Operationalising and mathematising agency

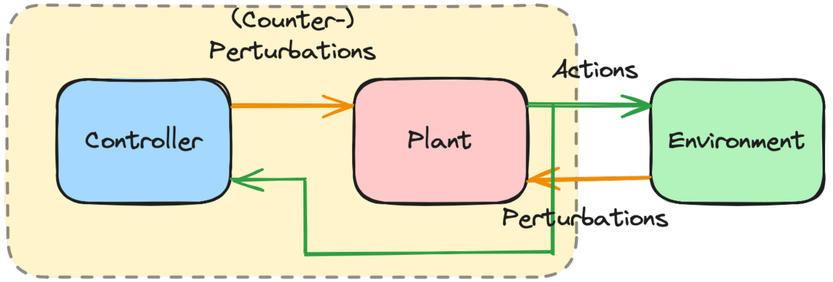
Challenges and opportunities

Manuel Baltieri - Milan, 20th February 2026

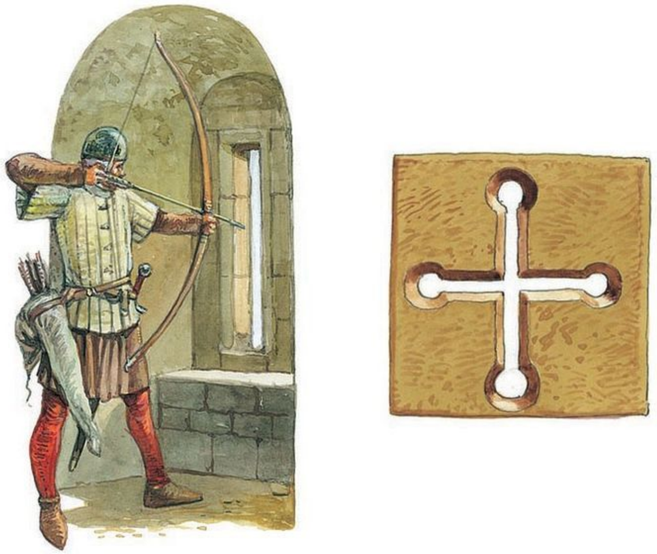
ARAYA

Contents

Artificial life as abstraction



- **A change in perspective**
  - From simulating life…
  - … to life within simulations
- **Abstracting evolution**
- **Abstracting self-sustaining systems (= agents)**



Disclaimer: I'm interested in **formalisations**, plenty of topics not discussed because of this

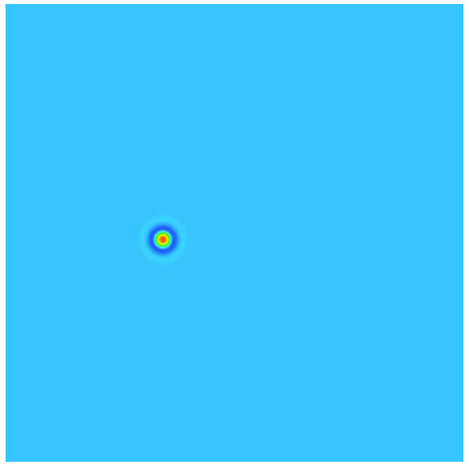
Defining life

NASA's working definition of life

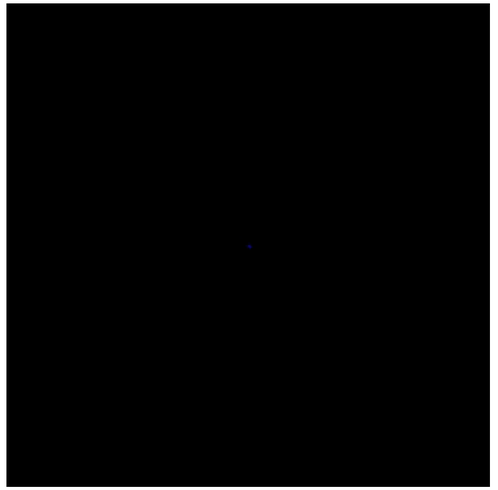
"A self-sustaining chemical system capable of undergoing Darwinian evolution."

Simulating (features of) life as we know it

"A self-sustaining chemical system capable of undergoing Darwinian evolution."



Gray-Scott model of reaction diffusion
https://www.youtube.com/watch?v=F5oKgVZ6bTk&ab_channel=TimHutton



Avida
https://www.youtube.com/shorts/ripHOtOG4TE

“Comparative biology”

“[W]e badly need a comparative biology. So far, we have been able to study only one evolving system, and we cannot wait for interstellar flight to provide us with a second. If we want to discover generalizations about evolving systems, we have to look at artificial ones.”

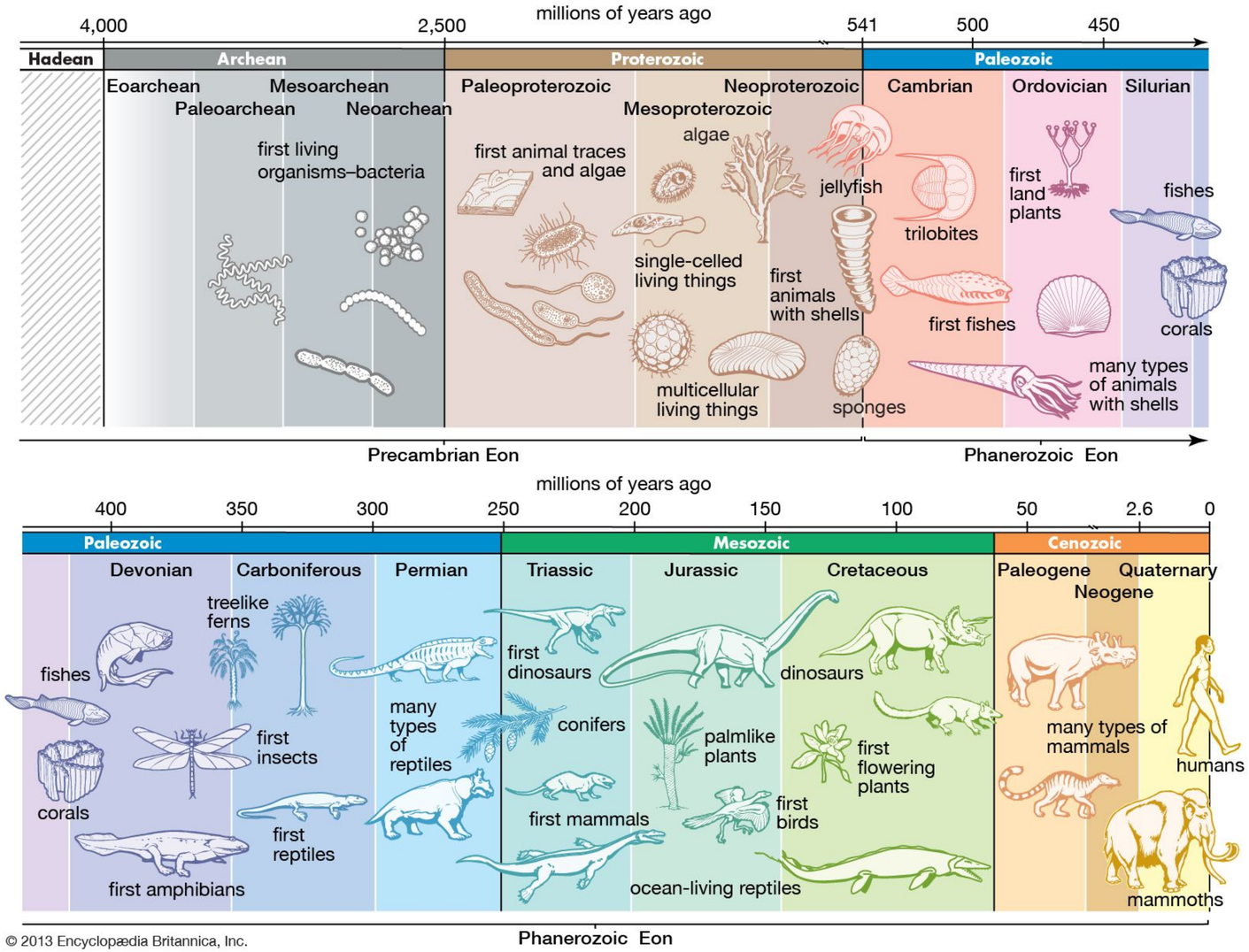
Maynard-Smith (1992)

