Peeling back the onion of Computer Vision



Kushagra Tiwary, PhD Student, MIT
kushagratiwary.com
October 26, 2026

Peeling Onions, Lilly Martin Spencer, ca. 1852

Goal #1 of this talk: Convince you that we have been slowly peeling back at the “onion of computer vision”

Each peel is a human-made abstraction to an underlying algorithm

Human-Abstraction 4:
Perception-Action loops *(current)*

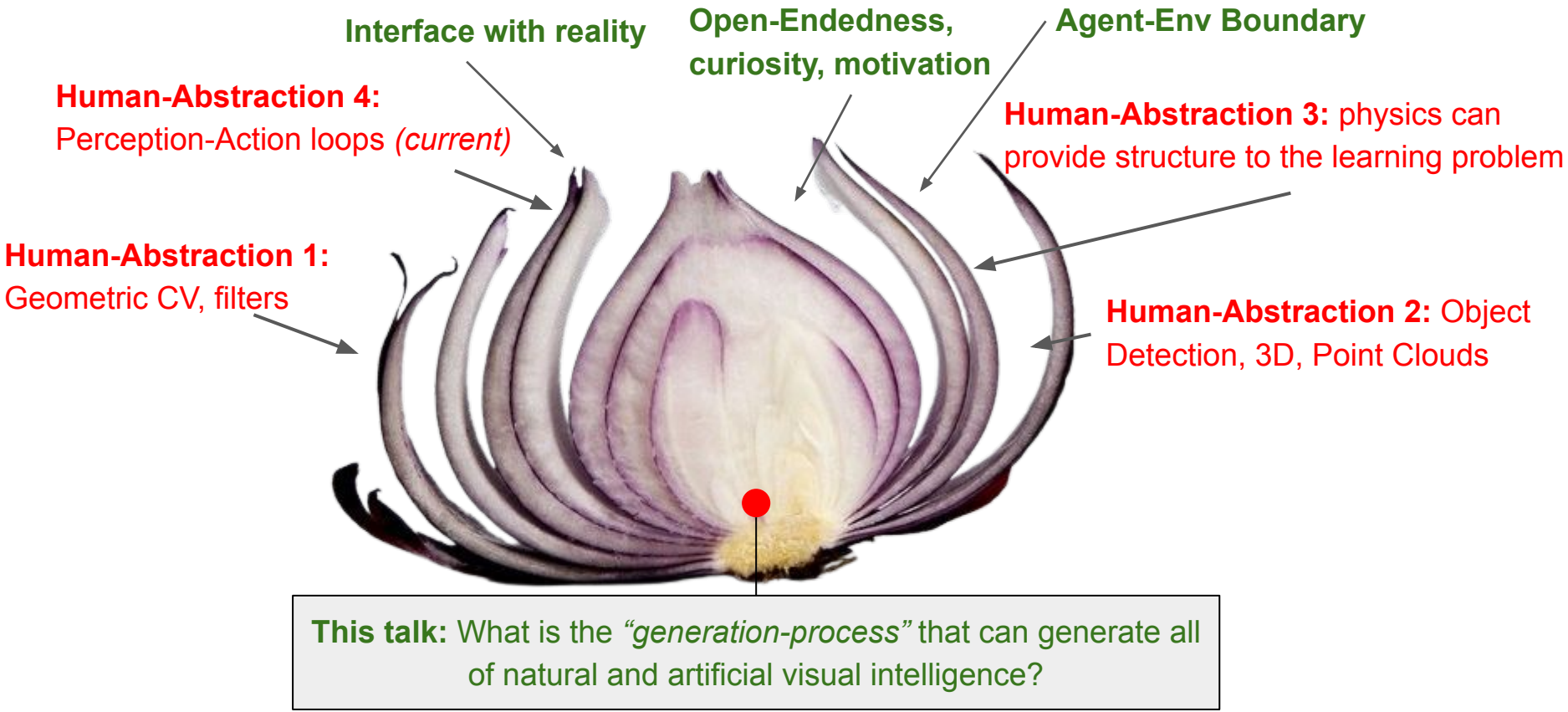
Human-Abstraction 3: physics can provide structure to the learning problem

