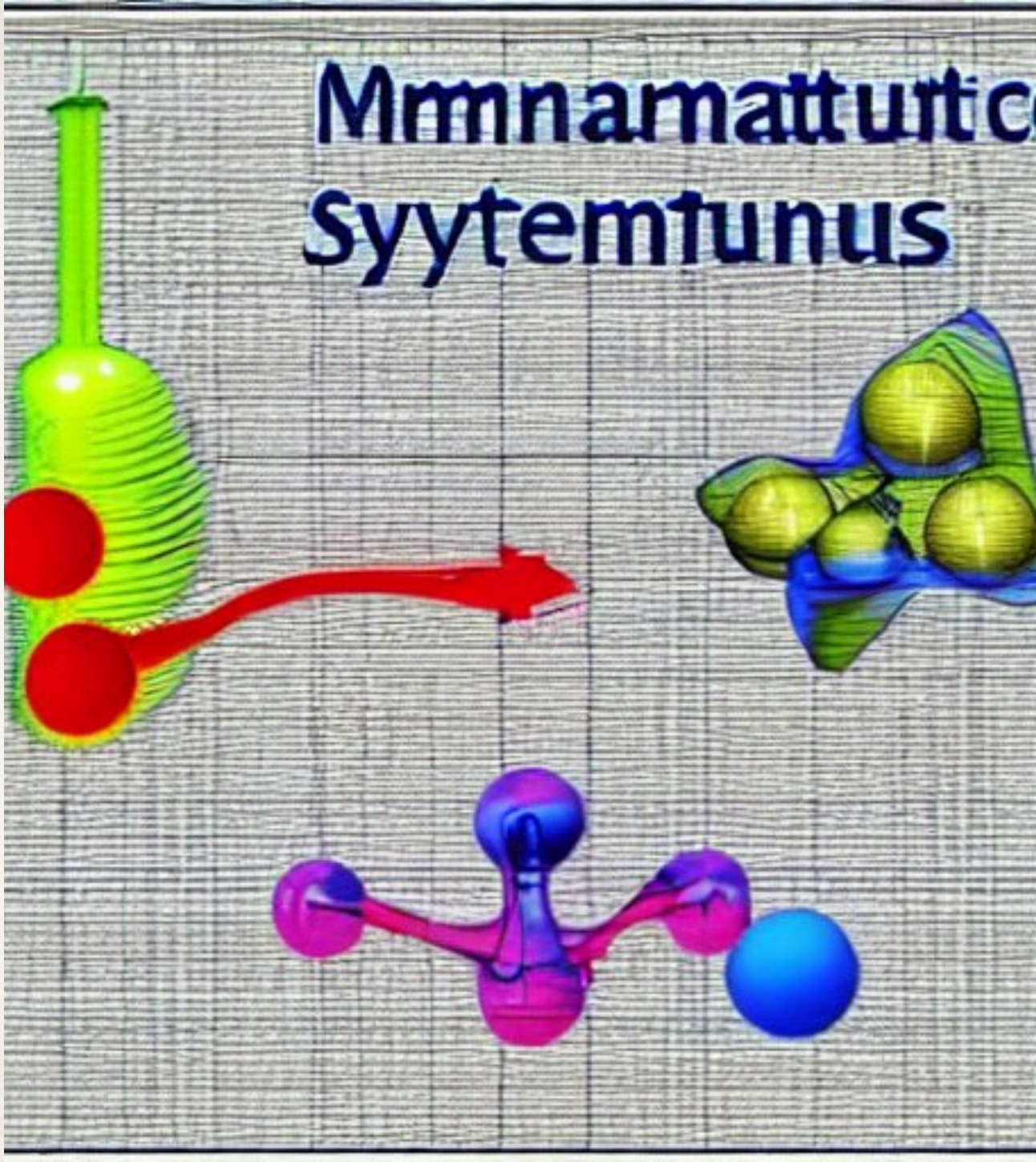
A relational theory of agency (and goals?)

Manuel Baltieri

12th March 2023

Japan AI Alignment Conference

ARAYA



About me

Working at 

Background: computer engineering, business administration | artificial life, computational neuroscience, machine learning, control theory | artificial life, applied c****ory theory

Interests: agent-environment interactions, definitions of agents

Co-organising ALIFE 2023 in Japan https://2023.alife.org/

- We'll host special session + workshop on "(In)Human Values and Artificial Agency" https://humanvaluesandartificialagency.com/ (Deepmind + University of Sussex)

Paper on "Hybrid Life" reviewing theories of systems and agents and different hybridisations between biological and artificial systems + reporting on 5 years of special sessions at ALIFE

Outline

- Motivation (based on my limited contact with AI Alignment)
- Informal definitions of agency / agents
- Formal theories of agency / agents
- A behavioural approach ("synthetic internal model principle"? WIP)
- Future work

