An experiment in Artificial Agency

Jean-François Cloutier
Research Fellow
Active Inference Institute

4th Applied Active Inference Symposium
November 2024
Updated April 21, 2025

What does it take, at a minimum, for an autonomous robot to learn to survive in a world it knows initially almost nothing about?

Can a robot act for its own reasons?

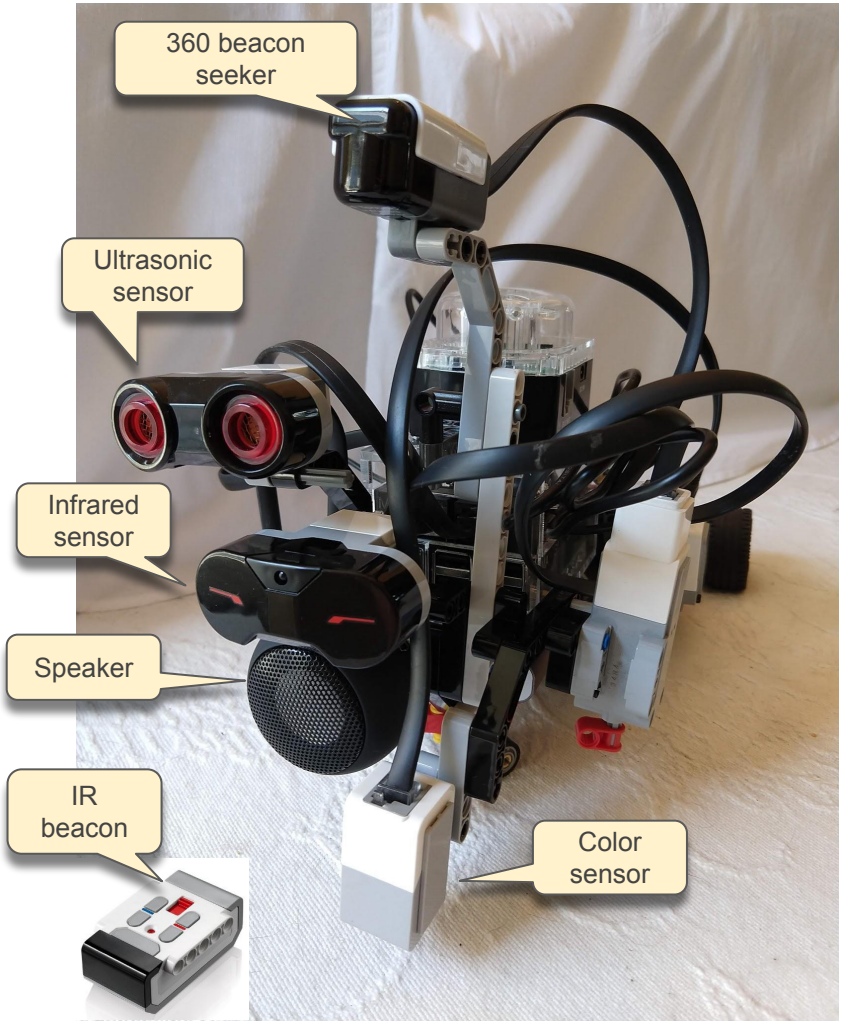
Can its agency be programmed?

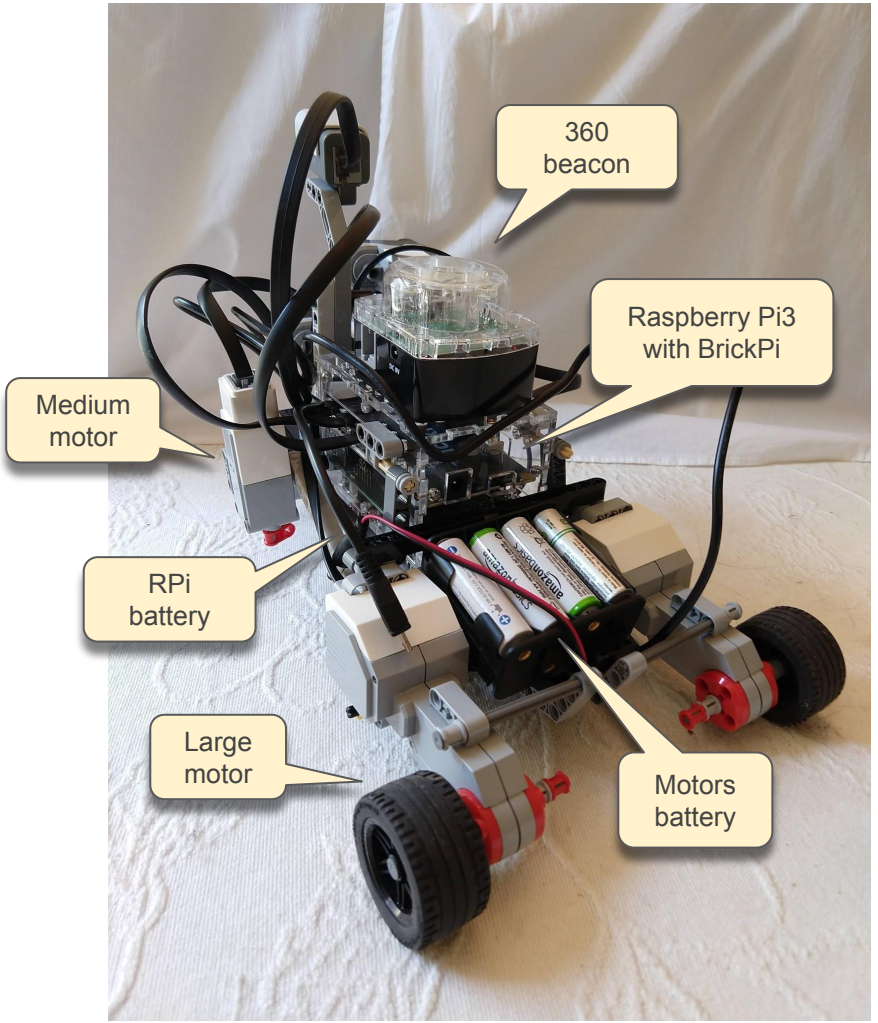
Would such a robot be an active inference agent?

One way to find out is to build an artificial agent

i.e. a sense-making, adaptive, self-directed robot

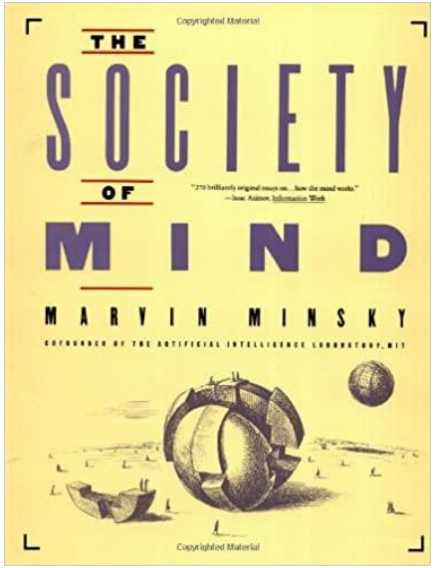
Three years ago…

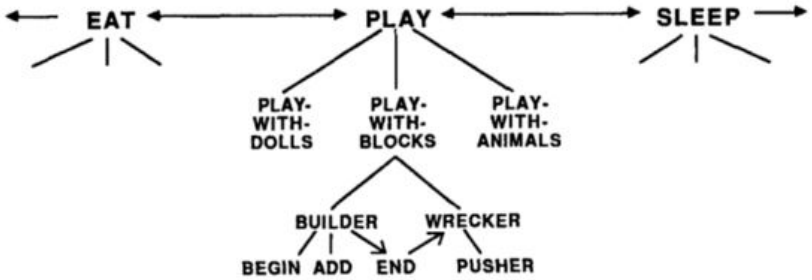

360 beacon seeker
Ultrasonic sensor
Infrared sensor
Speaker
IR beacon
Color sensor


360 beacon
Raspberry Pi3 with BrickPi
Medium motor
RPi battery
Large motor
Motors battery

Autonomous Lego robots

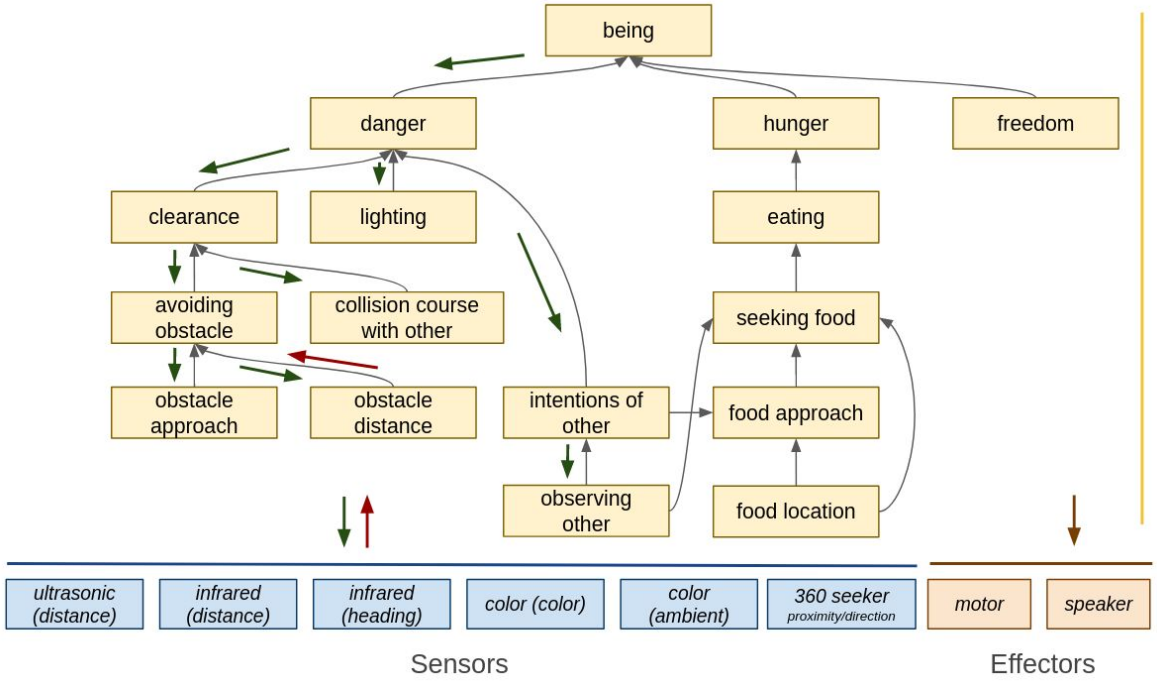
Cognition as collective intelligence

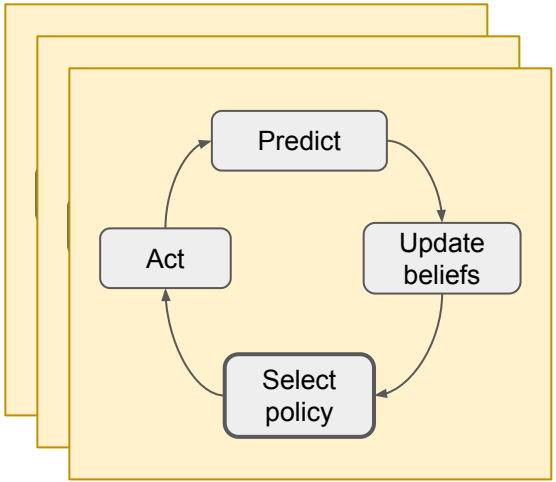




Behavior emerges from many simple actors interacting in simple ways

The robot's society of mind was predefined





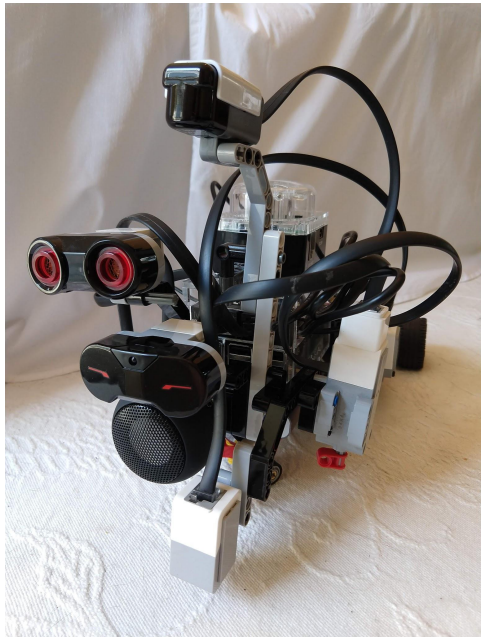
The robot acted for the *programmer's* reasons

Autonomy but no agency

2022

#sym_cog_robotics

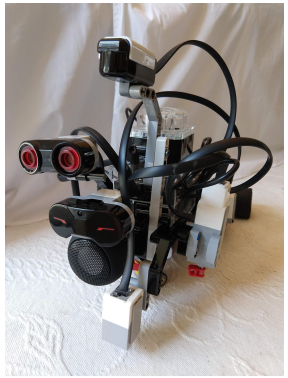
What if a robot started with only a rudimentary society of mind?