Human-Abstraction 2: Object Detection, 3D, Point Clouds

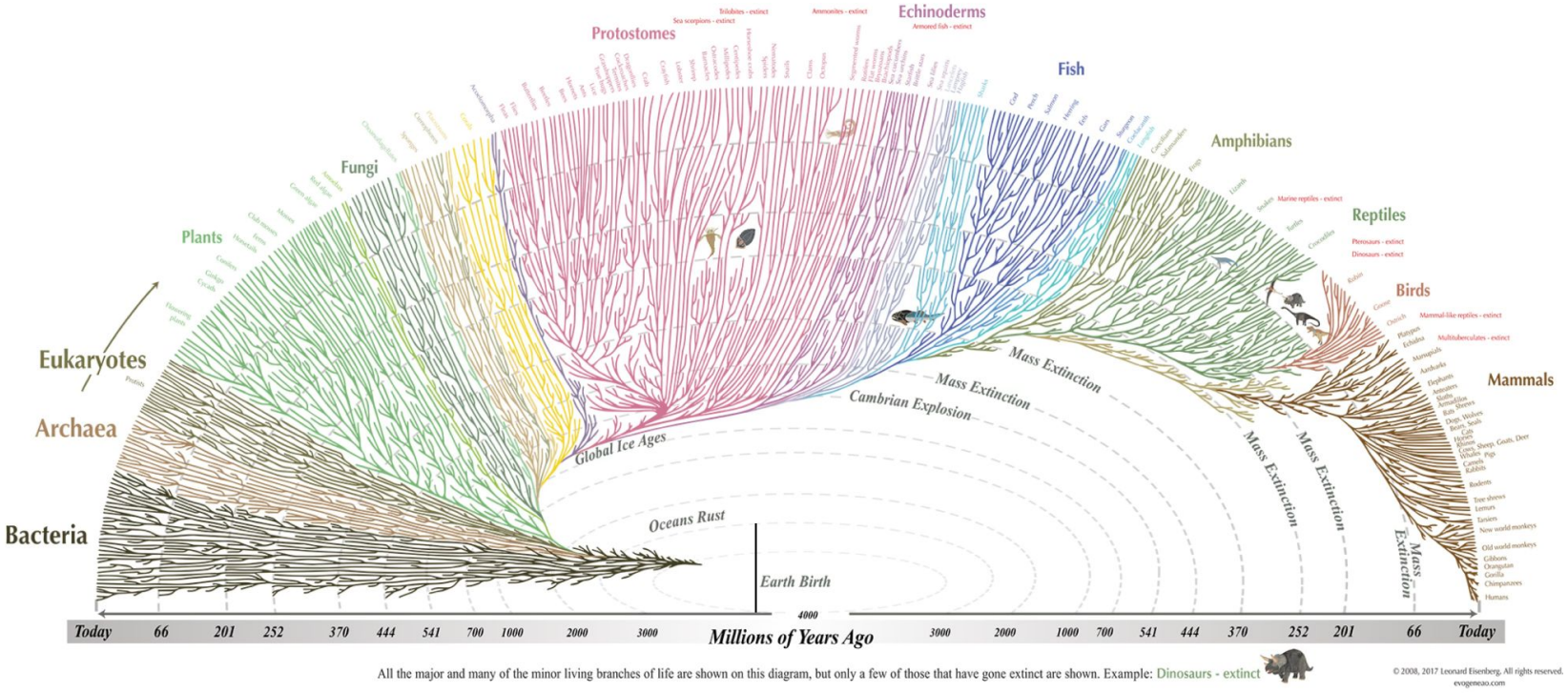
Human-Abstraction 1:
Geometric CV, filters



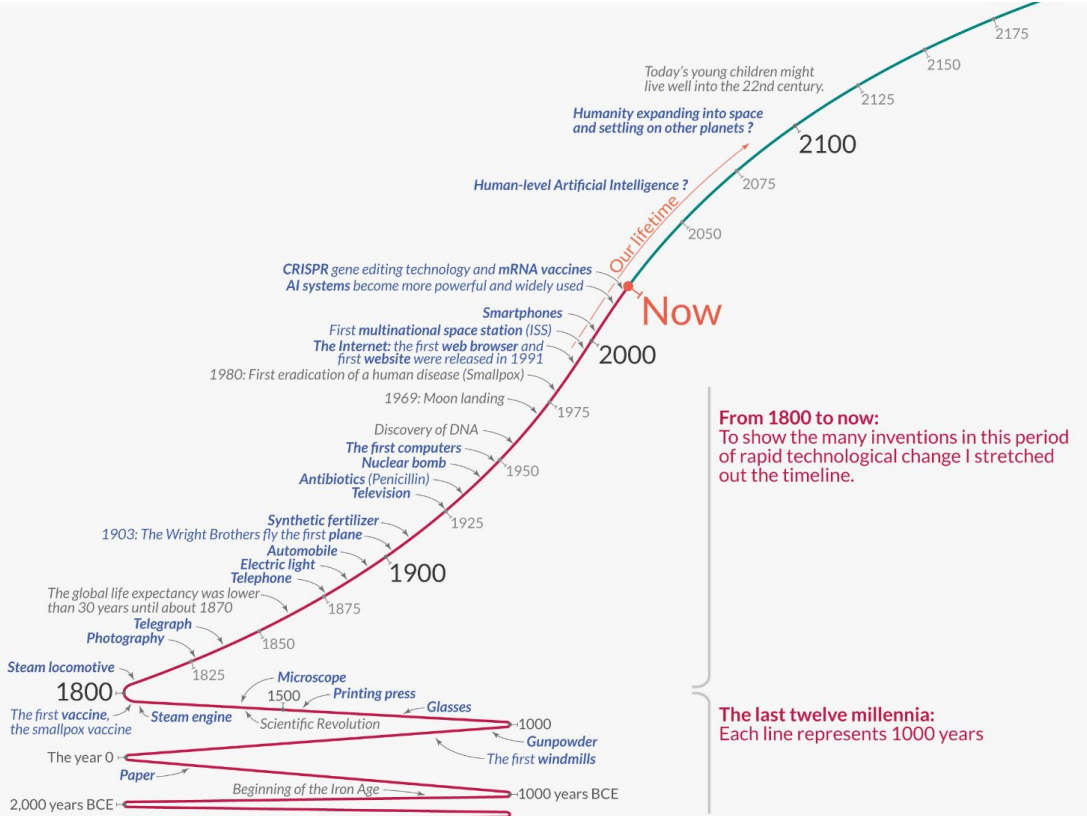
Goal #2 of this talk: Convince you that at the center of the peel lies an algorithm that generates all of vision



Biological evidence of this “generation-process”



Technological evidence of this “generation-process”



Max Roser, [Technology over the long run: zoom out to see how dramatically the world can change within a lifetime]

