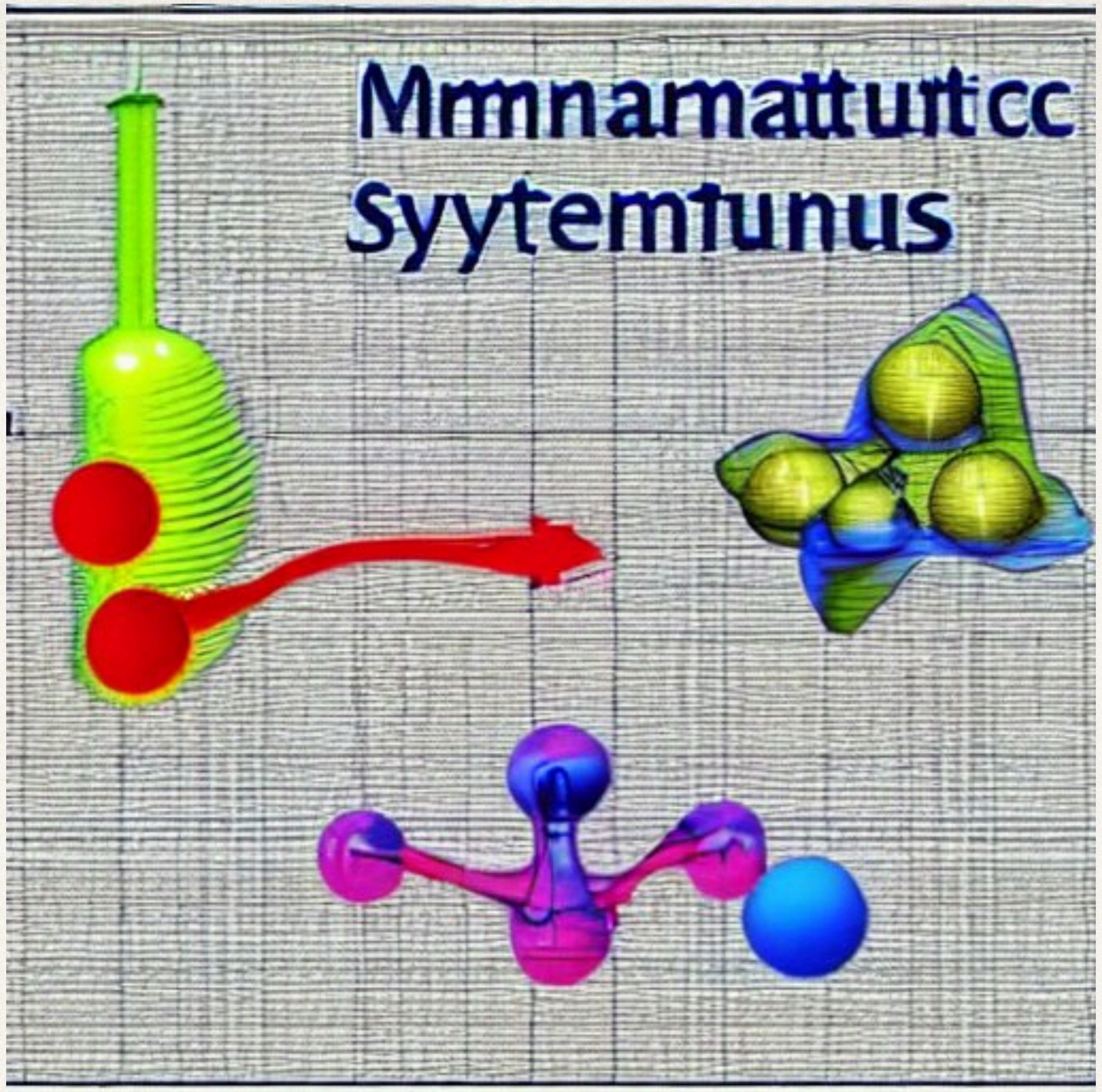


First-principles definitions of agents

Manuel Baltieri

9th Feb 2023

ARAYA

Outline

- Three classes of first-principles definitions of agents
  - Prediction-based
  - Causality-based
  - Relational
- Conclusion

Important note

The following frameworks have various goals:

- Defining individuality
- Defining autonomy
- Defining agency
- …

I will look at them as characterising agency/agents!

Prediction-based methods

Agency in the eye of the beholder

Prediction-based methods

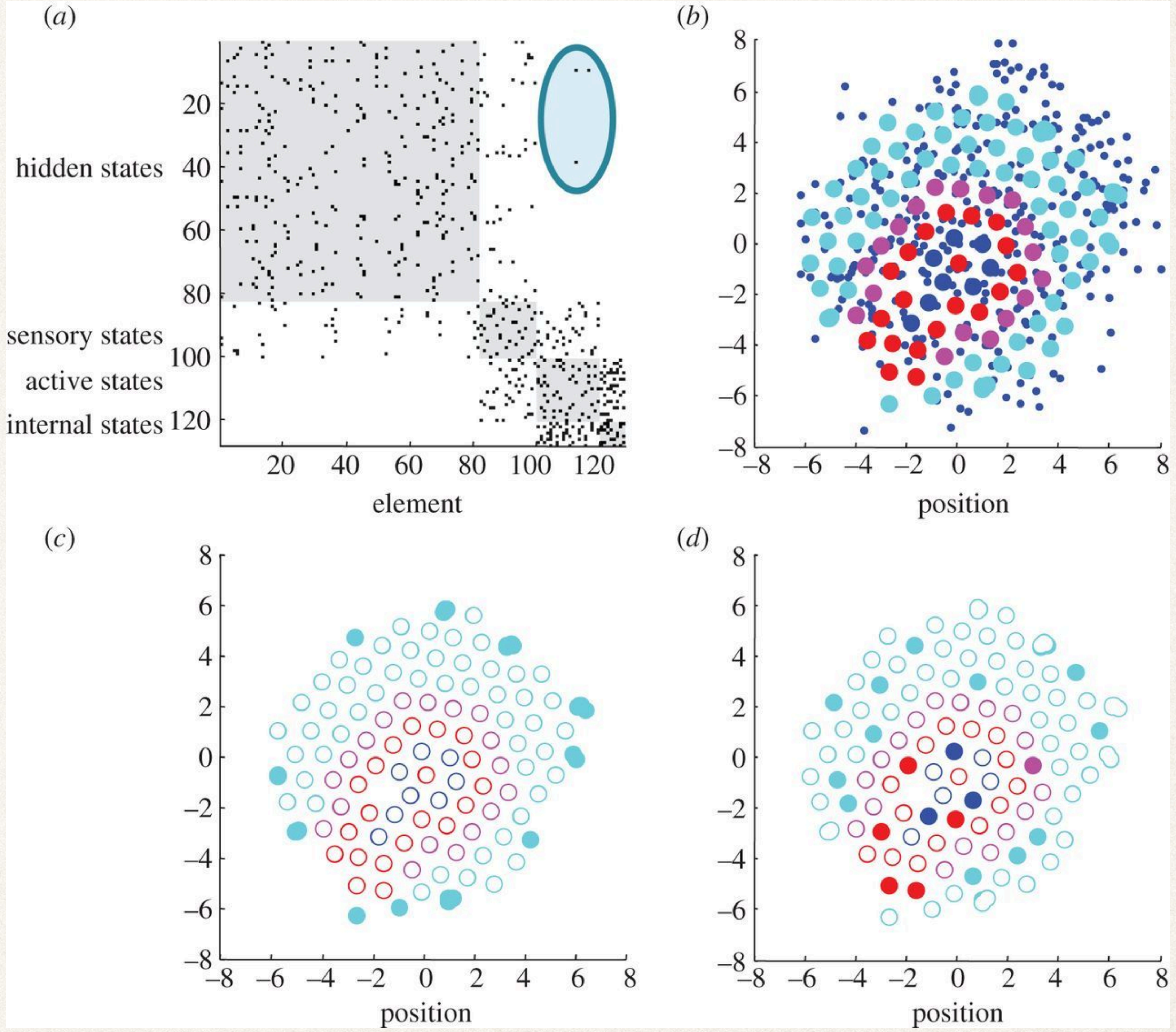
- **Main idea**: treating a system *AS-IF* it was an agent; this helps predicting its behaviour
- **Inspiration**: Dennet's intentional stance
- **Tools**: information theory, filtering theory, Bayesian inference, reinforcement learning, etc.

Examples:

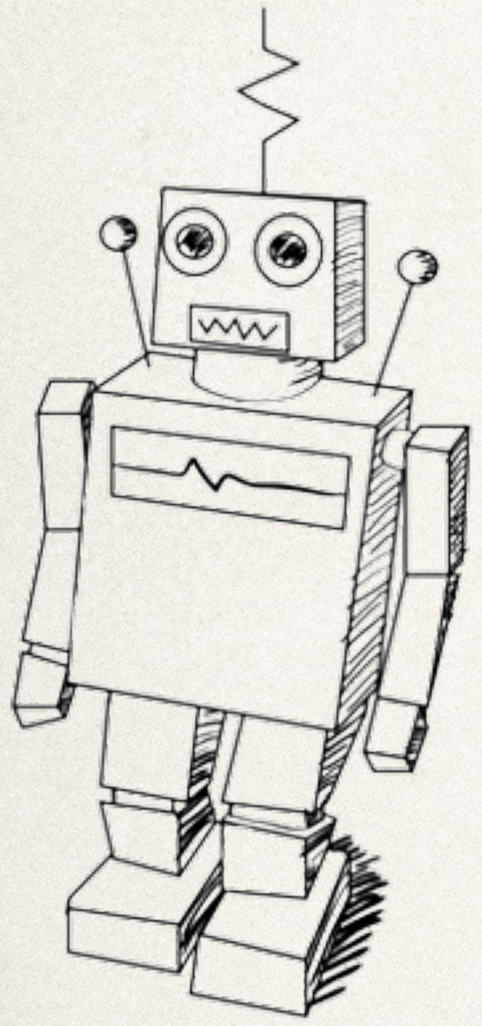
- The free energy principle
- The informational individual
- Behavioural compression

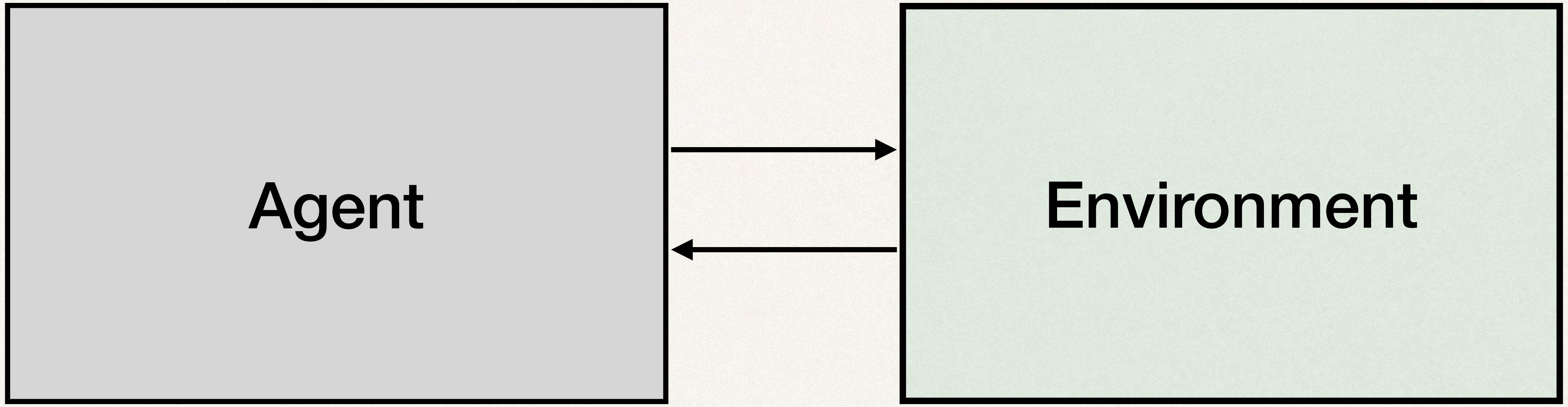
The free energy principle

- 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 veil that separates internal from external states



Friston, K. (2013). Life as we know it. *Journal of the Royal Society Interface*, *10*(86), 20130475.