Artificial Life: Life within a simulation

Lenia
https://www.youtube.com/watch?v=HT49wpyux-k&ab_channel=BertChan

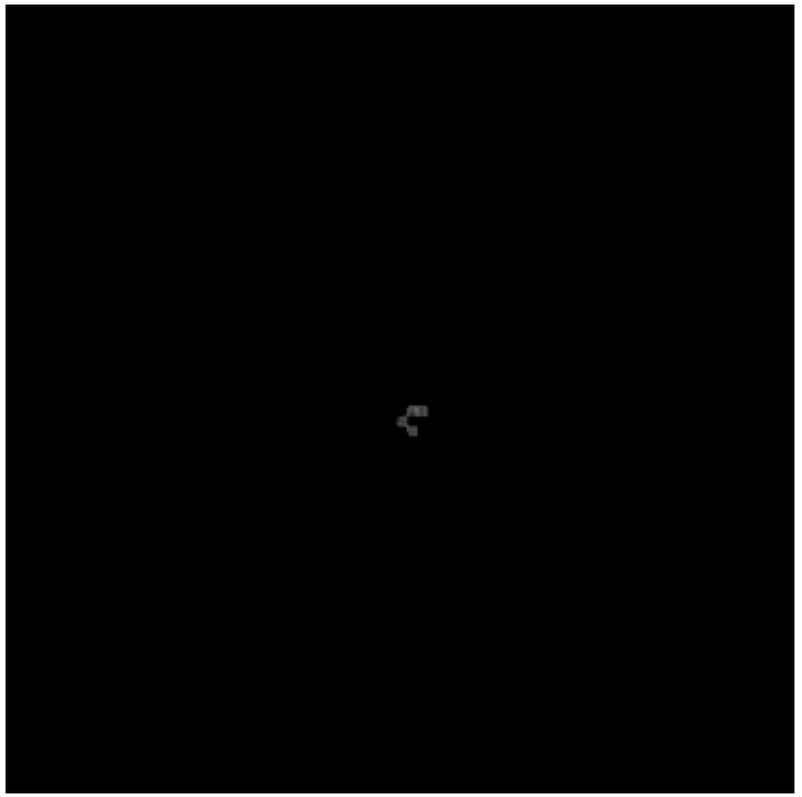
Abstracting evolution

Darwinian evolution



Digital evolution

Tierra, Avida, etc.

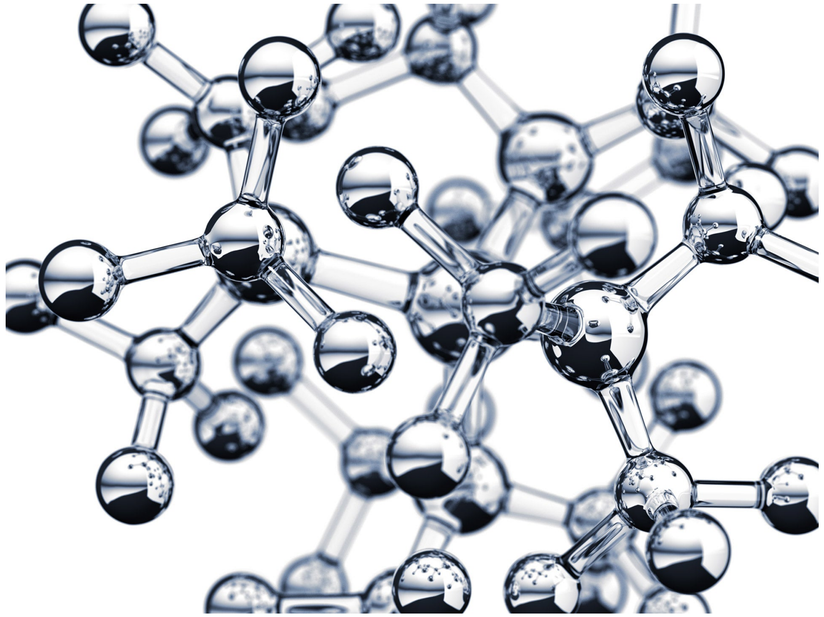


Avida
https://www.youtube.com/shorts/apANCQmPmL0

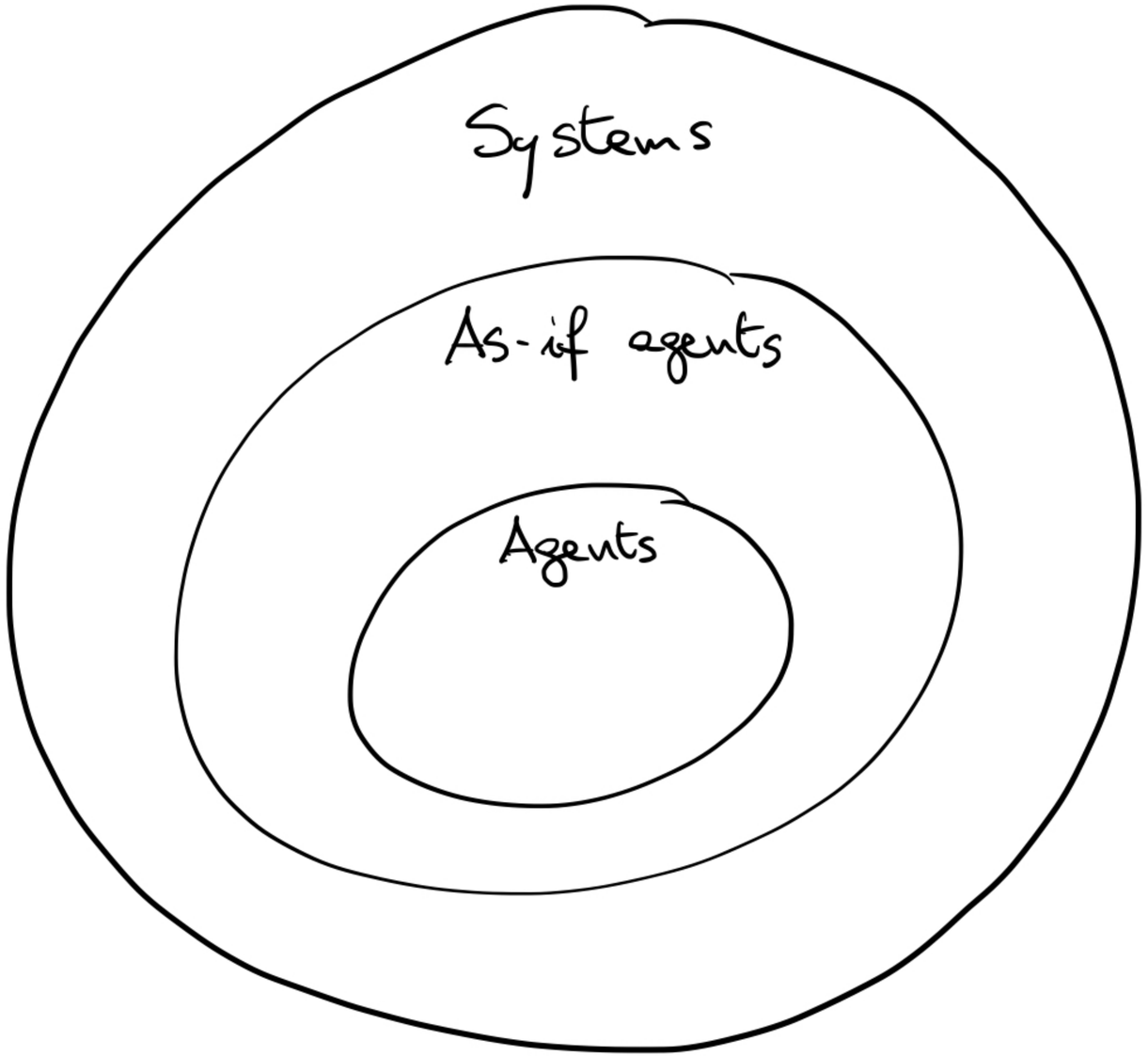
Abstracting self-sustaining systems

Agents - A stance

Intentional stance




#thedrummingpuppet
Mr Peewee
The Drumming
Puppet


Systems
As-if agents
Agents

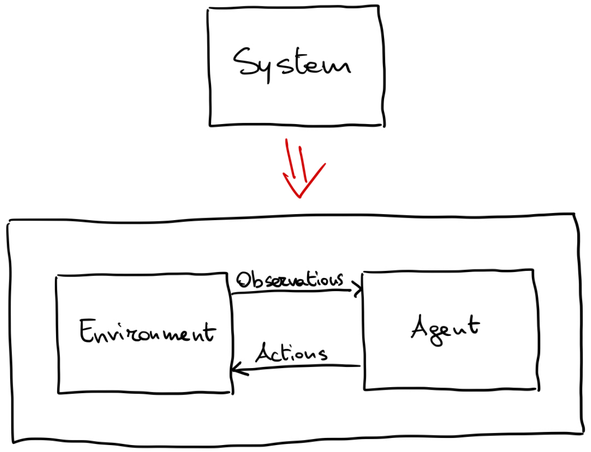
Agents - A conceptual grounding

Goal-directed systems acting in an environment

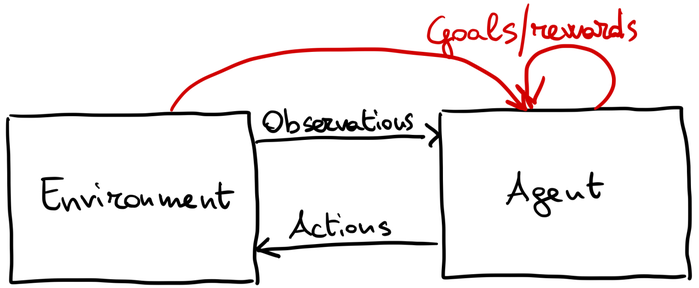
Following Barandiaran et al. (2009), we are interested in agents as systems with:

