An experiment in Artificial Agency

Jean-François Cloutier
Research Fellow
Active Inference Institute

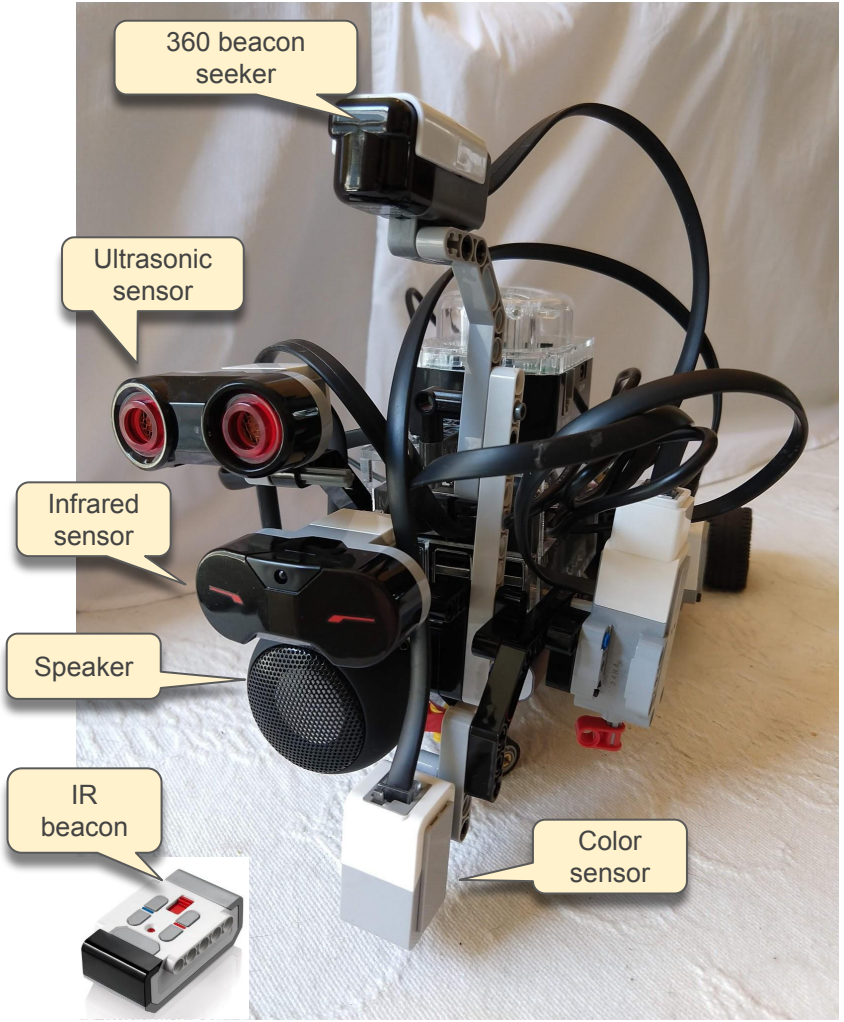
July 17, 2024

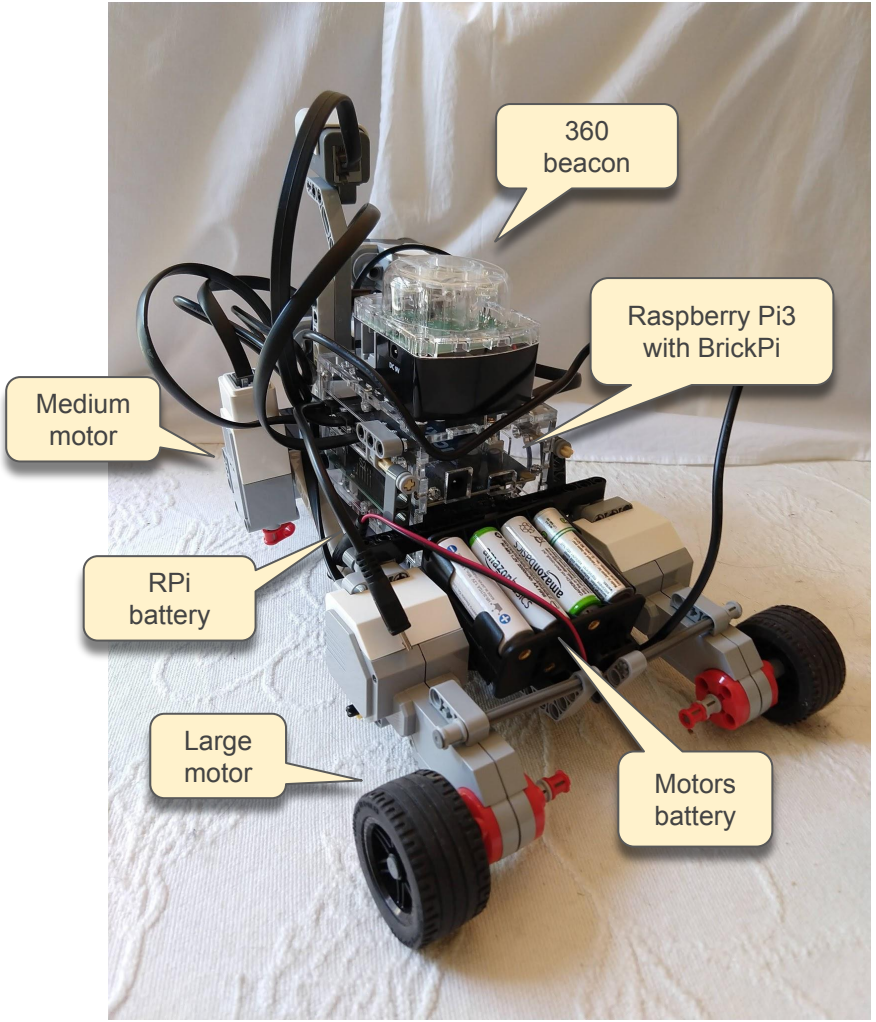
What does it take, at a minimum, for an autonomous robot to learn to survive in a world it knows initially almost nothing about?

Can a robot act for its own reasons?

Can agency be programmed?

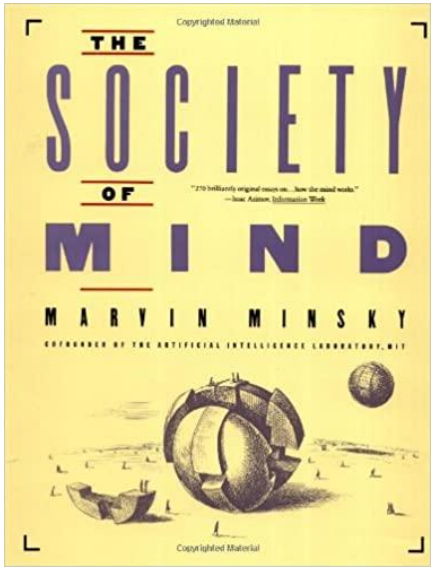
Three years ago…

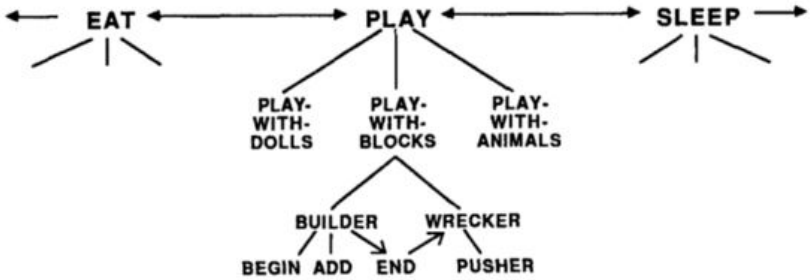

360 beacon seeker
Ultrasonic sensor
Infrared sensor
Speaker
IR beacon
Color sensor