Motivation

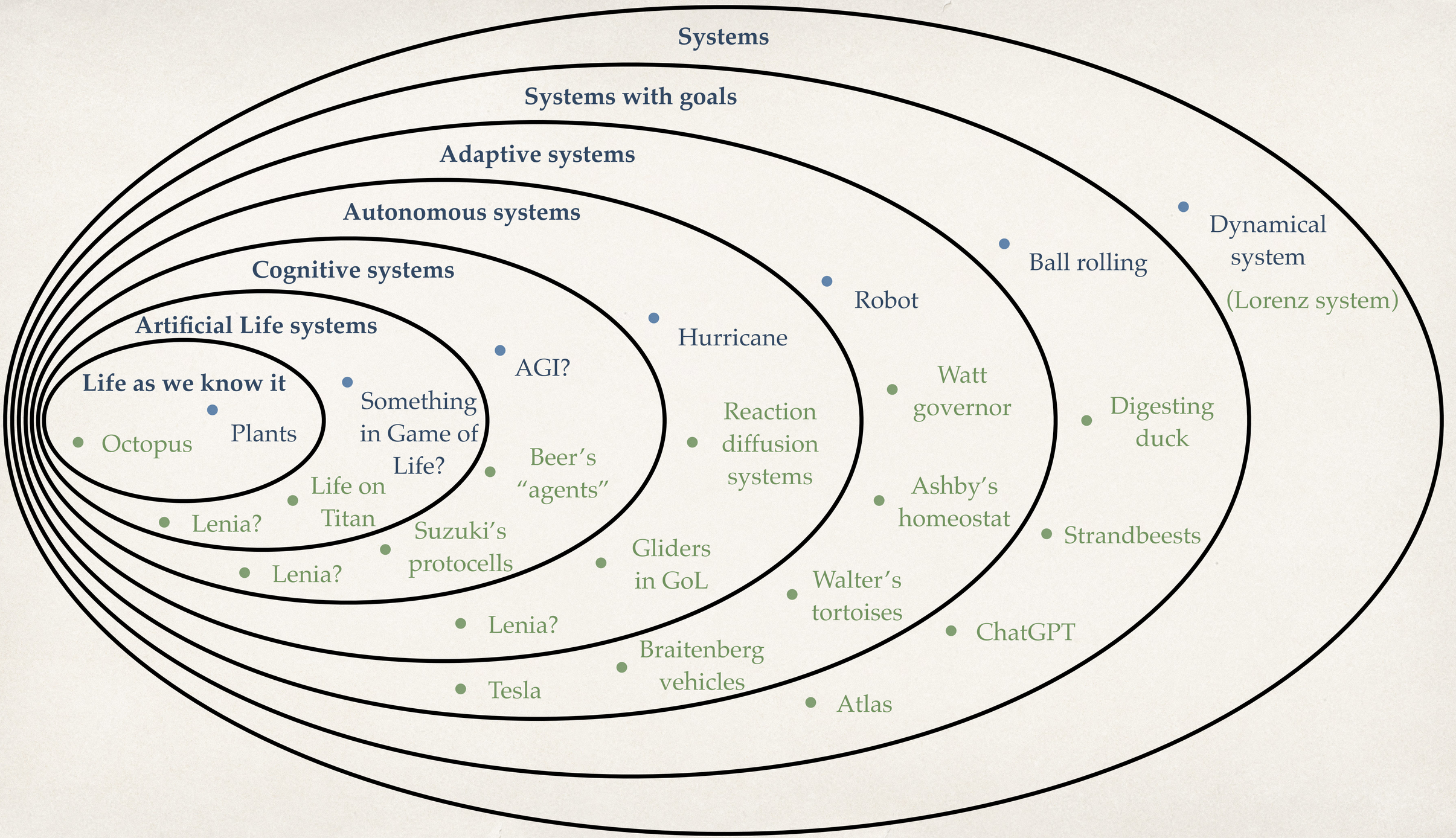
AI Alignment seems to deal with

- Systems possessing/being (partial?) models of some universe
- Goals of a system

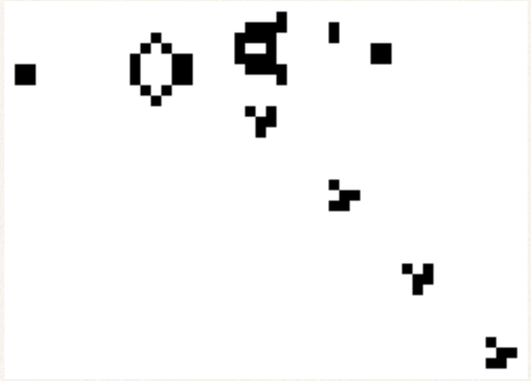
Some shared interests?

My informal definition of agency/agents

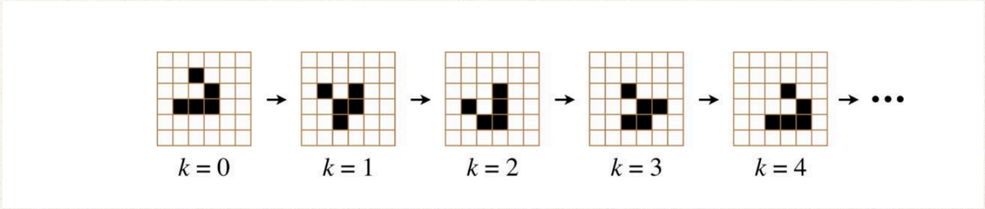
Agents, systems possessing agency, are generally understood as systems autonomously acting with a purpose, to achieve certain goals in an environment.