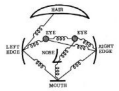
The dominant abstraction during my undergrad

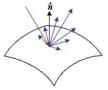
Human-Abstraction 2: Object Detection, 3D, Point Clouds

Human-Abstraction 1: Geometric CV, filters



The dominant abstraction in class















1 Introduction 1

What is computer vision? • A brief history • Book overview • Sample syllabus • Notation

2 Image formation 29

Geometric primitives and transformations • Photometric image formation • The digital camera

3 Image processing 99

Point operators • Linear filtering • More neighborhood operators • Fourier transforms • Pyramids and wavelets • Geometric transformations • Global optimization

4 Feature detection and matching 205

Points and patches • Edges • Lines

5 Segmentation 267

Active contours • Split and merge • Mean shift and mode finding • Normalized cuts • Graph cuts and energy-based methods

6 Feature-based alignment 309

2D and 3D feature-based alignment • Pose estimation • Geometric intrinsic calibration

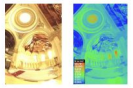
7 Structure from motion 343

Triangulation • Two-frame structure from motion • Factorization • Bundle adjustment • Constrained structure and motion

Human-made, *but useful,* abstractions to the vision problem















8 Dense motion estimation 381

Translational alignment • Parametric motion • Spline-based motion • Optical flow • Layered motion

9 Image stitching 427

Motion models • Global alignment • Compositing

10 Computational photography 467

Photometric calibration • High dynamic range imaging • Super-resolution and blur removal • Image matting and compositing • Texture analysis and synthesis

11 Stereo correspondence 533

Epipolar geometry • Sparse correspondence • Dense correspondence • Local methods • Global optimization • Multi-view stereo

12 3D reconstruction 57

Shape from X • Active rangefinding • Surface representations • Point-based representations • Volumetric representations • Model-based reconstruction • Recovering texture maps and albedos

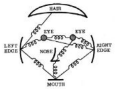
13 Image-based rendering 619

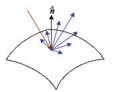
View interpolation • Layered depth images • Light fields and Lumigraphs • Environment mattes • Video-based rendering

14 Recognition 655

Object detection • Face recognition • Instance recognition • Category recognition • Context and scene understanding • Recognition databases and test sets

Became pretty obvious that these abstractions were no longer sufficient when I interned at a self-driving startup…













1 Introduction 1

What is computer vision? • A brief history • Book overview • Sample syllabus • Notation

2 Image formation 29

Geometric primitives and transformations • Photometric image formation • The digital camera

3 Image processing 99

Point operators • Linear filtering • More neighborhood operators • Fourier transforms • Pyramids and wavelets • Geometric transformations • Global optimization

4 Feature detection and matching 205

Points and patches • Edges • Lines

5 Segmentation 267

Active contours • Split and merge • Mean shift and mode finding • Normalized cuts • Graph cuts and energy-based methods

6 Feature-based alignment 309

2D and 3D feature-based alignment • Pose estimation • Geometric intrinsic calibration