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

Autonomous Lego robots

Cognition as collective intelligence

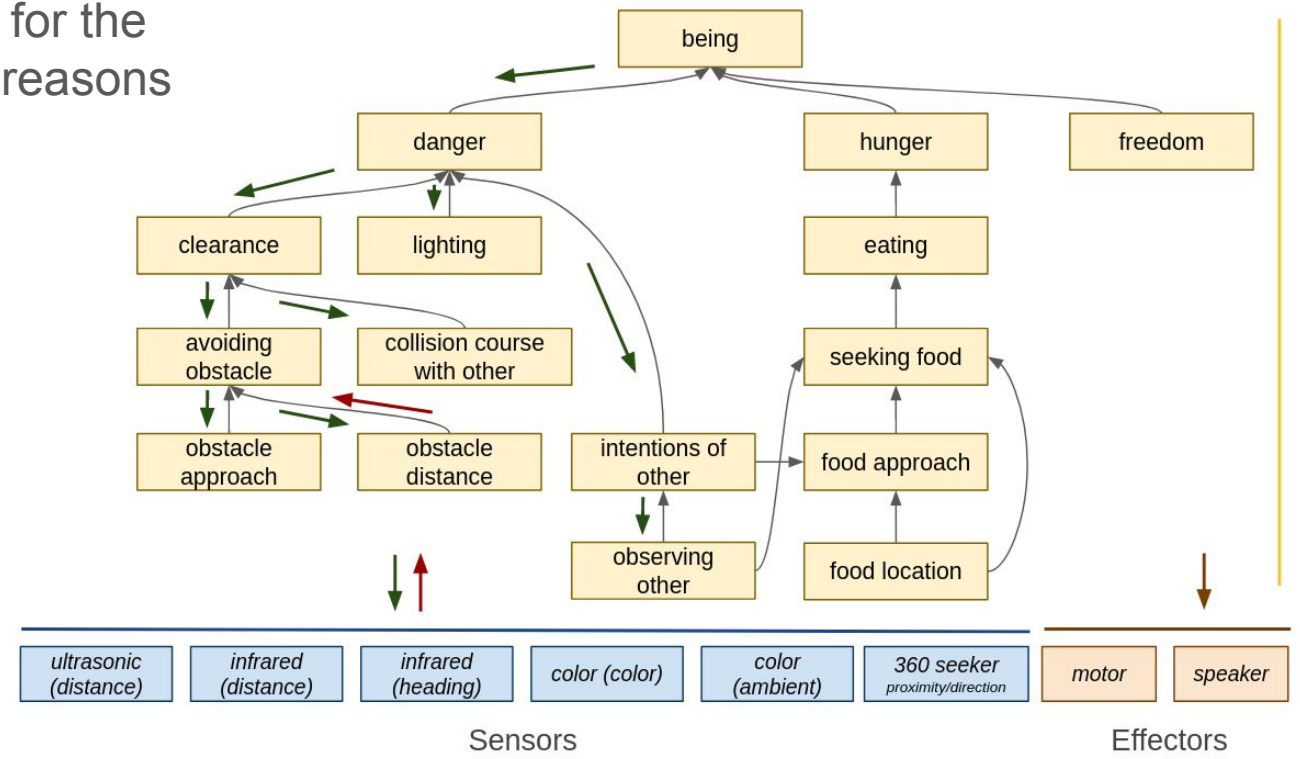




Behavior emerges from many simple actors interacting in simple ways

A robot's society of mind is predefined

The robot acts for the *programmer's* reasons

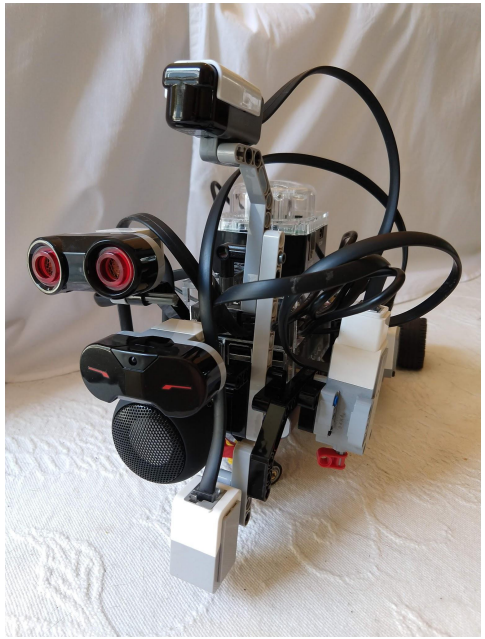


Autonomy but no agency

2022

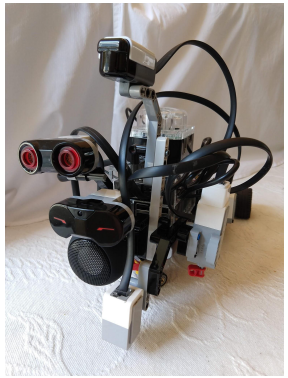
#cog_sym_robotics

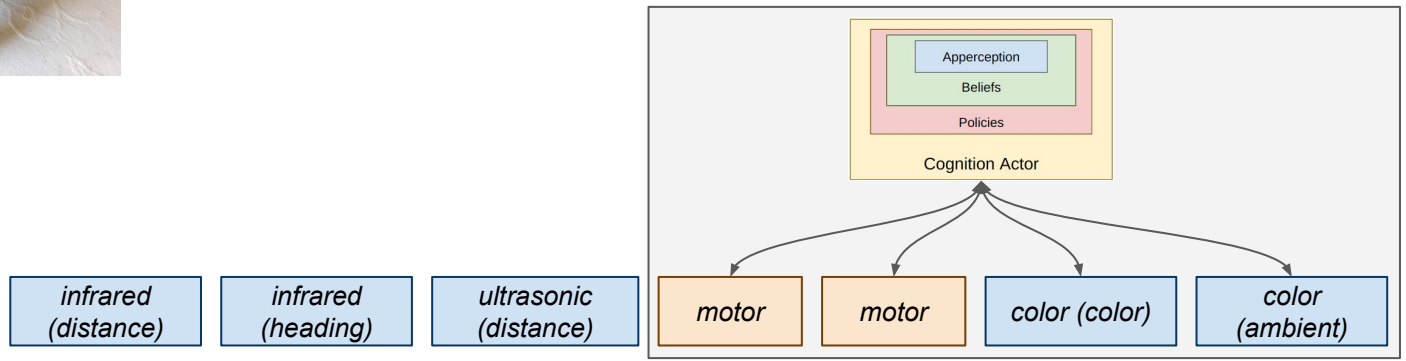
What if a robot started with only a rudimentary society of mind?



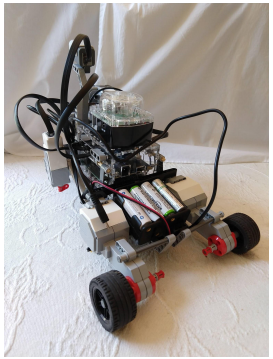
infrared (distance)	*infrared (heading)*	*ultrasonic (distance)*	*motor*	*motor*	*color (color)*	*color (ambient)*

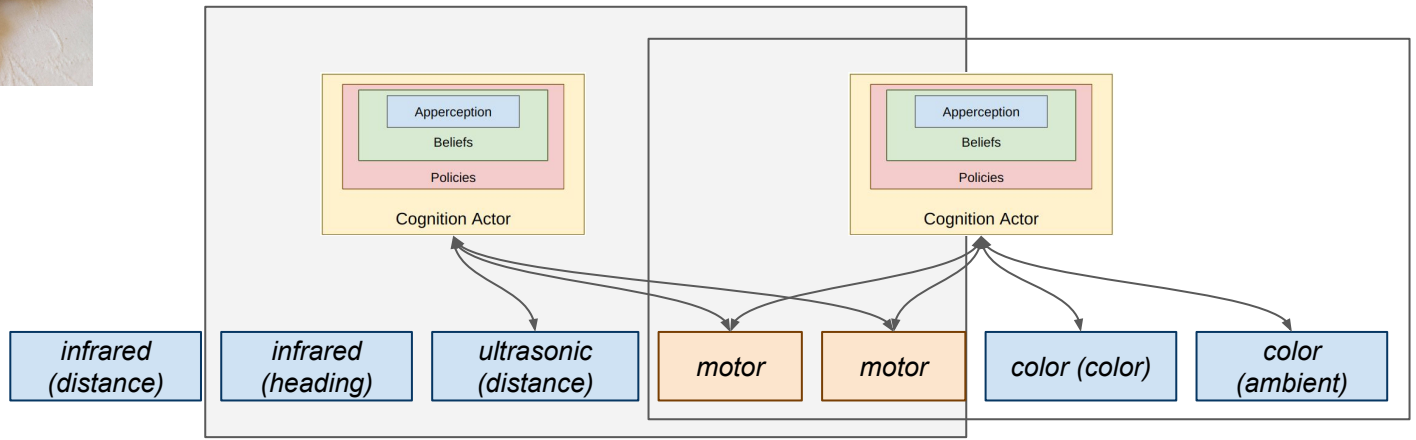
A society of mind would grow on its own through autonomous engagement



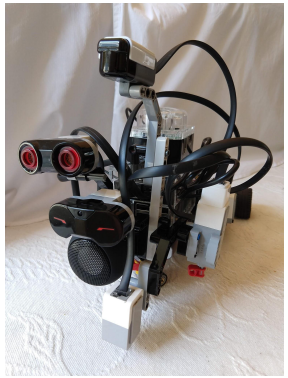


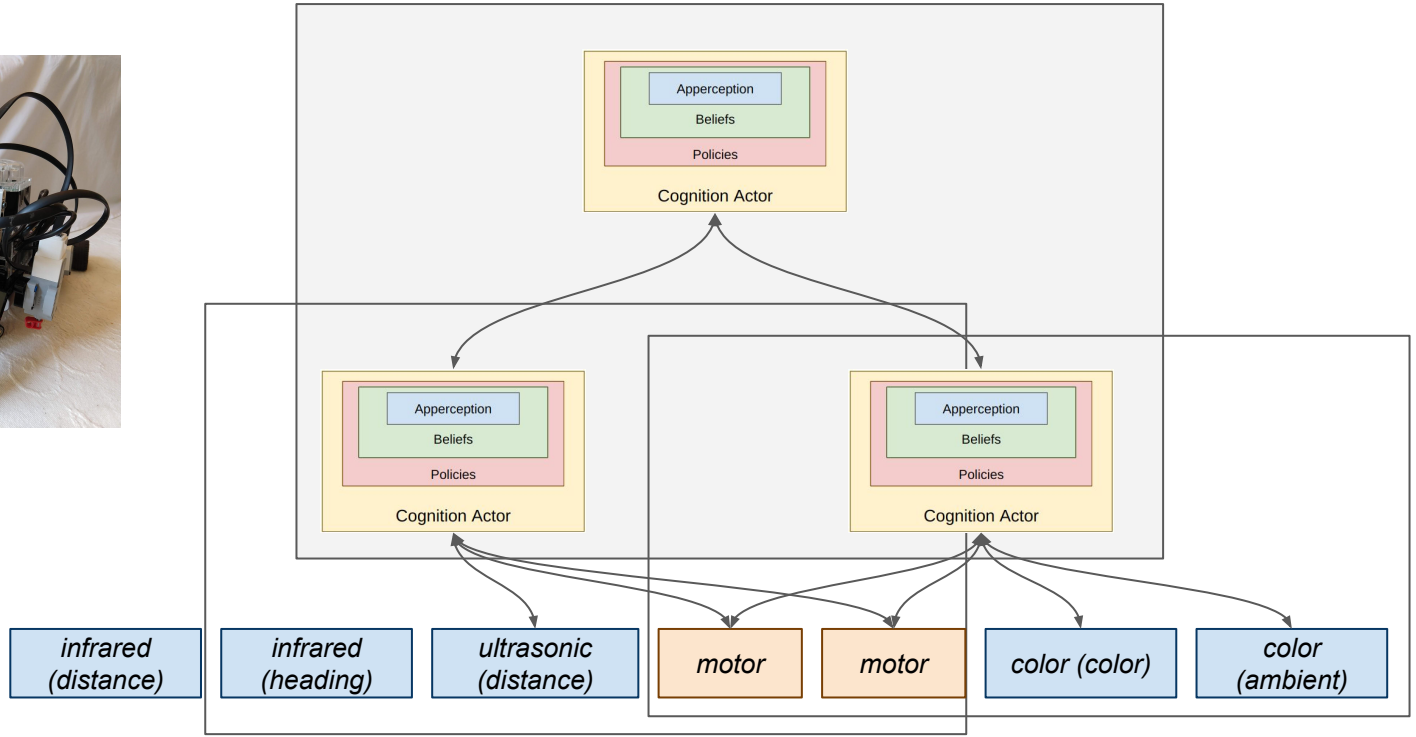
Each new Cognition Actor (CA) would select an umwelt of lower-level CAs