FEP: the agent performs (approximate) Bayesian inference on the environment

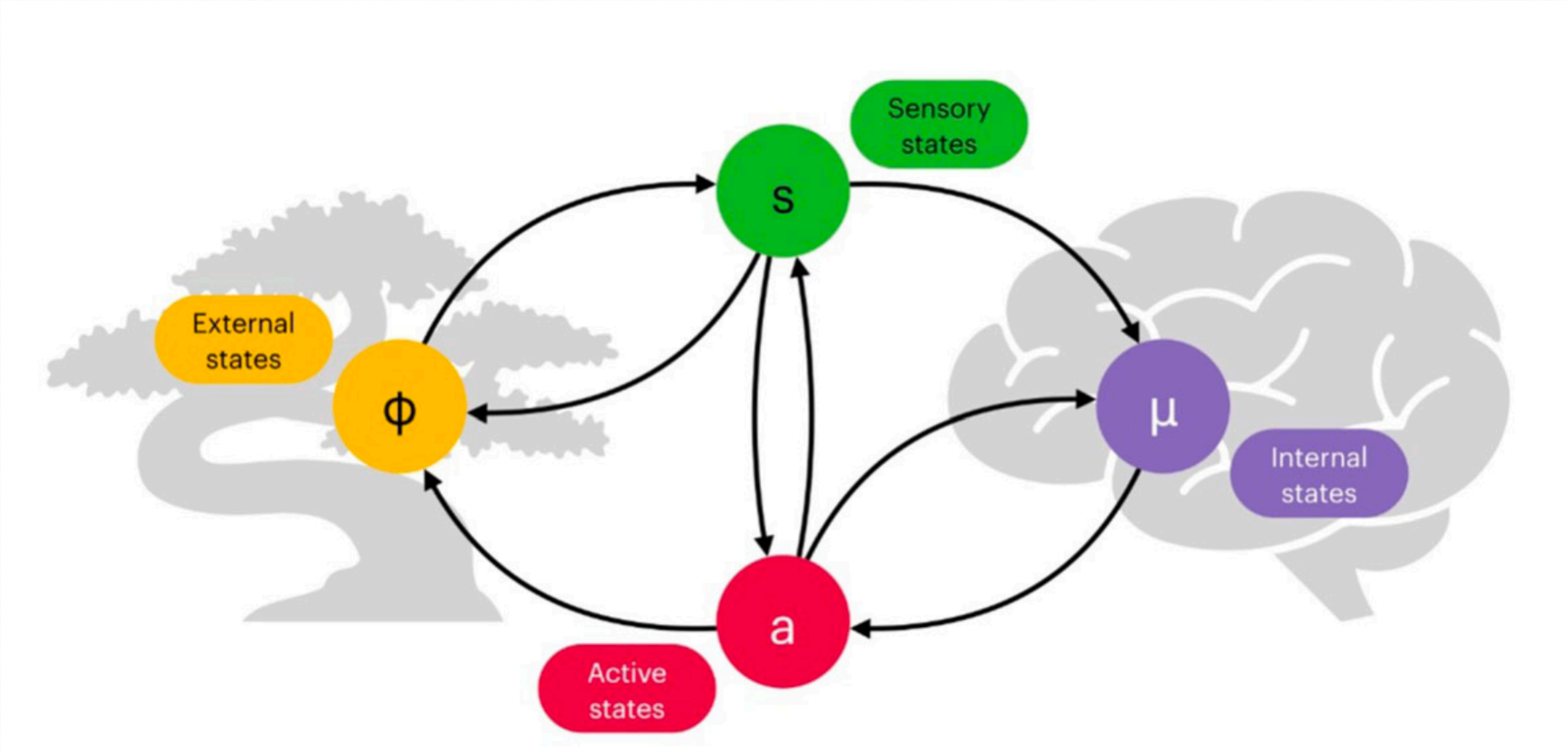
The FEP - pros and cons

Strengths

- Connections to biology, neuroscience, control theory, reinforcement learning and physics
- Agency at multiple scales (attempt of finding scale-free theory)

Limitations

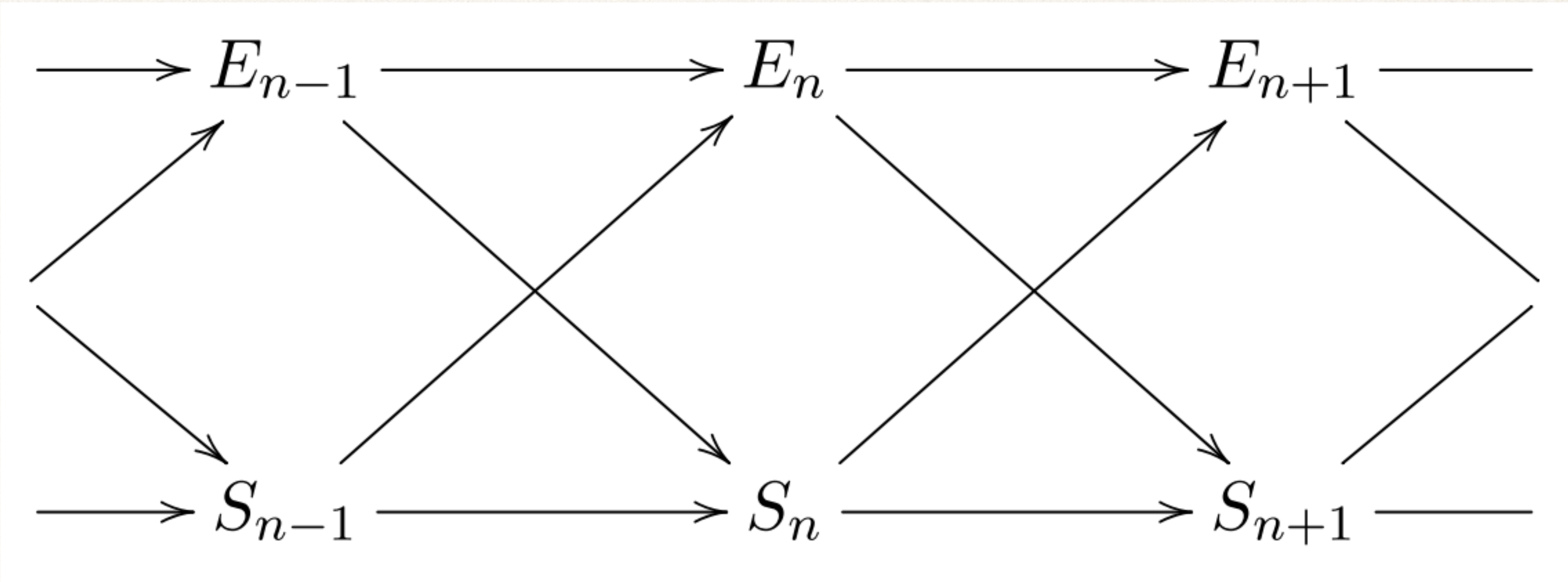
- Technically limited to stationary processes
- Ontological commitments (based just on probabilities?)
- Quantum systems are agents?



Bruineberg, J., Dołęga, K., Dewhurst, J., & Baltieri, M. (2021). The Emperor's New Markov Blankets. Behavioral and Brain Sciences, 45, e183.

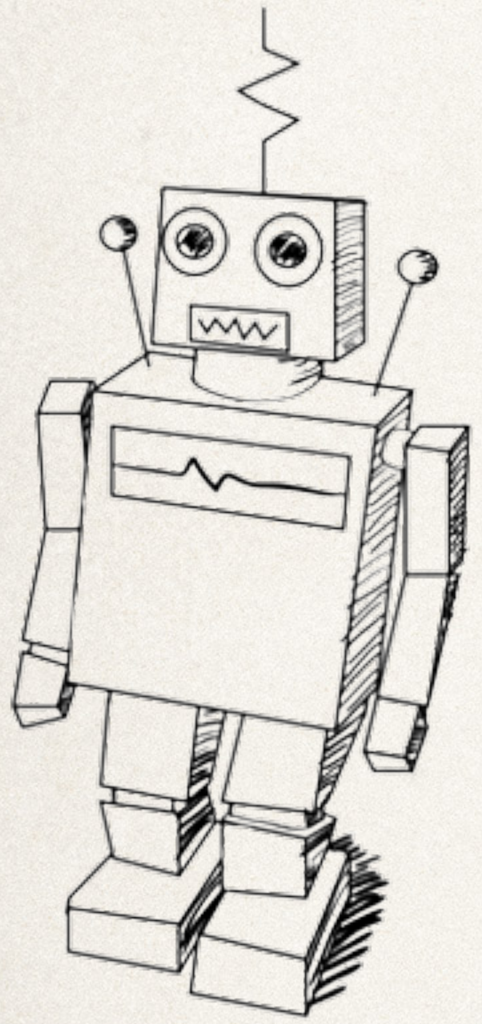
Information individual

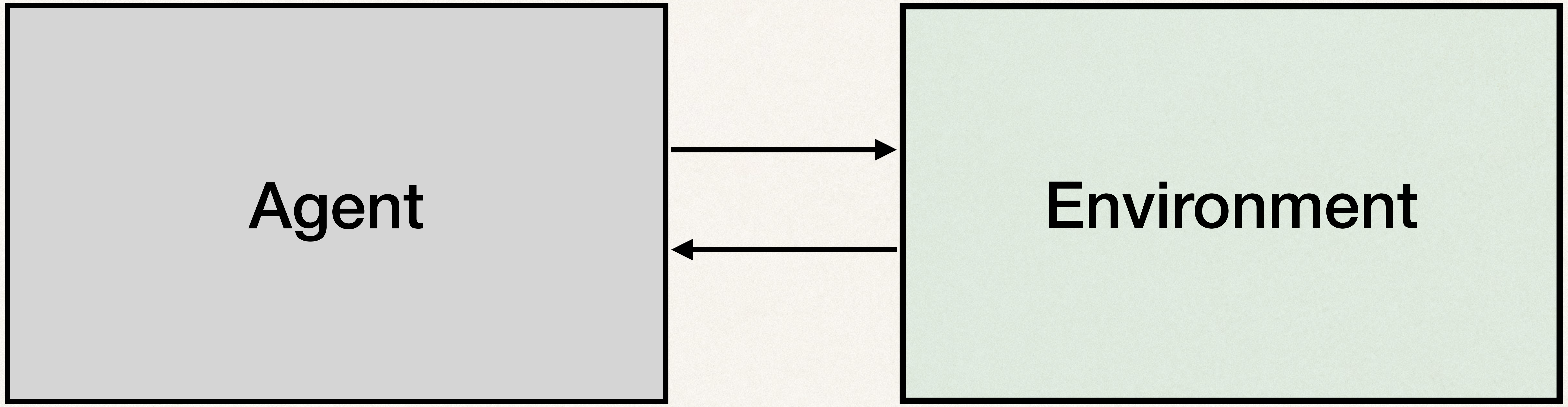
- Information individual: $I(S_n, E_n; S_{n+1})$, how much current state of agent + environment help predict next agent's state
- First decomposition:
$I(S_{n+1}; S_n) + I(S_{n+1}, E_n \mid S_n)$, predictive information of the agent + transfer entropy from environment
- Second decomposition:
$I(S_{n+1}; E_n) + I(S_{n+1}, S_n \mid E_n)$, complementary view



$$\begin{aligned} I(S_n, E_n; S_{n+1}) &= I(S_{n+1}; S_n) + I(S_{n+1}; E_n | S_n) \\ &= I(S_{n+1}; E_n) + I(S_{n+1}; S_n | E_n) \end{aligned}$$

Krakauer, D., Bertschinger, N., Olbrich, E., Flack, J. C., & Ay, N. (2020). The information theory of individuality. Theory in Biosciences, 139, 209-223.