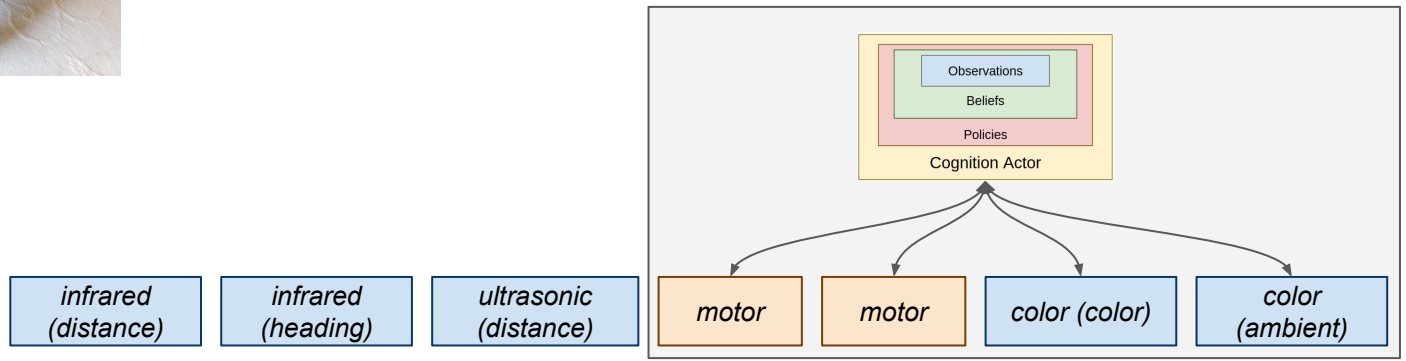


infrared (distance)	*infrared (heading)*	*ultrasonic (distance)*	*motor*	*motor*	*color (color)*	*color (ambient)*

The society of mind grows from experiencing autonomous engagement

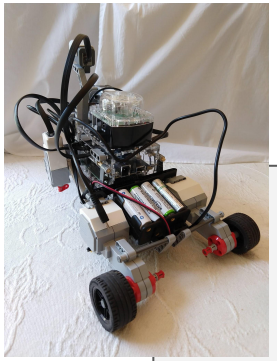
Cognition Actors are added to make sense of observations and actions

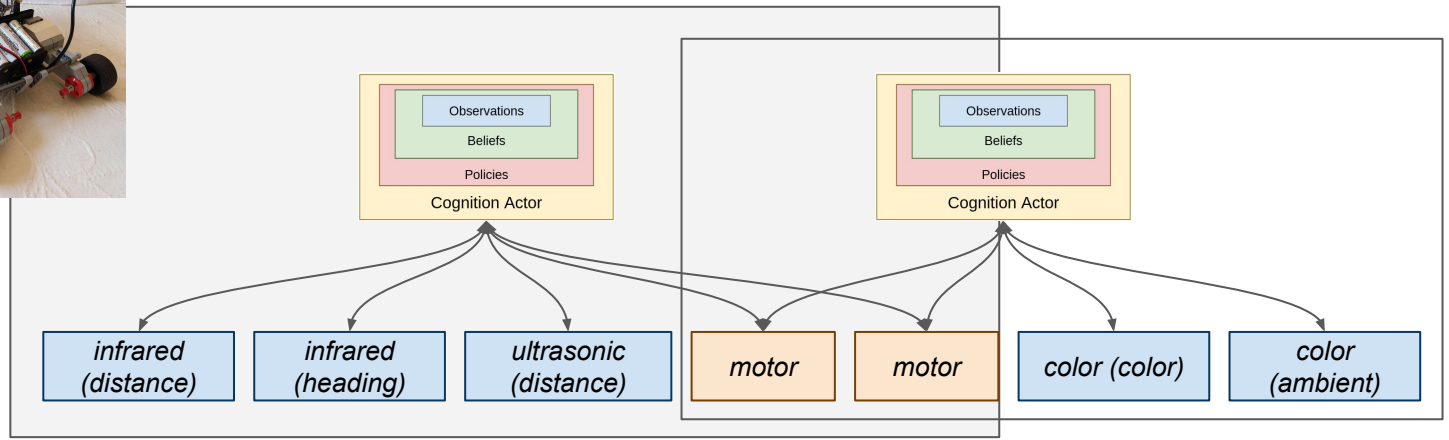




A new Cognition Actor (CA) selects an umwelt from CAs one level below

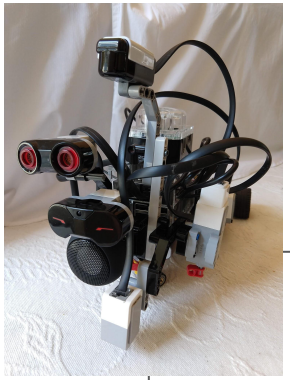
A level is complete when all CAs one level below belong to an umwelt

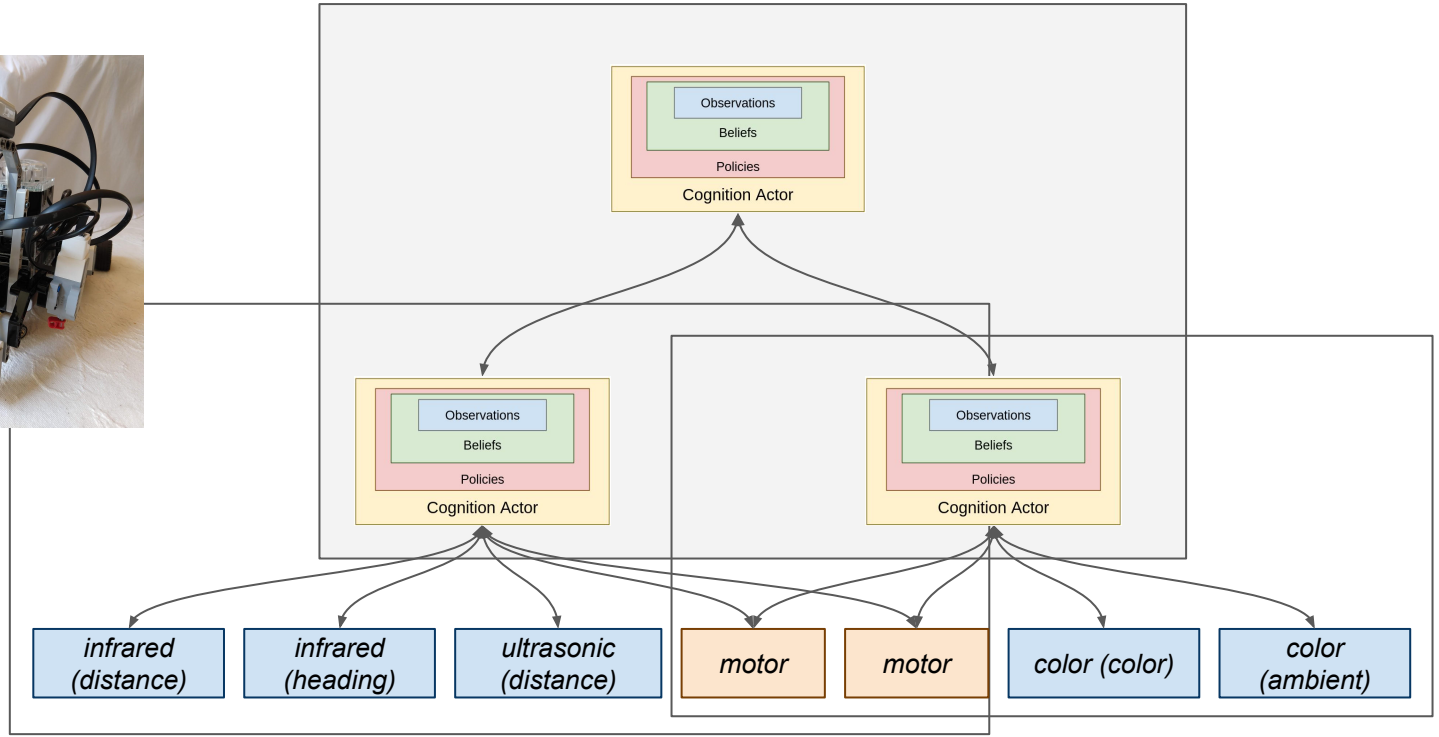




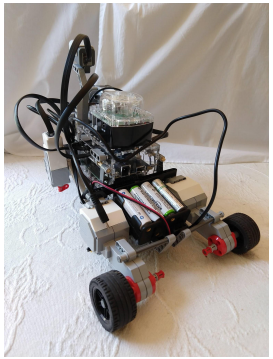
Hopelessly ineffective CAs are eventually removed and replaced