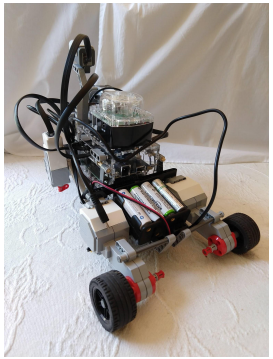


Cognition Actors would organize into an abstraction hierarchy

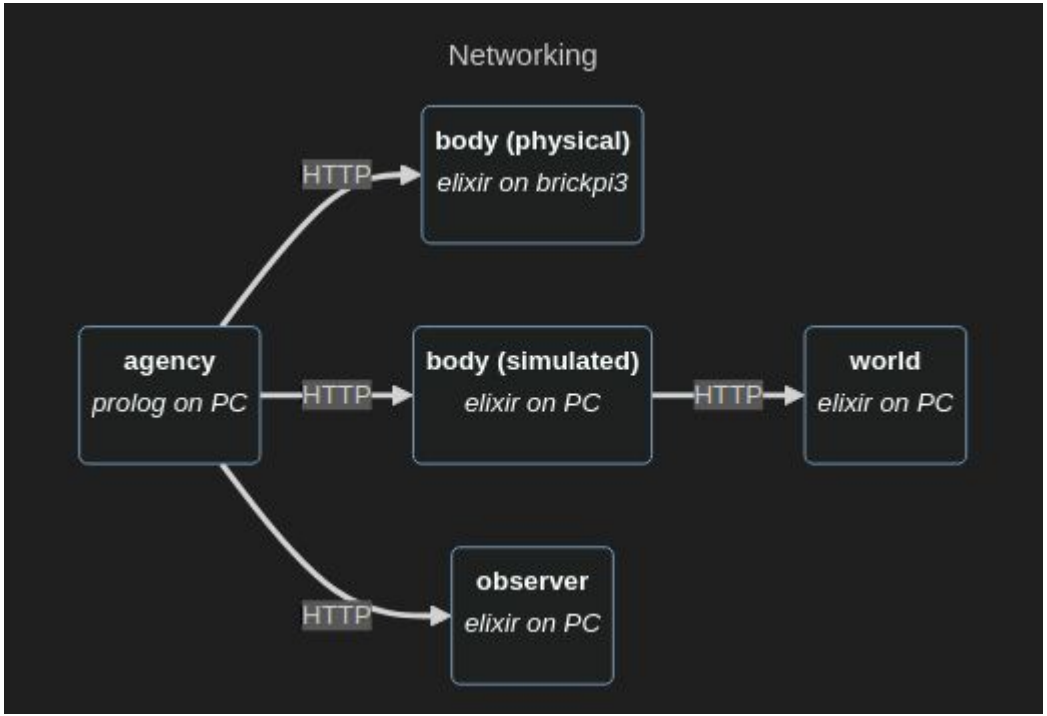




The society of mind would grow, shrink and evolve as needed to survive

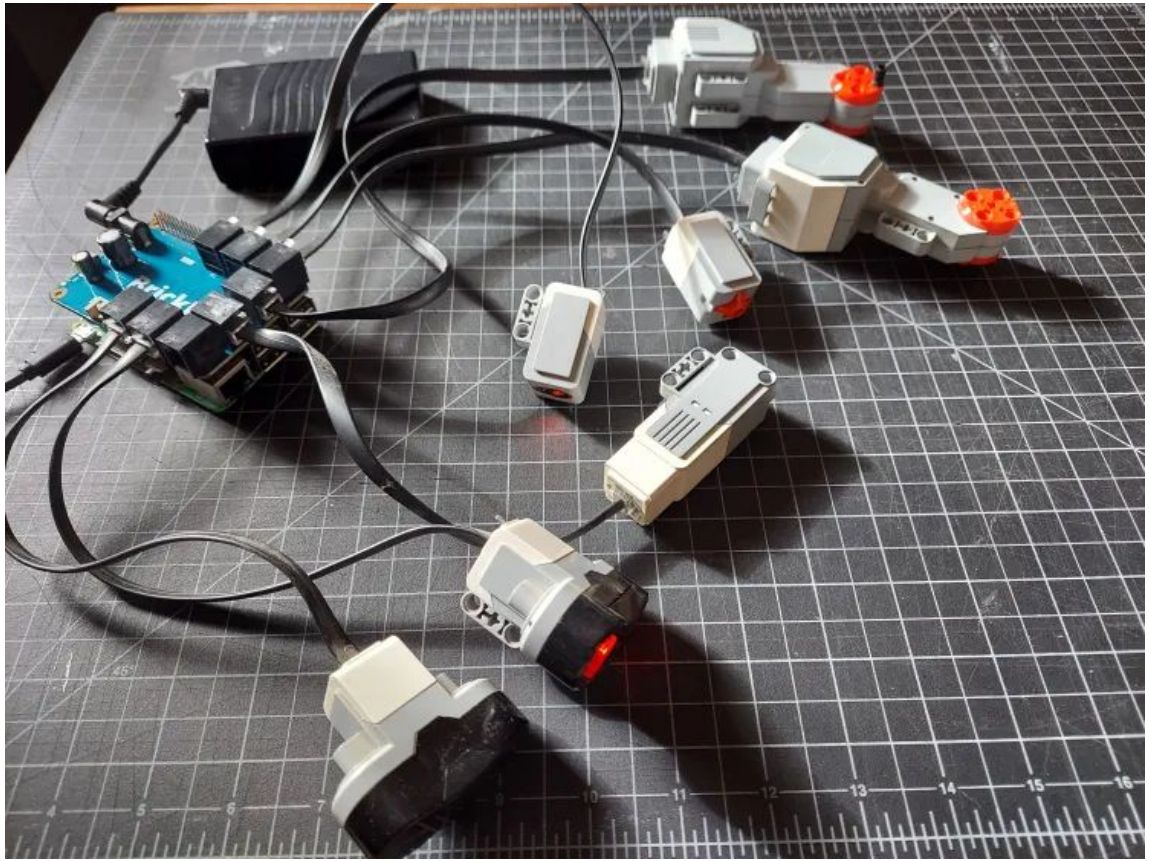


Work in progress: The Karma system

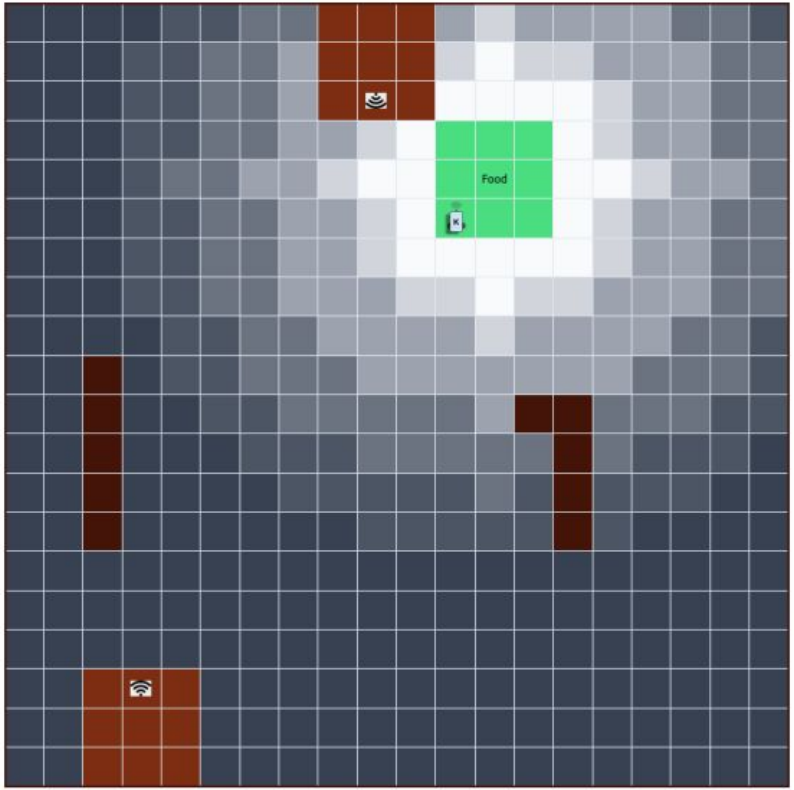


See https://github.com/jfcloutier/karma_system

Karma Body: Access to real and virtual sensors and effectors



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