- Individuality
- Goal-directedness/normativity
- Asymmetry

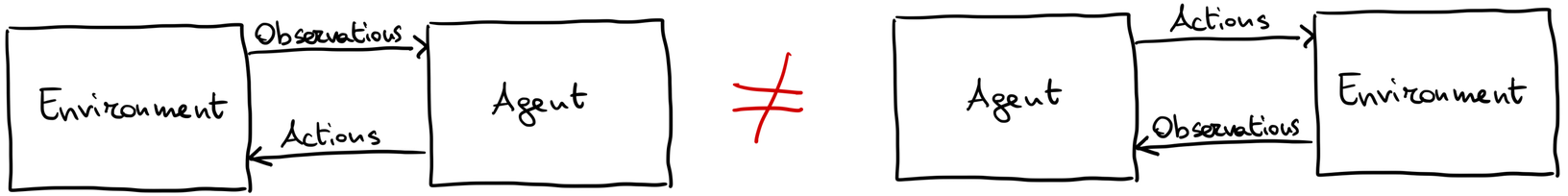
Individuality



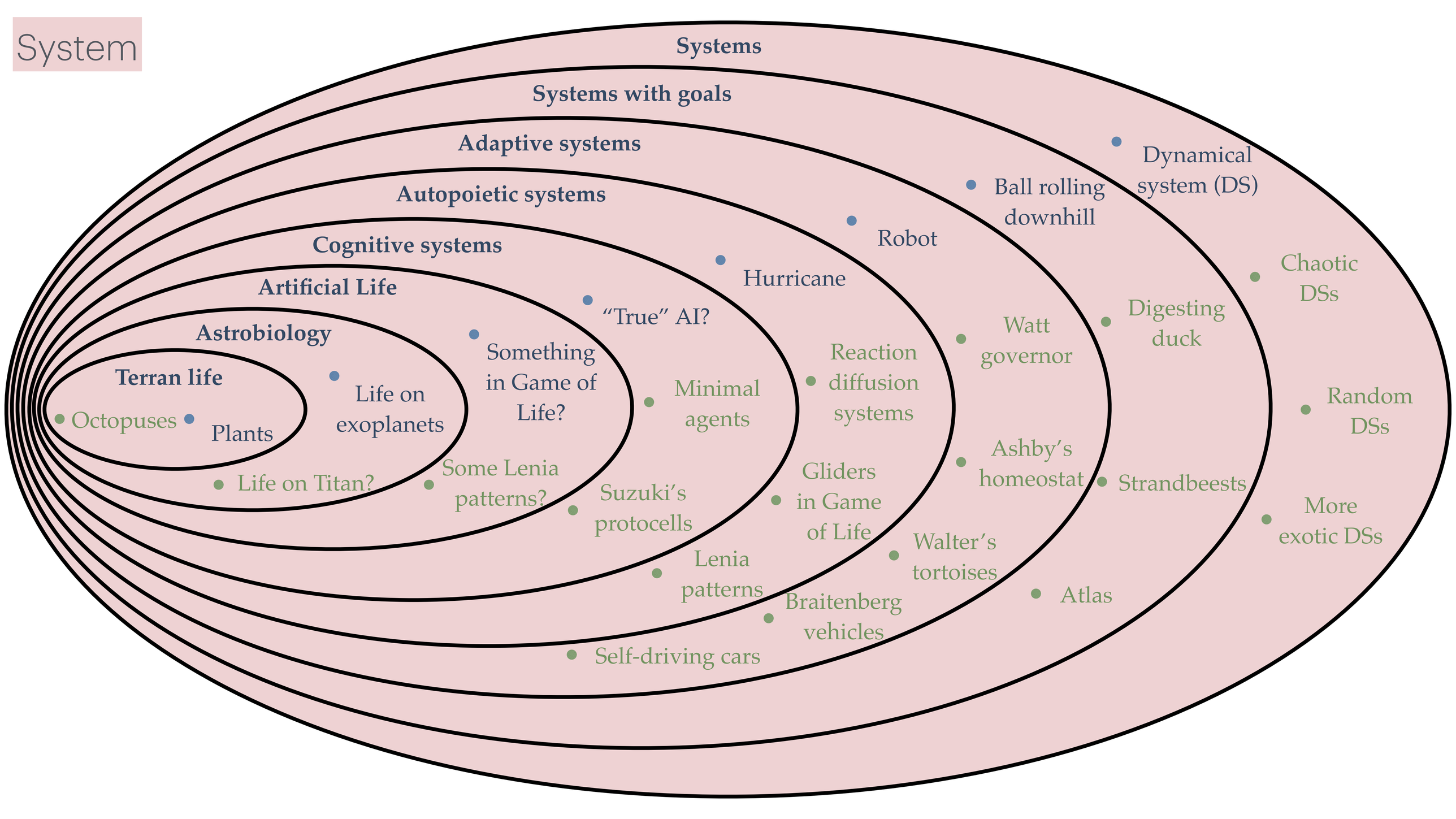
Goal-directedness



Asymmetry



System

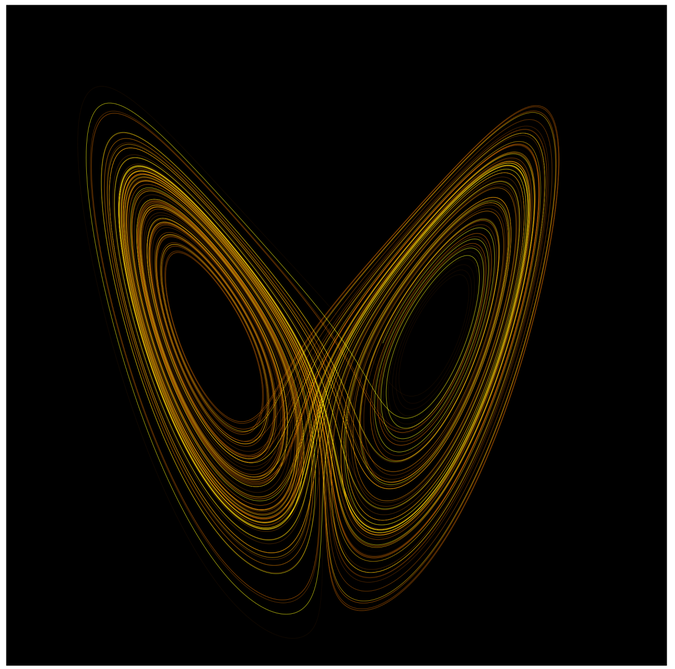

Systems
Systems with goals
Adaptive systems
Autopoietic systems
Cognitive systems
Artificial Life
Astrobiology
Terran life
Octopuses
Plants
Life on exoplanets
Something in Game of Life?
"True" AI?
Hurricane
Robot
Ball rolling downhill
Dynamical system (DS)
Chaotic DSs
Digesting duck
Watt governor
Reaction diffusion systems
Minimal agents
Random DSs
Ashby's homeostat
Strandbeests
More exotic DSs
Gliders in Game of Life
Walter's tortoises
Atlas
Braitenberg vehicles
Lenia patterns
Suzuki's protocells
Some Lenia patterns?
Life on Titan?
Self-driving cars

