The mathematics of systems with goals (agents?)

Manuel Baltieri

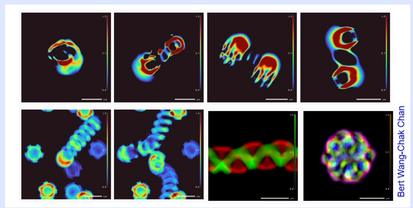
Araya Inc., Tokyo, Japan
Department of Informatics, University of Sussex, Brighton, UK

manuel_baltieri@araya.org

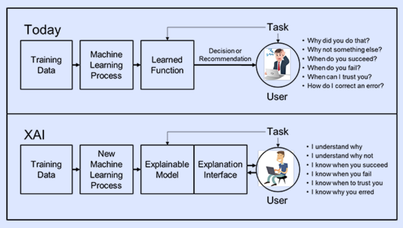
1.

Areas of interest

Life sciences: agents include living organisms and possibly related systems, defining agents can help studies of the *origins of life* and *life as it could be*.



Psychology and **neuroscience**: decision making and behavioural studies rely on knowing what agents do and what drives their *actions*, intentions and plans.



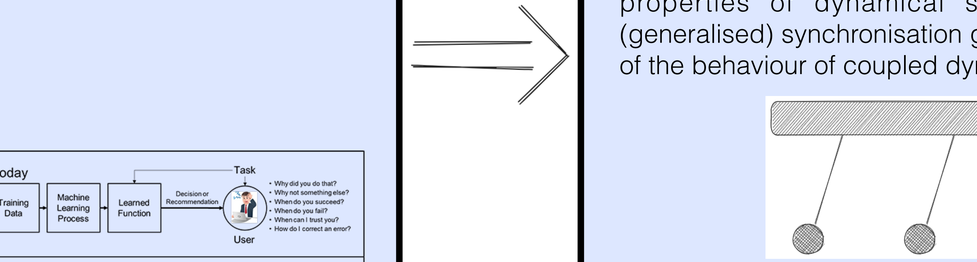
AI and **machine learning**: explainable AI and AI safety tackle complex AI/ML models to understand how they work and *interact* with users, so to generate interpretable output.

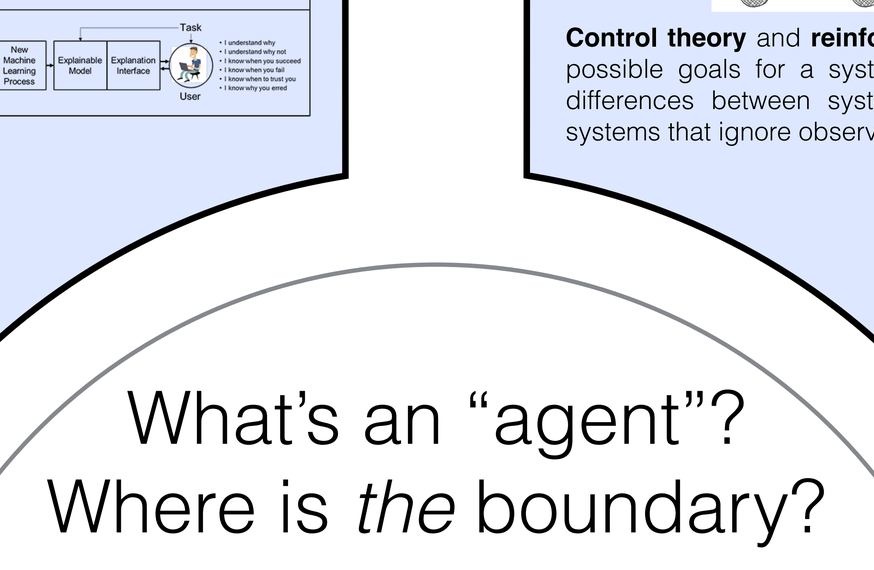
Robotics and engineering: self-driving cars, drones and robot assistants require a certain level of *autonomy* to handle complex real-world applications.