Cognition Actors form an abstraction hierarchy

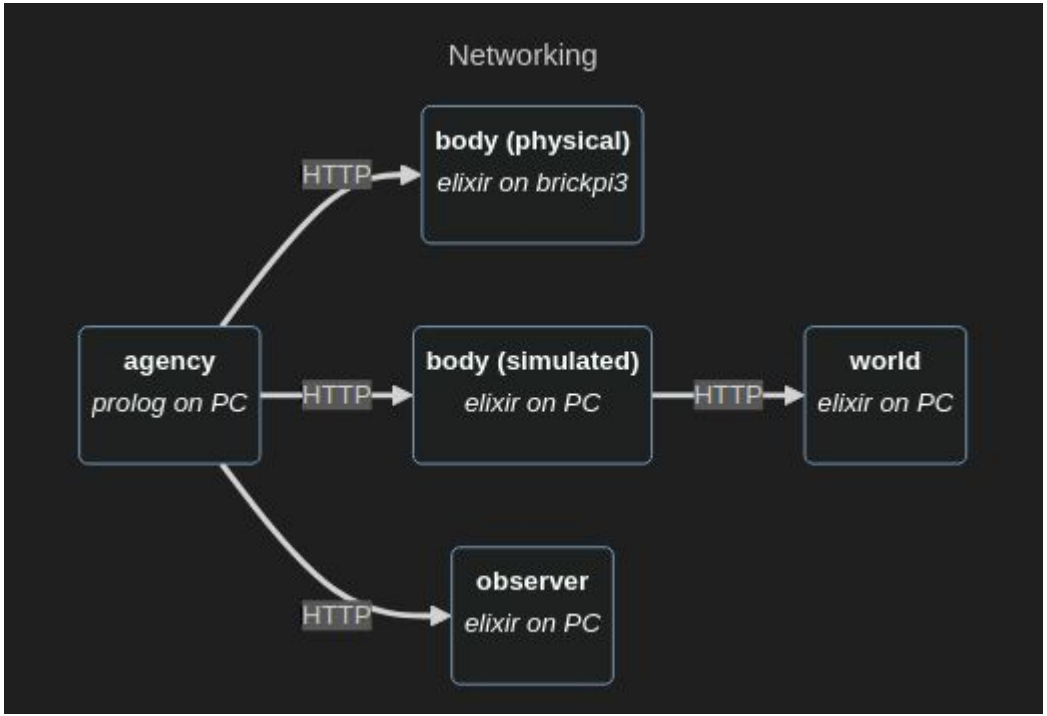




The society of mind grows, shrinks and evolves **as needed to survive**

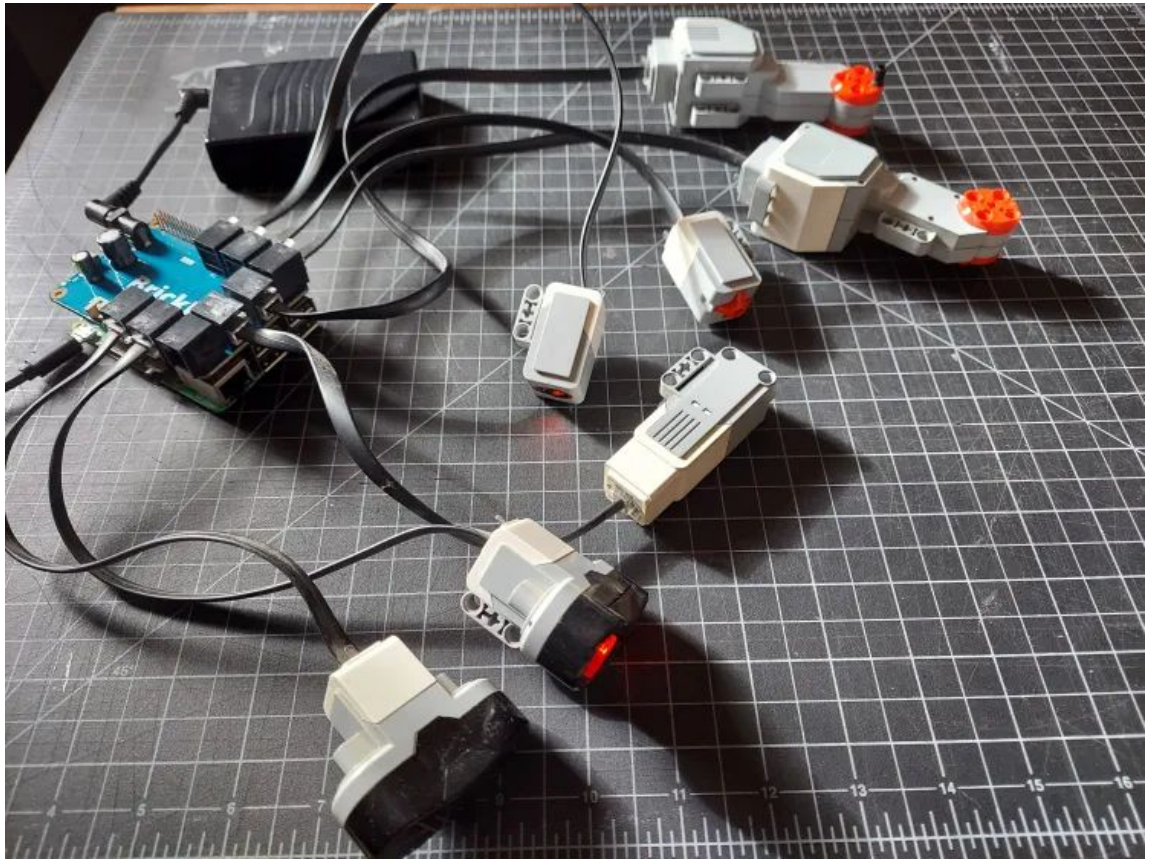


Work in progress: The Karma system

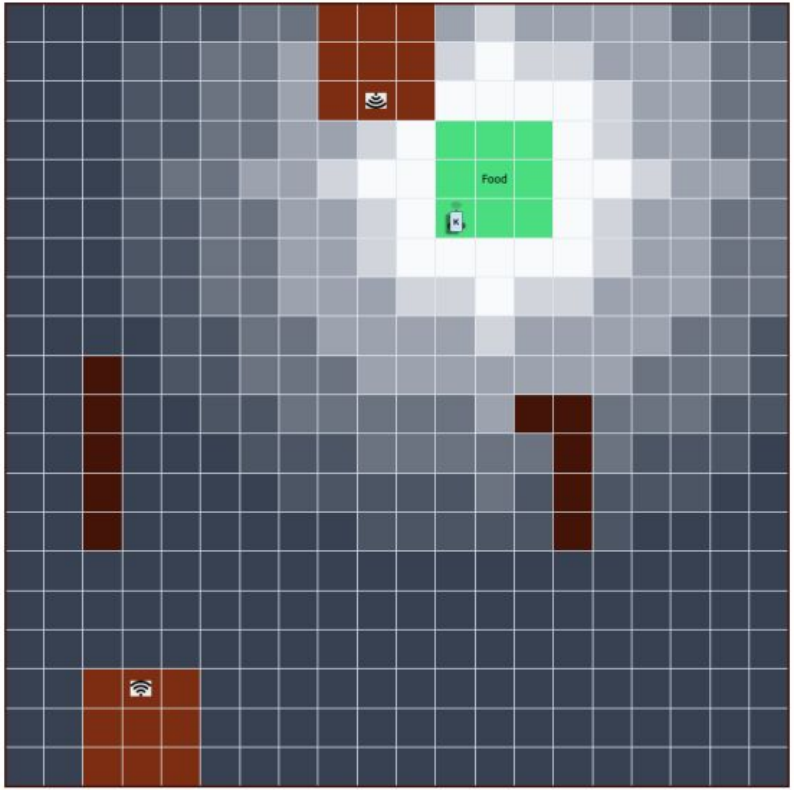


See https://github.com/jfcloutier/karma_system

Karma Body: Access to real and virtual sensors and effectors



Karma World: A generative process for the robot's virtual and real-life environments



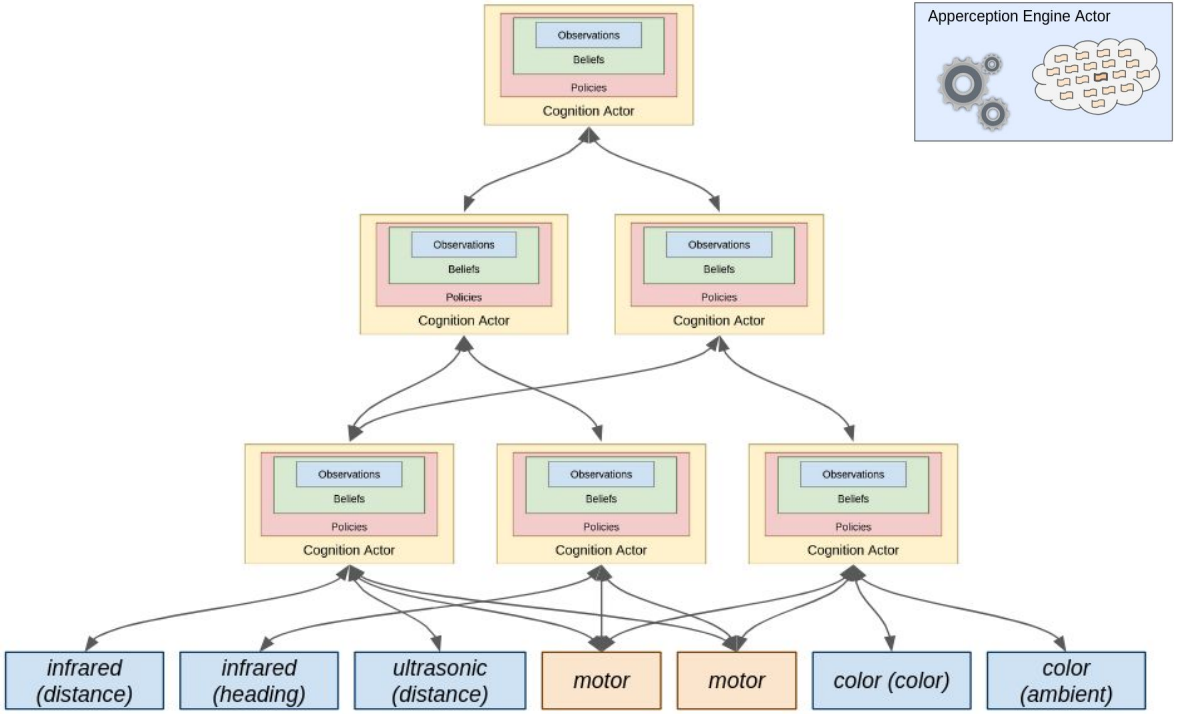
Karma Agency: A generative cognitive architecture for agency

Cognition Actors

Sensors

Effectors

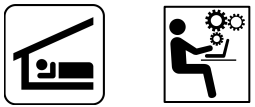
Apperception Engine



Wellbeing

Fullness 

Integrity 

Engagement 

A Cognition Actor (CA) strives to maximize its wellbeing

Low wellbeing levels indicate risk to its survival in the society of mind

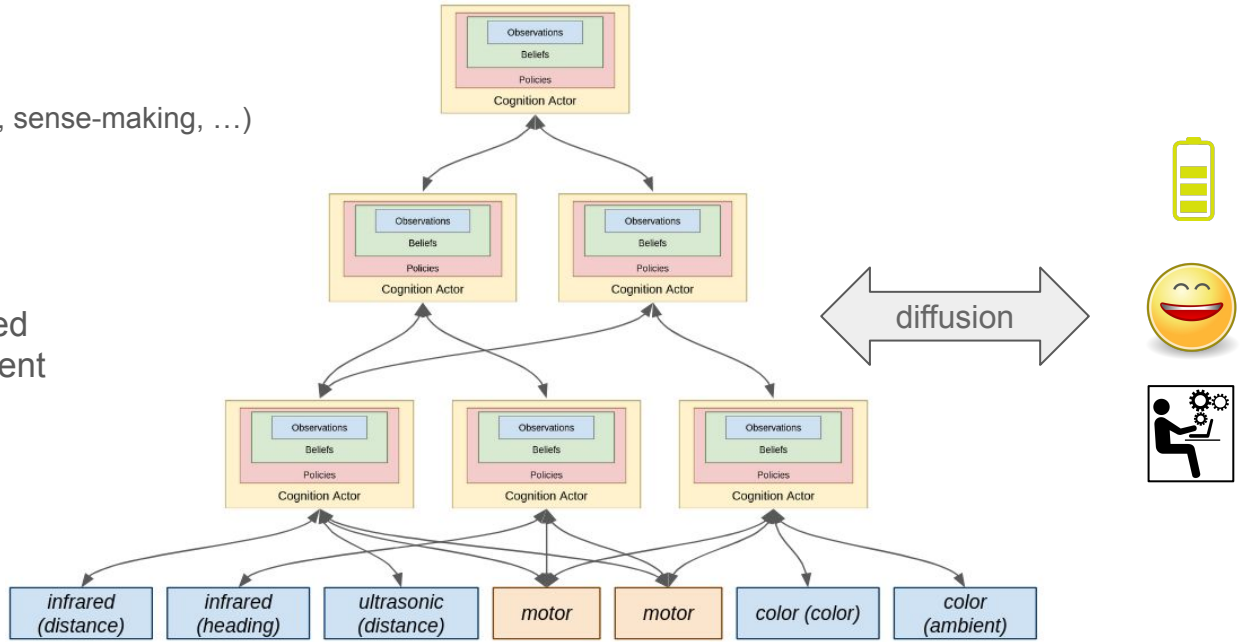
A CA signals its wellbeing for other CAs to integrate into their own

