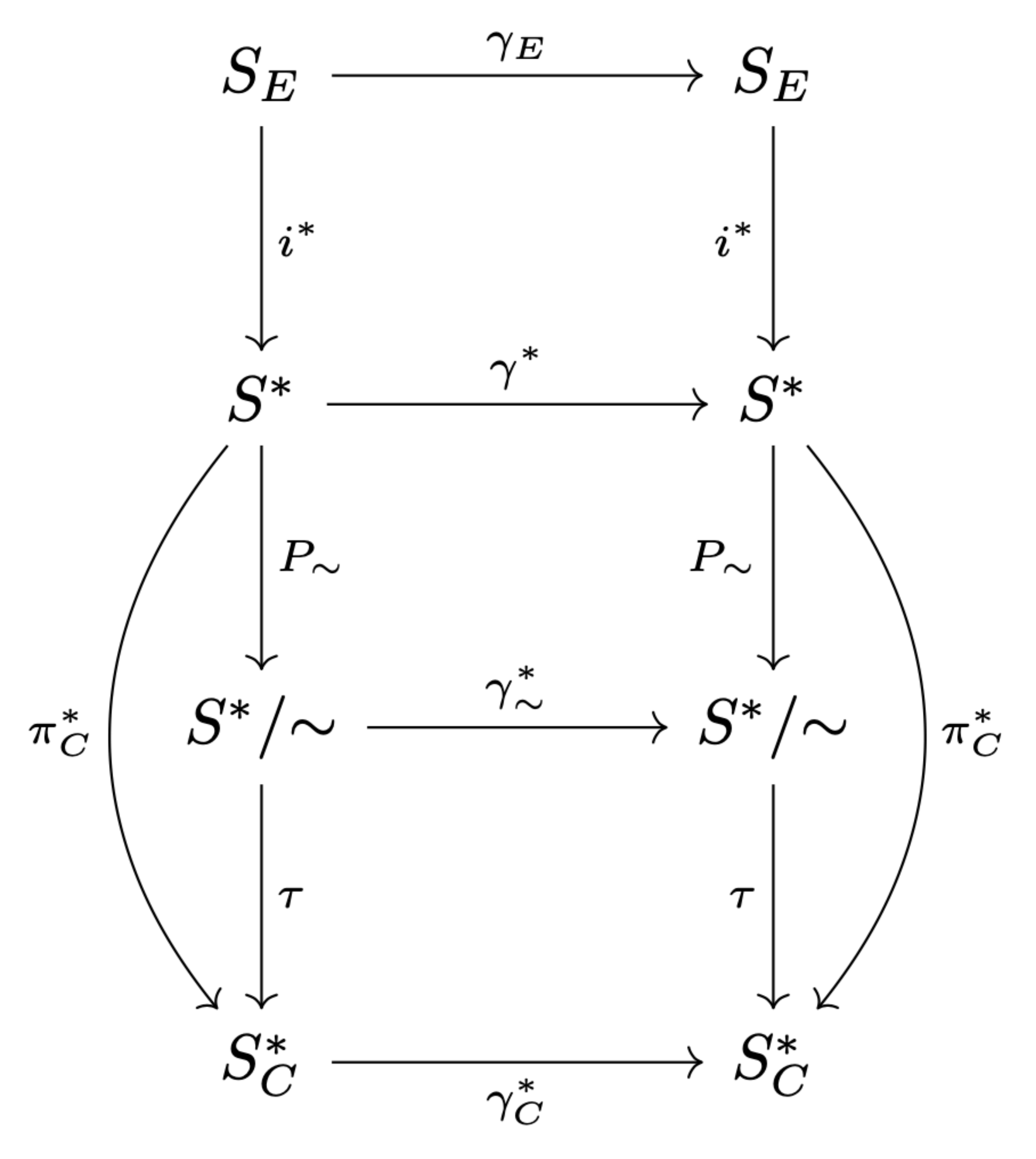
Internal model principle as a “mini” FEP