Information individual: an agent is good at predicting its future self

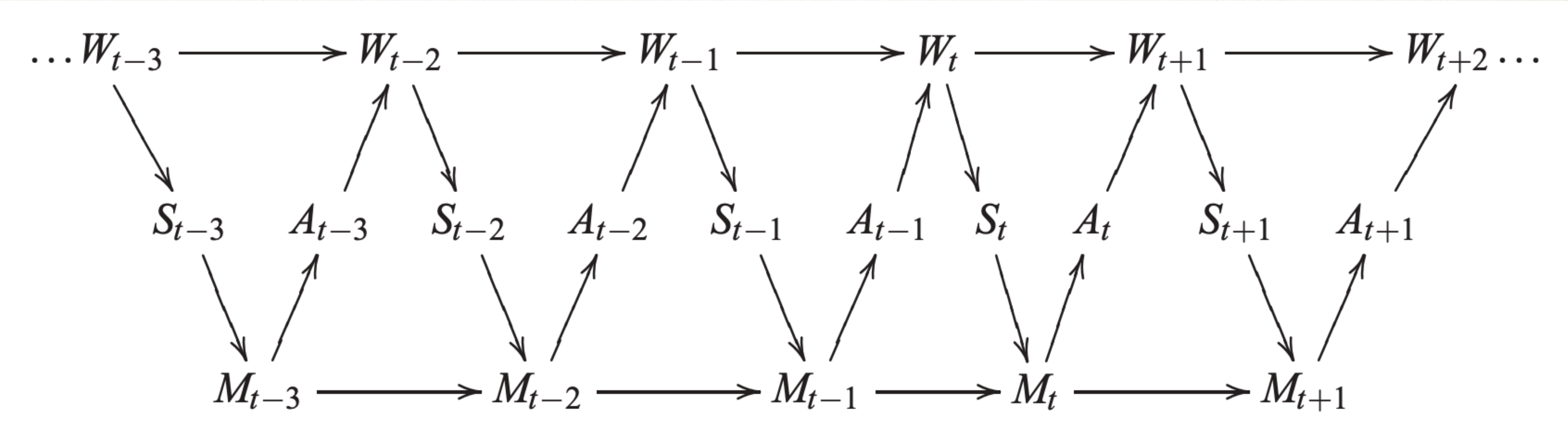
Information individual - pros and cons

Strengths

- No assumptions about the underlying physics
- Multi-level agency (group, individual, etc.)
- Challenging the role of boundaries (e.g., cell membrane)

Limitations

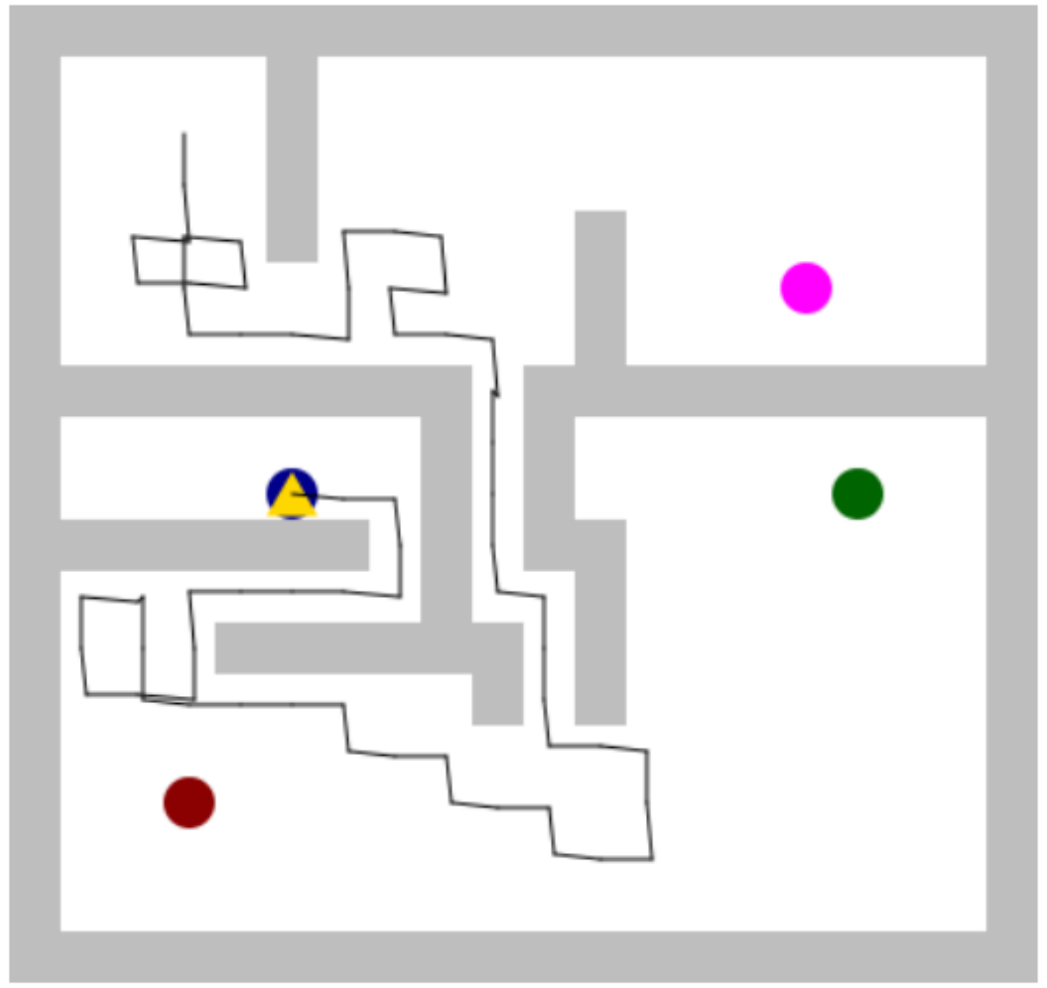
- No clear rule to generate agent-environment partitions
- No action
- Only for discrete time systems



Tishby, N., & Polani, D. (2011). Information theory of decisions and actions. Perception-action cycle: Models, architectures, and hardware, 601-636.

Behavioural compression

- A formalisation of the intentional stance (cf., a stone, a thermostat and a game-playing computer)
- Use inverse reinforcement learning to find the best possible goal for a system
- Use methods from algorithmic probability to find the simplest description of a trajectory of a system (no goals)
- Compare RL agents with policies for planning to reactive systems with step-by-step predictions



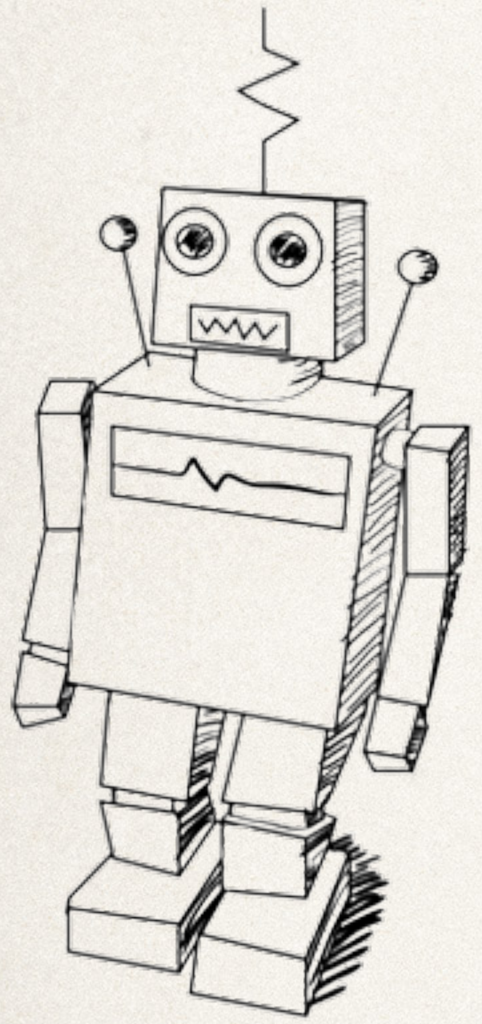
(a) Trajectory.

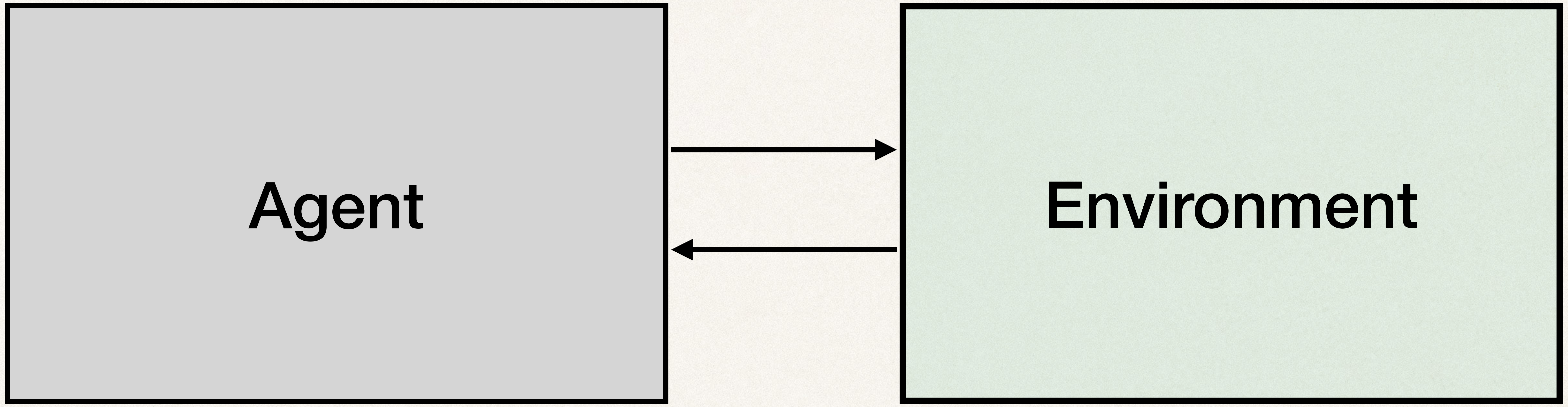
	Device	Agent
v_1	83.53	72.04
v_2	0.00	1.00
v_3	11.49	0.00

$$v_1 = -\ln P(yx_{1:T}|M_?)$$
$$v_2 = P(M_?|yx_{1:T})$$
$$v_3 = -\ln P(M_?|yx_{1:T})$$

(b) Posteriors of the device and agent mixtures.

Orseau, L., McGill, S. M., & Legg, S. (2018). Agents and devices: A relative definition of agency. arXiv preprint arXiv:1805.12387.





Behavioural compression: goals help predicting the long-term behaviour of a system

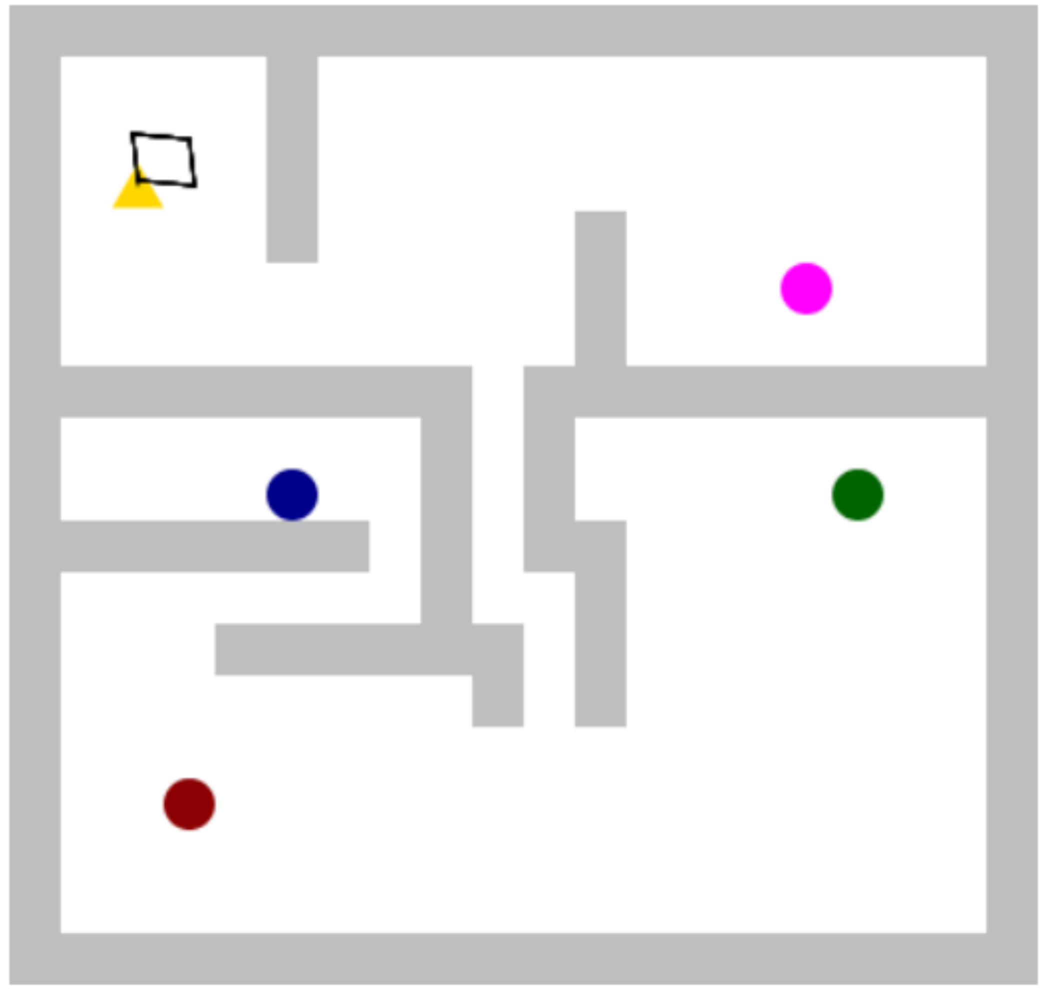
Behavioural compression - pros and cons

Strengths

- Formalising intentional stance
- Allows to consider different goals and compression strategies

Limitations

- Not clear what happens when behaviour can't be compressed
- Can't discriminate between agents, and systems behaving AS IF they were agents
- Unclear whether different choices (goals, predictors) would influence the final results



(a) Trajectory.