Wellbeing imparts normativity to beliefs, focuses attention, motivates action

Wellbeing measures are aggregated from wellness metrics

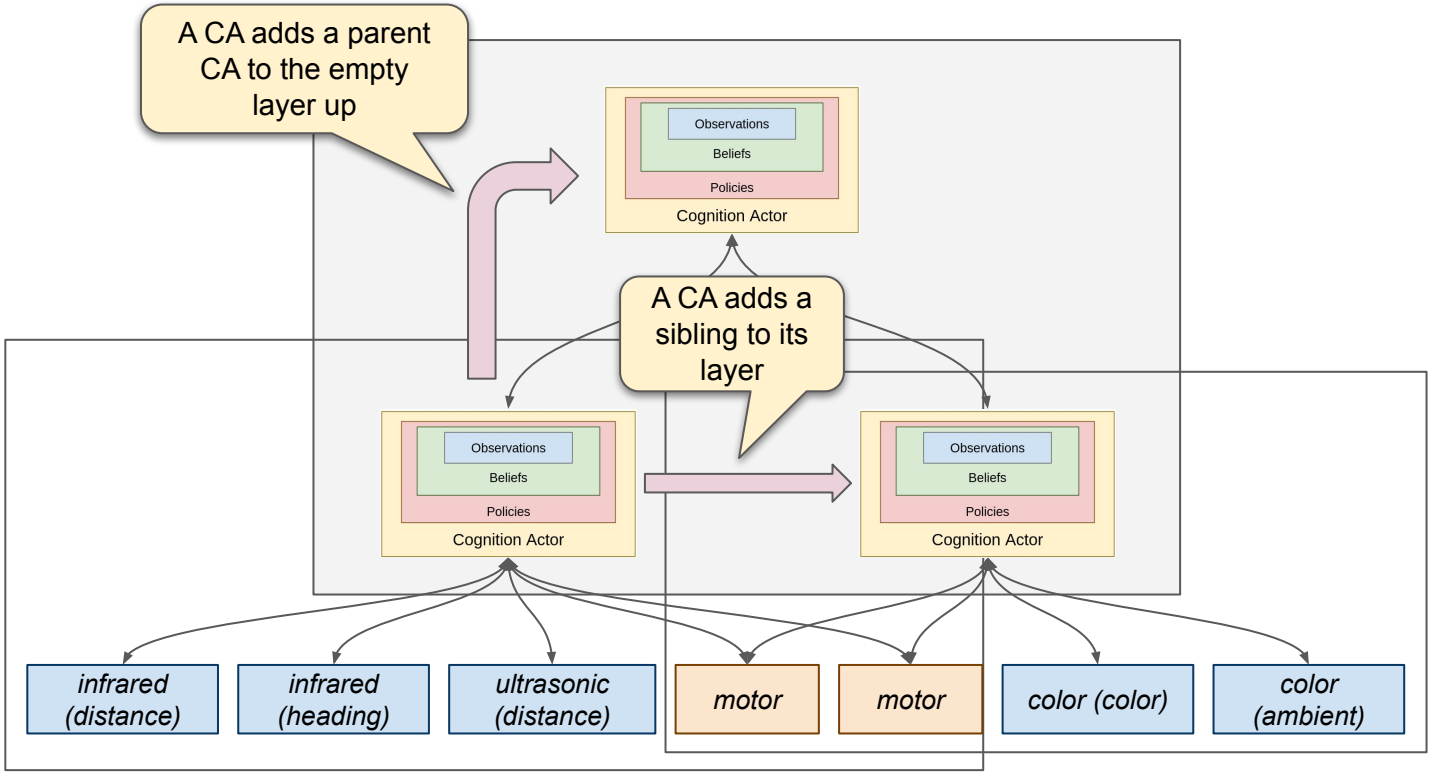
Metrics include:

- Work done (moving, sense-making, …)
- Food consumed
- Poison consumed
- Collision count
- Predictions made
- Predictions received
- Prediction errors sent
- Policies executed



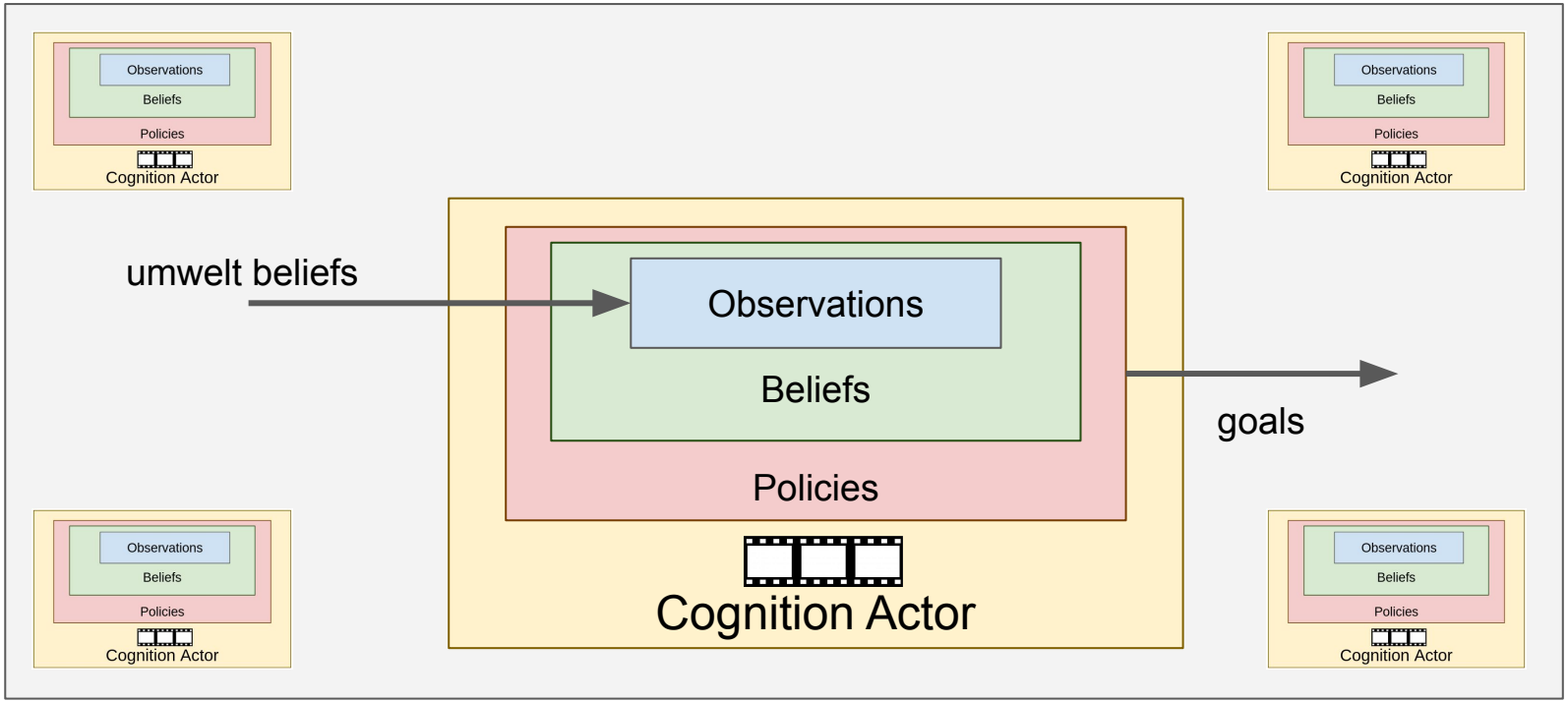
The wellbeing of a CA diffuses to its umwelt and its parent CAs, affecting their wellbeing by “osmosis”.
A CA can act as a “wellbeing capacitor”, accumulating and discharging wellbeing from and to its umwelt CAs

CAs add missing CAs and underperforming CAs remove themselves



Low wellbeing accelerates CA removal and slows down adding new CAs

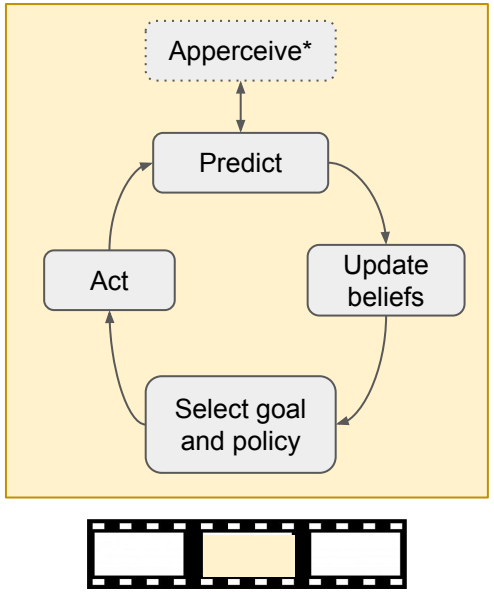
A CA observes, believes and acts within its umwelt one time frame at a time



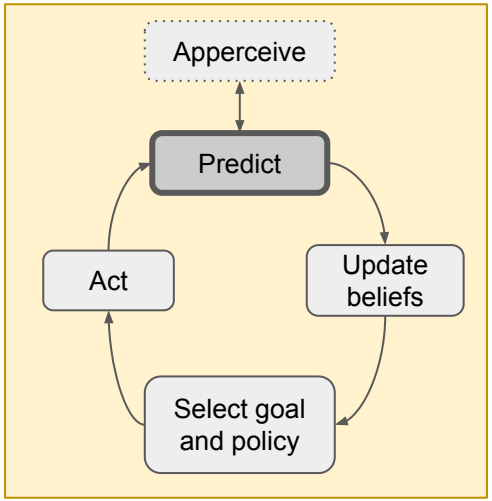
CA time frames vary in length and don’t overlap. The more abstract the CA, the longer its time frame

Within each time frame, a Cognition Actor…

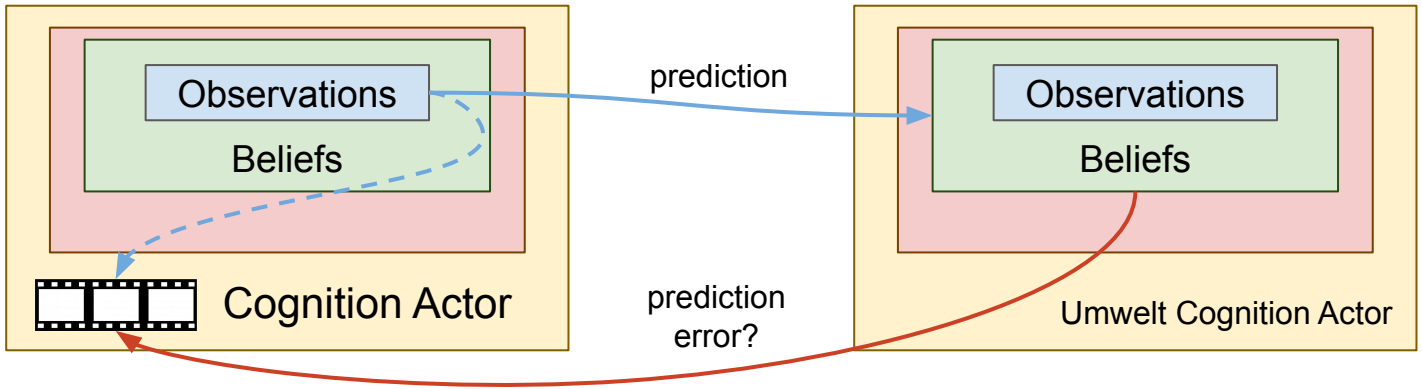
1. Predicts its umwelt from having made sense of it
2. Updates its beliefs by analyzing past and present observations
3. Selects a goal affecting a belief it intends to impact, constructs a policy and recruits its umwelt to realize it



* Apperception is the process of assimilating sensory information into a coherent unified whole


Apperceive
Predict
Update beliefs
Select goal and policy
Act

Observations = uncontested predictions + prediction errors

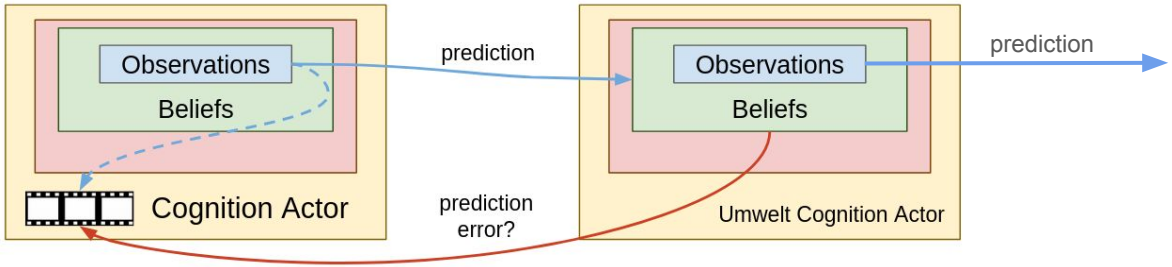


A CA observes the beliefs of the CAs in its umwelt. Not all observations receive the same attention