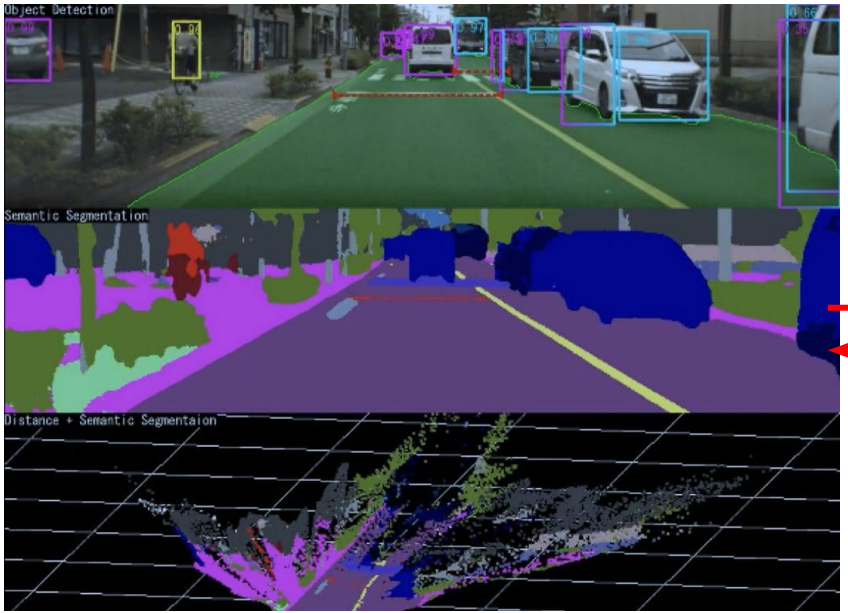
7 Structure from motion 343

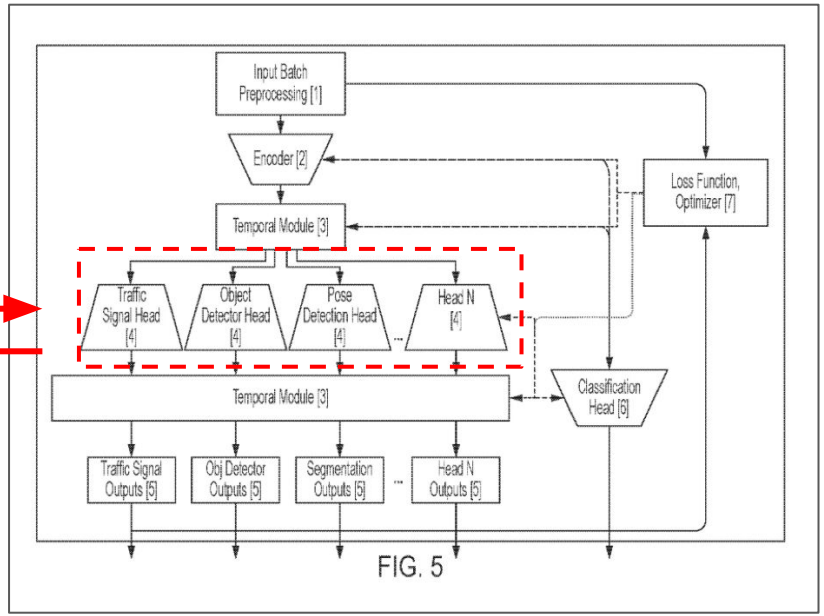
Triangulation • Two-frame structure from motion • Factorization • Bundle adjustment • Constrained structure and motion



What should the abstractions look like for self driving?

Instead use task based abstractions for self-driving





scale

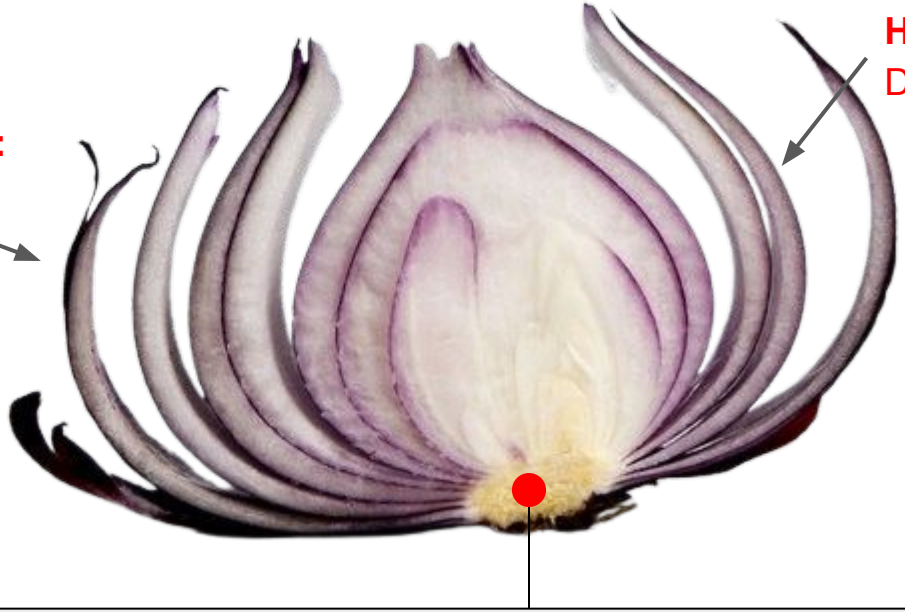
Software 2.0-type learning system

Tiwary, Almeida; Efficient detection of structure and status of traffic lights

Labeling isn't scalable, use light transport (physics) and self-supervised learning to support abstractions

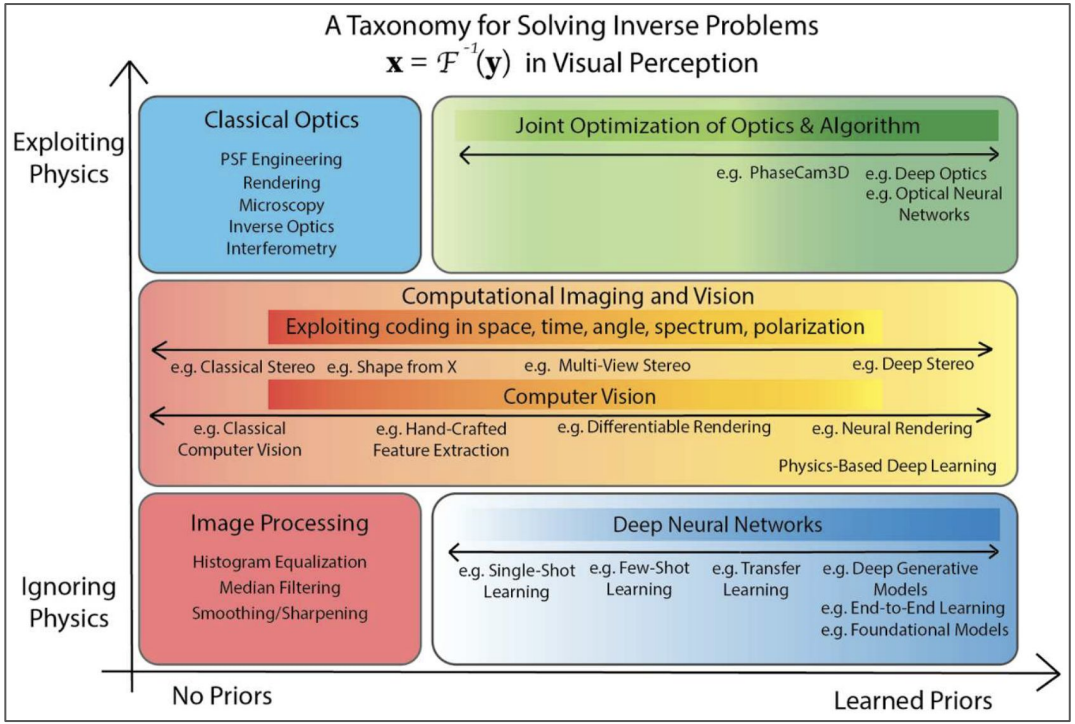
Human-Abstraction 1:
Geometric CV, filters