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

Is the glider an agent in the GoL?



https://en.wikipedia.org/wiki/Conway%27s_Game_of_Life



https://direct.mit.edu/artl/article/20/2/183/2768/The-Cognitive-Domain-of-a-Glider-in-the-Game-of

My view on formal theories of agency/agents

Three classes of first-principles definitions of agents

- Prediction-based (free energy principle, informational individual, behavioural compression)
- Causality-based (integrated information theory, semantic information, mechanised causal graphs)
- Relational (dynamical systems for agent-environment interactions, Bayesian interpretation map, synthetic internal model principle)

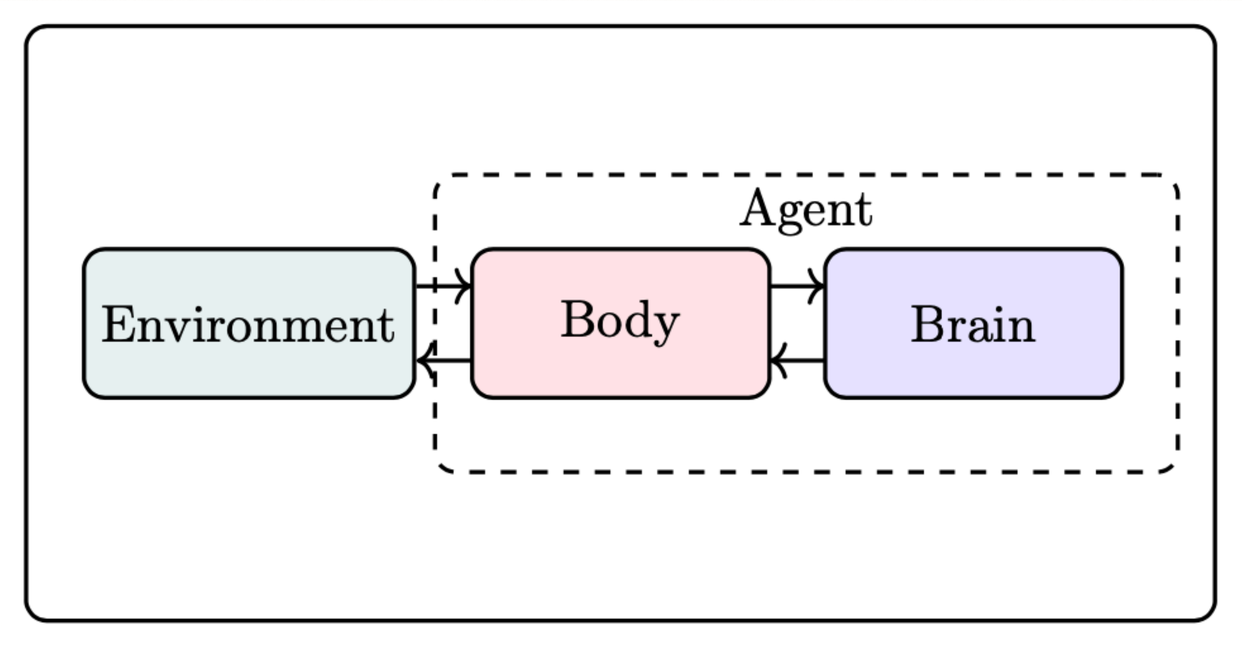
Synthetic internal model principle

0. Our objects of study are (coupled) dynamical systems

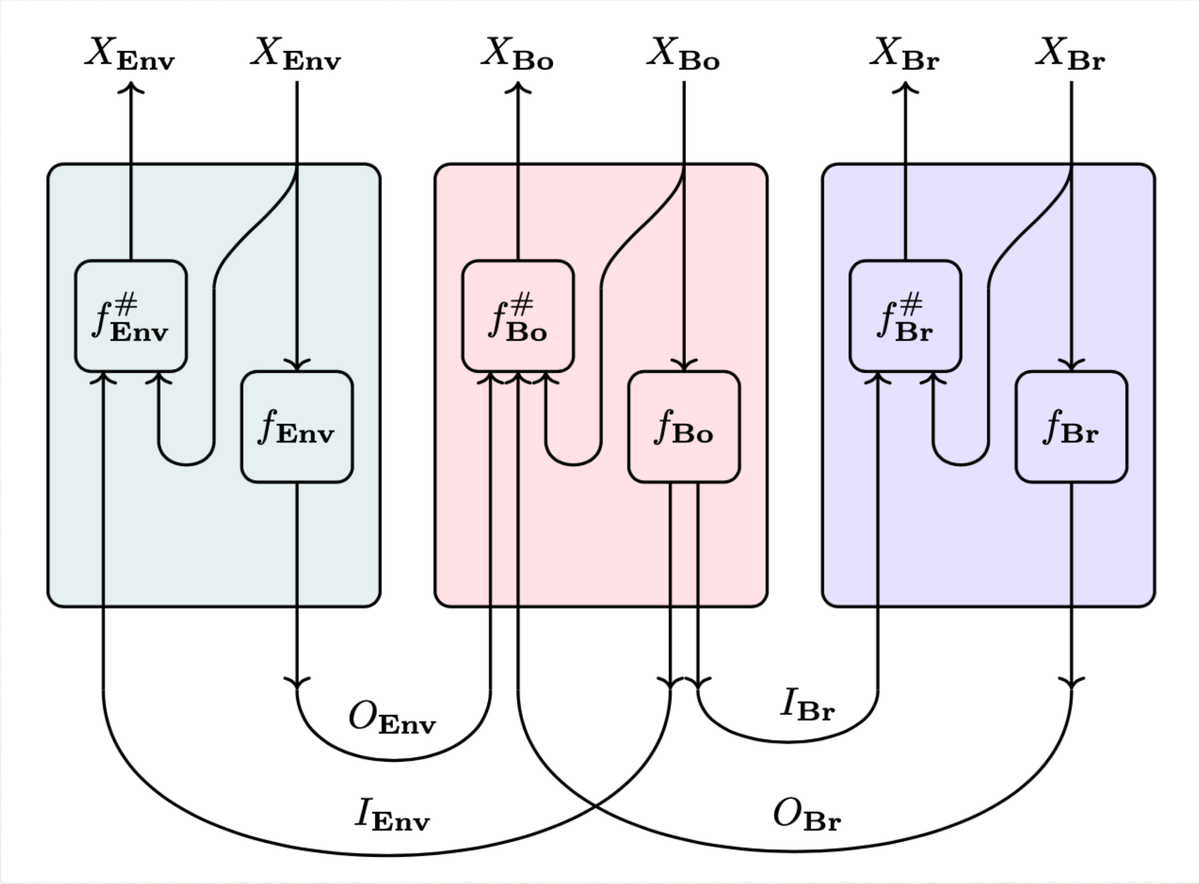
1. An agent, if present, can be decomposed into “brain” and “body”