Attention prioritizes which observations to update

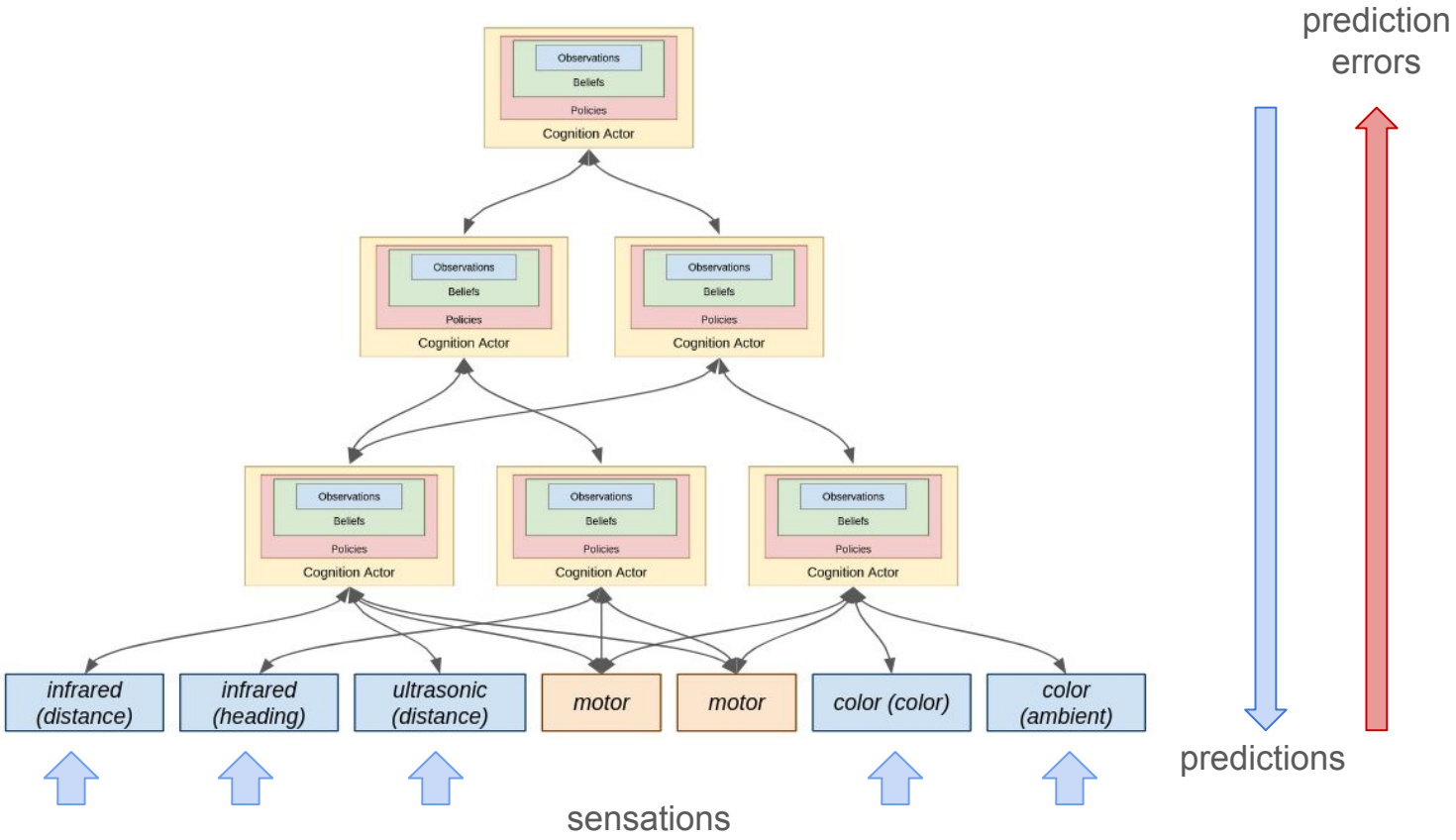
The CA prioritizes predicting its…

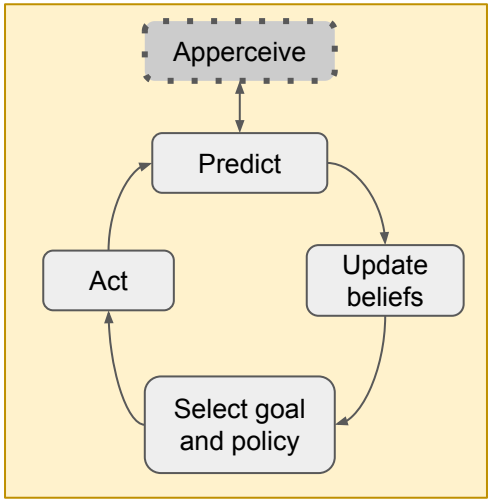
1. Observations supporting beliefs predicted by a parent CA
2. Observations supporting the strongest desirable and undesirable beliefs
3. Observations supporting weak or neutral beliefs
4. Remaining observations



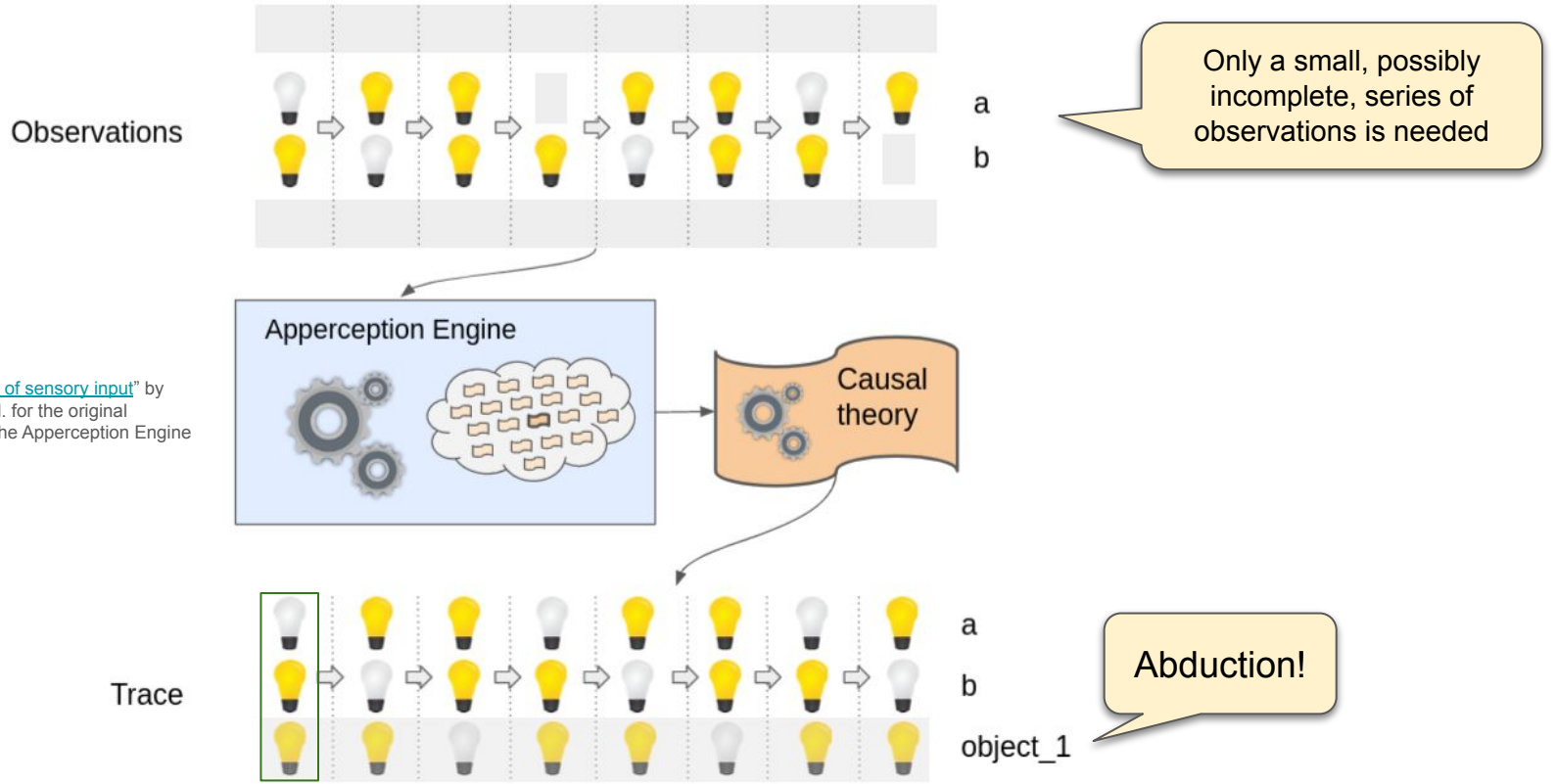
A CA may not have time to predict all incoming observations within a time frame