Human-Abstraction 2: Object Detection, 3D, Point Clouds



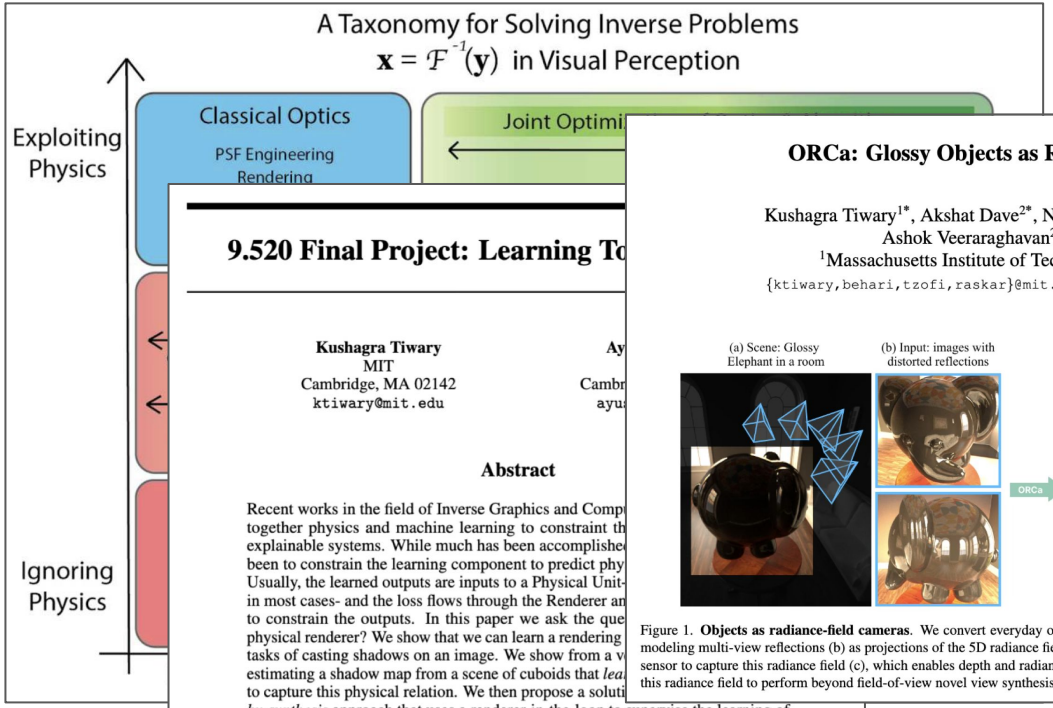
New abstraction (2020-2022): physics with machine learning can support abstractions

Hypothesis early in my PhD: physics with machine learning can support abstractions



Klinghoffer*, Somsaundam*, Tiwary*, Physics vs. Learned Priors: Rethinking Camera and Algorithm Design for Task-Specific Imaging, ICCP 2022

Hypothesis early in my PhD: Physics helps provide structure to the learning problem



9.520 Final Project: Learning To

Kushagra Tiwary
MIT
Cambridge, MA 02142
ktiwary@mit.edu

Ay
Cambr
ayu

Abstract

Recent works in the field of Inverse Graphics and Comp
together physics and machine learning to constraint th
explainable systems. While much has been accomplishe
been to constrain the learning component to predict phy
Usually, the learned outputs are inputs to a Physical Unit-
in most cases- and the loss flows through the Renderer an
to constrain the outputs. In this paper we ask the que
physical renderer? We show that we can learn a rendering
tasks of casting shadows on an image. We show from a v
estimating a shadow map from a scene of cuboids that *lea*
to capture this physical relation. We then propose a soluti
by-synthesis approach that uses a renderer-in-the-loop to supervise the learning of a renderer. Ultimately, we hope to use this method in important problems like 3D scene reconstruction, re-lighting and also provide important insight into the field of physics driven machine learning. The code is available here.

ORCa: Glossy Objects as Radiance-Field Cameras

Kushagra Tiwary[1*], Akshat Dave[2*], Nikhil Behari[1], Tzofi Klinghoffer[1],
Ashok Veeraraghavan[2], Ramesh Raskar[1]
[1]Massachusetts Institute of Technology, [2]Rice University
{ktiwary,behari,tzofi,raskar}@mit.edu, {akshat,vashok}@rice.edu

Figure 1. **Objects as radiance-field cameras**. We convert everyday objects with unknown geometry (a) into radiance-field cameras by modeling multi-view reflections (b) as projections of the 5D radiance field of the environment. We convert the object surface into a virtual sensor to capture this radiance field (c), which enables depth and radiance estimation of the surrounding environment. We can then query this radiance field to perform beyond field-of-view novel view synthesis of the environment (d).