Dynamical systems theory

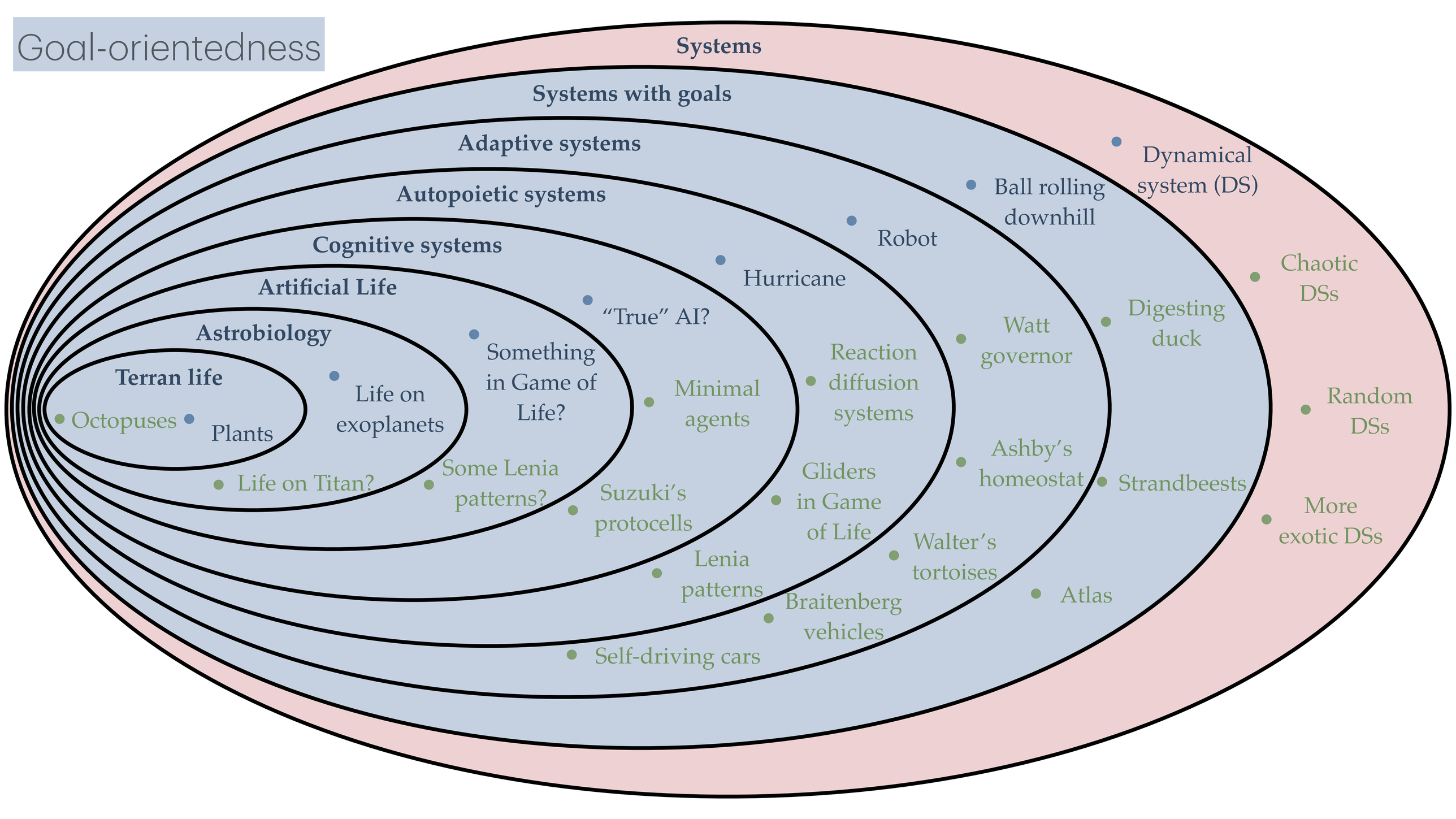
Definition.

A (closed discrete deterministic) dynamical system is a pair $(X, f : X \to X)$ where X is a set and $f : X \to X$ is a state-transition map.

Autonomous, open, continuous, non-deterministic, ... systems require some more machinery, but same idea.



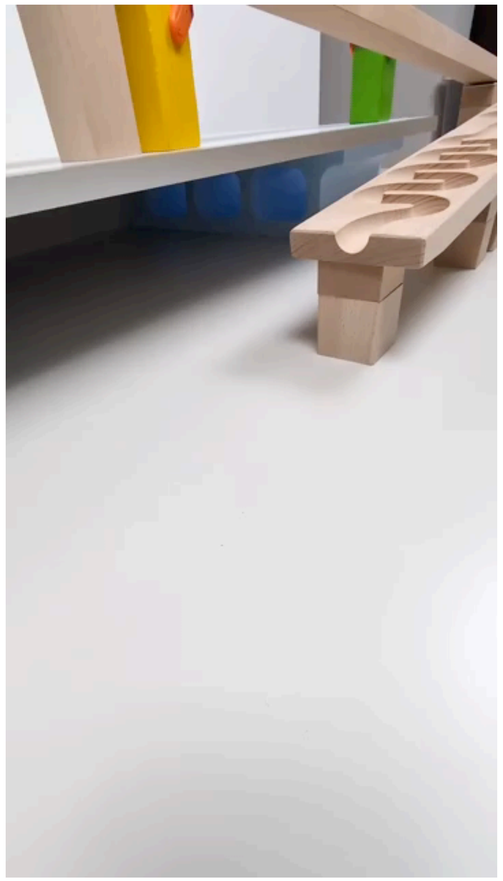
Goal-orientedness



Variational principles

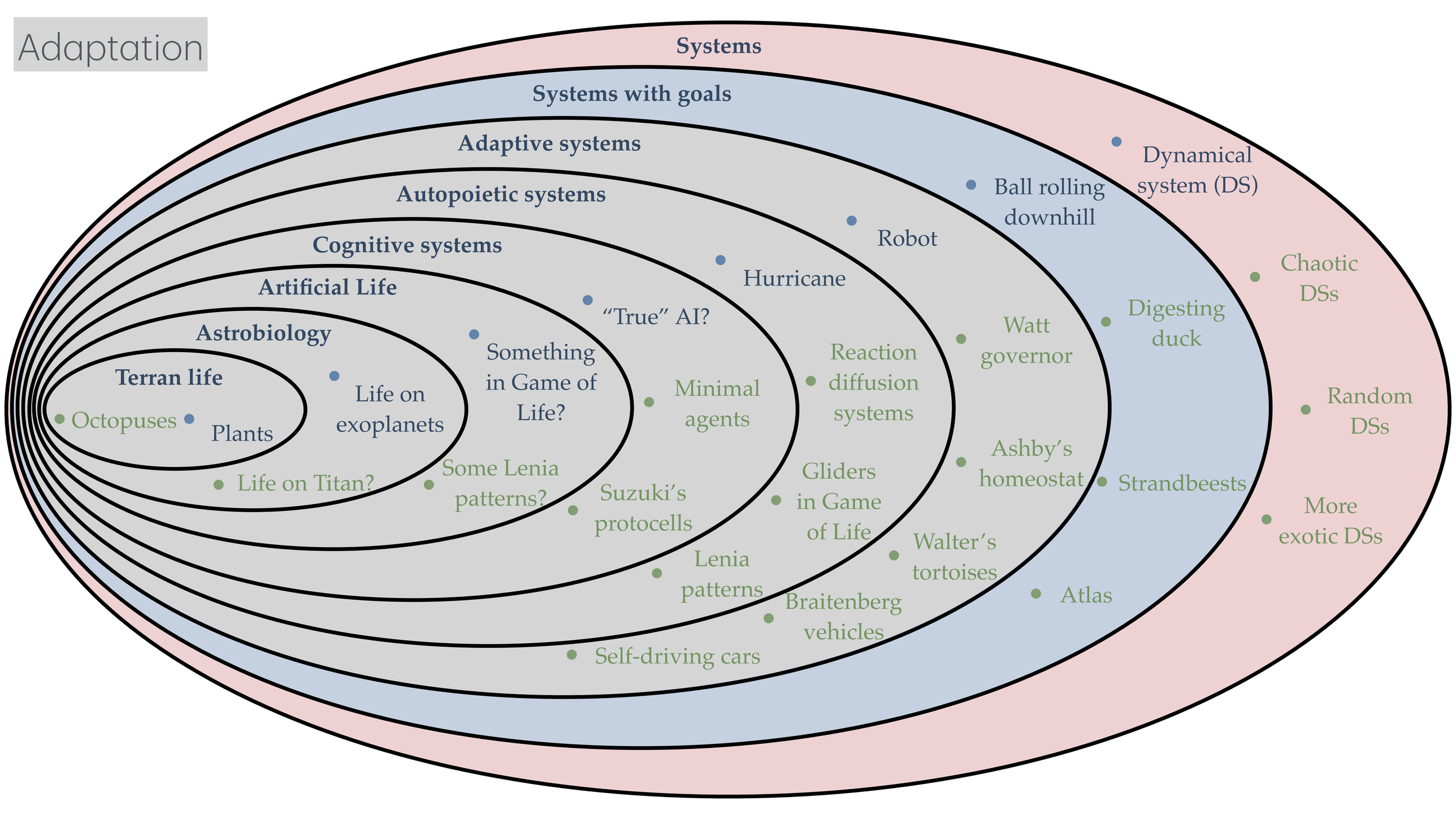
- Optimise function(al) to determine dynamics/behaviour of a system
- Goals as optimisation towards final state
- E.g. Hamilton's principle

$$\frac{\delta \mathcal{S}}{\delta \mathbf{q}(t)} = 0 \qquad \mathcal{S}[\mathbf{q}] := \int_{t_1}^{t_2} L(\mathbf{q}(t), \dot{\mathbf{q}}(t), t) dt$$



@MarbleASMRace
https://www.youtube.com/shorts/w5Fom-pdwNA

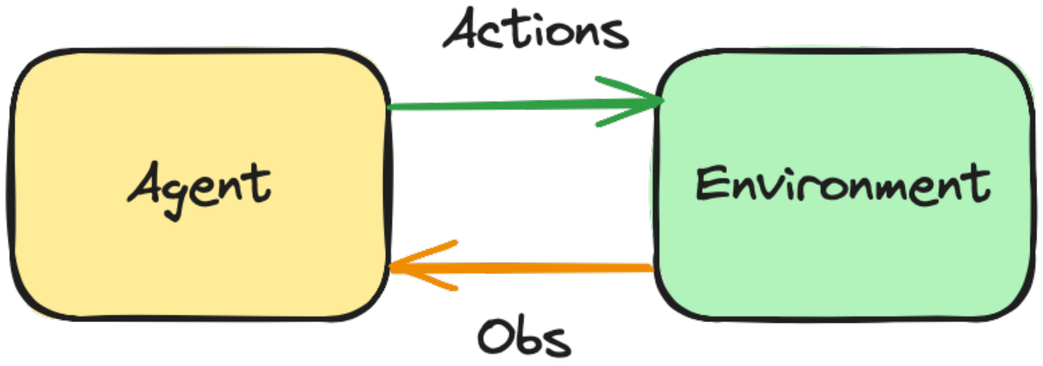
Adaptation


Systems
Systems with goals
Adaptive systems
Autopoietic systems
Cognitive systems
Artificial Life
Astrobiology
Terran life
Octopuses
Plants
Life on exoplanets
Life on Titan?
Something in Game of Life?
Some Lenia patterns?
"True" AI?
Minimal agents
Suzuki's protocells
Hurricane
Reaction diffusion systems
Gliders in Game of Life
Lenia patterns
Robot
Watt governor
Ashby's homeostat
Walter's tortoises
Braitenberg vehicles
Self-driving cars
Ball rolling downhill
Digesting duck
Strandbeests
Atlas
Dynamical system (DS)
Chaotic DSs
Random DSs
More exotic DSs