Predictive processing

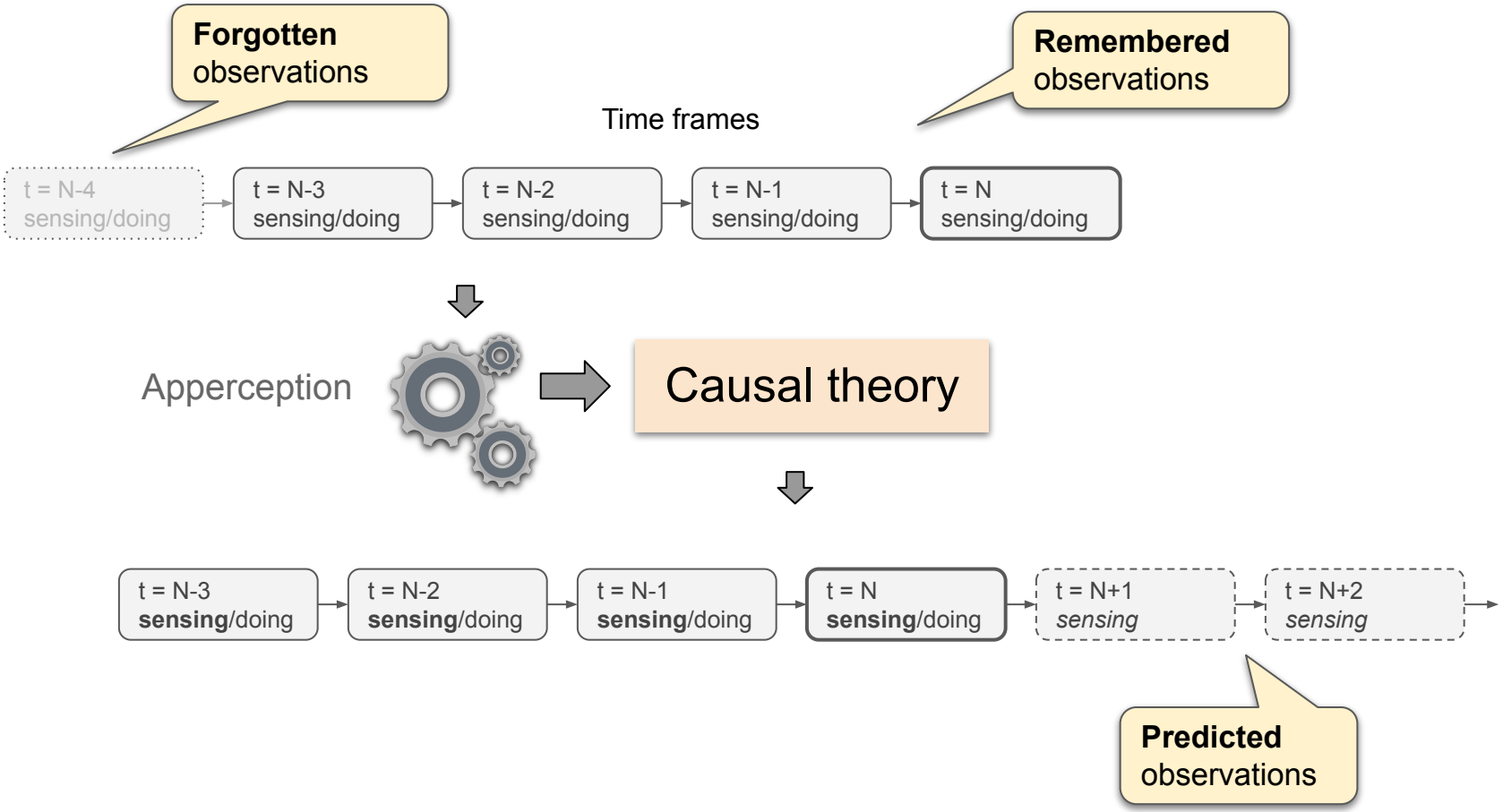



Apperceive
Predict
Update beliefs
Select goal and policy
Act

Making sense of observations in order to predict them

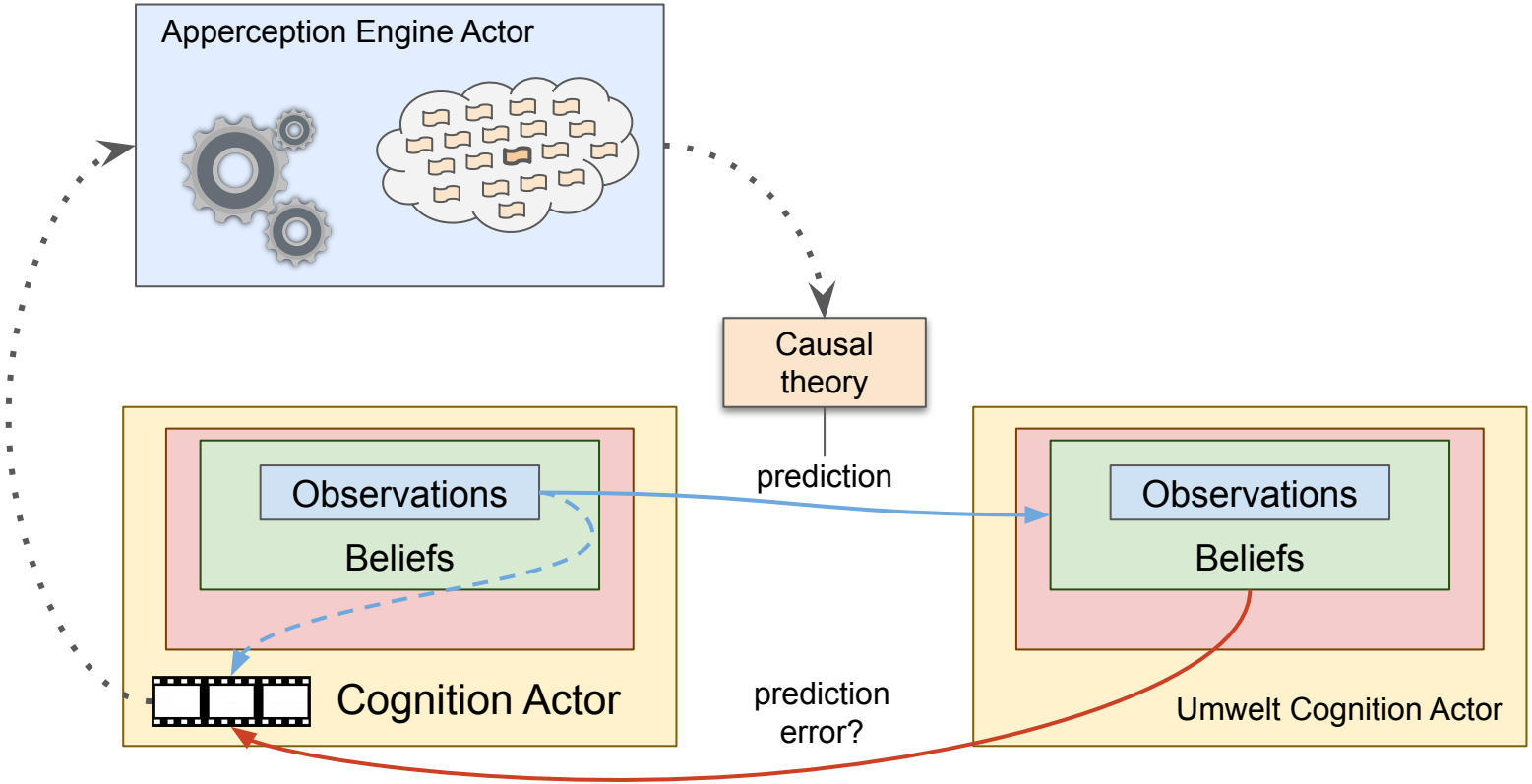


See "Making sense of sensory input" by Richard Evans et al. for the original implementation of the Apperception Engine

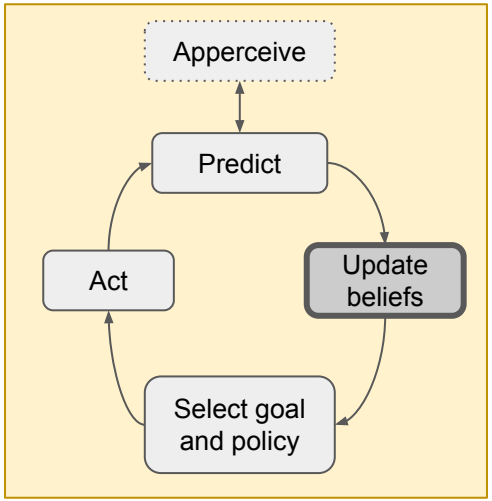


See https://zenodo.org/records/10325868

Apperception finds a causal theory that explains and predicts observations



Prediction errors undermine confidence in the causal theory


Apperceive
Predict
Update beliefs
Select goal and policy
Act

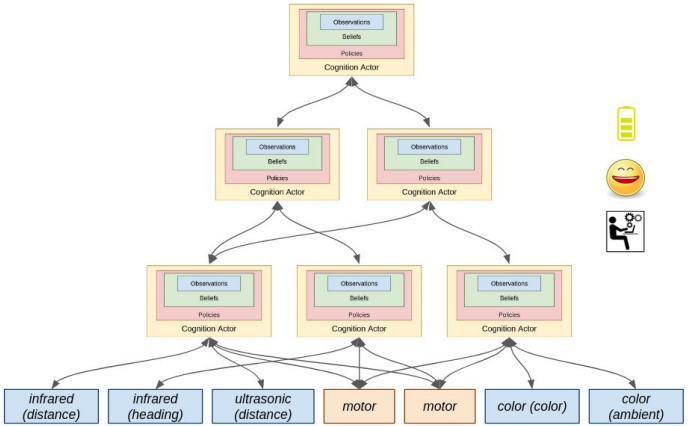
Belief updating

A CA updates its beliefs from analyzing remembered observations and actions taken

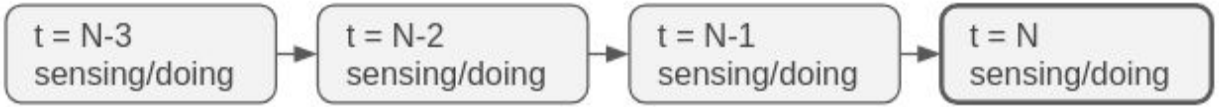
A belief is desirable or undesirable depending on how correlated it is with the CA's wellbeing

A belief is strongly held if

- it is supported by observations predicted by an accurate causal theory
- it persists over many time frames thanks to, or in spite of, actions taken



Kinds of beliefs



Abduction *I **assume** there's a **hidden** thing next to one I see*

Count *I am next to **2** things*

Trend *My distance to a thing is **getting smaller***

Ending *My getting closer to a thing **stopped***

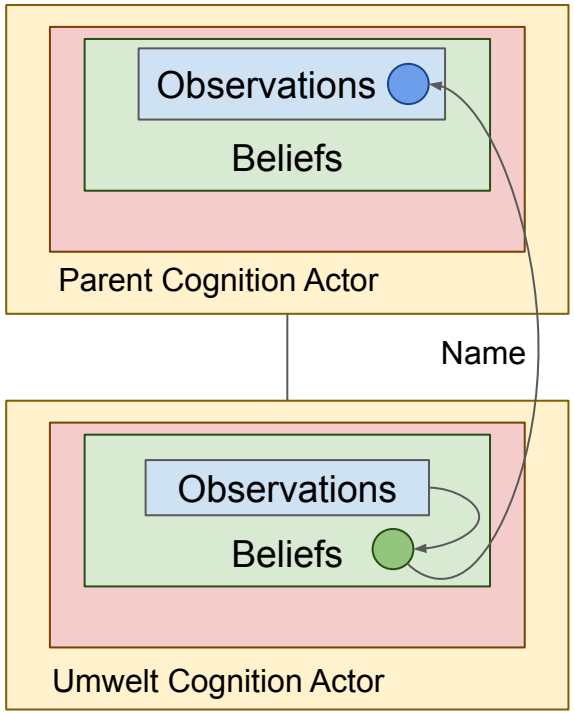
Attempt *I **tried** to stop my distance to a thing from getting smaller*

Naming beliefs

When a CA synthesizes a belief from analyzing its observations, it assigns the belief a globally unique predicate name

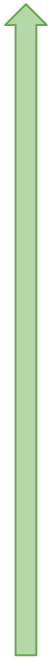
Parent CAs observe (predict) the beliefs of CAs in their umwelts, and reference them by name

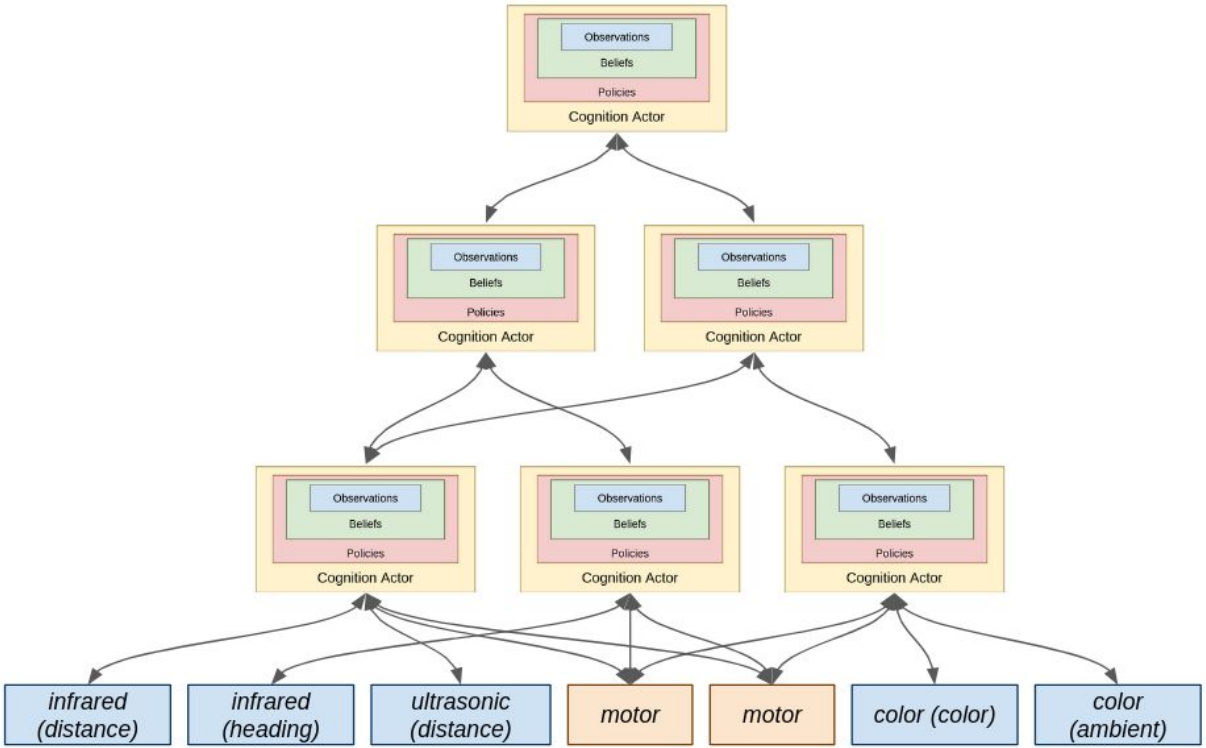
Only the CA holding the belief it named knows which of its own observations support the belief and how



One’s belief is another’s observation

beliefs about beliefs about beliefs…





(

Agency from symbolic reasoning

Agency actors run Prolog programs that do symbolic inferencing

A Cognition Actor invents symbols and synthesizes beliefs as symbolic statements

The Apperception Engine discovers *symbolic causal theories to be used as generative models by Cognition Actors*

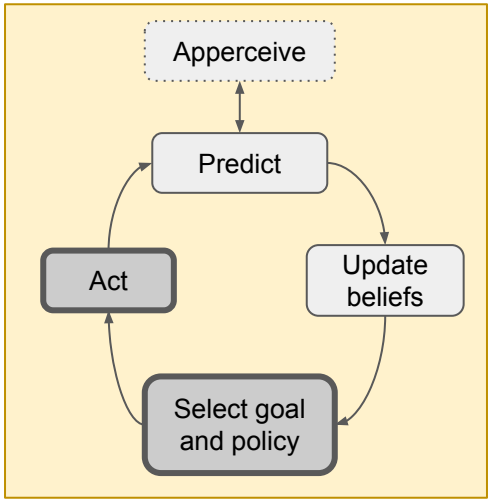
)

Precision and belief strength

The more accurate a causal theory (used as a generative model), the more precise its predictions

Precise predictions (if not contradicted by prediction errors) strengthen the beliefs the predicted observations support

The stronger beliefs are, the more attention they receive when selecting which ones to impact