	Device	Agent
v_1	18.01	37.48
v_2	1.00	0.00
v_3	0.00	19.40

$$v_1 = -\ln P(yx_{1:T} | M_?)$$
$$v_2 = P(M_? | yx_{1:T})$$
$$v_3 = -\ln P(M_? | yx_{1:T})$$

(b) Posteriors of the device and agent mixtures.

Orseau, L., McGill, S. M., & Legg, S. (2018). Agents and devices: A relative definition of agency. arXiv preprint arXiv:1805.12387.

Prediction-based methods

Advantages

- They take into account the role of observers
- Agnostic about the underlying system (we just need some notion of information)
- Generally flexible enough to consider multiple scales

Disadvantages

- Observer-dependent measures (agency is nothing else?)
- Different existing measures of information
- Can't distinguish between "real" and "as-if" agents

Causality-based methods

Agency as a property intrinsic to a system

Causality-based methods

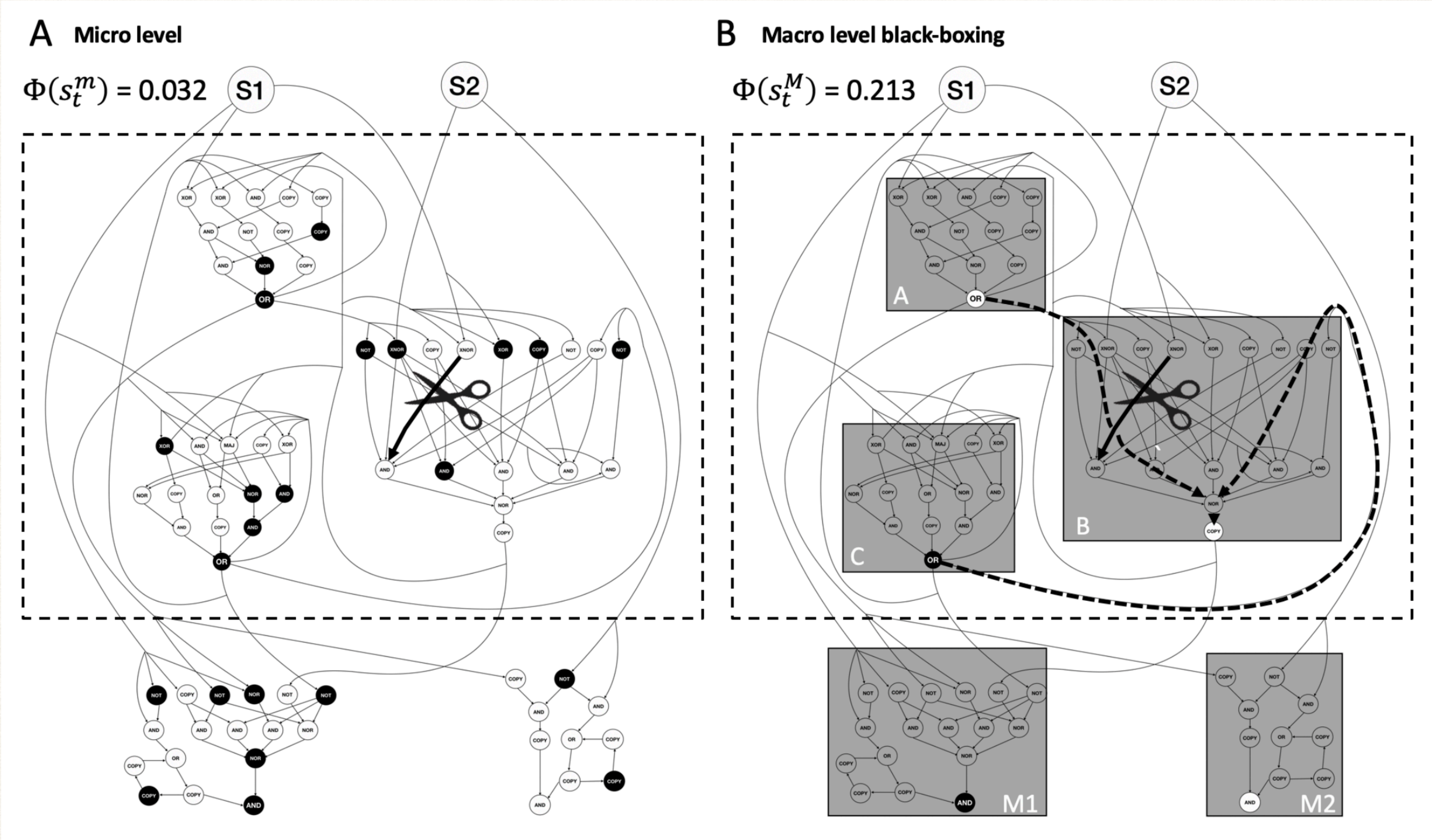
- **Main idea**: actions are causal
- **Inspiration**: Davidson's "causalism", mental states cause actions in the world; Pearl causality
- **Tools**: do calculus, information theory, Bayesian networks, etc.

Examples:

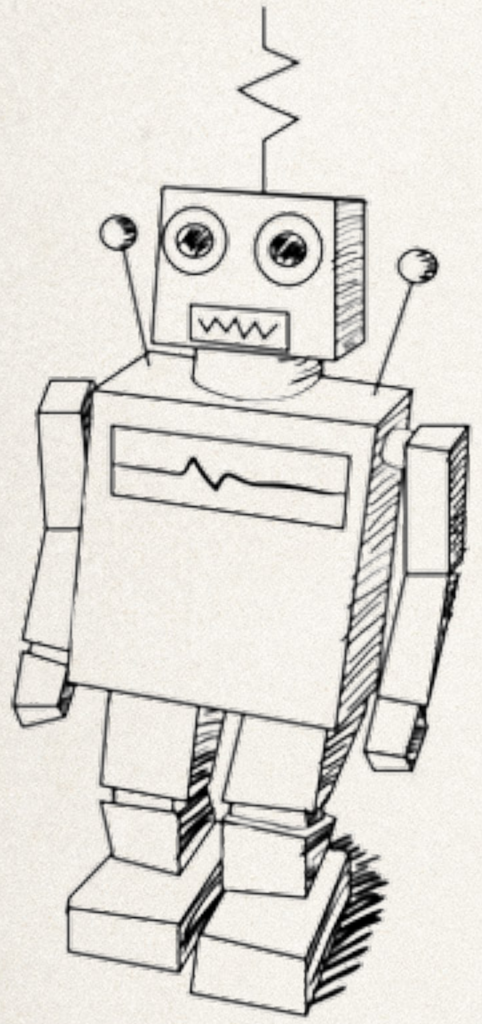
- Integrated information theory
- Semantic information
- Mechanised causal graphs

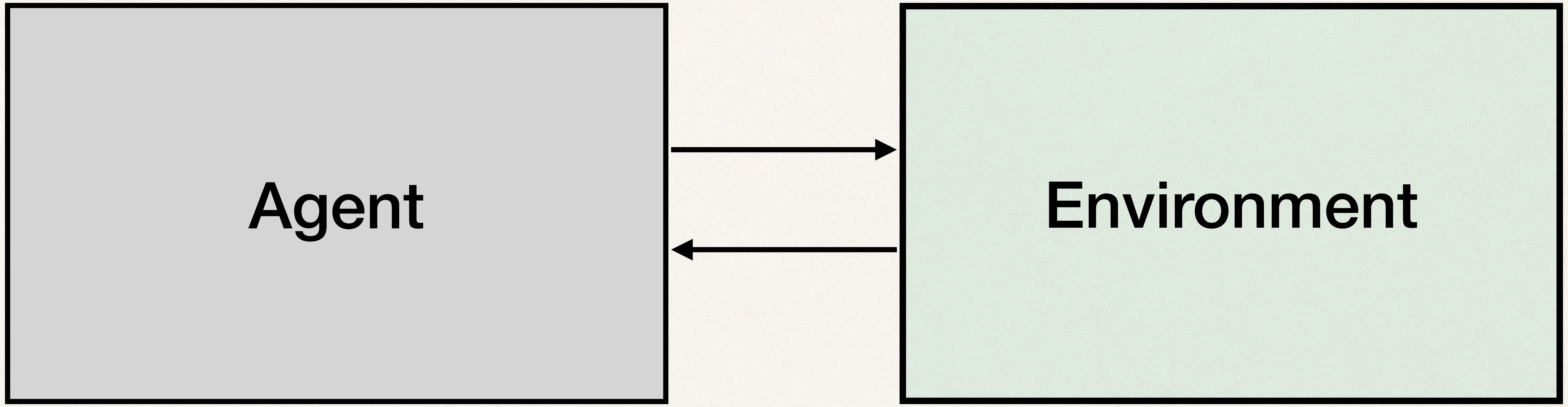
Integrated information theory

- A foundational theory of consciousness (originally), also used for agency recently
- IIT quantifies the “intrinsic irreducibility” of a system: how much cause-effect power of the whole cannot be reduced to its parts
- The highest intrinsic irreducibility of all possible levels determines the system of interest (conscious system, agent, etc.)



Albantakis, L., Massari, F., Beheler-Amass, M., & Tononi, G. (2021). A macro agent and its actions. In Top-Down Causation and Emergence (pp. 135-155). Cham: Springer International Publishing.





IIT: agents are emergent, they can’t be explained
with a reductionist approach

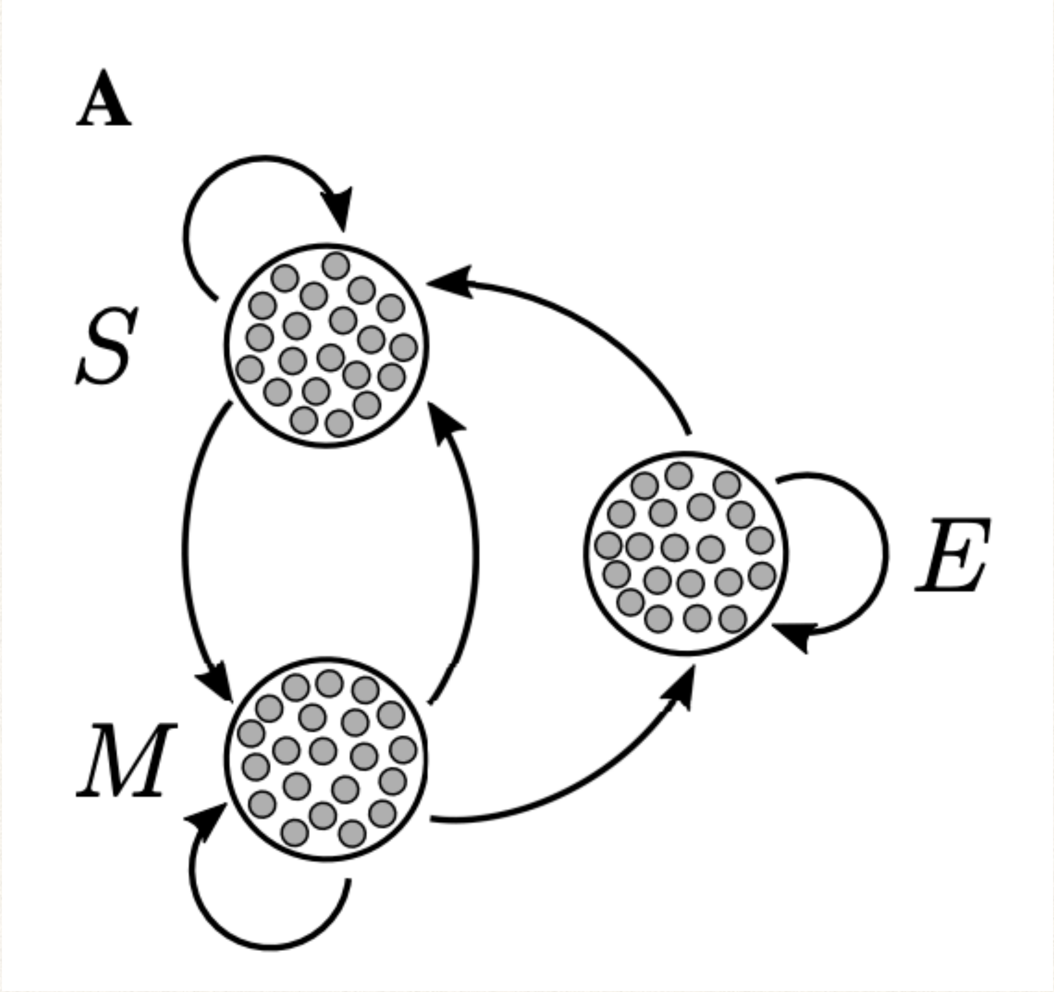
IIT - pros and cons

Strengths

- Formalising agency discovery at multiple scales
- Formalising a way to consider emergence vs. reductionist explanations
- Causal vs observational analysis