2. We want to understand if/how/why agents (brains) model their environment

Brain-body-environment systems



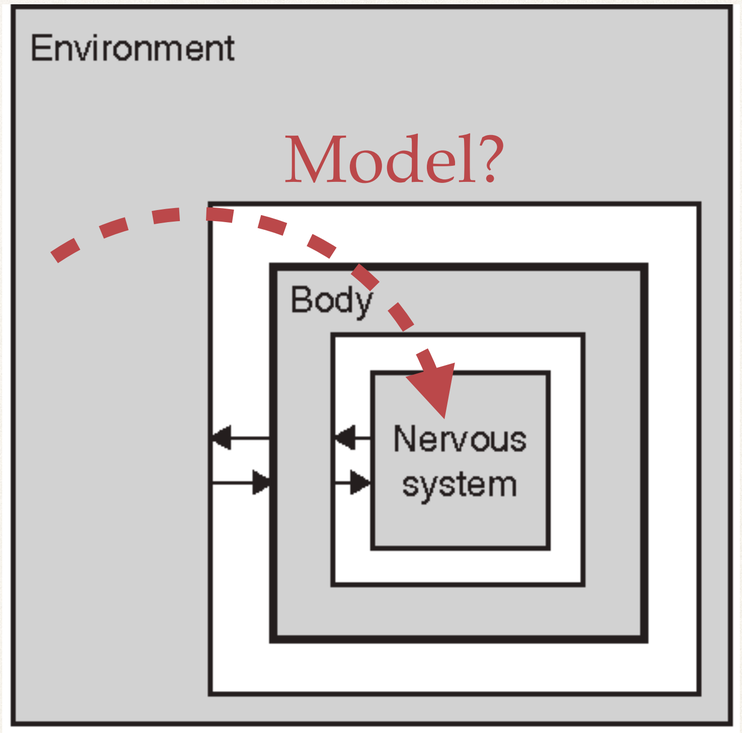
Showing hidden states



$$o_t = f(x_t)$$

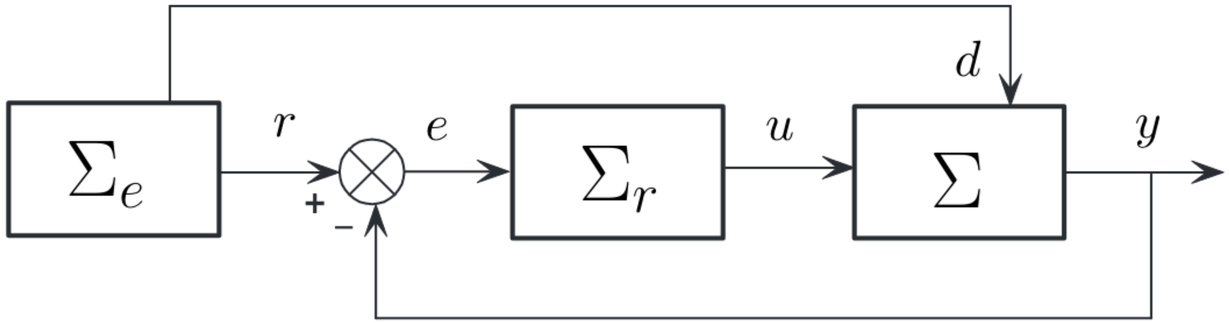
$$x_{t+1} = f^{\#}(x_t, i_t)$$

Relations between hidden states?



Beer, R. D. (2008). The dynamics of brain–body–environment systems: A status report. Handbook of Cognitive Science, 99-120.