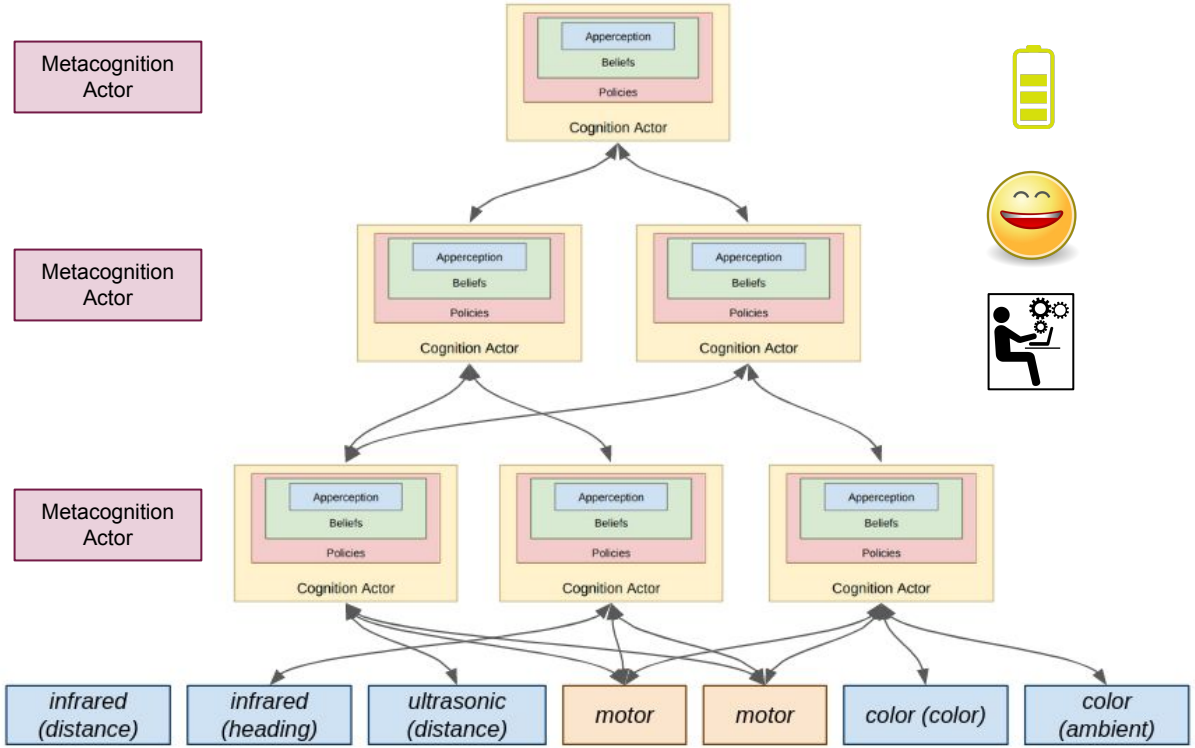


Karma Agency: A generative process for the robot's society of mind

Wellbeing

Metacognition Actors

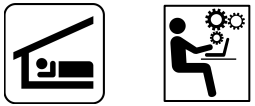
Cognition Actors



Wellbeing

Fullness 

Integrity 

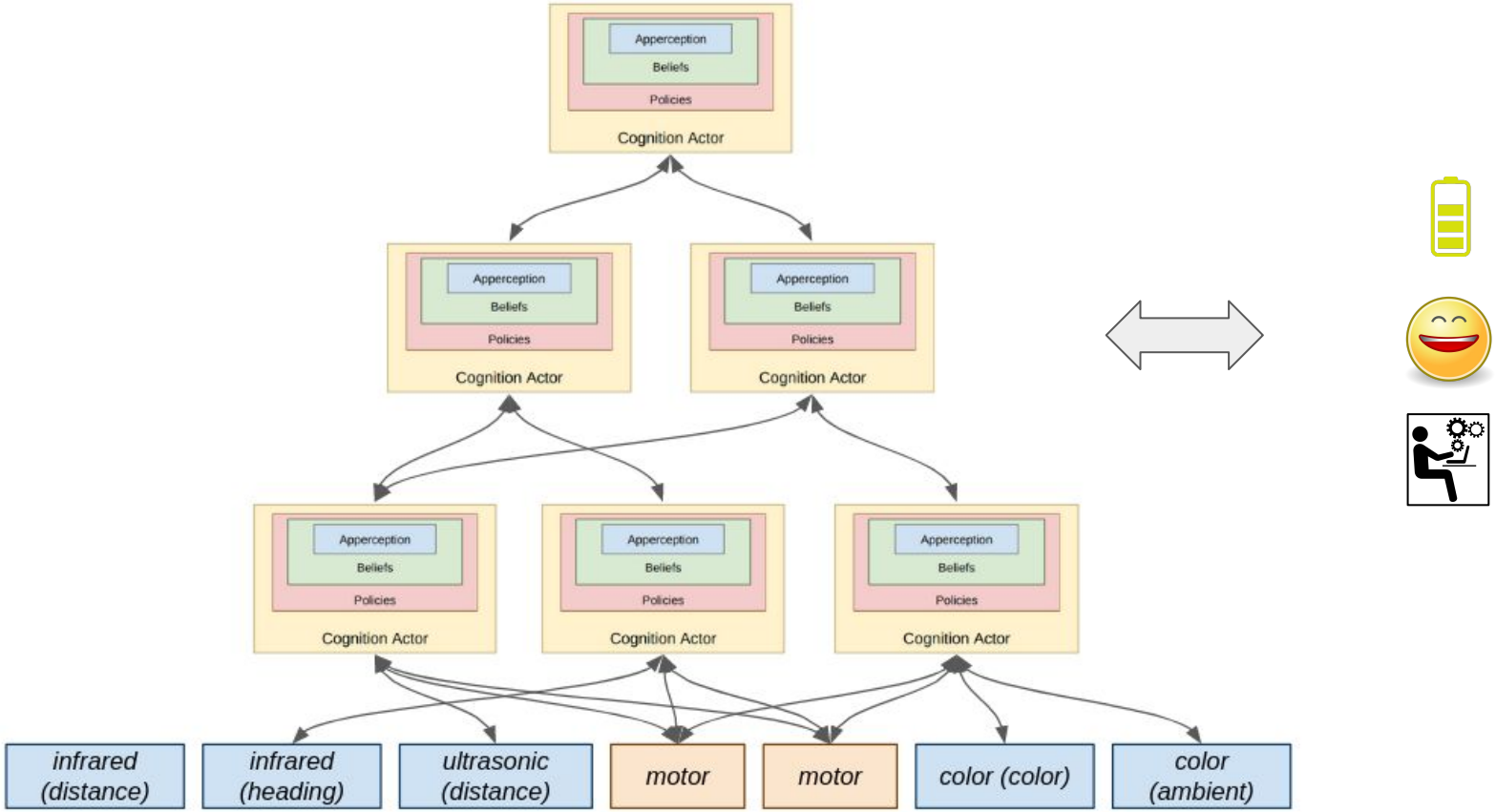
Engagement 

The robot strives to maximize predefined wellbeing measures

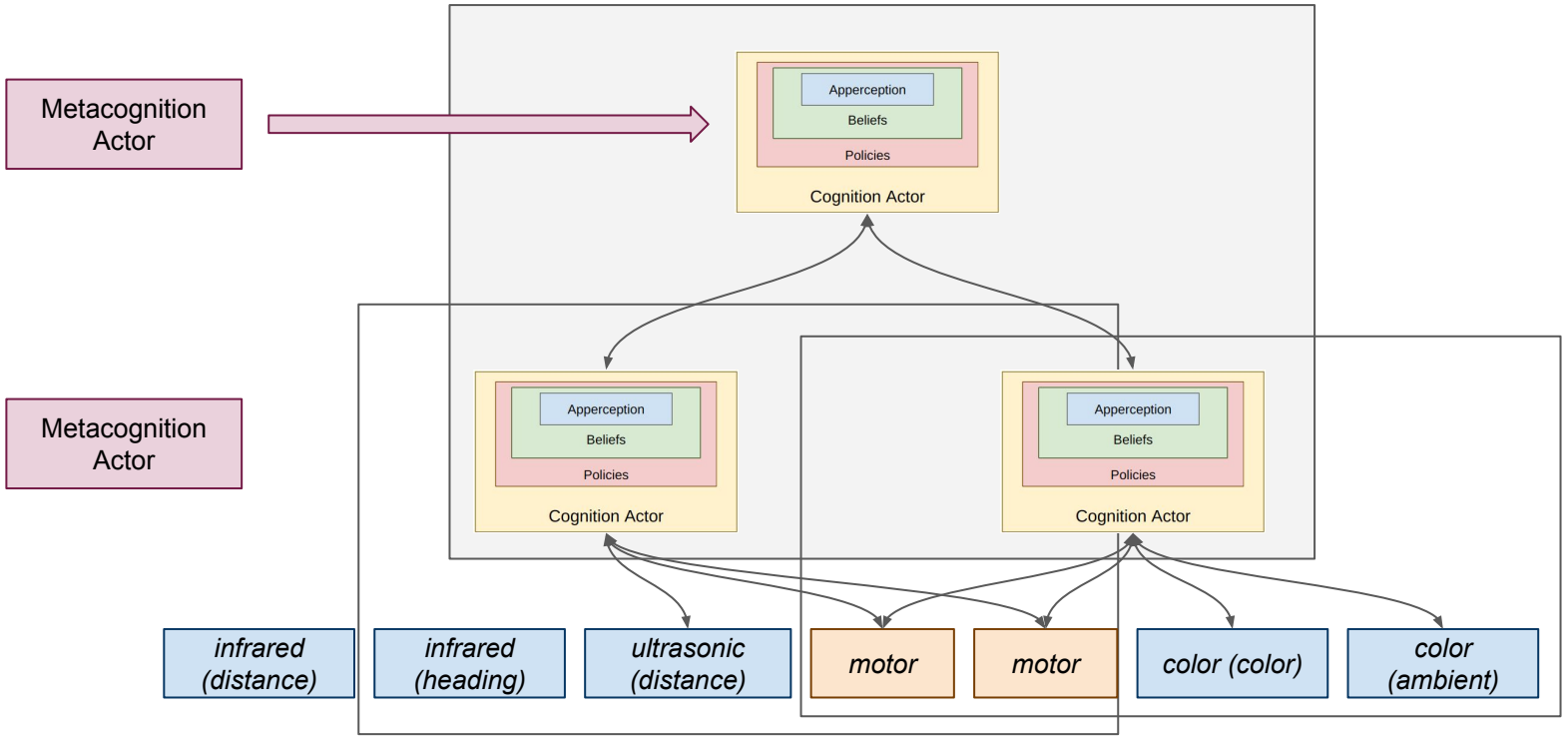
Low measures signal risk to the survival of the society of mind