Towards Learning Neural Representations from Shadows

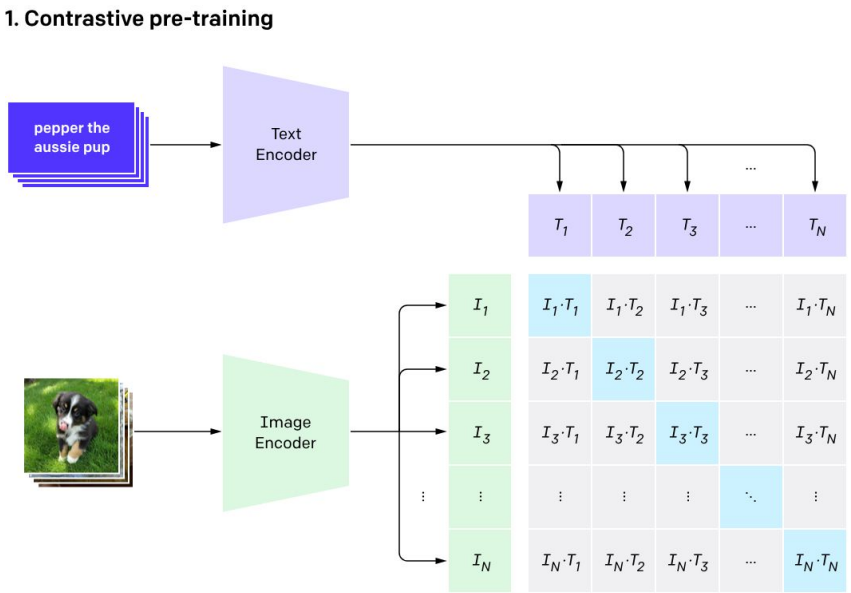
Kushagra Tiwary *, Tzofi Klinghoffer *, and Ramesh Raskar

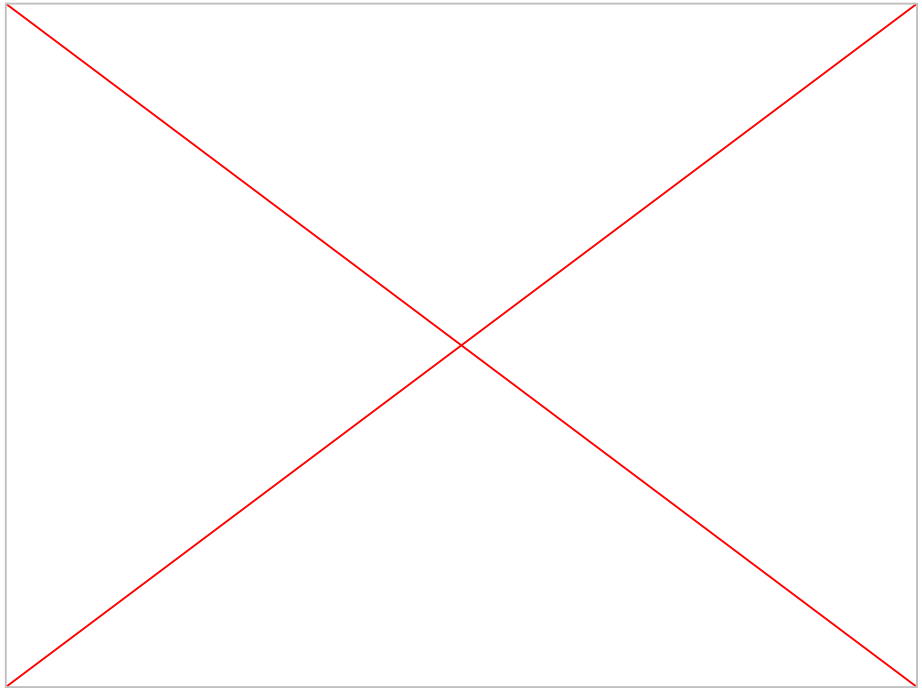
Massachusetts Institute of Technology
{ktiwary, tzofi, raskar}@mit.edu

hod that learns neural shadow fields which
ons that are *only* learnt from the shadows
traditional shape-from-shadow (SfS) algo-
from shadows, they assume a fixed scan-
alize to complex scenes. Neural rendering
d, rely on photometric consistency between
ore physical cues such as shadows, which
valuable information about the scene. We
owerful cue that can constrain neural scene
and even outperform NeRF to reconstruct
Ve propose a graphics-inspired differentiable
hadows with volumetric rendering, predict-
be compared to the ground truth shadow.
v maps, we show that neural rendering can
ate coarse geometry. Our approach reveals
n be used to estimate geometry using dif-
ng. Moreover, our framework is highly gen-
gside existing 3D reconstruction techniques
ometric consistency.

ations, Differentiable Rendering, 3D Scene
Shadows, Volume Rendering

Evidence early in my PhD: ~~Physics~~ **scale** ~~helps~~ provides structure to the learning problem *(physics <-> data are interchangeable)*





Back to work…



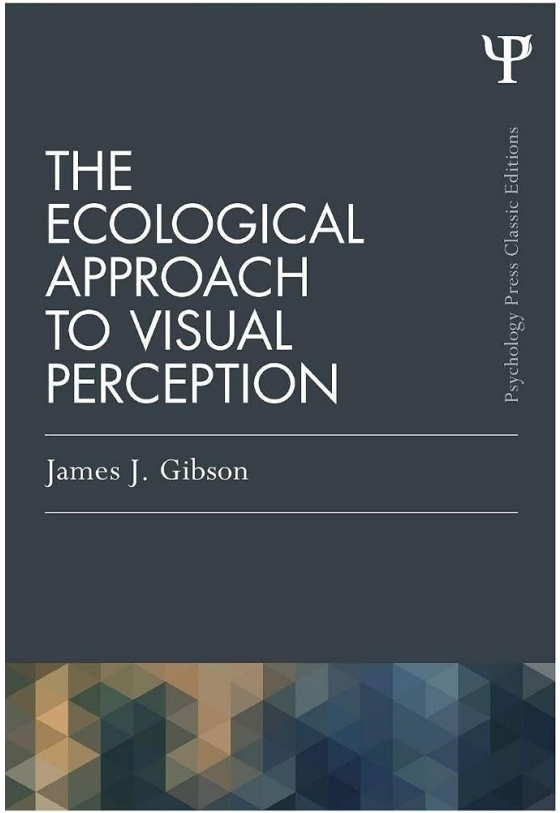
Human-Abstraction 3: physics can provide structure to the learning problem