A standard control architecture



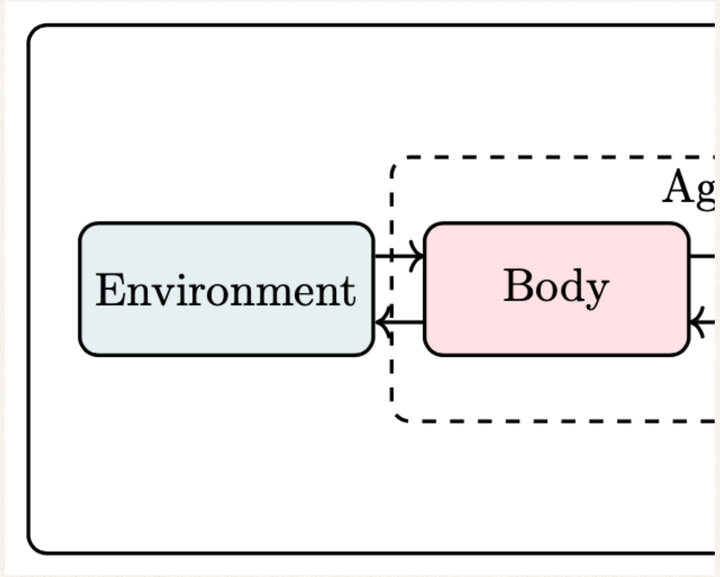
Marro, G. (2008). The geometric approach to control: a light presentation of theory and applications. In Control Science Evolution (pp. 157-204). CNR Publications.

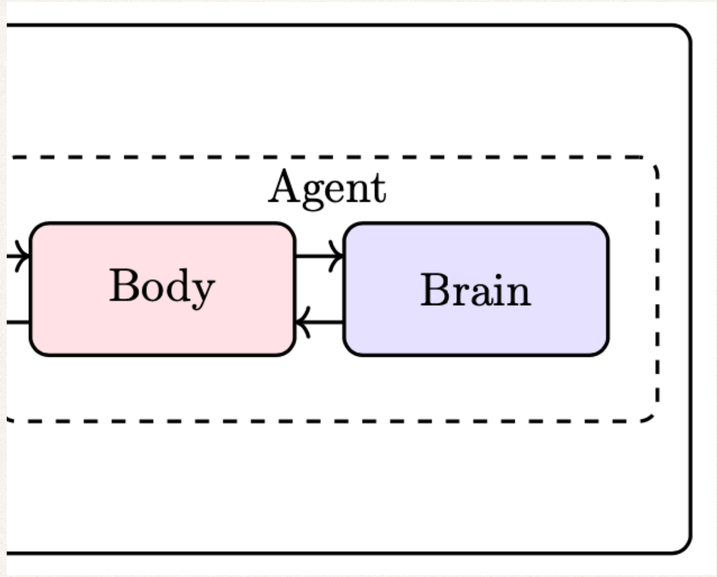
Systems with models

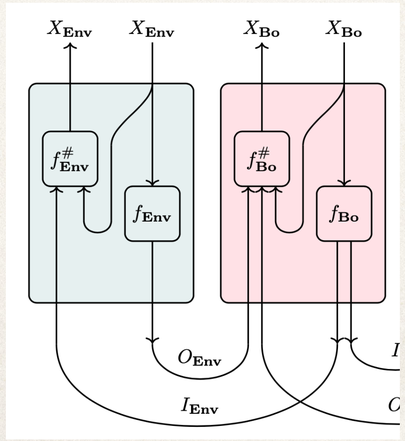
Systems that control a set of variables (= have goals) against disturbances from the environment, must contain/have/be a model of their environment

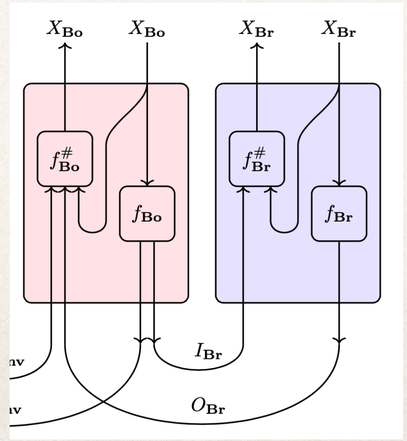
- Law of requisite variety
- Good regulator theorem
- Internal model principle
- Free energy principle
- …