Limitations

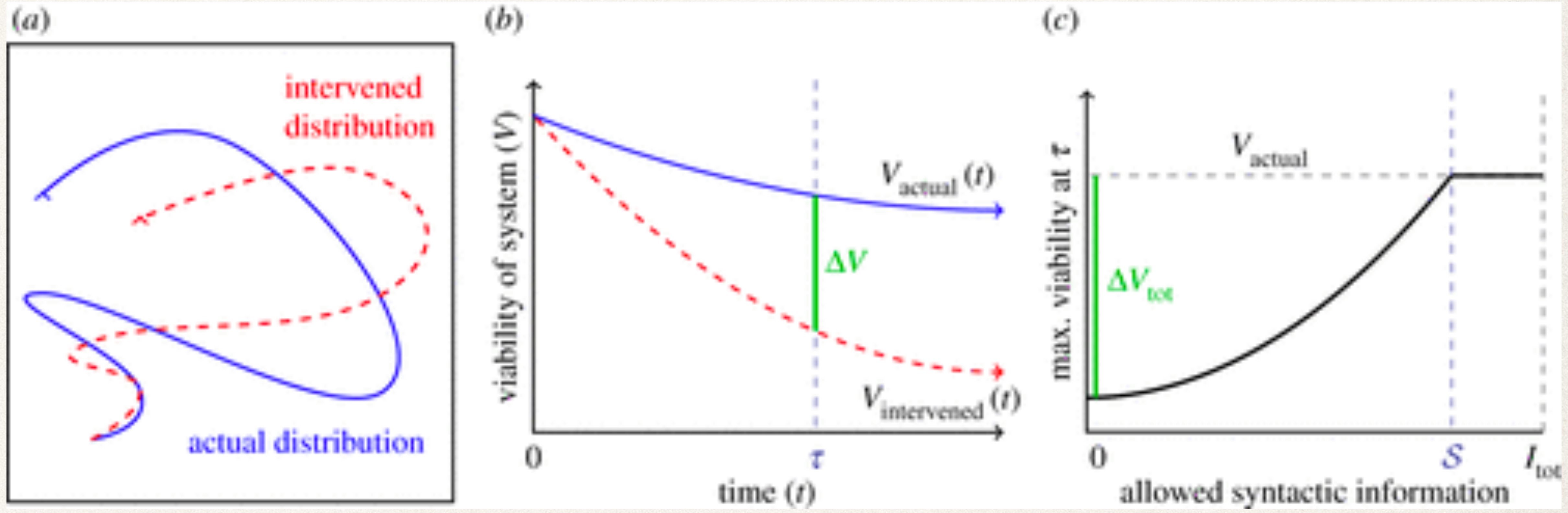
- Agency (consciousness) only defined at a single scale
- General, potential issues with IIT (4.0 version just came out to fix some, or add more?)
- Only for discrete time systems



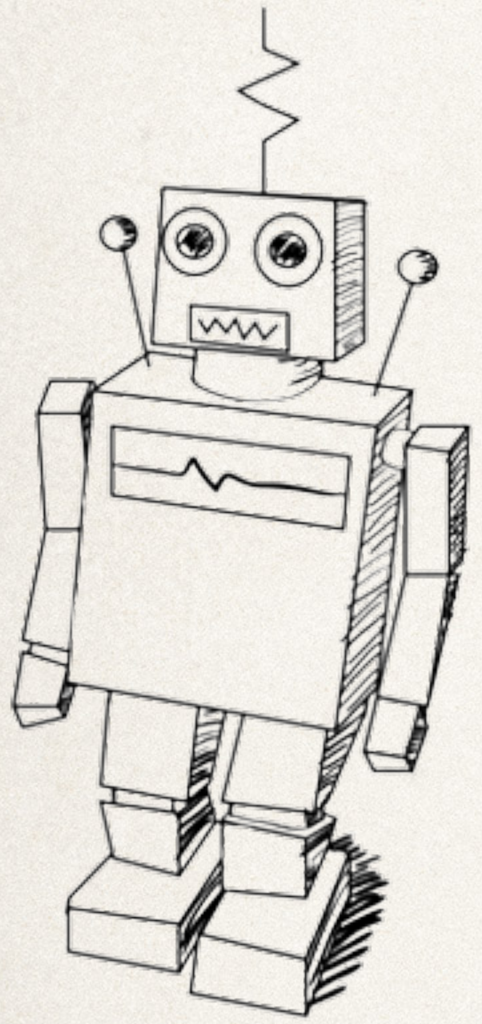
Aguilera, Miguel, and Ezequiel A. Di Paolo. "Integrated information and autonomy in the thermodynamic limit." ALIFE 2018: The 2018 Conference on Artificial Life. MIT Press, 2018.

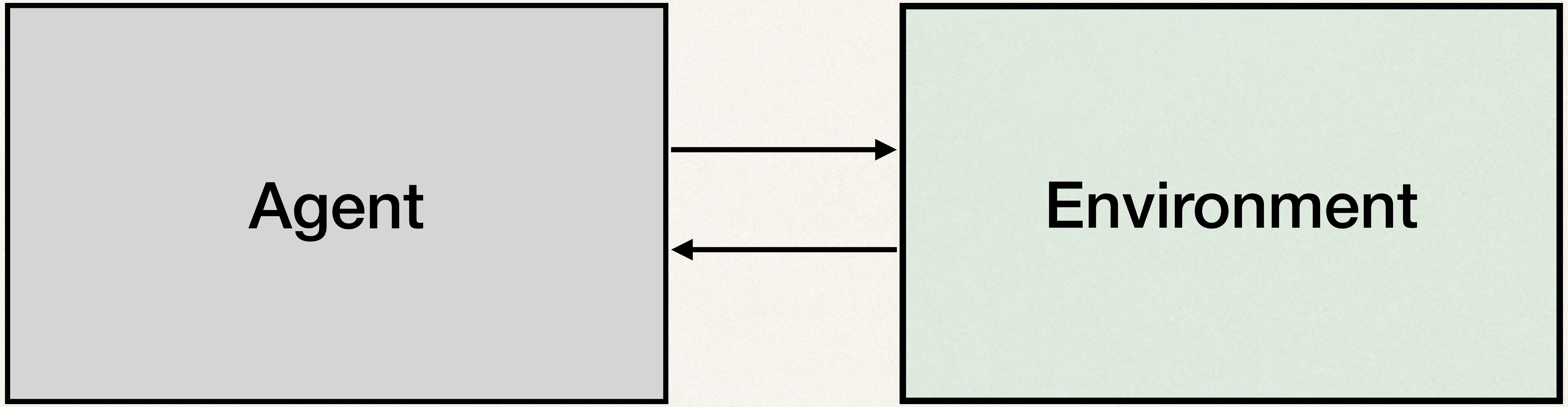
Semantic information

- Shannon information captures correlations, cannot be used to quantify value (semantics)
- Semantic information is a way to provide non-correlational ("scrambled") proxies of standard information measures after choosing a particular goal (e.g., survival)
- Store semantic information: "scrambled" mutual information between agent and environment
- Observed semantic information: "scrambled" transfer entropy between environment and agent



Kolchinsky, A., & Wolpert, D. H. (2018). Semantic information, autonomous agency and non-equilibrium statistical physics. Interface focus, 8(6), 20180041.





Semantic info: agents are systems with a high degree of stored semantic information and observed semantic information

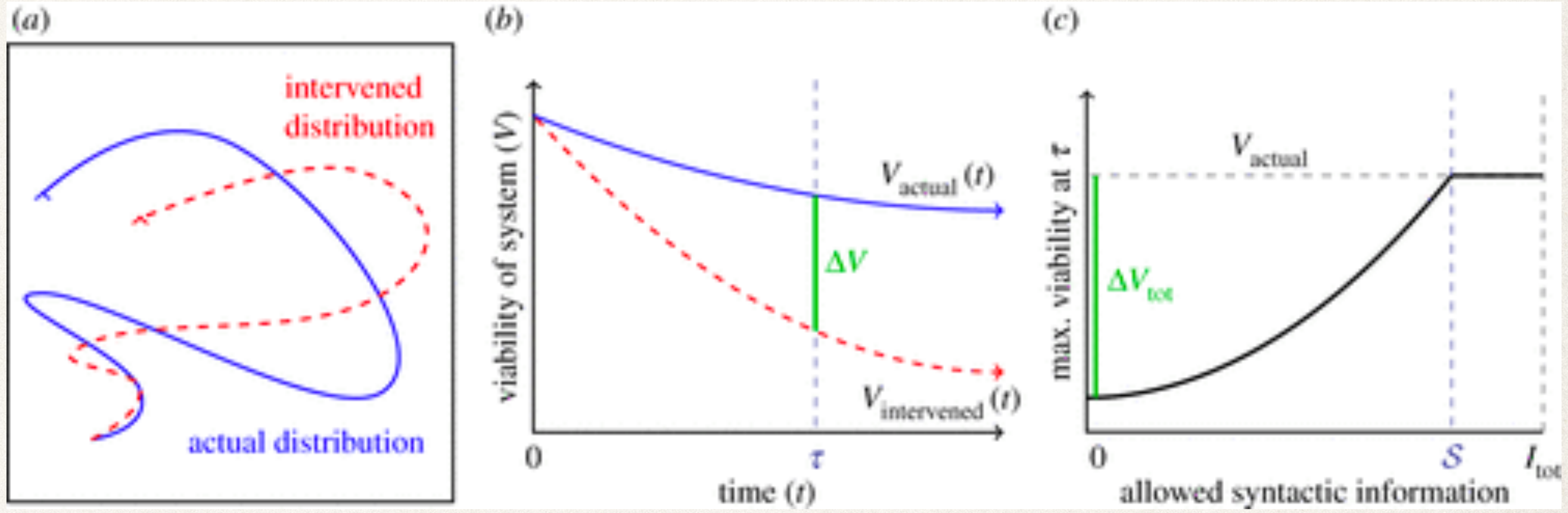
Semantic information - pros and cons

Strengths

- Goal-agnostic theory (we can swap "survival" with something else)
- Clarifying causal vs observational analysis

Limitations

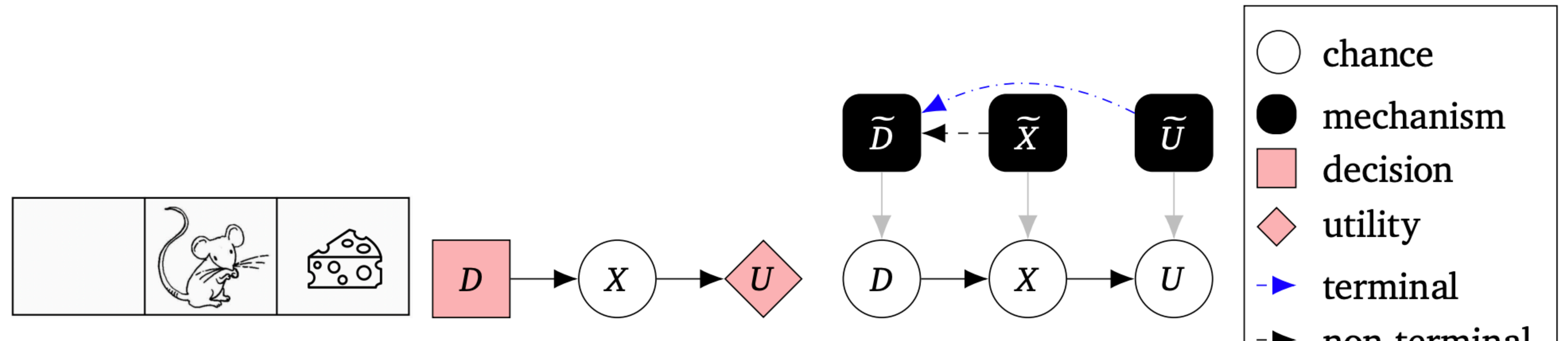
- Doesn't settle on a specific goal for agents
- Assuming we know the agent



Kolchinsky, A., & Wolpert, D. H. (2018). Semantic information, autonomous agency and non-equilibrium statistical physics. Interface focus, 8(6), 20180041.