Optimal control theory

Optimal control/reinforcement learning

- "Physics with inputs", or rather *action policies*
- Optimise function(al) to determine dynamics/behaviour of a system, **but**
- **Systems have inputs**
- Inputs, action policies, to a physical system that can be "chosen" by a *controller/agent*



Homeostasis - Feedback

Heterostasis, homeodynamics, homeokinesis, homeorhesis, etc.

IEEE CSS

Received 4 October 2022; accepted 29 January 2023. Date of publication 10 February 2023;
date of current version 9 March 2023. Recommended by Senior Editor Prof. Wei Zhang.

Digital Object Identifier 10.1109/OJCSYS.2023.3244089

The Internal Model Principle for Biomolecular Control Theory

ANKIT GUPTA AND MUSTAFA KHAMMASH (Fellow, IEEE)

(Position Paper)

Department of Biosystems Science and Engineering, ETH-Zürich, 4058 Basel, Switzerland

CORRESPONDING AUTHOR: MUSTAFA KHAMMASH (e-mail: mustafa.khammash@bsse.ethz.ch)

ABSTRACT The well-known Internal Model Principle (IMP) is a cornerstone of modern control theory. It stipulates the necessary conditions for asymptotic robustness of disturbance-prone dynamical systems by asserting that such a system must embed a subsystem in a feedback loop, and this subsystem must be able to reduplicate the dynamic disturbance using only the regulated variable as the input. The insights provided by IMP can help in both designing suitable controllers and also in analysing the regulatory mechanisms in complex systems. So far the application of IMP in biology has been case-specific and ad hoc, primarily due to the lack of generic versions of the IMP for biomolecular reaction networks that model biological processes. In this short article we highlight the need for an IMP in biology and discuss a recently developed version of it for biomolecular networks that exhibit maximal Robust Perfect Adaptation (maxRPA) by being robust to the maximum number of disturbance sources.

Robust perfect adaptation in bacterial chemotaxis through integral feedback control

Tau-Mu Yi*†, Yun Huang†‡, Melvin I. Simon*§, and John Doyle‡

*Division of Biology 147-75 and ‡Department of Control and Dynamical Systems 107-81, California Institute of Technology, Pasadena, CA 91125

Contributed by Melvin I. Simon, February 7, 2000

Integral feedback control is a basic engineering strategy for ensuring that the output of a system robustly tracks its desired value independent of noise or variations in system parameters. In biological systems, it is common for the response to an extracellular stimulus to return to its prestimulus value even in the continued presence of the signal—a process termed adaptation or desensitization. Barkai, Alon, Surette, and Leibler have provided both theoretical and experimental evidence that the precision of adaptation in bacterial chemotaxis is robust to dramatic changes in the levels and kinetic rate constants of the constituent proteins in this signaling network [Alon, U., Surette, M. G., Barkai, N. & Leibler, S. (1998) *Nature (London)* 397, 168–171]. Here we propose that the robustness of perfect adaptation is the result of this system possessing the property of integral feedback control. Using techniques from control and dynamical systems theory, we demonstrate that integral control is structurally inherent in the Barkai–Leibler model and identify and characterize the key assumptions of the model. Most importantly, we argue that integral control in some form is necessary for a robust implementation of perfect adaptation. More generally, integral control may underlie the robustness of many homeostatic mechanisms.

neering systems has a large theoretical literature that began with electrical network design (14). Quantitative application of engineering robustness methods in molecular biology began with studies of biosynthetic pathways (15), although robustness of biological responses as a selective property in evolution was emphasized qualitatively even earlier (16).

In an elegant study, Barkai and Leibler investigated the robustness of perfect adaptation in bacterial chemotaxis (17). They constructed a two-state model (active or inactive) of the receptor complex (receptor + CheA + CheW); the system output, modulated by ligand binding and methylation, was the concentration of active receptor complexes. In this model, perfect adaptation was the intrinsic property of the connectivity of the signaling network and did not require specific values for the kinetic rate constants or concentrations of the constituent enzymes. Alon *et al.* elegantly provided experimental evidence for the robustness of perfect adaptation to parameter changes when they demonstrated exact adaptation even when the levels of the chemotactic proteins were varied dramatically (18). In this work, we have reexamined these findings from the perspective of robust control theory, which has allowed us to analyze in a more

ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Systems & Control Letters 50 (2003) 119–126

SYSTEMS & CONTROL LETTERS

www.elsevier.com/locate/sysconle

Adaptation and regulation with signal detection implies internal model

Eduardo D. Sontag*,[1]

Department of Mathematics, Rutgers University, New Brunswick, NJ 08903, USA

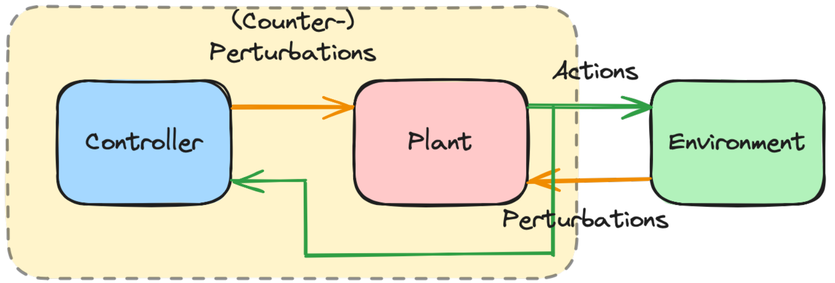
Received 14 June 2002; accepted 12 December 2002

Abstract

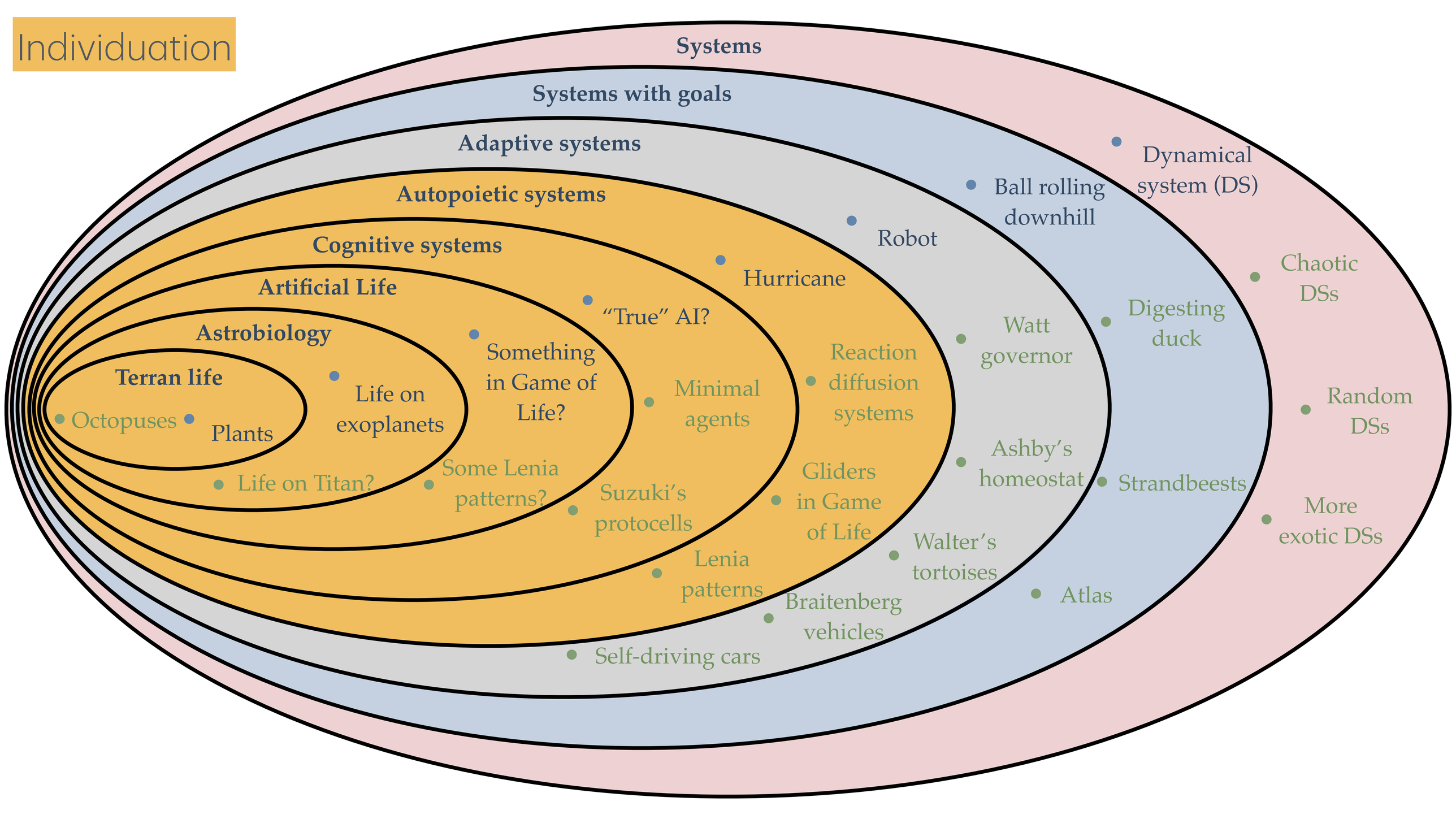
This note provides a simple result showing, under suitable technical assumptions, that if a system Σ adapts to a class of external signals $\mathscr{U}$, in the sense of regulation against disturbances or tracking signals in $\mathscr{U}$, then Σ must necessarily contain a subsystem which is capable of generating all the signals in $\mathscr{U}$. It is not assumed that regulation is robust, nor is there a prior requirement for the system to be partitioned into separate plant and controller components. Instead, one assumes that a "signal detection" property holds.