Wellbeing imparts normativity to beliefs, focuses attention, motivates action

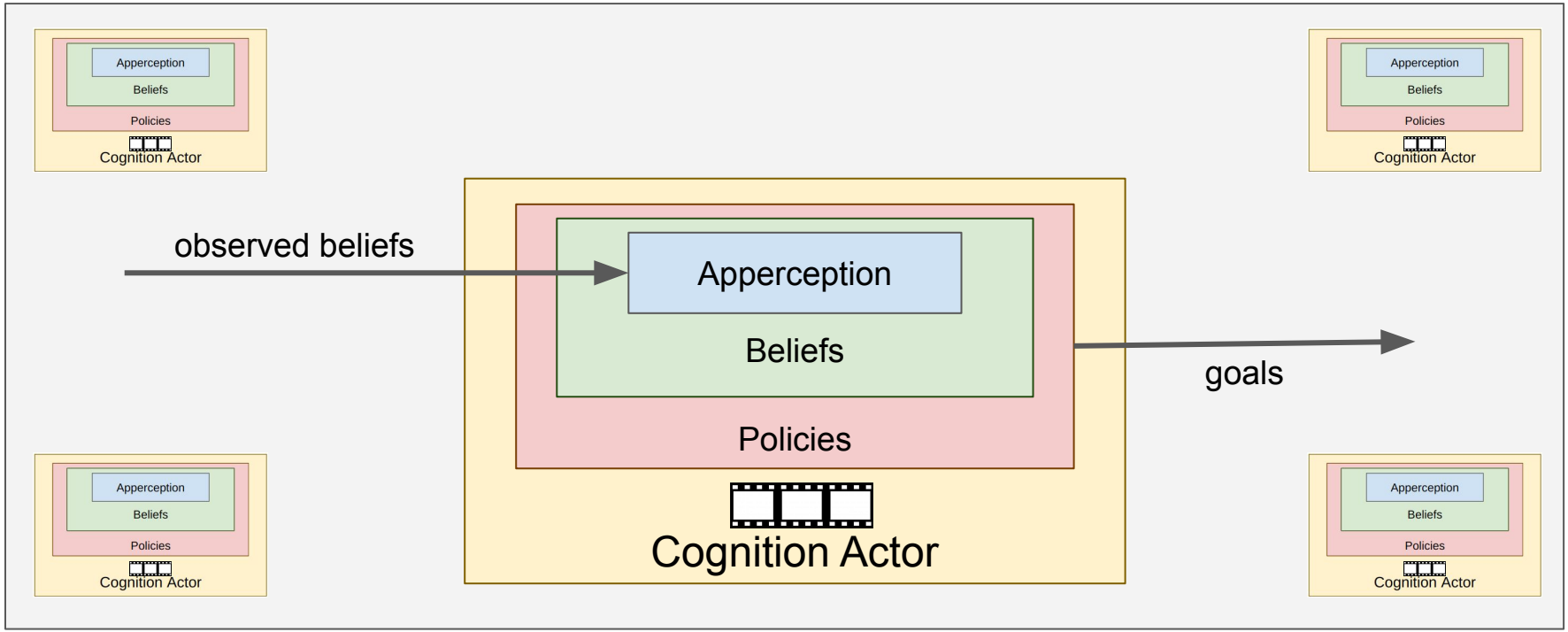
Wellbeing aggregates metrics from all Cognition Actors



Metacognition Actors add and remove Cognition Actors at their level

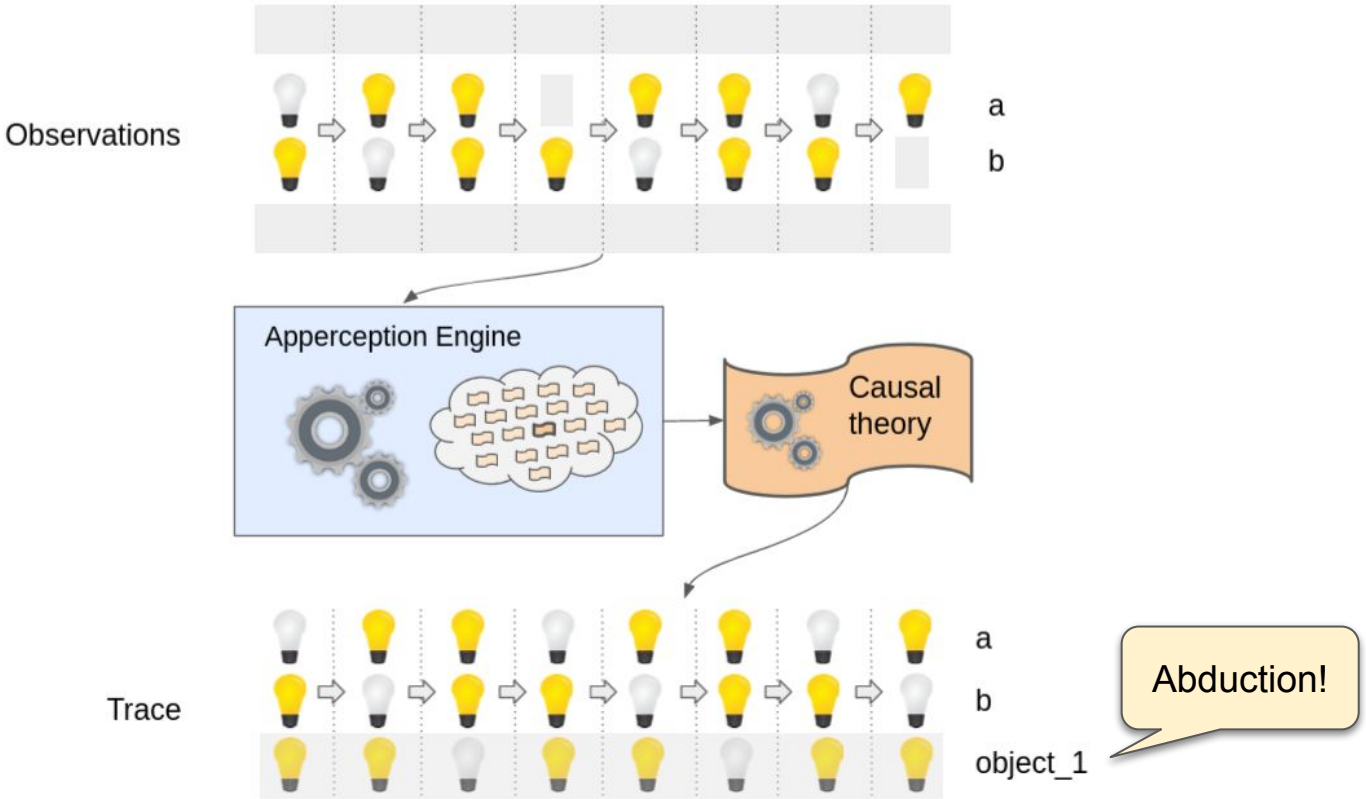


Cognition Actors observe, believe and act within their umwelts one time frame at a time



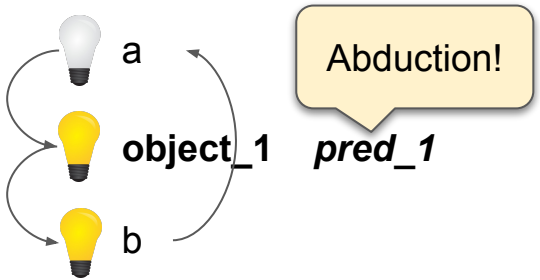
The more abstract the CA, the longer its time frame

Apperception is making sense of *direct* observations



A discovered causal theory

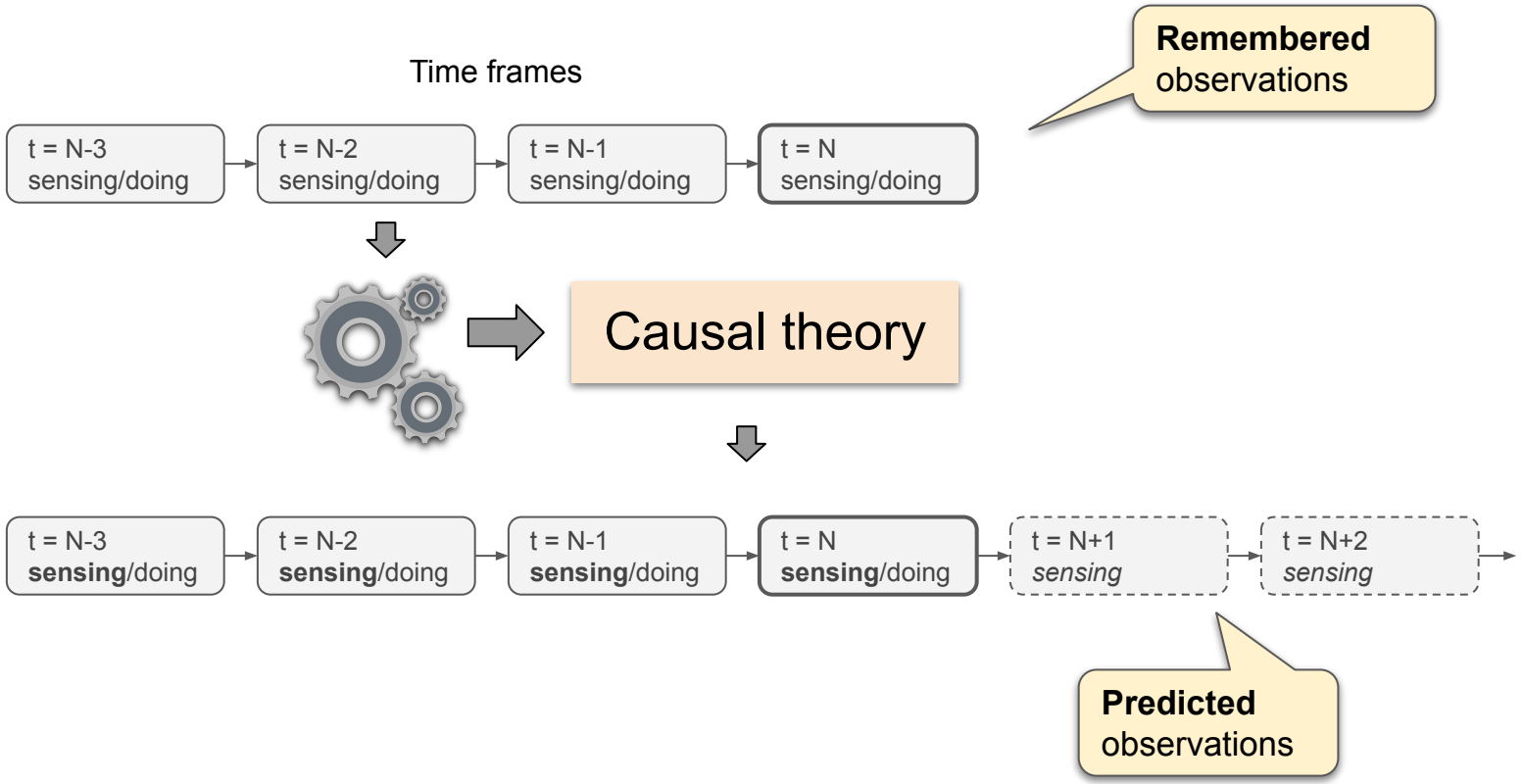
```
rating:100-22

found_time:64

static_constraints:[one_related(pred_1)]

static_rules:[on(_A, true)-[pred_1(_B, _A), on(_B, false)]]

causal_rules:[on(_C, false)-[pred_1(_C, _D), on(_D, false)]]

initial_conditions:[pred_1(object_1, b), pred_1(b, a), pred_1(a, object_1), on(object_1, true), on(b, true)
on(a, false)]
```