Fully observable environment





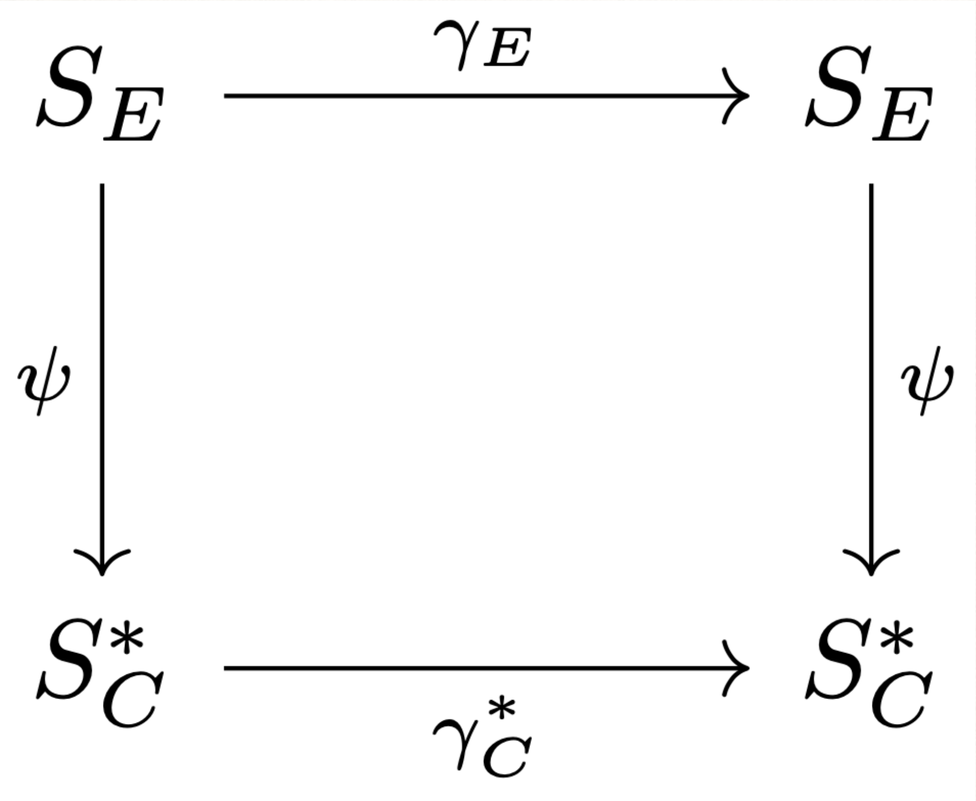
Partially observable environment

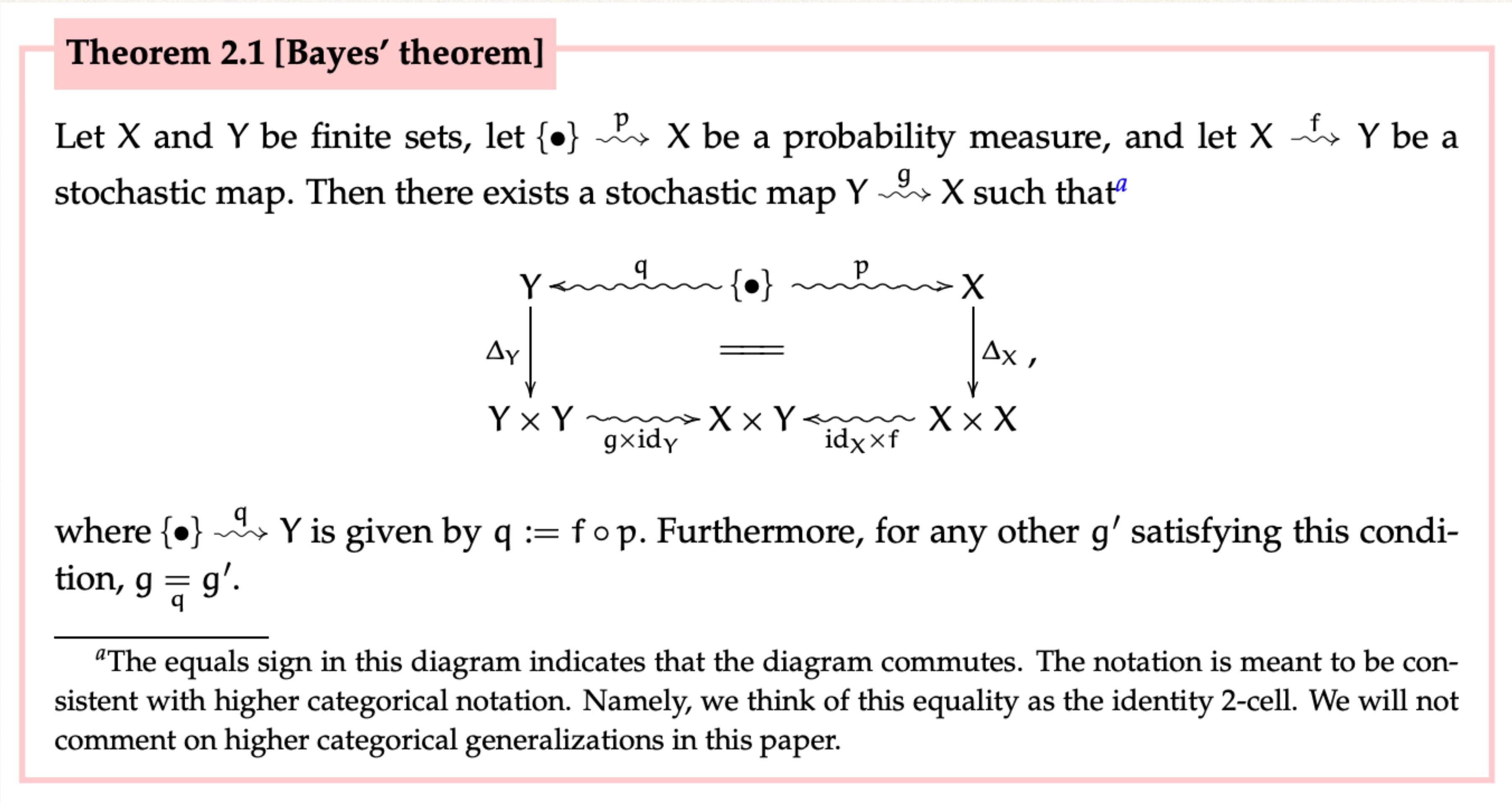


WIP: From the IMP to Bayesian inference

Bayes theorem as a consistency equation…

… with dynamics





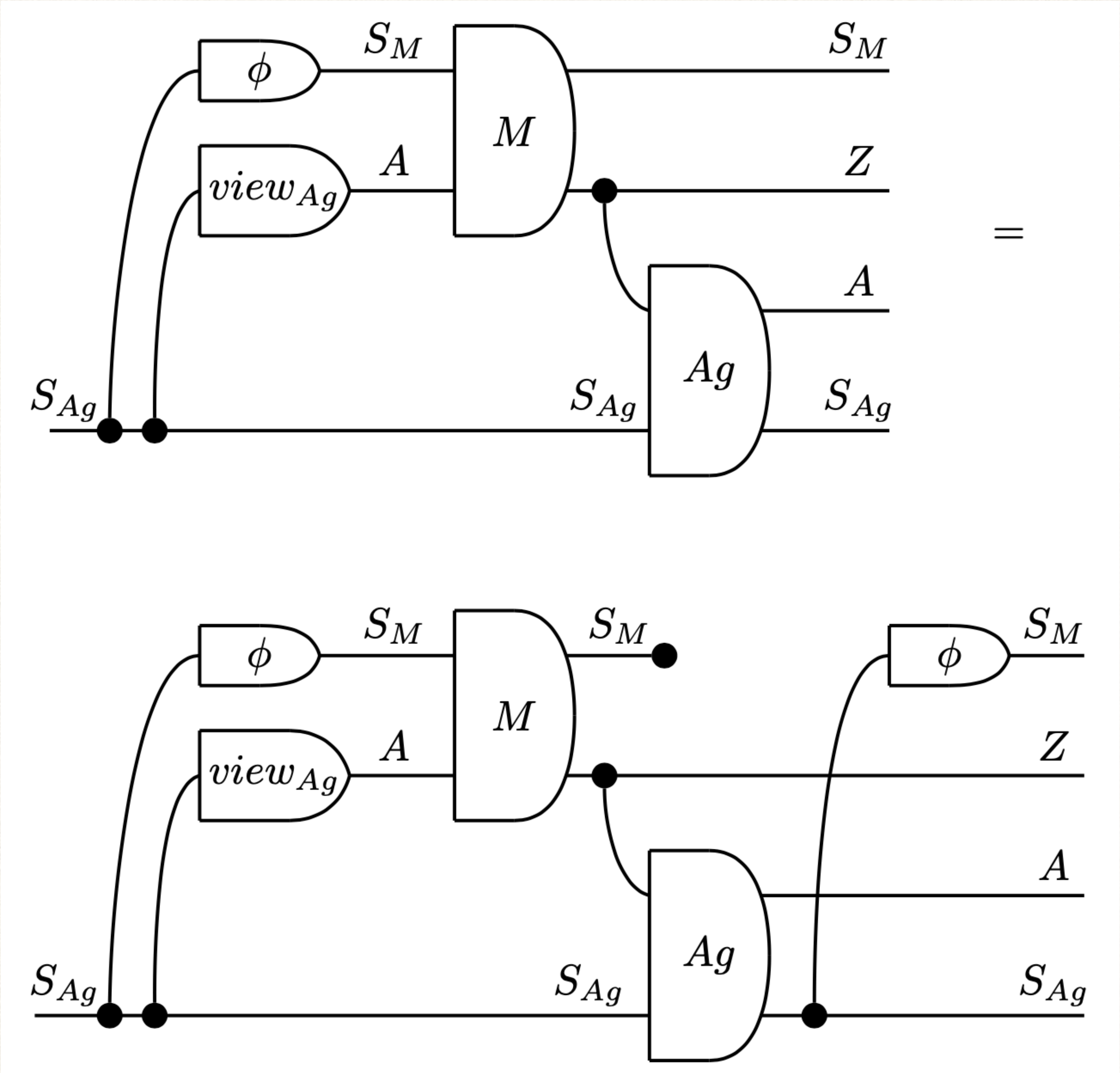
Theorem 2.1 [Bayes' theorem]

Let X and Y be finite sets, let $\{\bullet\} \overset{p}{\rightsquigarrow} X$ be a probability measure, and let $X \overset{f}{\rightsquigarrow} Y$ be a stochastic map. Then there exists a stochastic map $Y \overset{g}{\rightsquigarrow} X$ such that[a]

$$\begin{array}{ccccc} Y & \overset{q}{\leftsquigarrow} & \{\bullet\} & \overset{p}{\rightsquigarrow} & X \\ \downarrow \Delta_Y & & = & & \downarrow \Delta_X \\ Y \times Y & \underset{g \times \mathrm{id}_Y}{\rightsquigarrow} & X \times Y & \underset{\mathrm{id}_X \times f}{\leftsquigarrow} & X \times X \end{array},$$

where $\{\bullet\} \overset{q}{\rightsquigarrow} Y$ is given by $q := f \circ p$. Furthermore, for any other g' satisfying this condition, $g \underset{q}{=} g'$.

[a] The equals sign in this diagram indicates that the diagram commutes. The notation is meant to be consistent with higher categorical notation. Namely, we think of this equality as the identity 2-cell. We will not comment on higher categorical generalizations in this paper.