- Homeostasis implies **internal models**

- In biology "perfect adaptation", in control theory "internal model principle", etc.



Individuation



Mathematical frameworks for agency

Agents - A mathematical grounding

- Relational approaches for individuality
- Prediction-based methods for goal-directedness
- Causality-based methods for asymmetry

PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Research

Article submitted to journal

Subject Areas:
artificial life, cognitive science, artificial intelligence

Keywords:
agency, individuality, normativity, asymmetry

Author for correspondence:
Manuel Baltieri
e-mail: manuel_baltieri@araya.org

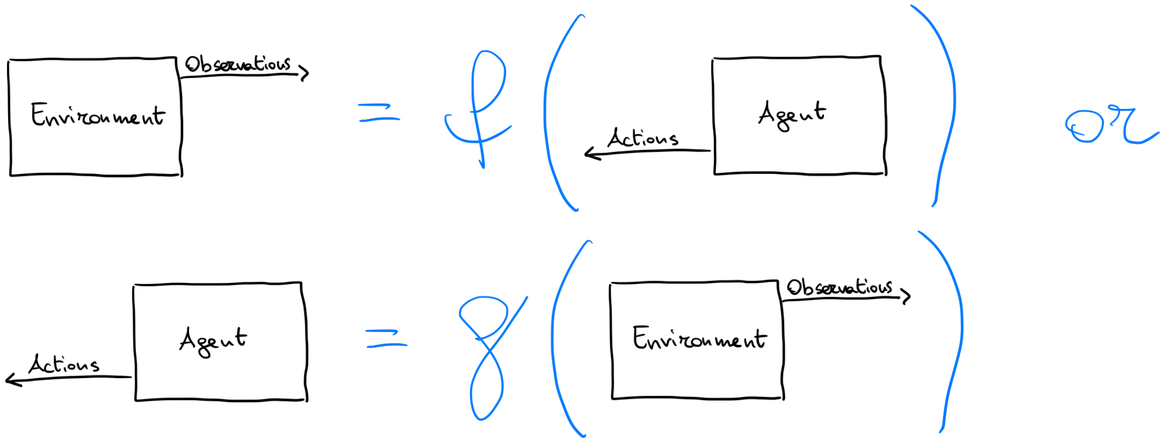
Mathematical approaches to the study of agents

Manuel Baltieri[1,2] and Keisuke Suzuki[3]

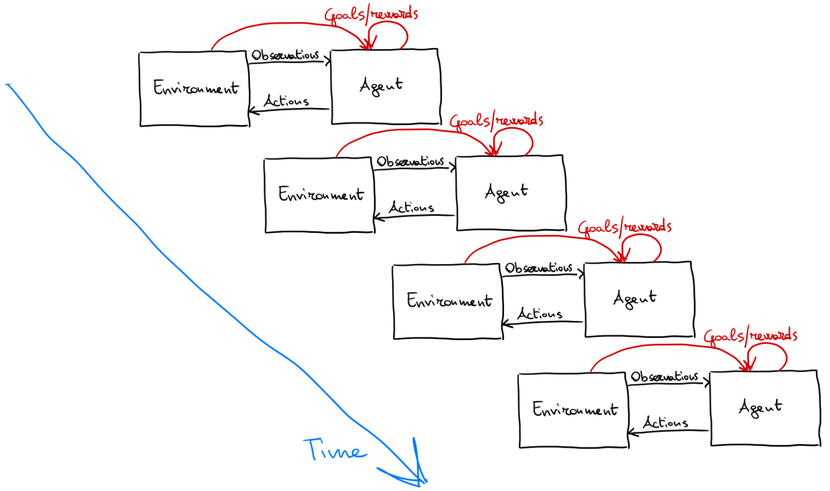
[1] Araya Inc., Tokyo, Japan
[2] School of Engineering and Informatics, University of Sussex, Brighton, UK
[3] Center for Human Nature, Artificial Intelligence, and Neuroscience (CHAIN) Hokkaido University, Sapporo, Japan

The definition of life remains one of science's most profound challenges, with contemporary approaches usually focusing on two research programmes: Darwinian evolution and self-maintenance in chemical systems. While evolution has been successfully abstracted and mathematically modelled, the concept of a self-sustaining system has so far resisted a comparable level of formalisation. This paper tackles this challenge by reframing the concept of self-sustaining system within a more abstract framework to study *agents*: goal-directed systems acting in an environment. We build on an existing conceptual framework comprising three requirements for agents: individuality, normativity (or goal-directedness), and interactional asymmetry. We then provide a systematic analysis, under a unified notation, of several mathematical approaches aiming to formalise these requirements, including the free energy principle, integrated information theory and dynamical systems. Unlike this conceptual framework, which commits to an intrinsic perspective on agency, we commit to a less ontologically committed *as-if* stance. Using this, we discuss links between identity and normativity, and a way to understand actions as if they were produced by causal interventions. Taken together, our systematic analysis clarifies the limitations of current proposals and reveals how they can work synergistically within a unified, mathematical account of agency across natural and artificial domains.