Apperceive
Predict
Update beliefs
Select goal and policy
Act

A Cognition Actor acts via policies it formulates

A CA strives to invalidate unpleasant beliefs and to validate pleasant beliefs, prioritizing strongly held beliefs

A CA impacts a belief it holds by promoting or inhibiting observations that led to it

These observations are themselves beliefs of umwelt CAs

A CA formulates a policy as a list of goals for the umwelt CAs, each goal a desired impact to an umwelt belief

The goals listed in a policy are themselves realized by policies

A CA tells umwelt CAs which of their beliefs to validate/invalidate (i.e. goals)

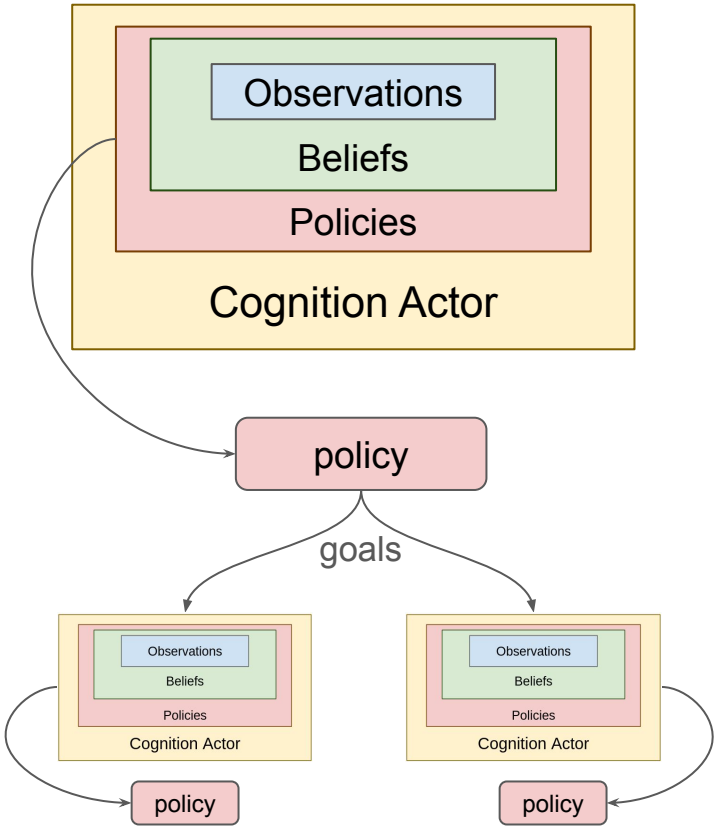
An umwelt CA formulates its own policy to achieve a goal received “from above”

A received goal has priority over the CA’s own goals

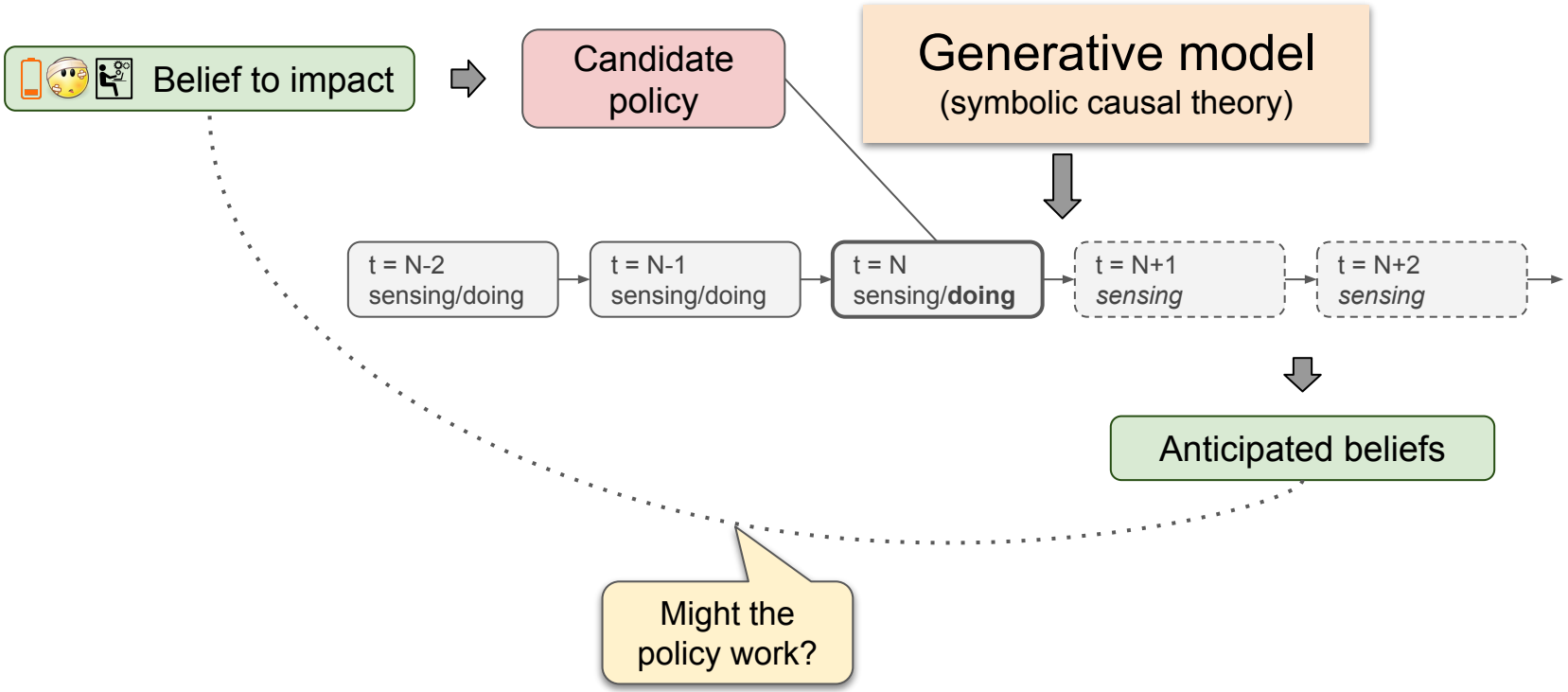
A CA can execute only one policy at a time

A CA determines policy success or failure from noticing changes to its beliefs

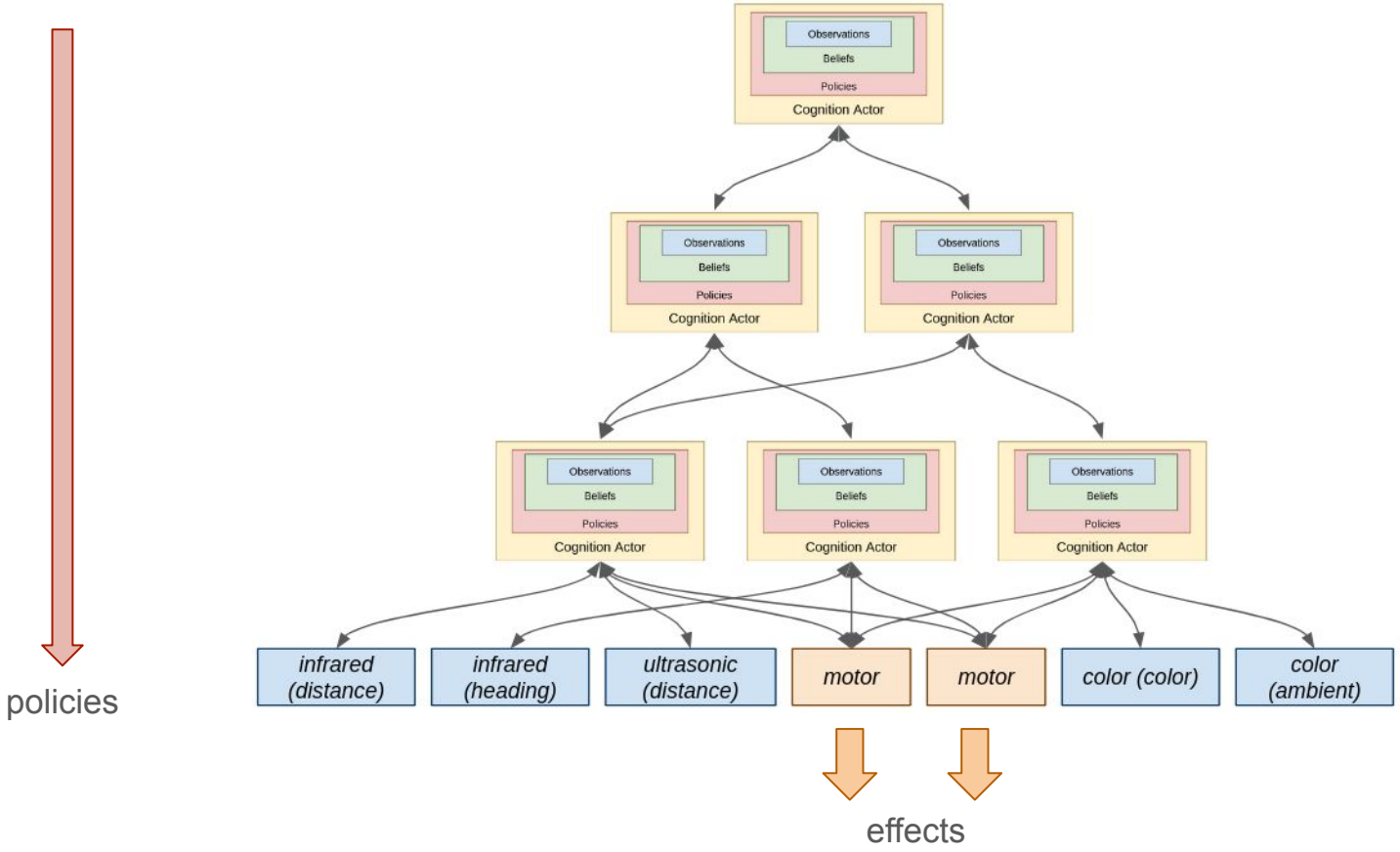
A policy known to work is reused (habits)



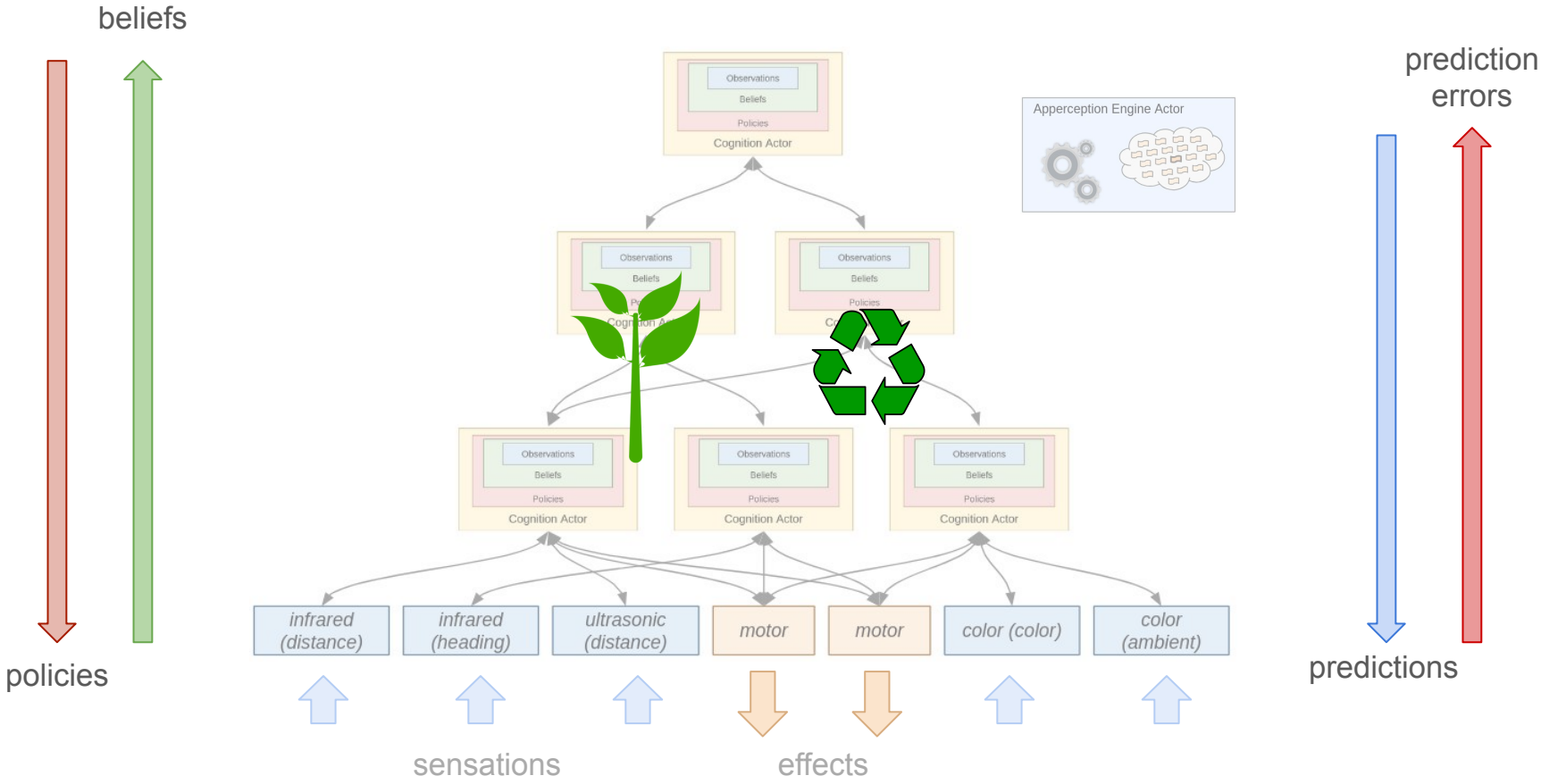
The CA's causal theory guides the formulation of its policies