Advantages:

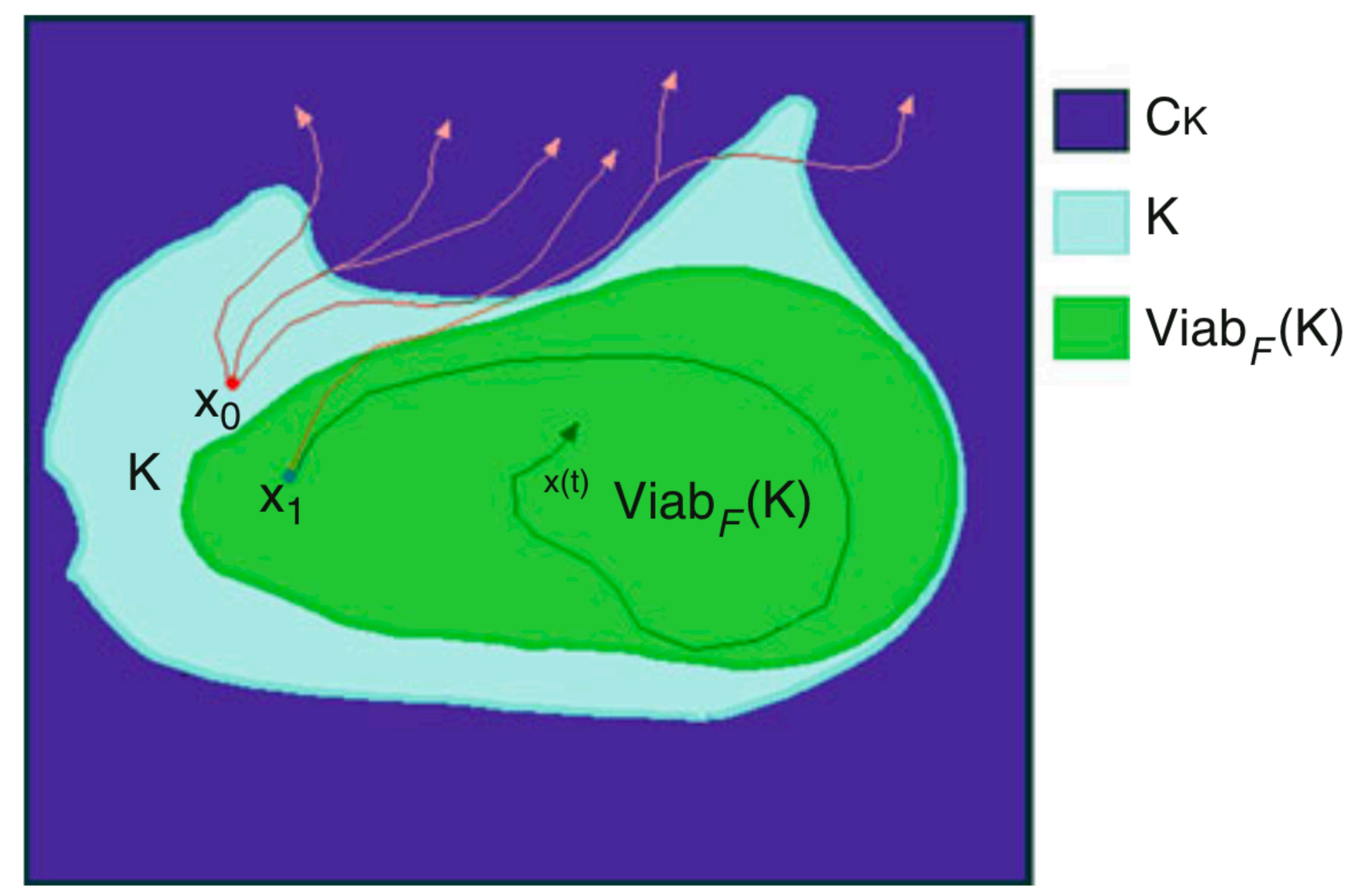
- discrete time
- no measure theory (possibilistic setup, see next slide)
- nice (I think) graphical language
- straightforward to abstract (=/= generalise)
- recovering FEP (not actinf) from abstraction of this idea