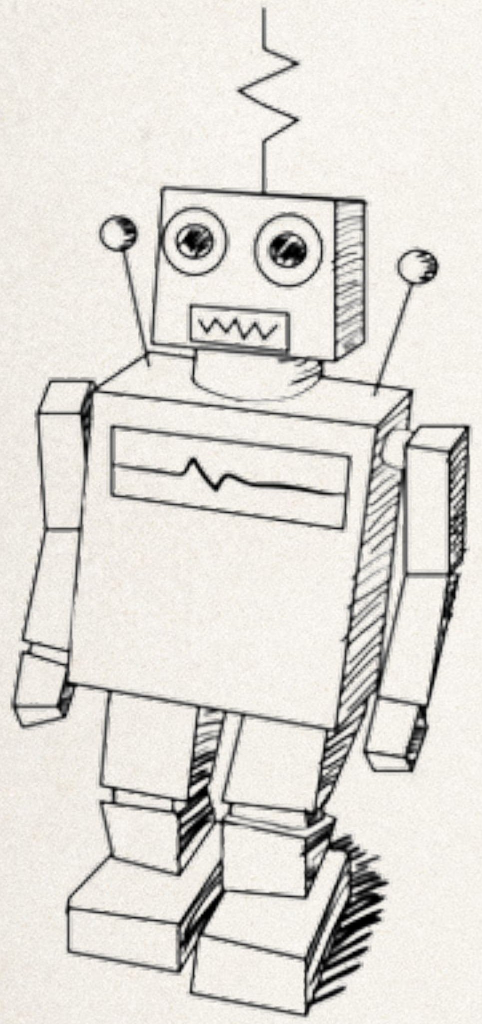
Mechanised causal graphs

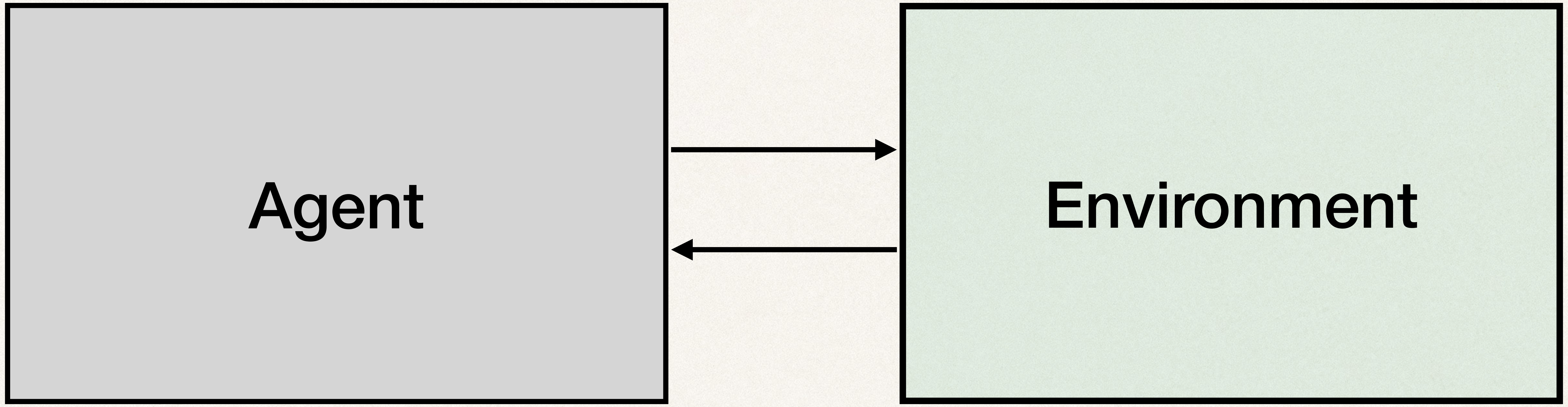
- Augment a graph reprinting decisions with "mechanisms"
- Mechanisms are parameters of "objective" variables
- Dependencies on mechanisms generally "reverse" causal chain (they need <u>interventional data</u>)



(a) Gridworld (b) Game graph, $\mathcal{G}$ (c) Mech. causal graph, $\mathcal{C}$

Kenton, Z., Kumar, R., Farquhar, S., Richens, J., MacDermott, M., & Everitt, T. (2022). Discovering Agents. arXiv preprint arXiv:2208.08345.





Mechanised causal graphs: “agents are systems that would adapt their policy if their actions influenced the world in a different way”

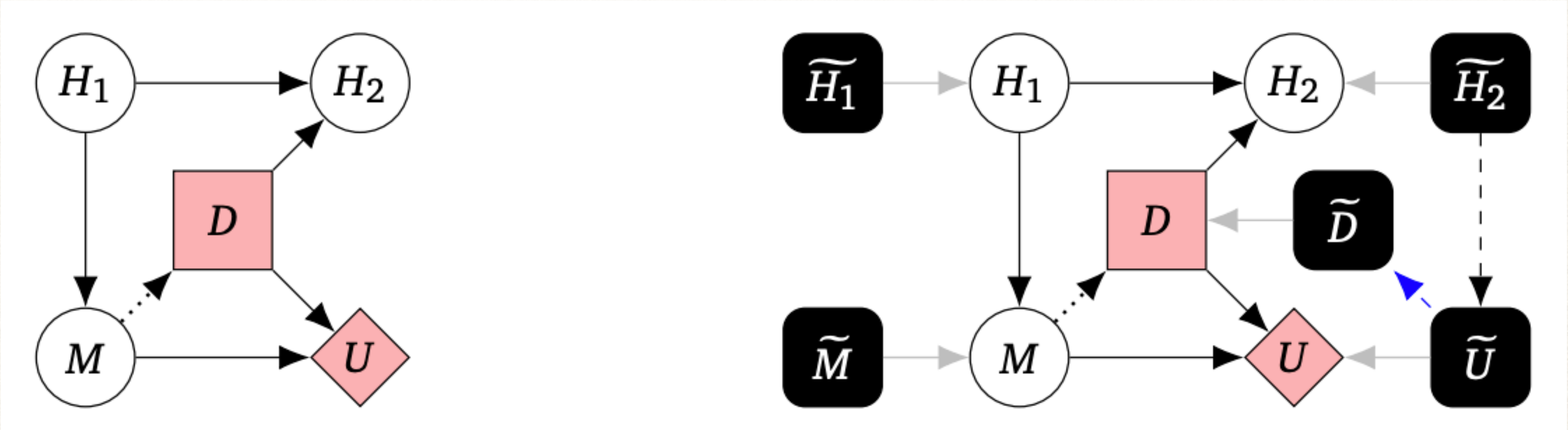
Mechanised causal graphs - pros and cons

Strengths

- Fully "Pearlian" description of agency
- Can pick up some interesting features of agency boundaries (what needs to be included and change given a certain mechanism, e.g., learning in RL)

Limitations

- Variables are chosen a-priori
- Interventions in the real world are difficult to obtain
- Are "soft" interventions (repeated observations) required for agency?



Kenton, Z., Kumar, R., Farquhar, S., Richens, J., MacDermott, M., & Everitt, T. (2022). Discovering Agents. arXiv preprint arXiv:2208.08345.

Causality-based methods

Advantages

- They take into account intrinsic notions of agency
- Pearl's causality is arguably the best account of causality we have
- Observer-independent

Disadvantages

- Is action necessarily related to causality?
- Choice of variables in causal models is somewhat subjective (cf. micro-state of a system)
- Actually, not really clear if these methods are observer-independent..

Relational methods

Agency with respect to something

Relational methods

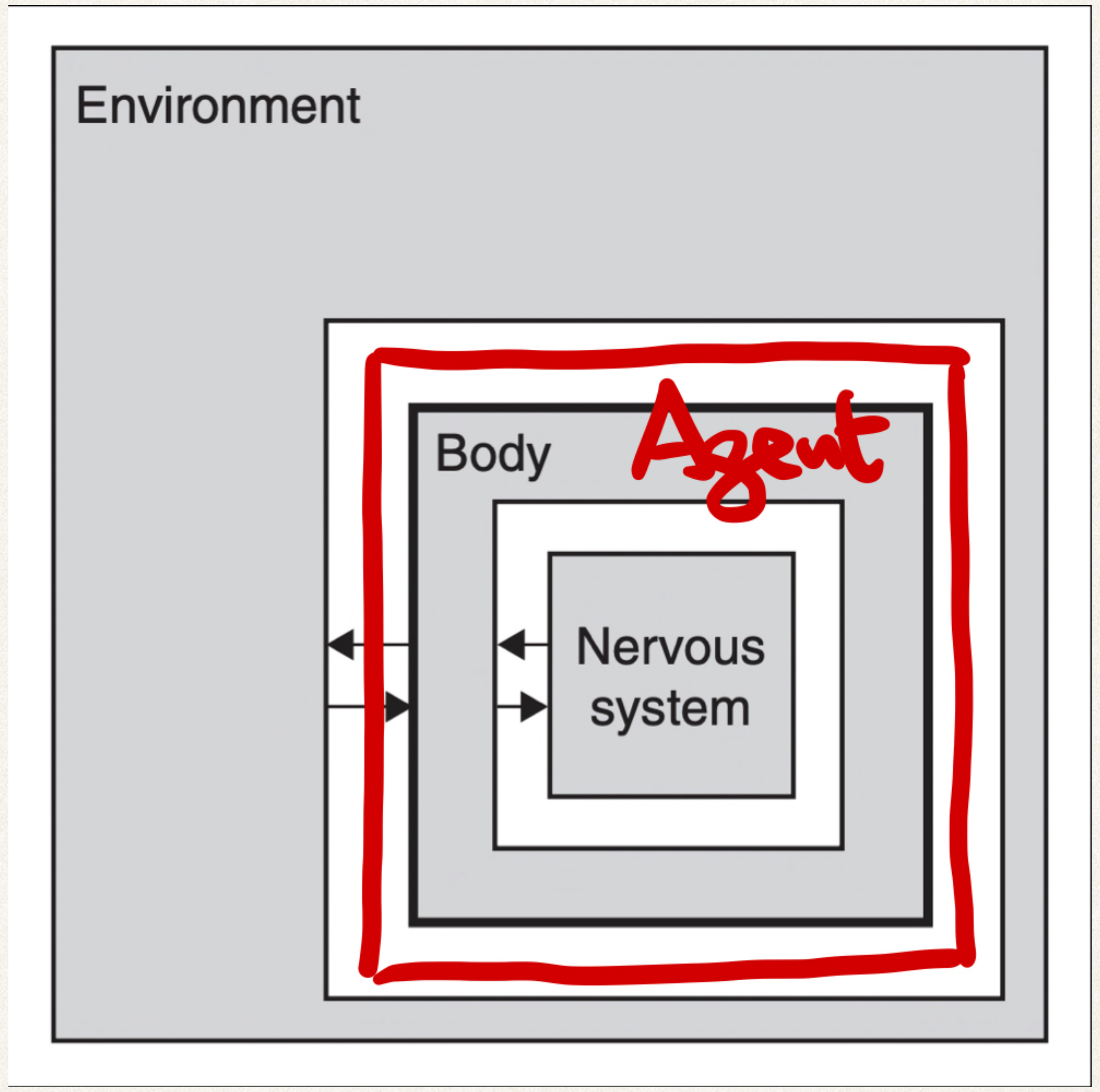
- **Main idea**: agency is in the way a system relates to other systems (environment, observer, other systems)
- **Inspiration**: cybernetics, Beer’s work
- **Tools**: dynamical systems theory, systems theory, category theory, etc.