Looking at pair-wise relations

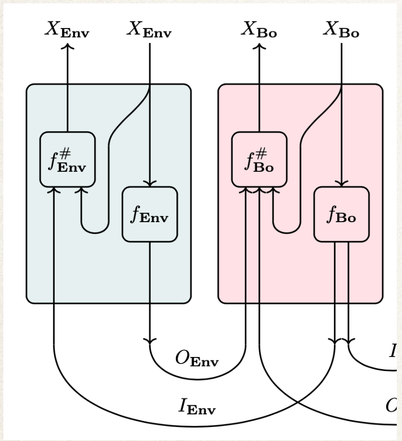


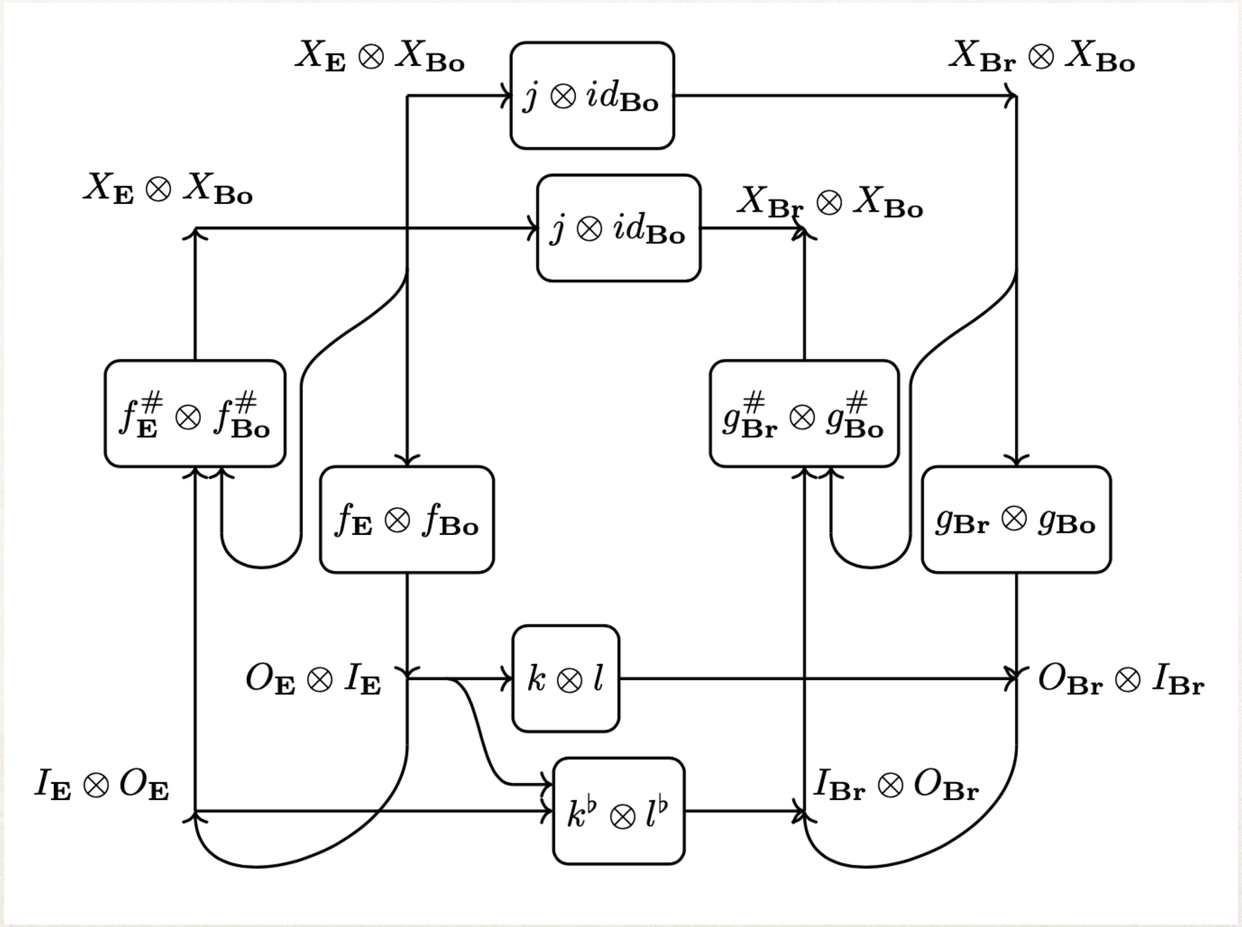


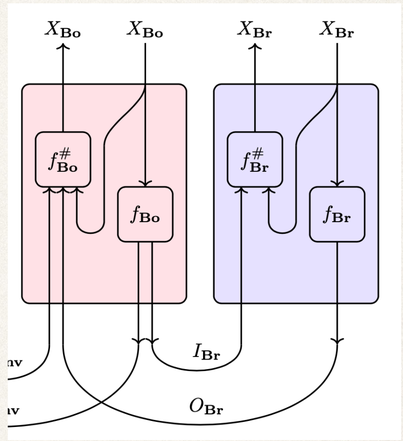




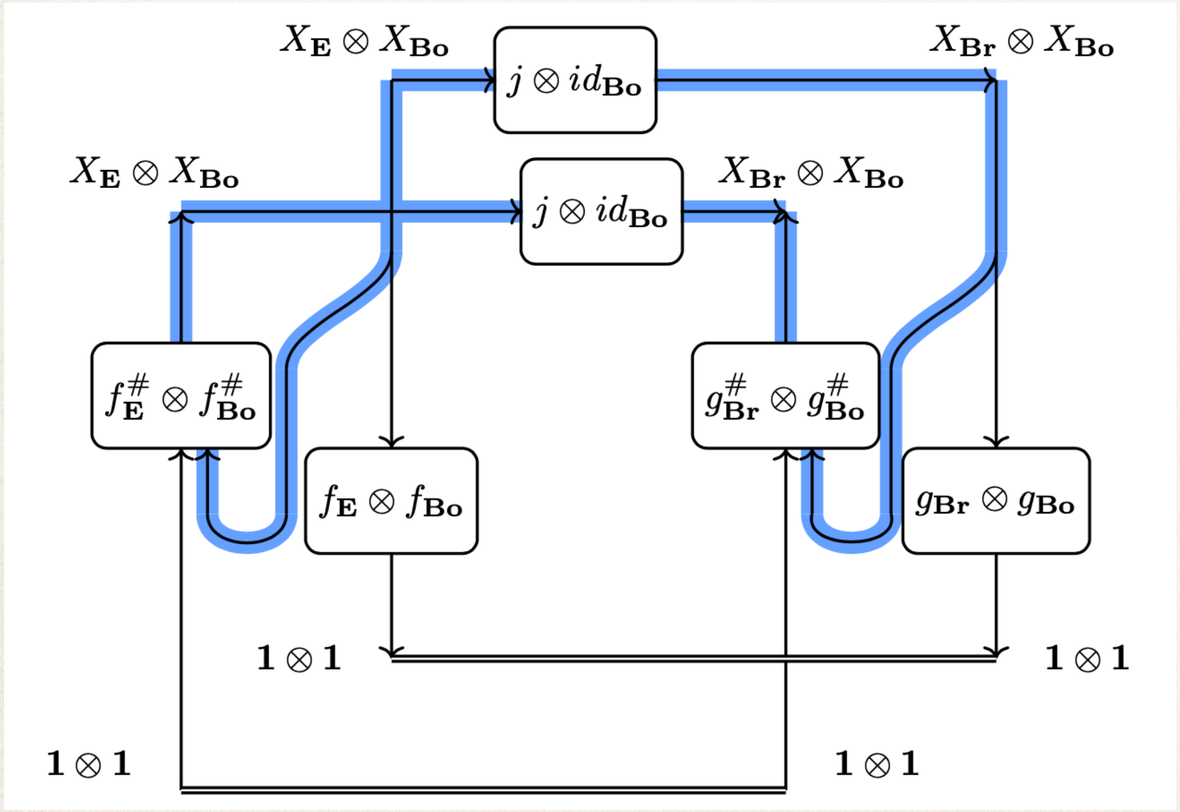
Looking at relations between relations



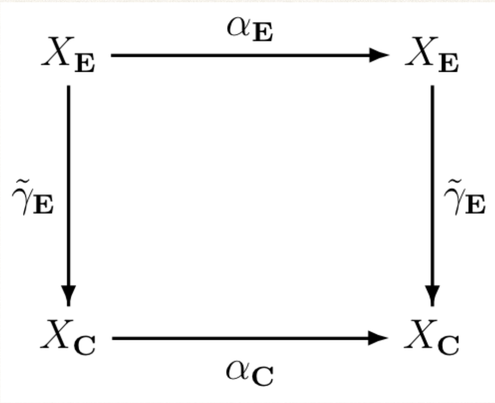




Recovering the internal model principle



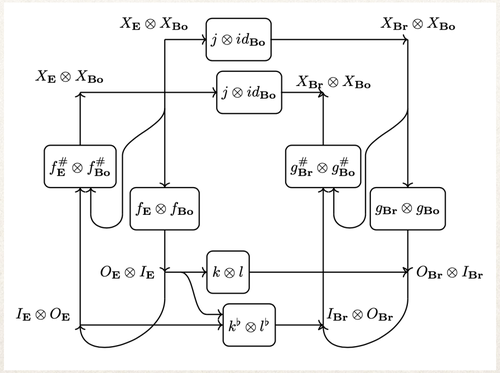
cf., transpose and rename functions/variables

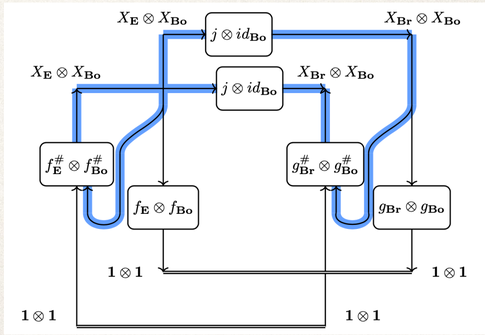


Wonham, W. M., & Cai, K. (2019). Supervisory control of discrete-event systems.