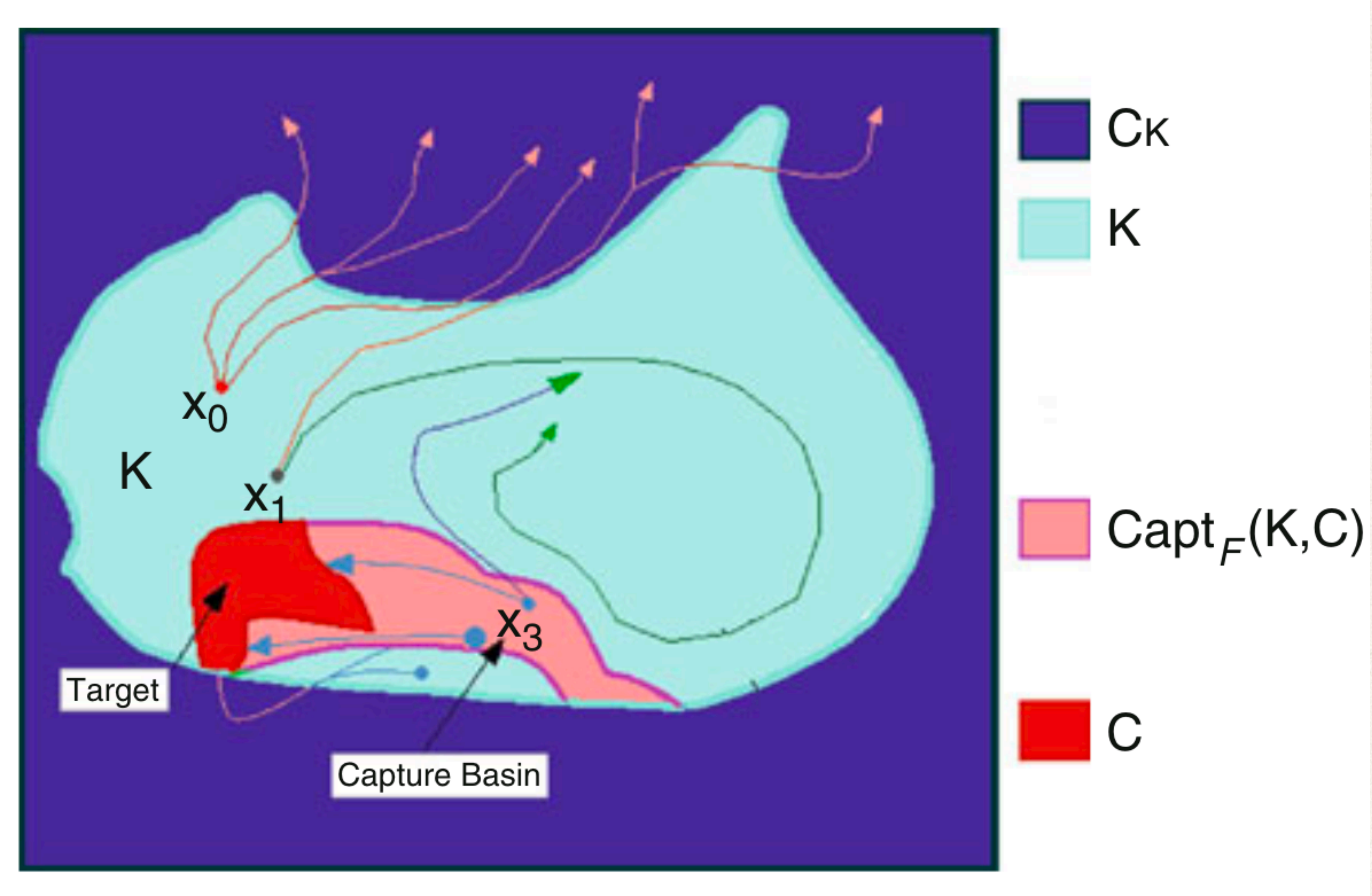
Viability theory

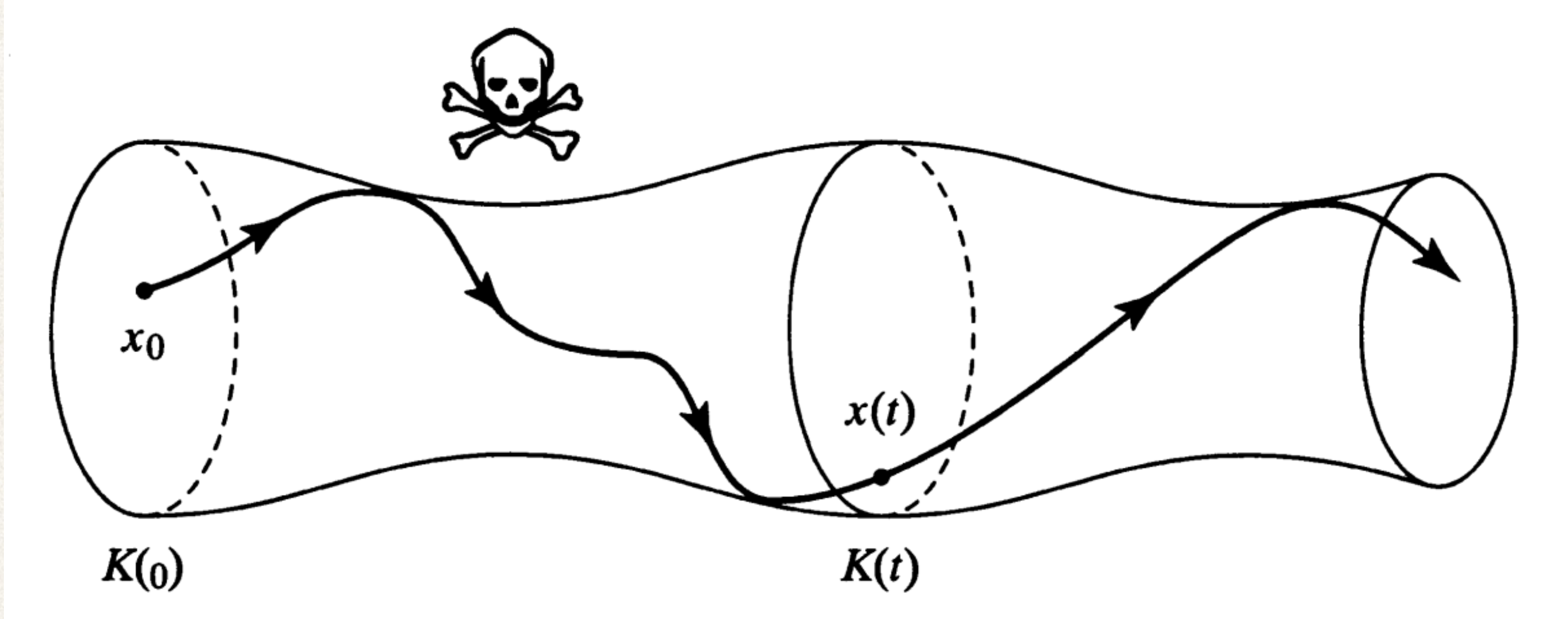
- Study of the possibilistic (non-deterministic but **<u>not</u>** stochastic) evolution of systems with restrictions on which parts of a state-space they can inhabit
- Quite useful to study what systems meet the criteria to have an internal model
- Used in biology, control, economics and other areas but rather niche

Viability theory (maths)

- For the maths-oriented mind: dynamical systems defined using multi-valued functions ("set valued analysis") with (co)restrictions ("viability")
- For the cat-theory-oriented mind: dynamical systems living in the Kleisly category of the nonempty powerset monad on **FinSet** with with (co)restrictions (of interest is also **Smooth**)