Examples:

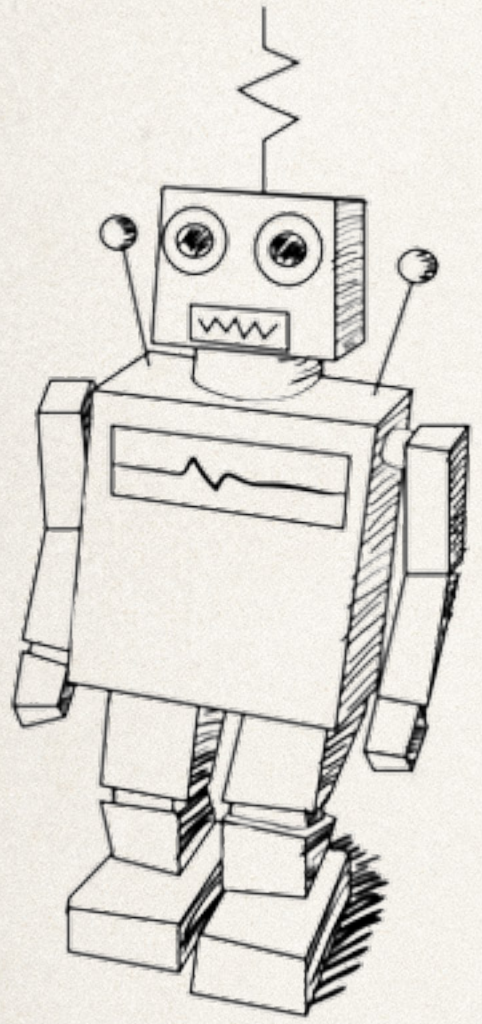
- Dynamical systems for agent-environment interactions
- Bayesian interpretation map
- Categorical agent-environment interactions

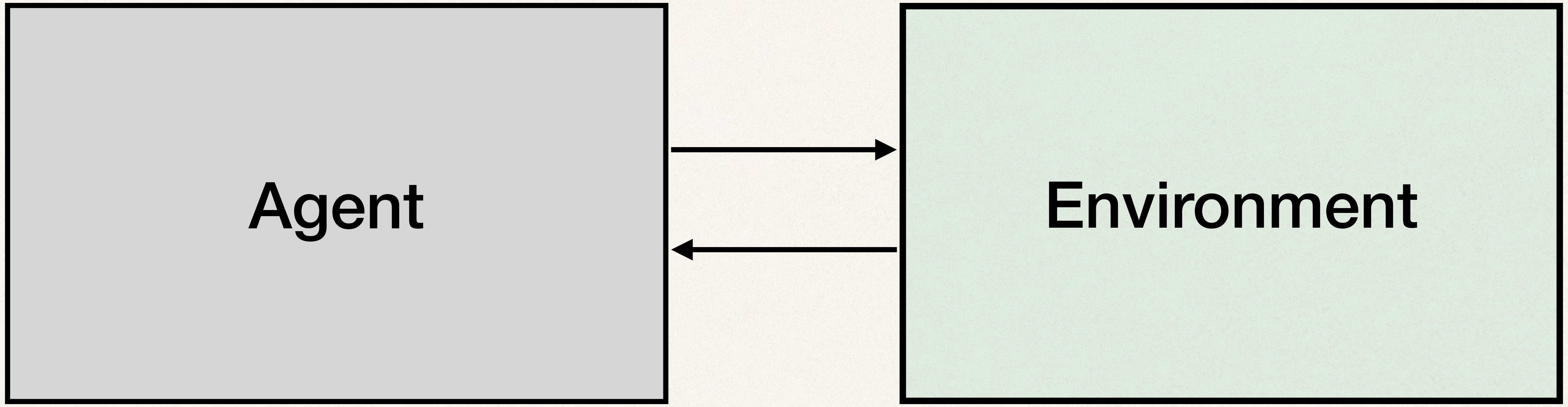
Dynamical systems for agent-environment interactions

- Take 3 dynamical systems and couple them
- Define adaptive fit: "An ~~animal~~ [agent] is adaptively fit to an environment only so long as it maintains its trajectory within this constraint volume [= the agent's existence in state-space] despite the perturbations that it receives from its environment."



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





Beer's DSs: agents are systems that adaptively fit their environment

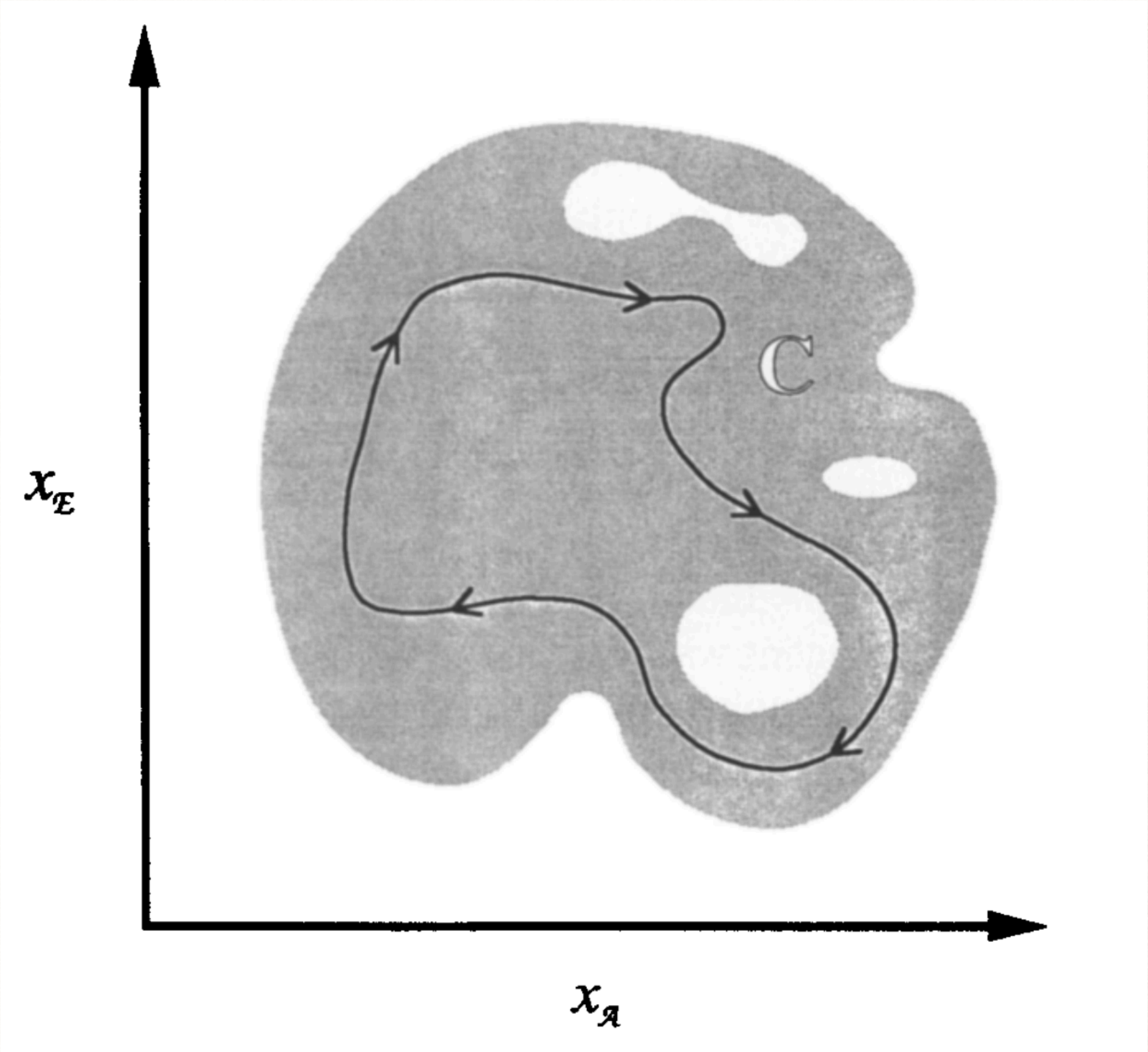
DSs for agent-environment interactions - pros and cons

Strengths

- Few assumptions
- Environment plays a role (adaptive fit)

Limitations

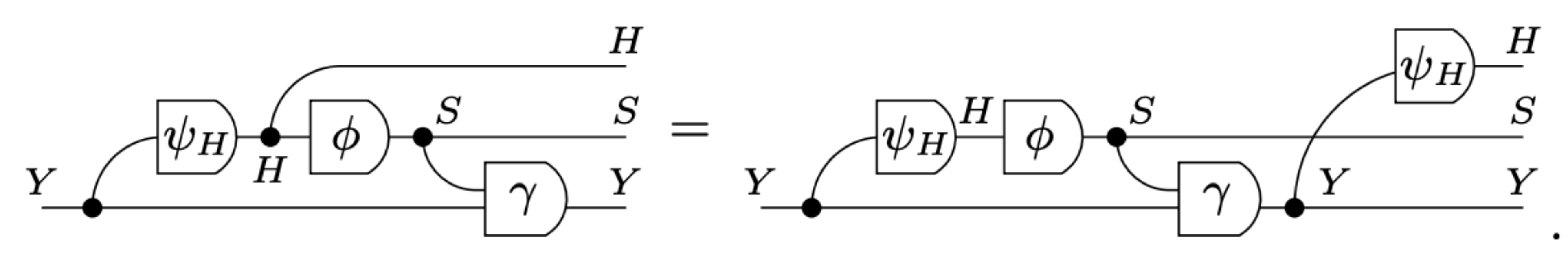
- "Adaptive fit" is never formalised
- Systems other than agents might show "adaptive fit"?
- Agents are assumed?

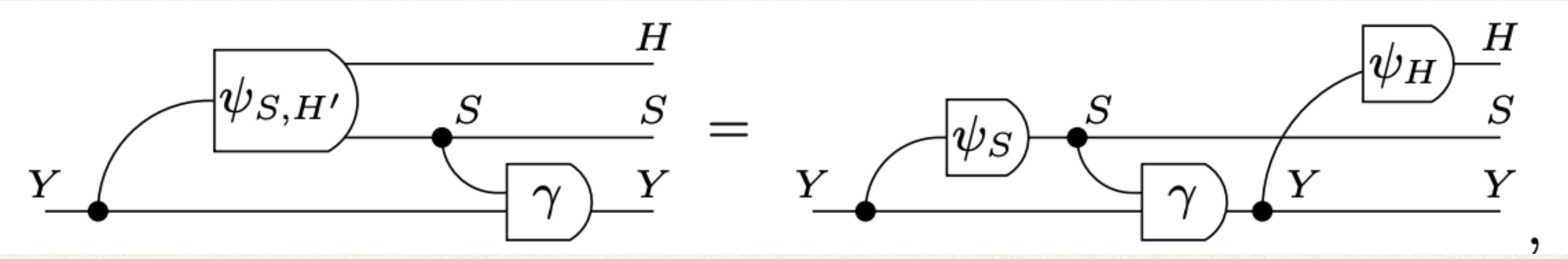


Beer, R. D. (1995). A dynamical systems perspective on agent-environment interaction. Artificial intelligence, 72(1-2), 173-215.

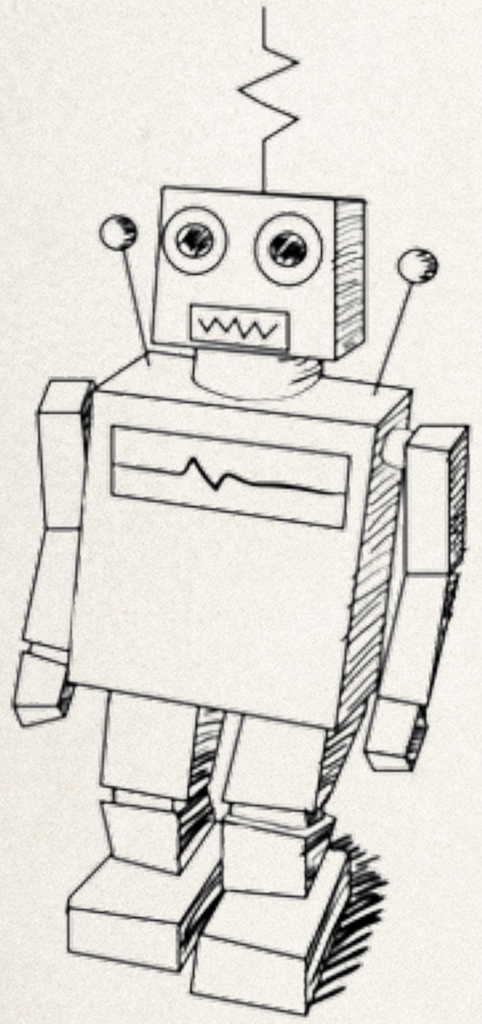
Bayesian interpretation map

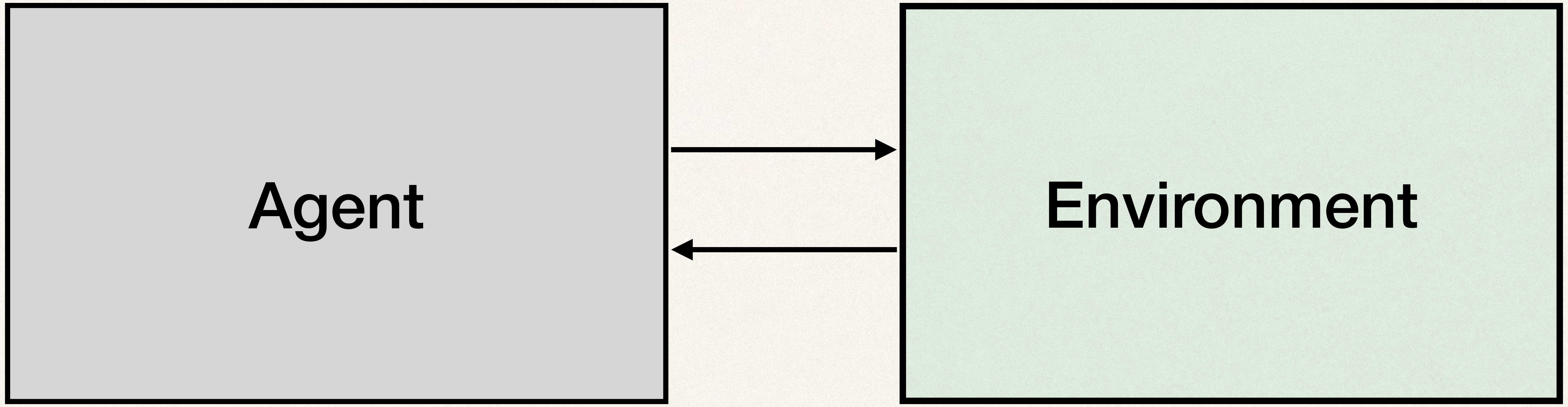
- Take a dynamical system and build an "interpretation map"
- An interpretation map is a function that maps states of the system to "beliefs" as probability measures
- Perform Bayesian inference / filtering on these probabilities and check if this process is consistent with dynamical system evolution