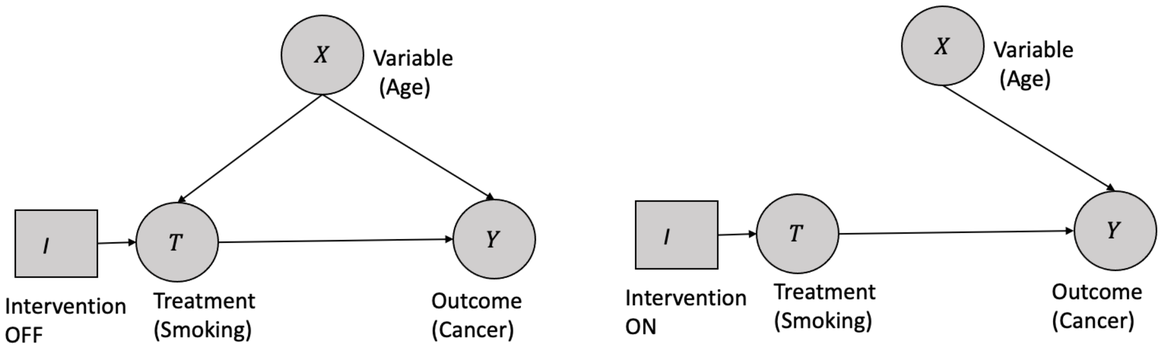
Individuality

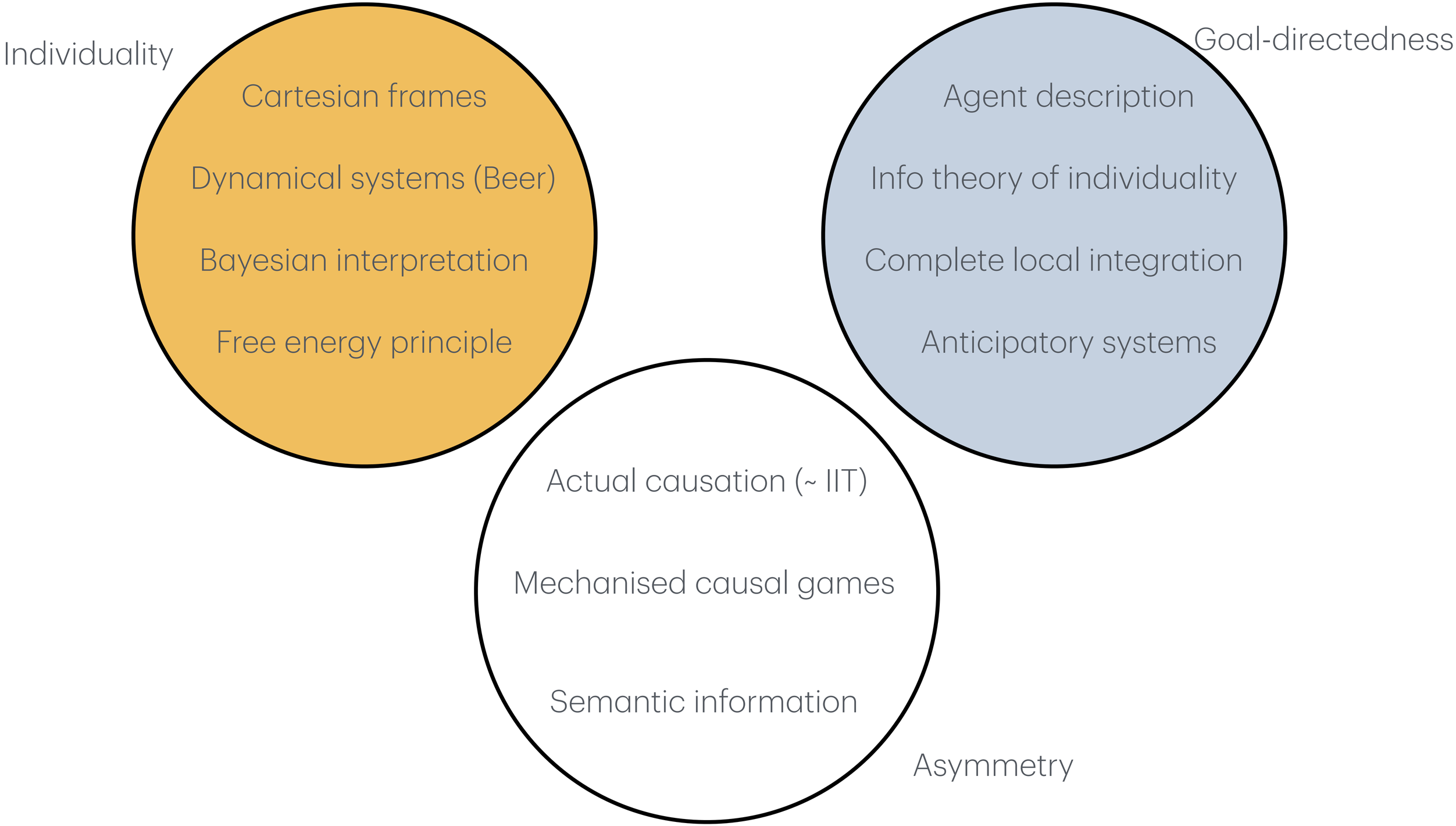


Goal-directedness



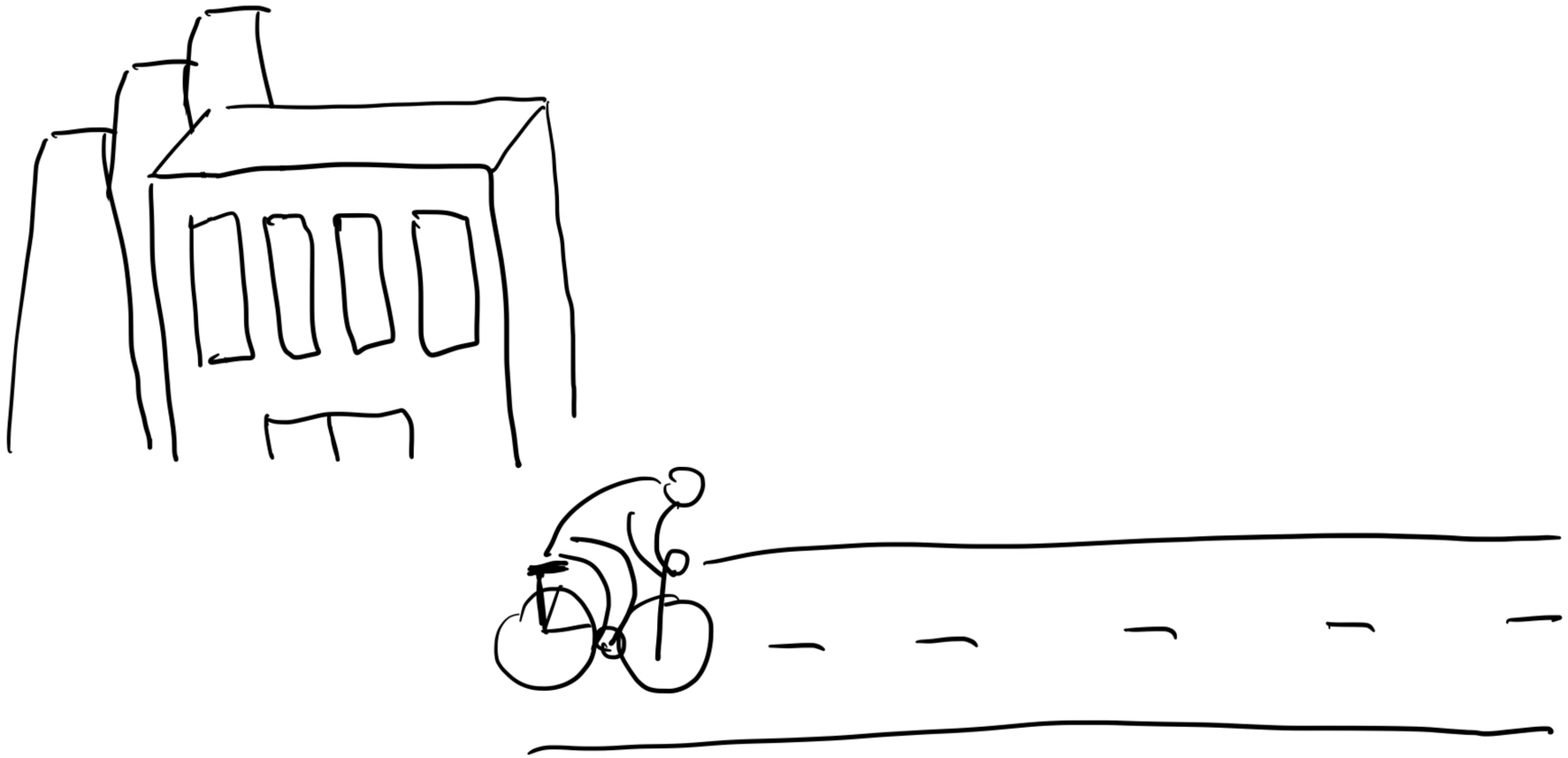
Asymmetry




Individuality
Cartesian frames
Dynamical systems (Beer)
Bayesian interpretation
Free energy principle
Goal-directedness
Agent description
Info theory of individuality
Complete local integration
Anticipatory systems
Actual causation (~ IIT)
Mechanised causal games
Semantic information
Asymmetry

Individuality <-> Goal-directedness

Drawing the boundary



Symmetry and as-if causality

Ant gardens

Acknowledgments

People: Keisuke Suzuki, Fernando Rosas, Nathaniel Virgo, Martin Biehl, Matteo Capucci, Alexander Boyd, Franz Nowak, David Hyland, Filippo Torresan.

Funding:

EDA-Robots +

Funded by

Advanced Research + Invention Agency

ARIA

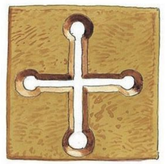
MOONSHOT RESEARCH & DEVELOPMENT PROGRAM

ARAYA RESEARCH

Internal models imply individuation?

Reverse the implication

- Not all internal models are good candidates, some are *too simple*
- No full definition of individuation yet, but “agent” is a good first step
- Agents are “goal-directed autonomous systems that interact with its environment”



Analogy for “too simple” = we somehow know all the movements of guards inside a castle without observing them properly

Internal models in optimal control

From Baltieri et al. (2025), these are *too simple*

Definition II.3 (Map of systems)**.** Let $\mathsf{X} : \mathbf{Sys}\binom{I}{X}$ and $\mathsf{X}' : \mathbf{Sys}\binom{I'}{X'}$ be systems. A *map of systems* $f : \mathsf{X} \to \mathsf{X}'$ is comprised of two parts:

1) a *map on states*, given by a function

$$f_{\mathsf{s}} : X \to X', \tag{2}$$

2) a *map on inputs*, given by a function

$$f_{\mathsf{i}} : X \times I \to I', \tag{3}$$

such that the following diagram commutes:

$$\begin{array}{ccc} X \times I & \xrightarrow{(\pi_X \fatsemi f_{\mathsf{s}}, f_{\mathsf{i}})} & X' \times I' \\ {\scriptstyle \mathsf{upd}_{\mathsf{X}}}\downarrow & & \downarrow{\scriptstyle \mathsf{upd}_{\mathsf{X}'}} \\ X & \xrightarrow{f_{\mathsf{s}}} & X' \end{array} \tag{4}$$

meaning that, for every $x \in X, i \in I$, the following equation is satisfied:

$$f_{\mathsf{s}}\big(\mathsf{upd}_X(x,i)\big) = \mathsf{upd}_{X'}\big(f_{\mathsf{s}}(x), f_{\mathsf{i}}(x,i)\big). \tag{5}$$

Definition II.7 (Model)**.** Given systems $\mathsf{X} : \mathbf{Sys}\binom{I}{X}$ called the *referent* and $\mathsf{M} : \mathbf{Sys}\binom{J}{M}$ called the *referrer*, a *model* is a map $\mu : \mathsf{X} \to \mathsf{M}$ such that

1) its part on states $\mu_{\mathsf{s}} : X \to M$ is surjective, and
2) its part on inputs $\mu_{\mathsf{i}}(x, -) : I \to J$ is surjective for each $x \in X$.