A light is related to one and only one other light via pred_1

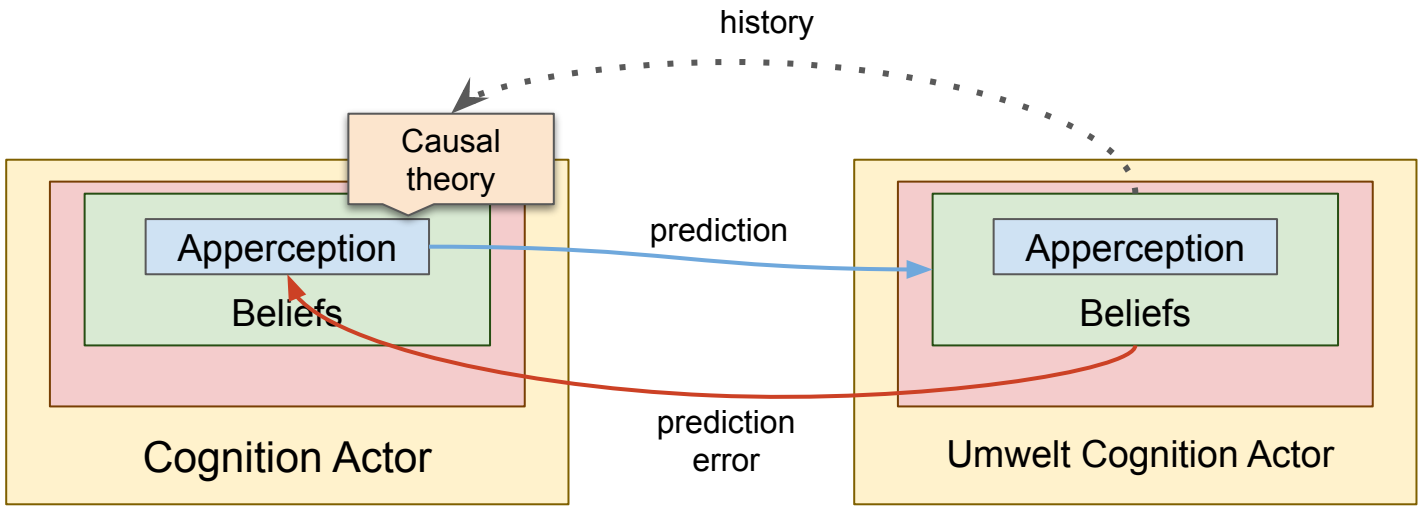
A light is on whenever another that is pred_1 to it is off

A light turns off if it is pred_1 to another light that was off

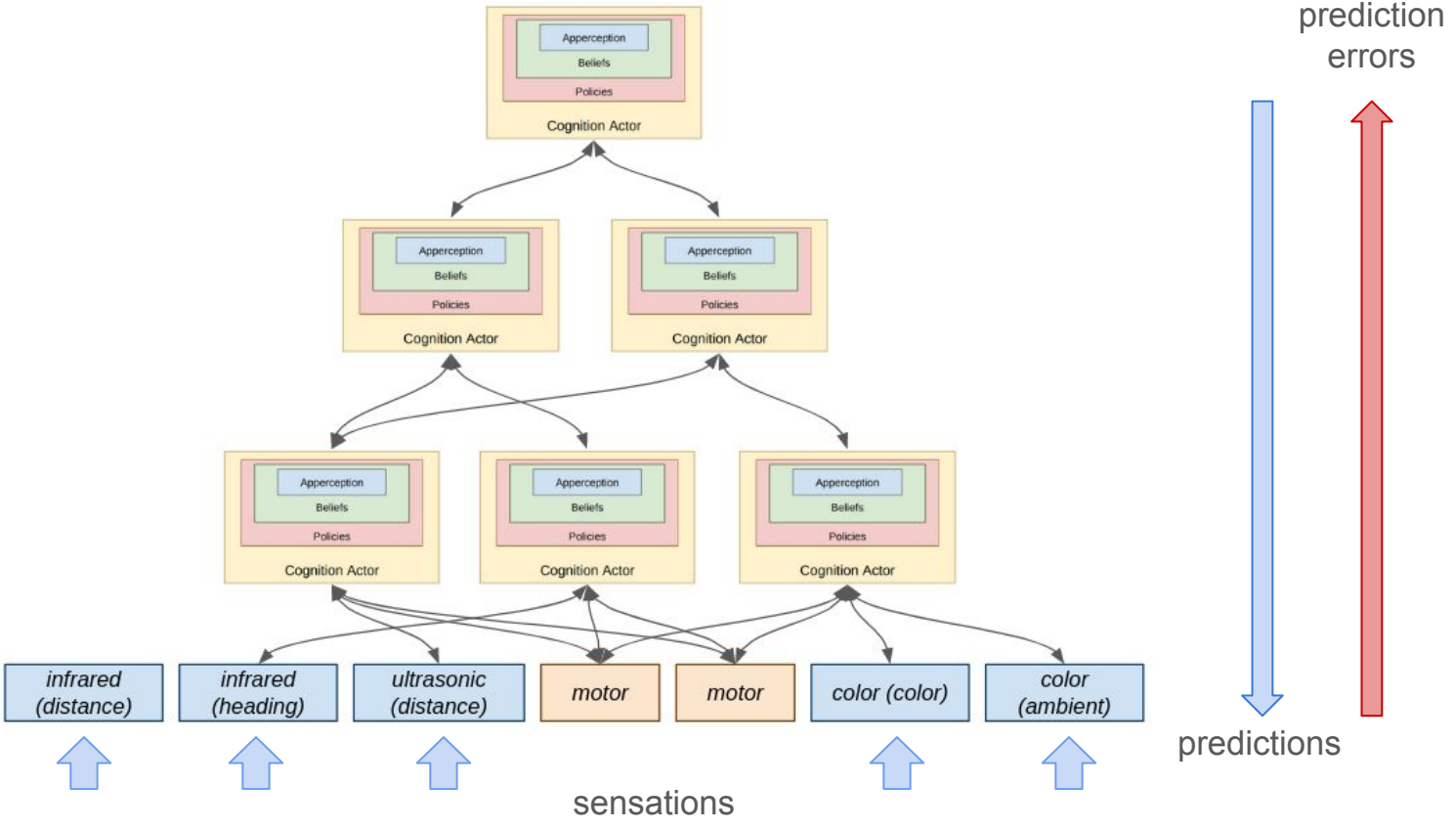


See https://zenodo.org/records/10325868

observations = uncontested predictions + prediction errors



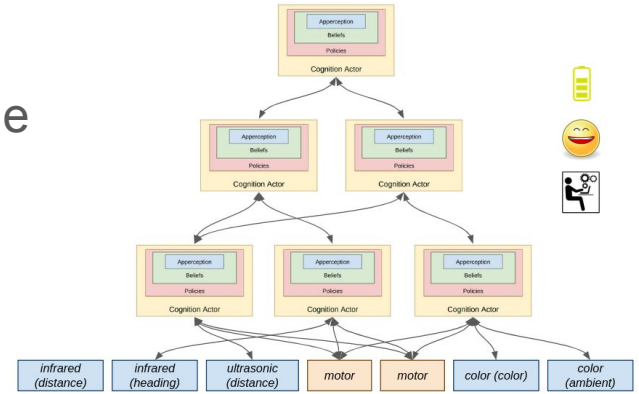
Predictive processing



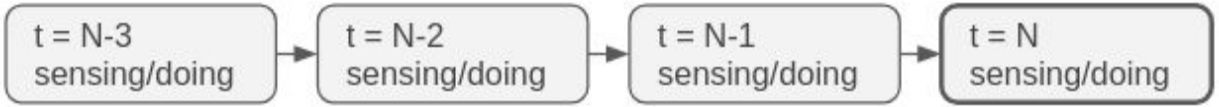
A Cognition Actor updates its beliefs in each time frame

From the analysis of its observations and from remembered attempts to change its own beliefs