Virgo, N., Biehl, M., & McGregor, S. (2022). Interpreting Dynamical Systems as Bayesian Reasoners. In Machine Learning and Principles and Practice of Knowledge Discovery in Databases



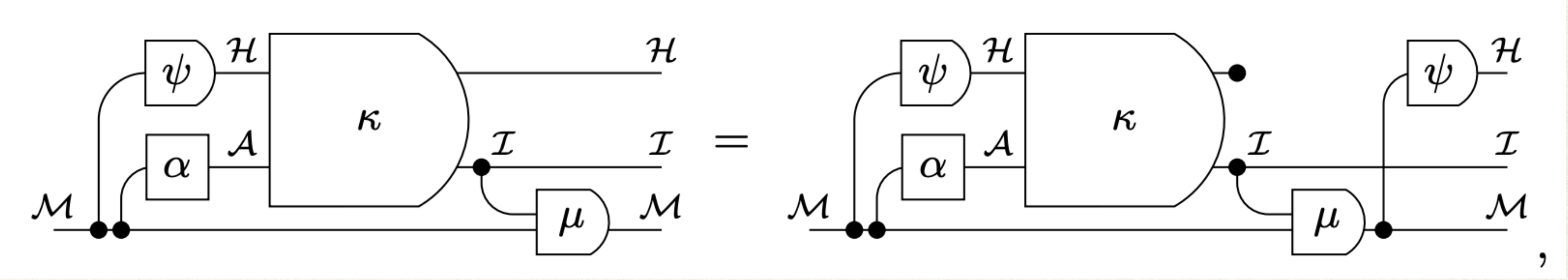


Bayesian interpretation map: if a system can be interpreted as performing Bayesian inference with some arbitrary model, then it's an agent

Bayesian interpretation map - pros and cons

Strengths

- Very general (category theory)
- Doesn't need to assume accurate knowledge of the environment (can fail causality test and still be an agent)



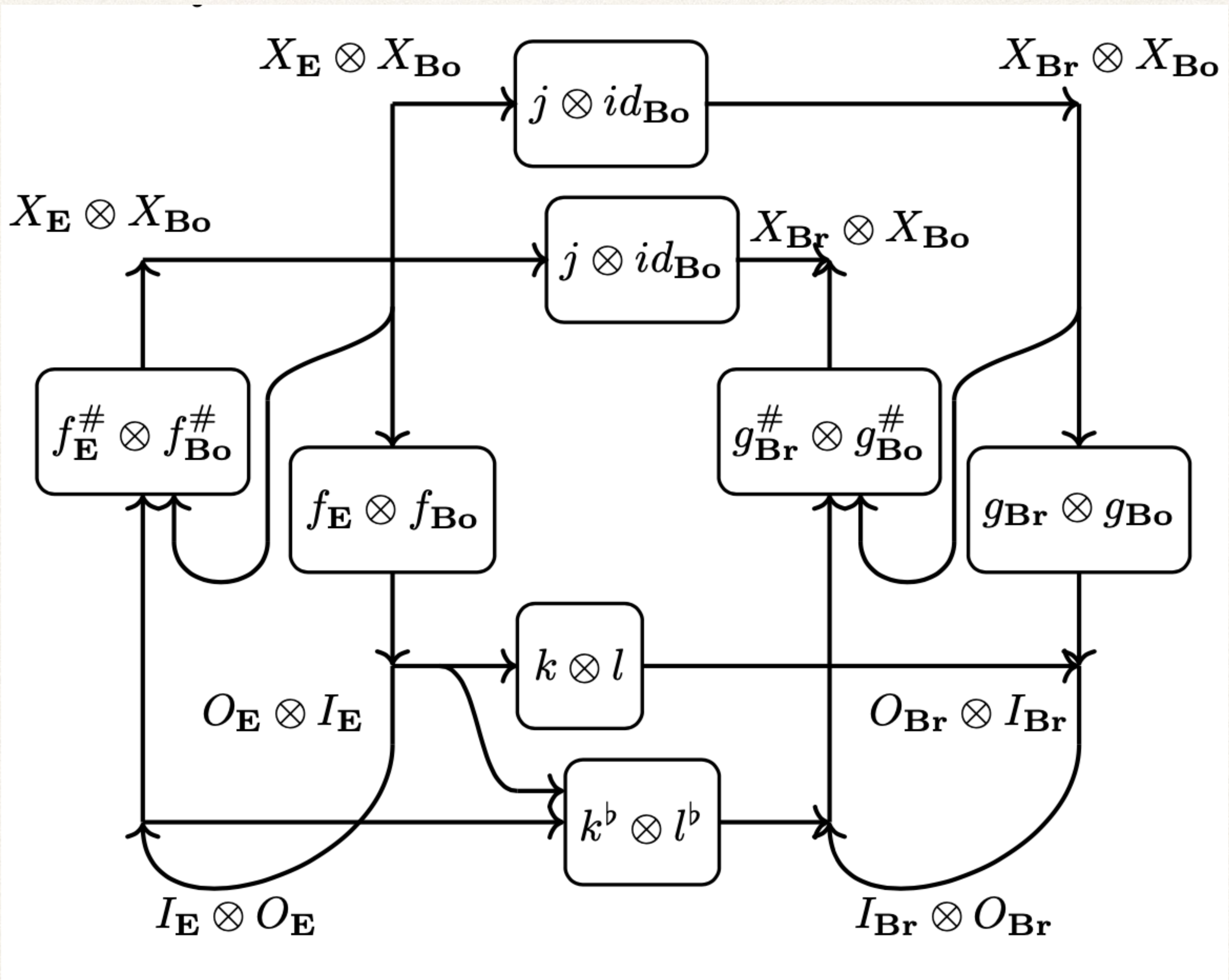
Limitations

- Doesn't consider real environment
- Including systems other that agents? (Thermostats, controllers, etc.)

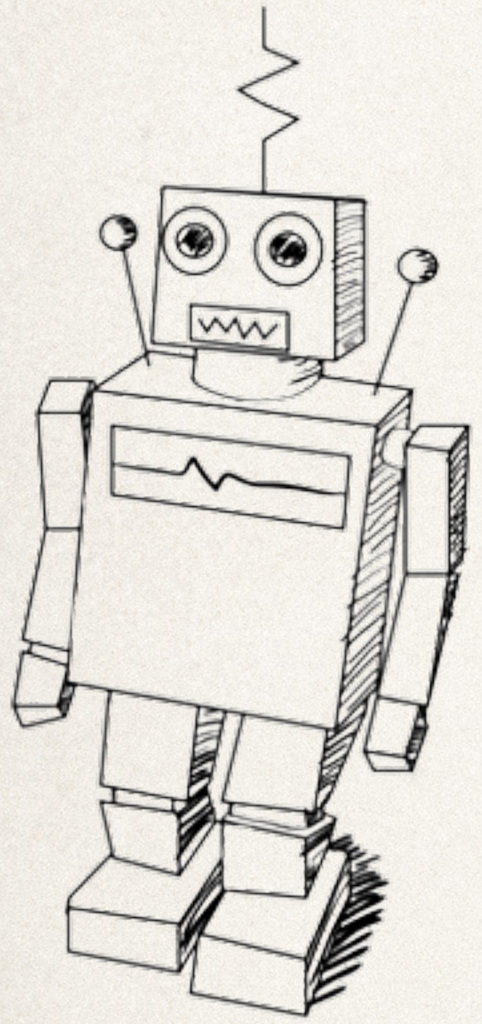
Biehl, M., & Virgo, N. (2022). Interpreting systems as solving POMDPs: a step towards a formal understanding of agency. *arXiv preprint arXiv:2209.01619*.

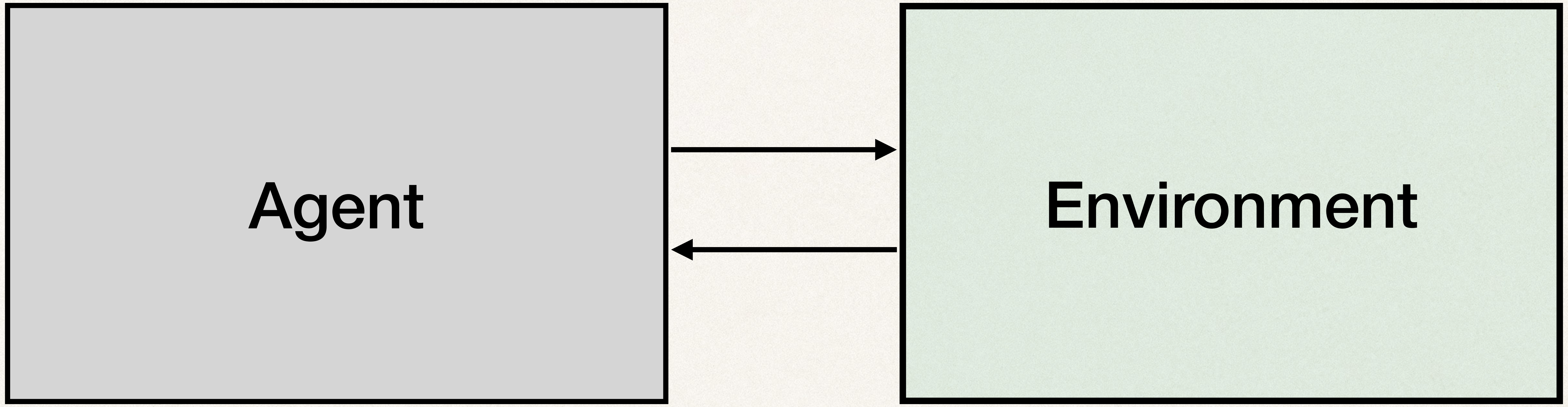
Categorical agent-environment interactions

- Take Beer's account, study physical connections among brain-body-environment
- Formalise "adaptive fit" using internal model principle from control theory (cf. law of requisite variety/good regulator theorem in cybernetics)
- Study "higher order" functional relations between brain and environment (via the body)



Baltieri, M. In progress.





Categorical agent-environment interactions: if there is a map between brain and environment while brain and environment are both physically connected to the body, we have a proto-agent

Categorical agent-environment interactions - pros and cons

Strengths

- Very general (category theory)
- It generalises FEP, Beer's proposal, probably other information-based ideas, partially Biehl + Virgo

Limitations

- Including systems other that agents?
- No account of causality
- Limited to "real" environments unlike Biehl + Virgo

$$\begin{tikzcd}
X_{\mathbf{E}} \arrow[r, "\alpha_{\mathbf{E}}"] \arrow[d, "j"'] & X_{\mathbf{E}} \arrow[d, "j"] \\
X_{\mathbf{Br}} \arrow[r, "\alpha_{\mathbf{Br}}"'] & X_{\mathbf{Br}}
\end{tikzcd}$$

Baltieri, M. In progress.

Relational methods

Advantages

- Most general domain of applications (physical vs non-physical, sets vs. graphs vs. probabilities, etc.)
- Agency as a relational property, we can in principle add other agents, observers, etc. and account for how they affect agency
- Few assumptions (due to their generality)

Disadvantages

- Too general to say something practically useful?
- Not obvious how causal claims could be considered in this class of approaches
- Hard to know if we are capturing specifically a notion of agency or something else (maybe related to it)

Conclusion

- Prediction-based methods: agency in the eye of the beholder
- Causality-based methods: agency as a property intrinsic to a system
- Relational methods: agency with respect to something

What should your definition of agency include?

(Can you define agency?)