Advantages

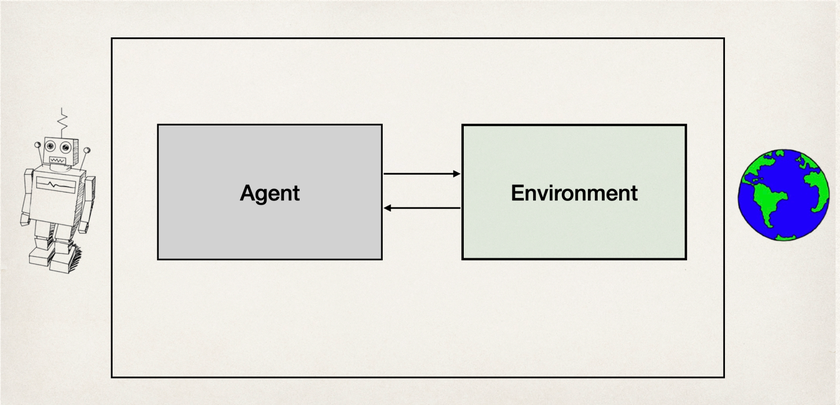
- Can be generalised to other objects (sets, probabilities, manifolds, topologies (?), continuous-time stochastic processes (?), …) more systematically
- Can be applied to other fields (robotics, ML, neuroscience, physics) more formally
- Helped me find a HUGE assumption that makes me question the whole framework (it's open loop? —> hopefully it can be fixed)
- Formal connection to code implementations (WIP)
- …