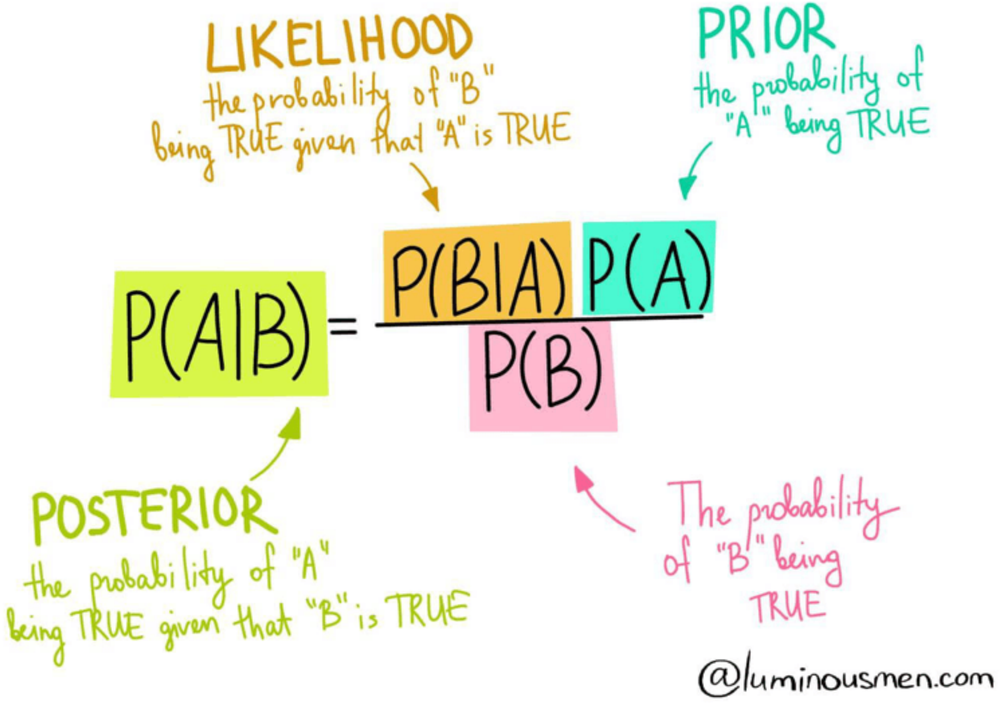
Remark II.8. When applied to autonomous systems, a model reduces to the definition implicit in [27], [28], namely a surjective map of states commuting with the dynamics:

$$\begin{array}{ccc} X & \xrightarrow{\mu_{\mathsf{s}}} & M \\ {\scriptstyle \mathsf{upd}_{\mathsf{X}}}\downarrow & & \downarrow{\scriptstyle \mathsf{upd}_{\mathsf{M}}} \\ X & \xrightarrow[\mu_{\mathsf{s}}]{} & M \end{array} \tag{8}$$

This is because the map on inputs of a model is necessarily of the form $X \times 1 \to 1$ for autonomous systems, due to the surjectivity condition. Therefore systems that model autonomous systems must also be autonomous, explaining the need for Assumption 3 later on.

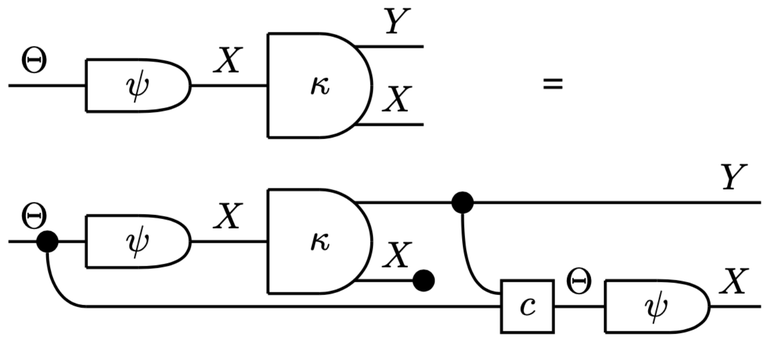
Bayesian models or interpretations

Systems that perform Bayesian updates “in the wild” are agents



Conjugate priors for Bayesian filtering

—> There exists a map c such that

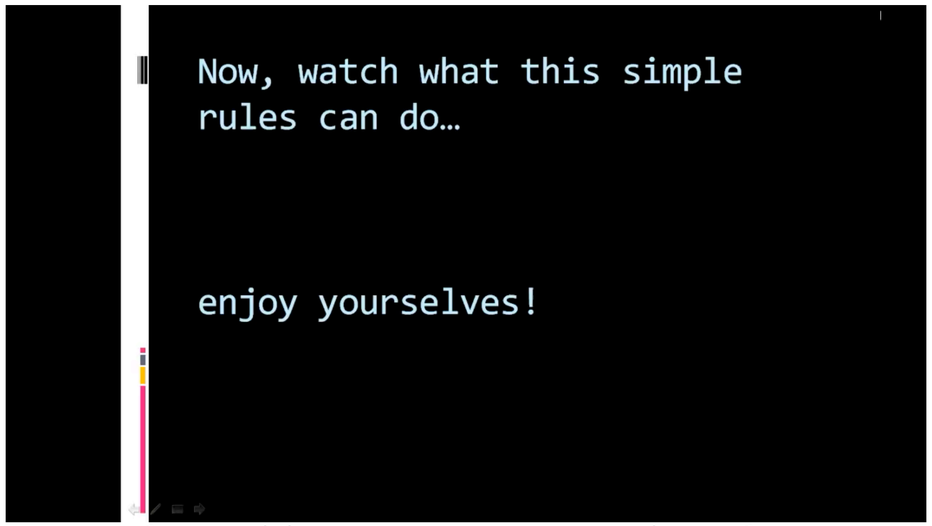


See Virgo et al. (2021), Biehl & Virgo (2022), Baltieri et al. (2025), these are **not** *too simple*

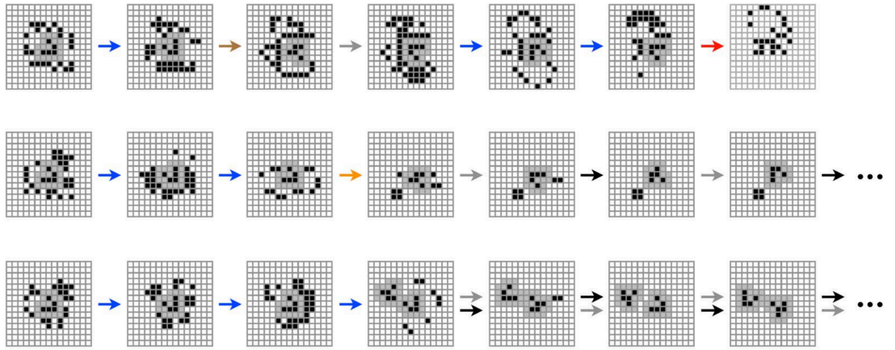
Alternative views on abstractions

Randall Beer's work on gliders in the Game of Life

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



https://www.youtube.com/watch?v=C2vgICfQawE&t=218s&ab_channel=RationalAnimations



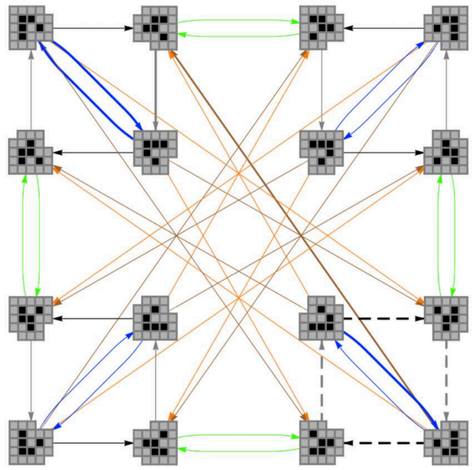


Figure 10. Three examples of structurally coupled interactions from a preliminary study of (1,2)-environments. The path through a glider's cognitive domain corresponding to the uppermost interaction is shown at the bottom, with dashed lines indicating the path the glider would have followed in isolation and thick lines indicating its structurally coupled path.

Beer (2014)

Summary

- Life *within* a simulation
- Abstracting self-sustaining systems
- We can do maths on some things, not on others yet
- To understand individuation, we define “agents”
- Agents as systems that can be interpreted as using Bayesian models to updates their states