Philosophy:studies of cognition and mind for autonomous, and perhaps conscious, agents.

Other areas: economics, policy making, sociology.

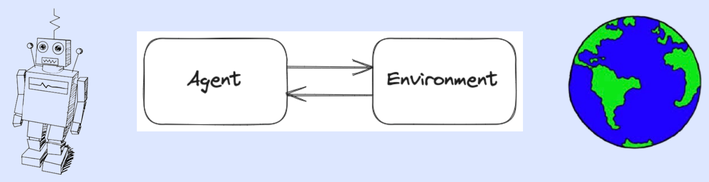
What's an "agent"? Where is *the* boundary?



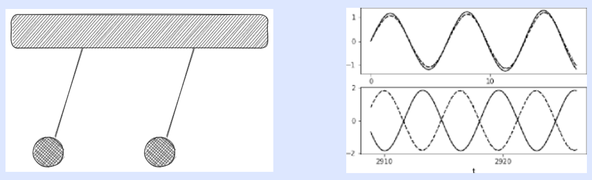


2.

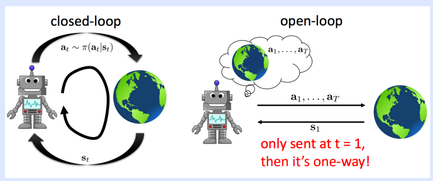
Agents interact with their environment



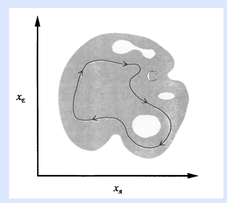
Dynamical systems theory: attractors give us a way to study invariant properties of dynamical systems and their long-term behaviour, (generalised) synchronisation gives us ways to describe particular instances of the behaviour of coupled dynamical systems.



Control theory and **reinforcement learning**: regulation condition(s) define possible goals for a system, open vs. closed loop control show us the differences between systems that are paired to their environment and systems that ignore observations from the environment.

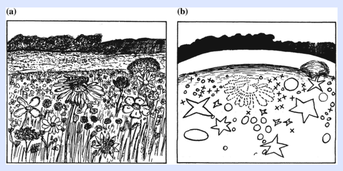


Artificial life: adaptive fit defines a subspace of the cartesian product of agent and environment as the space of *vaiability* for an agent-environment system, anything outside of it cannot be an agent.



3.

Agents are systems with models.



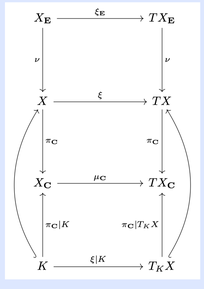
Bayesian inference: Bayesian brain hypothesis is a statement about brains as systems performing Bayesian inference.

ρ : observation
x : size of the item

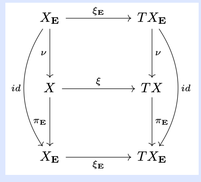
$$P(x \mid \rho) = \frac{P(\rho \mid x)P(x)}{P(\rho)}$$

Theoretical biology and **physics**: the free energy principle + active inference are trying to formalise these statements into a working theory of the origins of life and more importantly, a definition of agents (but see Bruineberg et al., 2021).

Control theory: good regulator theorem and internal model principle are results from the control theory literature motivating the idea that a system, A, controlling another system B contains/is a model of system B.



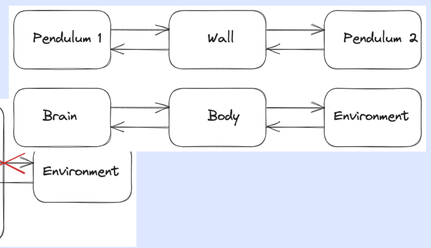
such that



4.

Agents: systems that interact with the environment they model

Challenges: 2. doesn't tell us (formally) the differences between pendulums and agents, 3. doesn't work (?) with closed-loop systems and relies on assumptions of stationarity.



Results (WIP):