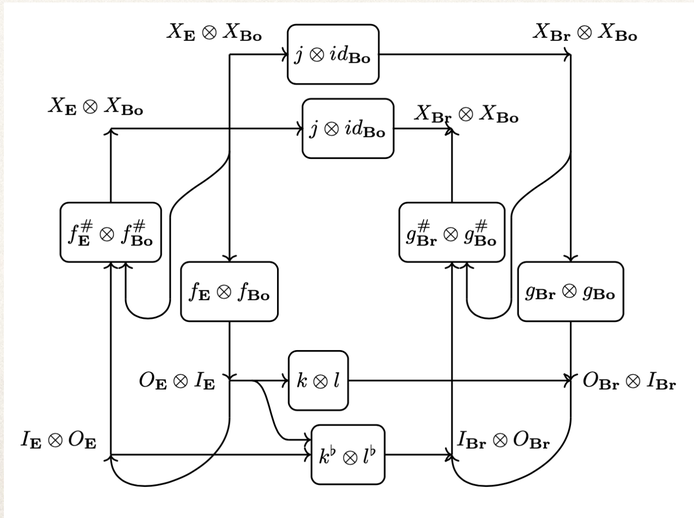




Summary

- Agents as systems with models and goals
- Models should not be taken for granted, yet (NB: it doesn't mean there can't be systems with models doing cool things)
- If models can be assumed, there is a behavioural approach (sketched here) to open black boxes given certain patterns at the interface(s)

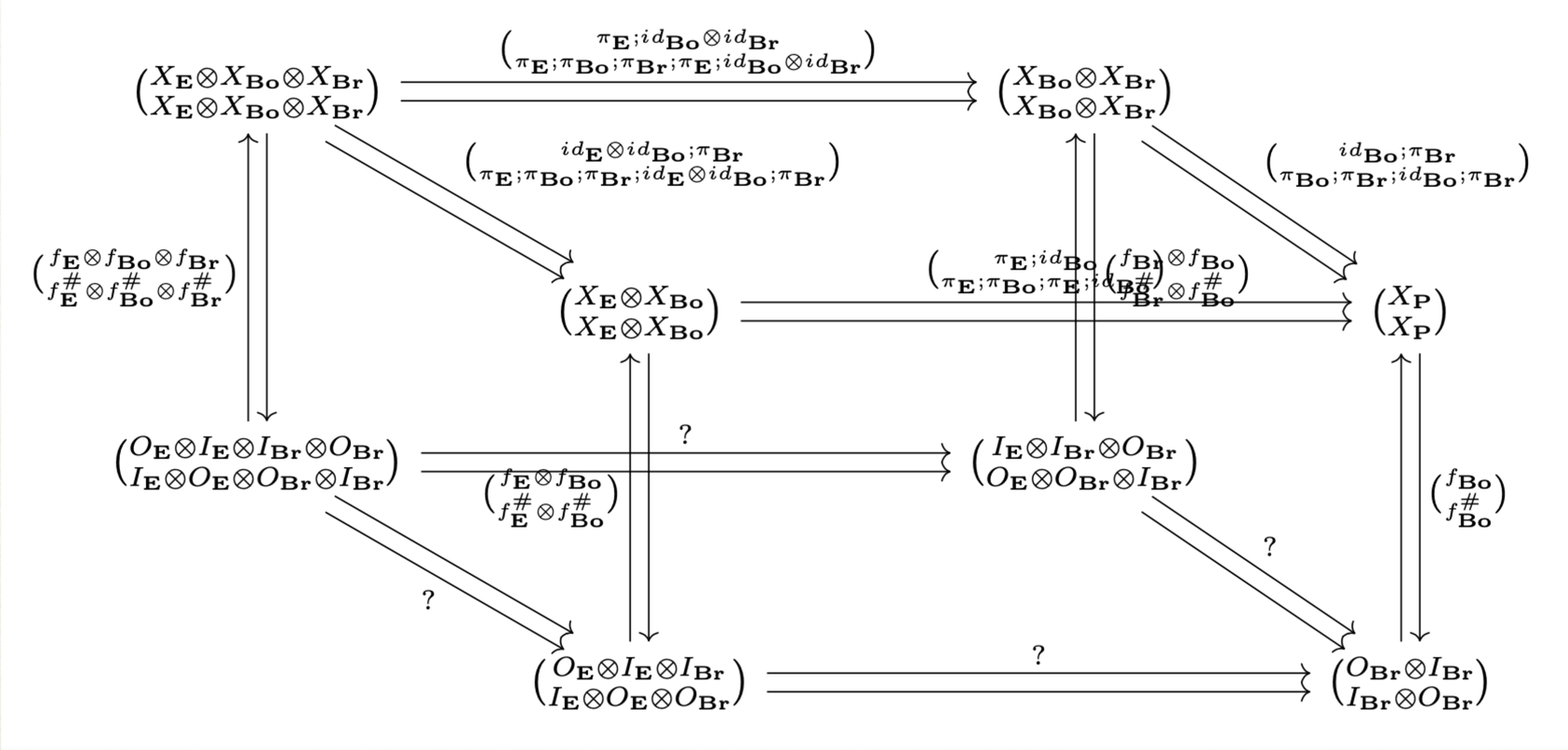




Next steps

- Check similar (?) approach in relational biology
- Internal model principle allowing an agent's actions
- Proof of existence of a environment-brain map behind this result (seems to work too well with sets, requires extra assumptions on manifolds, maybe it doesn't work with something else)
- Include goals

Existence of environment-brain map



Goals

- Either a cube like the above (but with a different third dimension) if we take goals to be objects of the “same kind”
- Or parametrisations of the previous objects, but if so they should somehow also be changing due to an agent’s presence