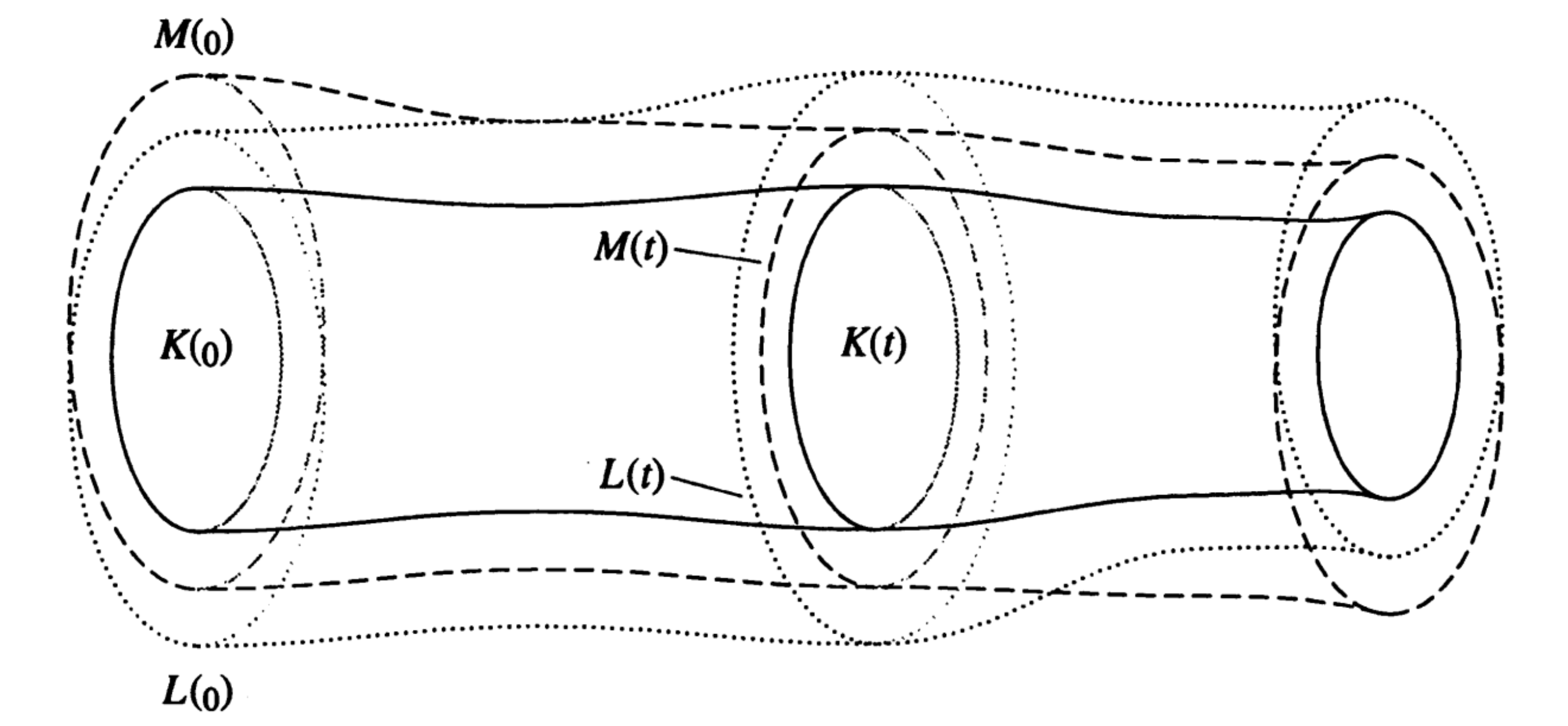


FIGURE 0.2. Tubes satisfying the intersectability property $L(t) \cap M(t) \neq \emptyset$ and the confinement property $K(t) \subset L(t) \cap M(t)$.