Beliefs are pleasant or unpleasant depending on the robot's wellbeing



Kinds of beliefs



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

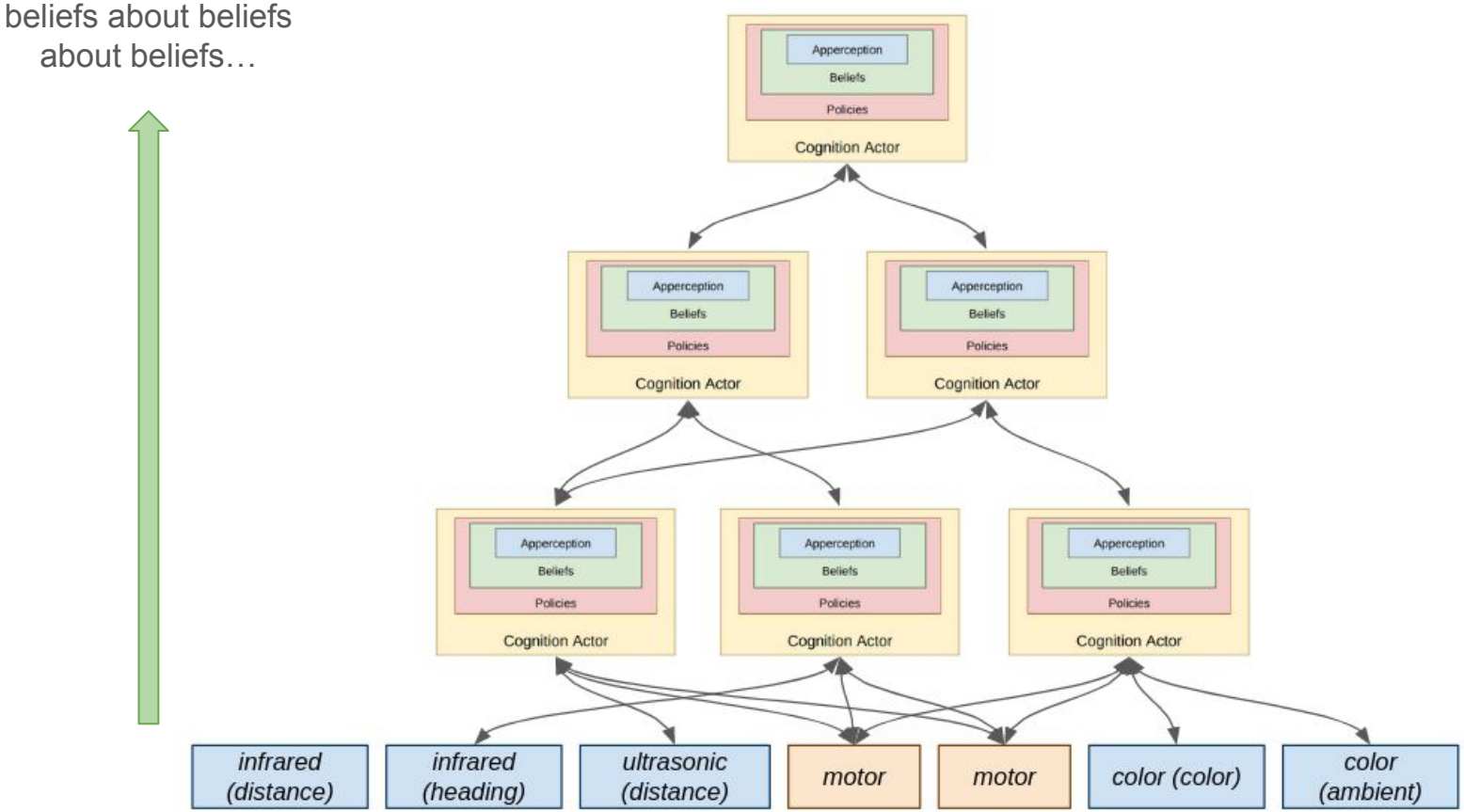
Count *I am next to **2** things*

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

Ending *Getting closer to a thing **stopped***

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

One Cognition Actor's belief is another's observation



A Cognition Actor acts via policies it formulates

A CA strives to invalidate its unpleasant beliefs and validate its pleasant beliefs

A CA impacts a belief it holds by impacting observations that led to this belief

The observations are beliefs of umwelt CAs

A CA formulates a policy as a list of goals

Each goal is a desired impact to an umwelt belief

The goals listed in a policy are themselves realized by policies

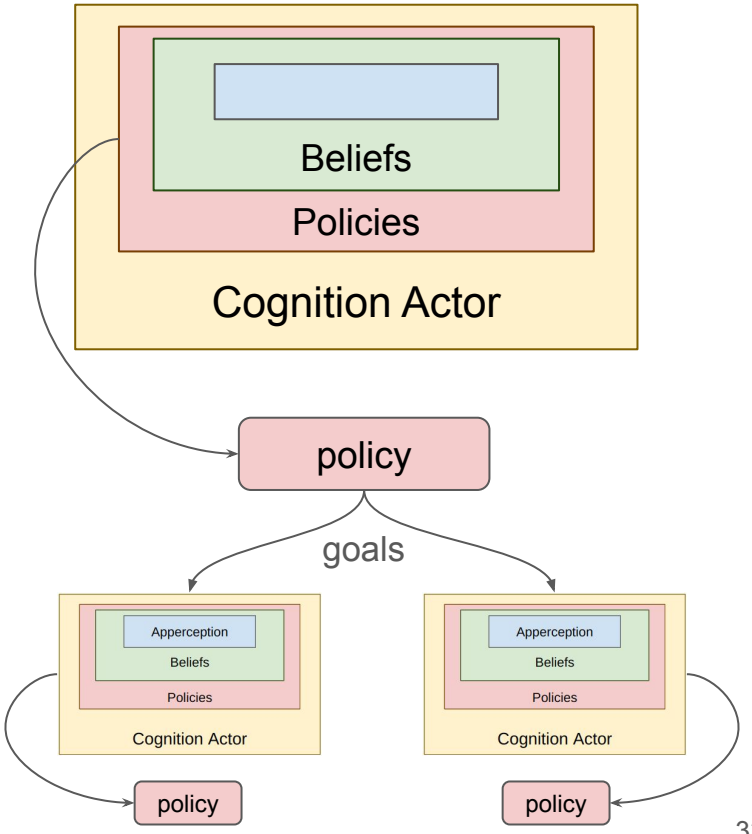
A CA tells its umwelt CAs which of their beliefs to validate/invalidate (goals)

Each umwelt CA formulates its own policy to achieve a received goal

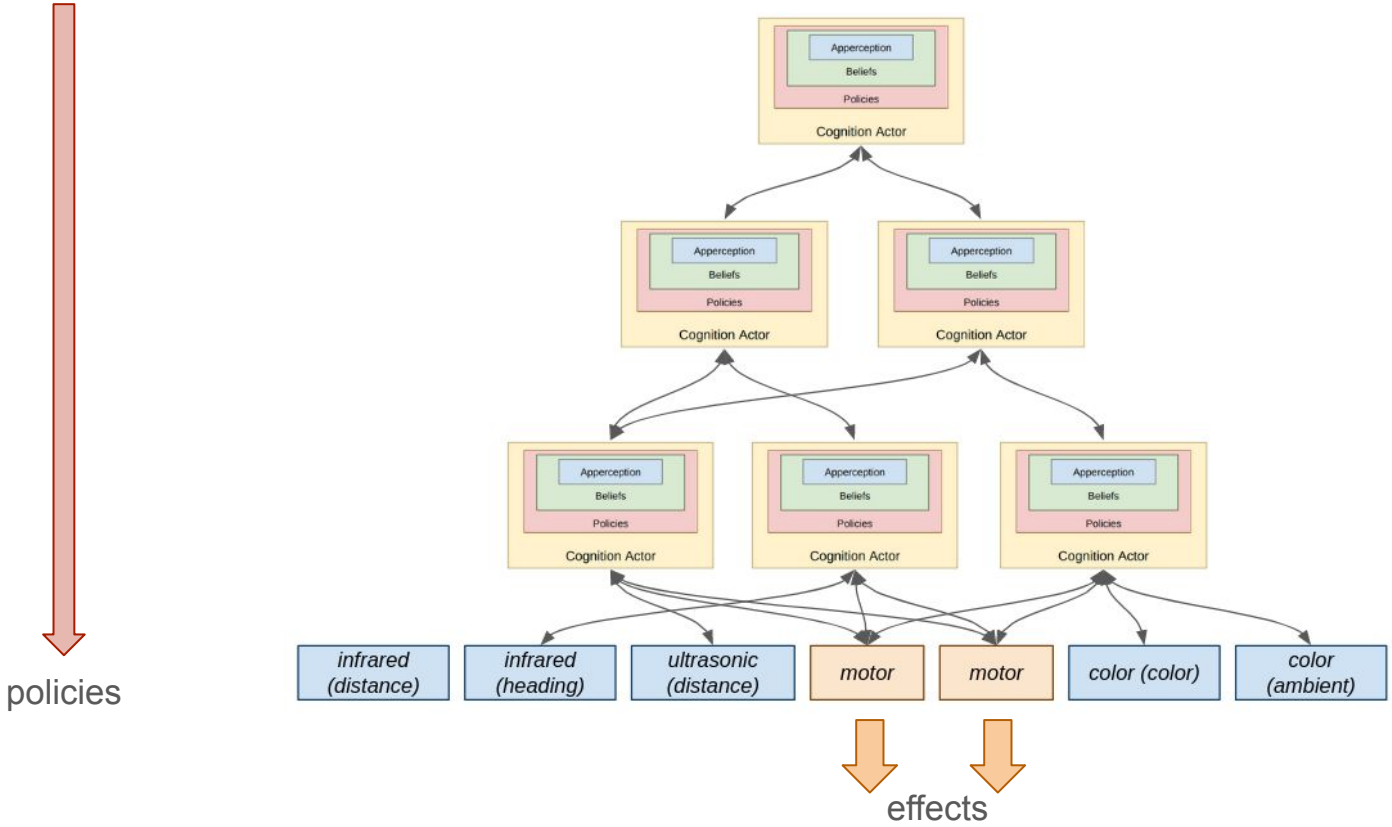
A CA can execute only one policy at a time