Human-Abstraction 2: Object Detection, 3D, Point Clouds

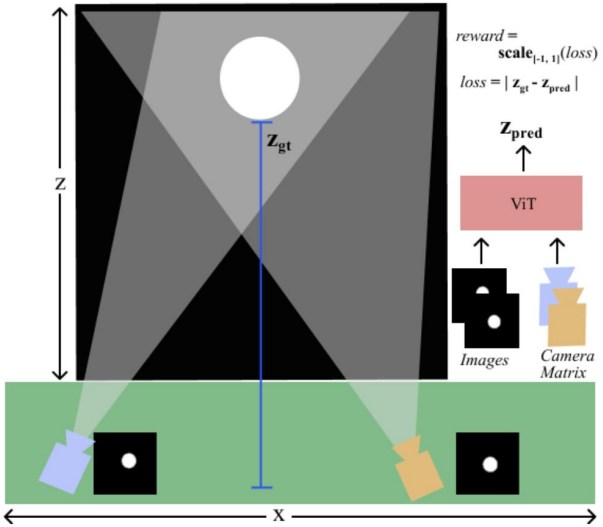
My Abstraction at the time: Perception-Action loops

Some historical perspective on Perception-Action loops…

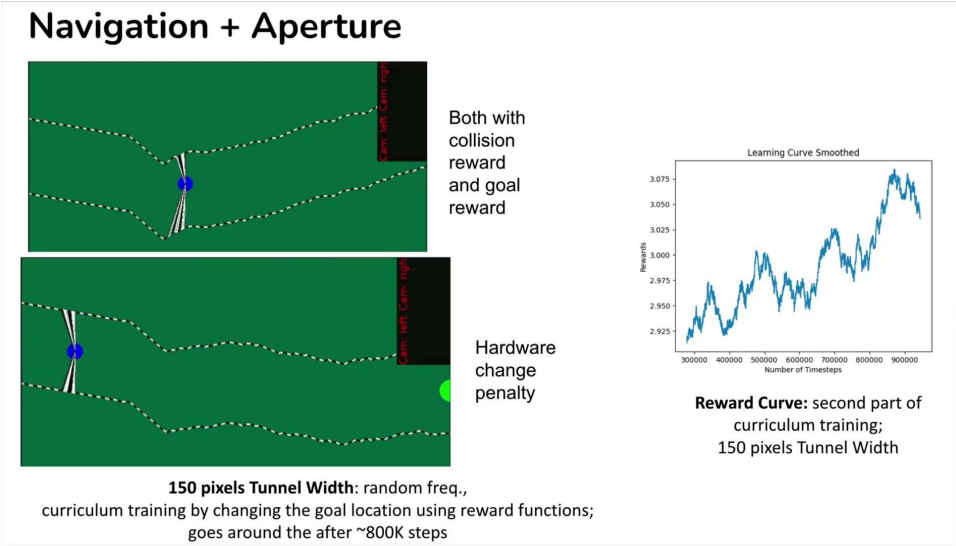
- **Affordances:** Actionable possibilities offered by the environment to an animal and are perceived directly based on the relationship between the physical properties of the body and the environment
  - e.g., a chair affords sitting; a staircase affords climbing
- **The Active Visual System:** Vision does not happen inside a detached brain or passive eyeball. Perception belongs to an active, whole body moving through space



Simple task: Can we “discover” basic vision algorithms like nature?