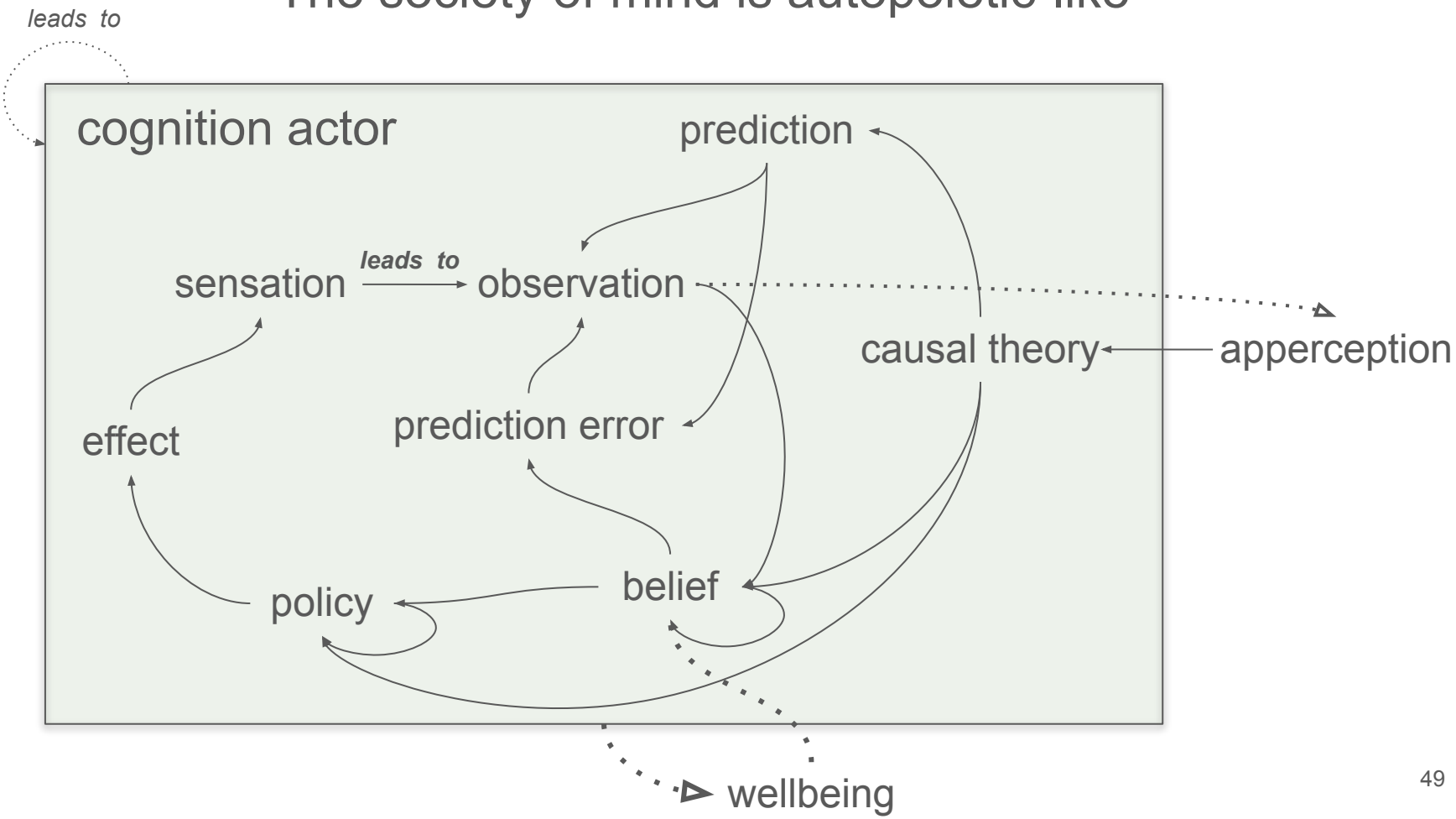
Action is policies all the way down



Agency emerges from operational closure



The society of mind is autopoietic-like



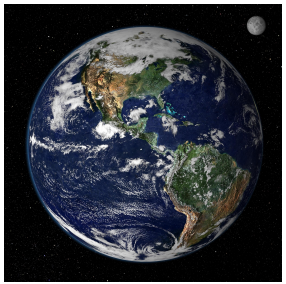
Philosophical stances taken

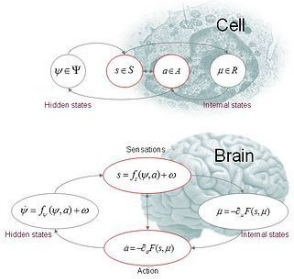
Philosophical stances are “**pragmatically justified perspectives or ways of seeing the world**”

Boucher, S.C. What is a philosophical stance? Paradigms, policies and perspectives. Synthese 191, 2315–2332 (2014). https://doi.org/10.1007/s11229-014-0400-y

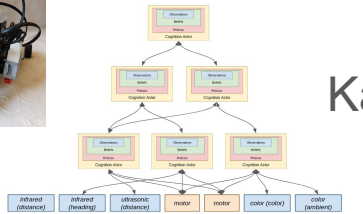
Active Inference is a *normative* framework

*To persist, an agent must behave **as if** it acts to minimize surprisal*

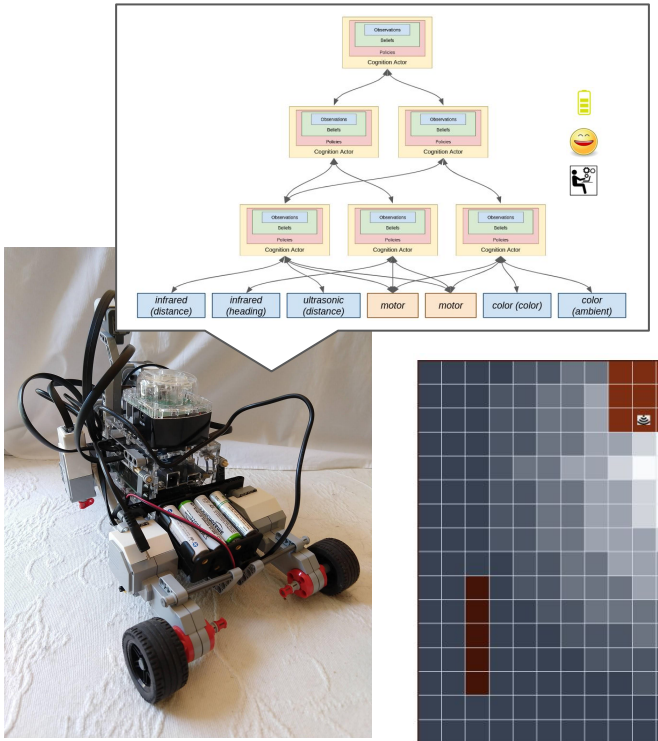




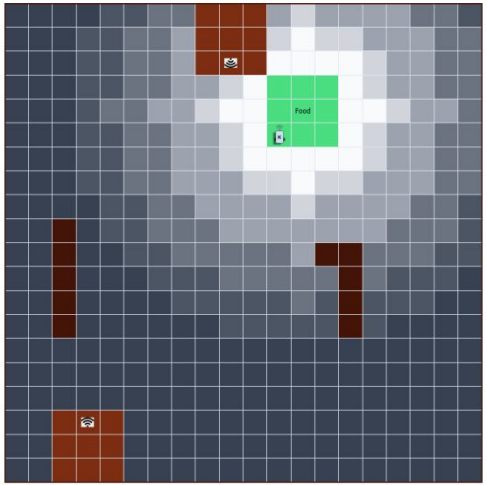
FEP



Karma



The implementation of both agent and environment is a system of generative *processes*



Active Inference *analysis* applies to the robot's observable behaviors and states

Cognition arises from an embodied agent engaging its environment

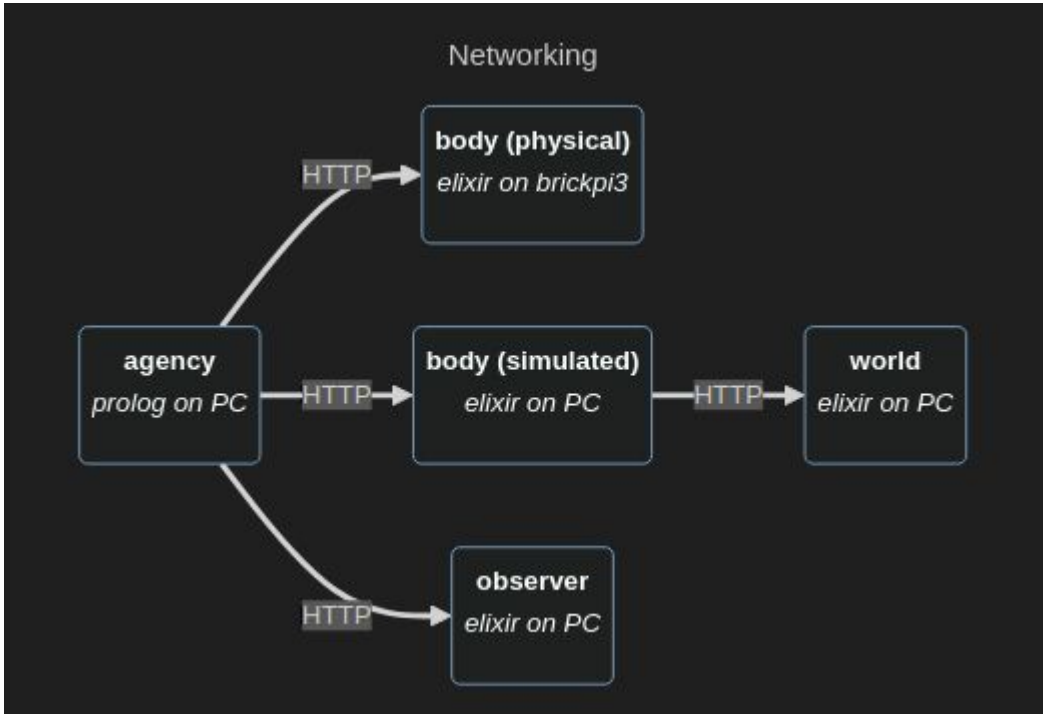
Sense making is the discovery of unified causal theories

Meaning is grounded in the agent's drive to survive

Intelligence emerges from interactions between a collective of parts

The parts constrain the whole and the whole constrains the parts

Implementing artificial agency



Progress report

Phase 1

- Cognitive architecture ✔
- Apperception Engine ✔
- Karma Body and World ✔
- Karma Agency (Wellbeing, Cognition Actor…)

Phase 2

- Robot redesign
- Real-life test environment
- Karma Observer

Phase 3

- Data gathered from runs (RL and virtual)
- Active Inference analysis

jf.cloutier@activeinference.institute

#sym_cog_robotics on the AII’s Discord

https://github.com/jfcloutier/karma_system