A CA determines policy success or failure from changes in its beliefs

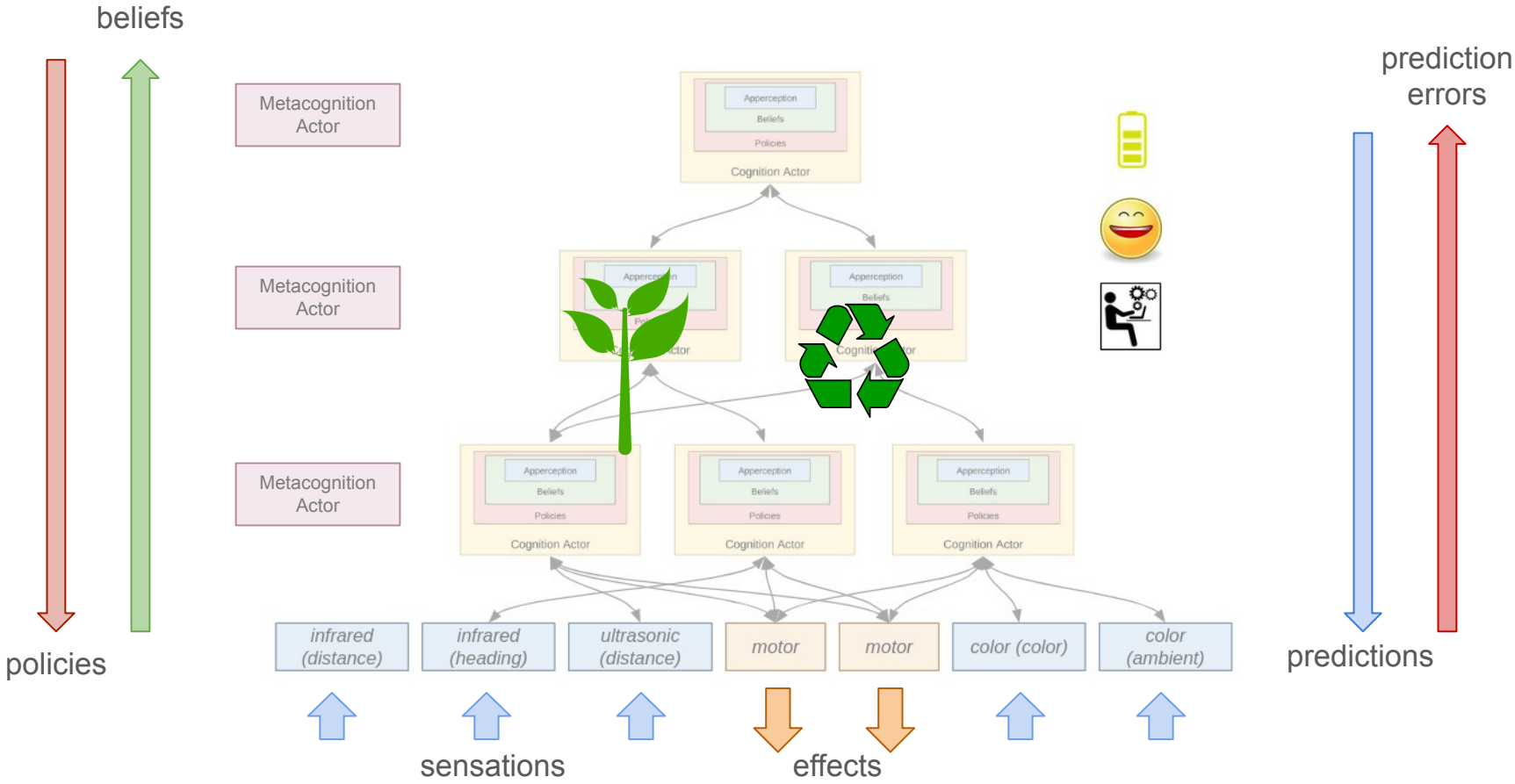
A policy known to work is reused (habits)



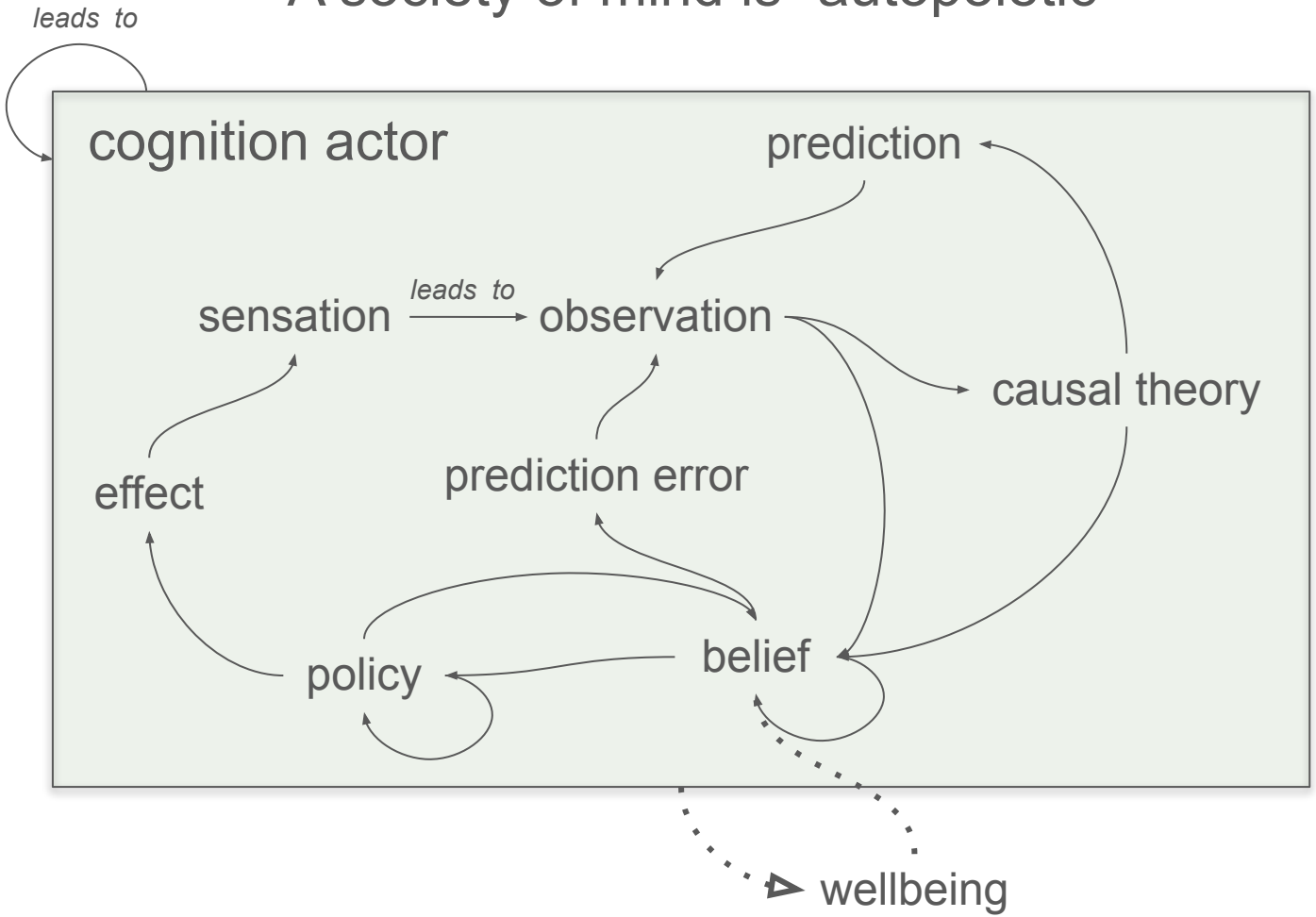
Action is policies all the way down



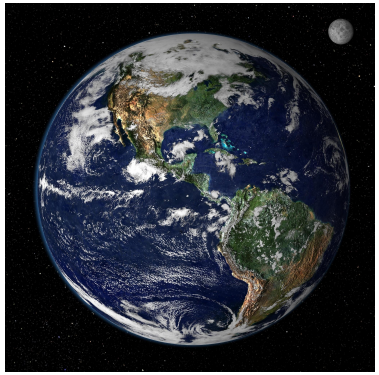
Agency emerges from operational closure

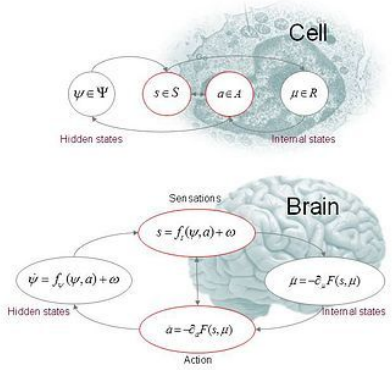


A society of mind is “autopoietic”

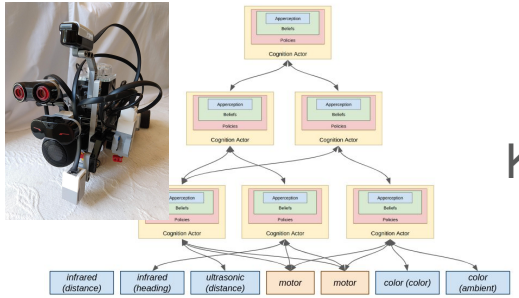


Map vs territory

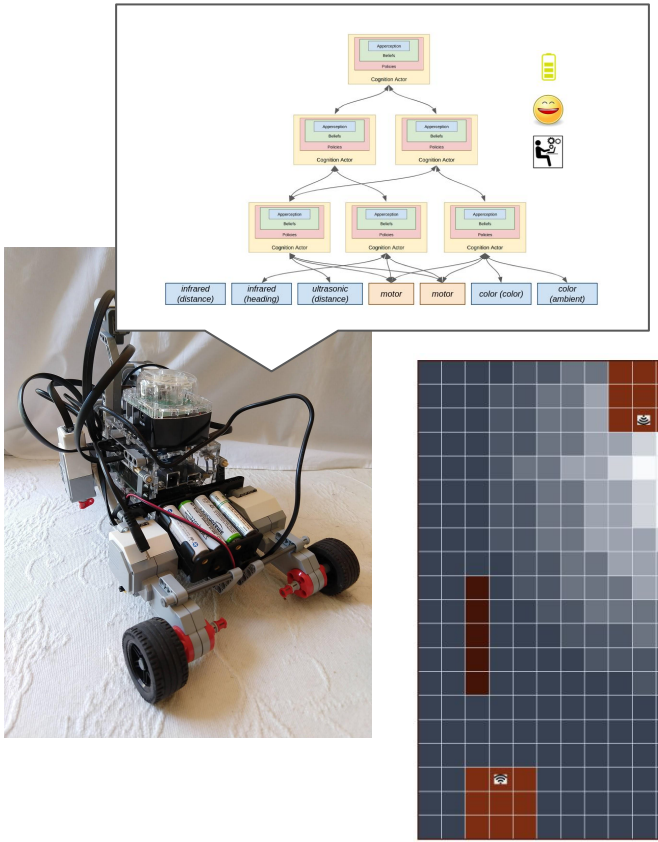




FEP



Karma



I am programming generative *processes*

Active Inference analysis with its generative *models* etc. comes after

To do

Implement Agency’s Wellbeing, Metacognition Actor, Cognition Actor…

Implement Karma Observer

Build a new robot

Setup real-life test environment

Gather data from runs (RL and virtual)

Carry out FEP analysis on the data