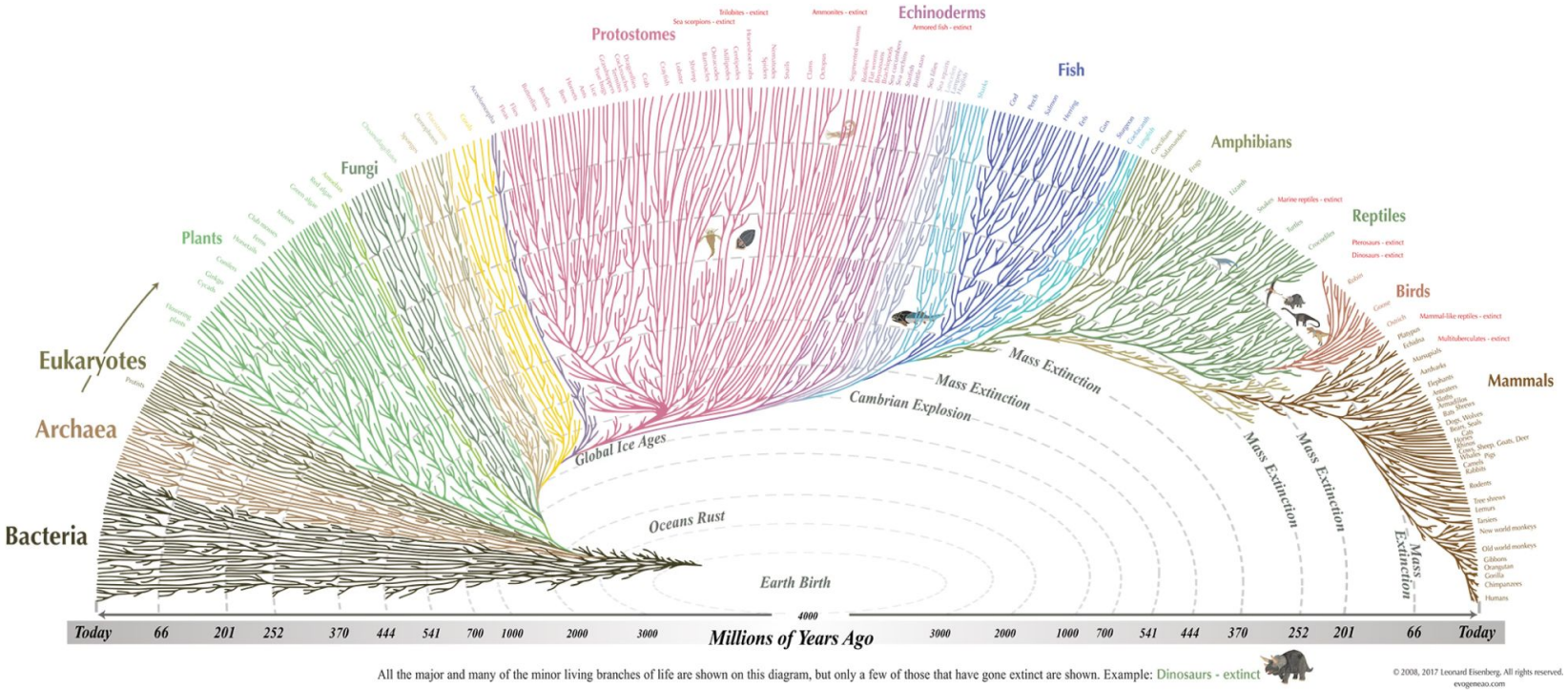
RL agent discovers stereo-depth algorithm without any monocular cues



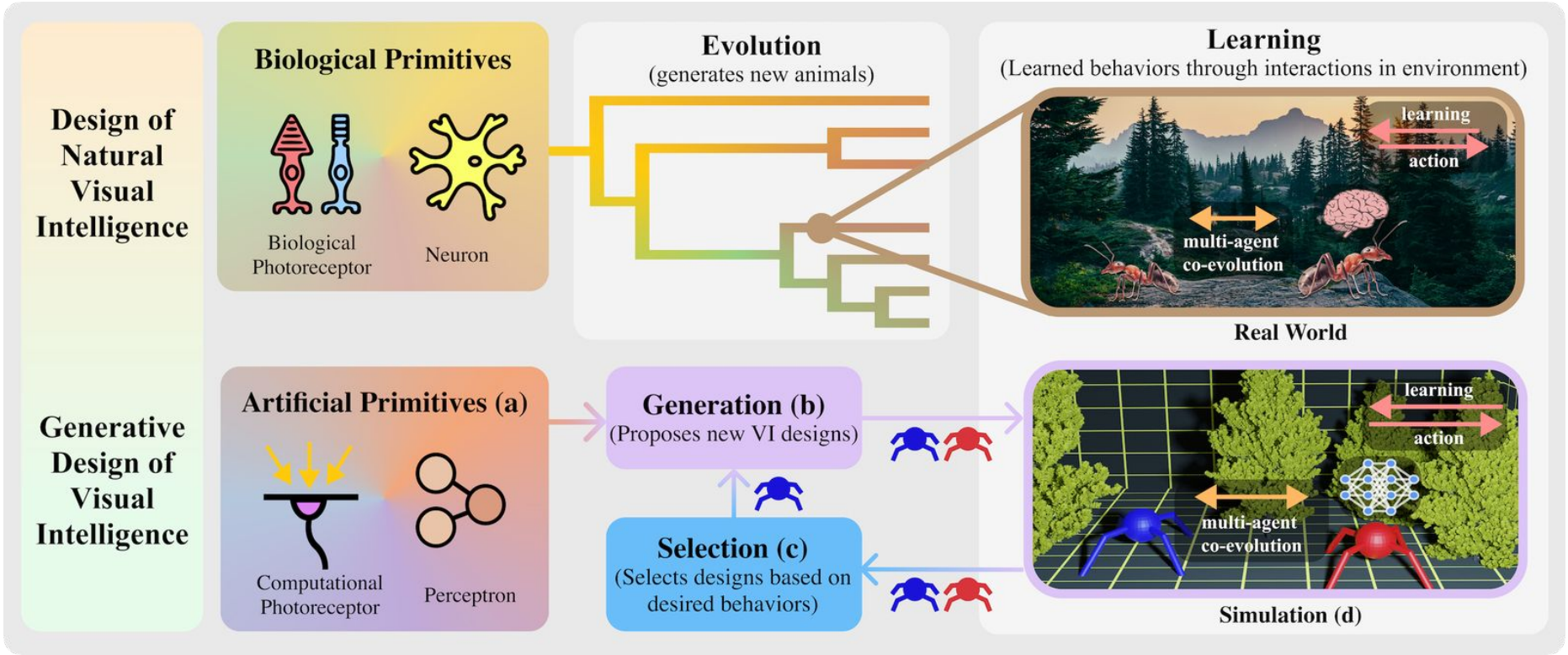
RL agent discovers a generalized algorithms to navigate mazes (navigation using aperture and optic flow algorithm)

Klinghoffer*, Tiwary*, et. al, DISeR: Designing Imaging Systems with Reinforcement Learning, ICCV

More Challenging Task: Can we re-discover visual intelligence across Marr's levels?

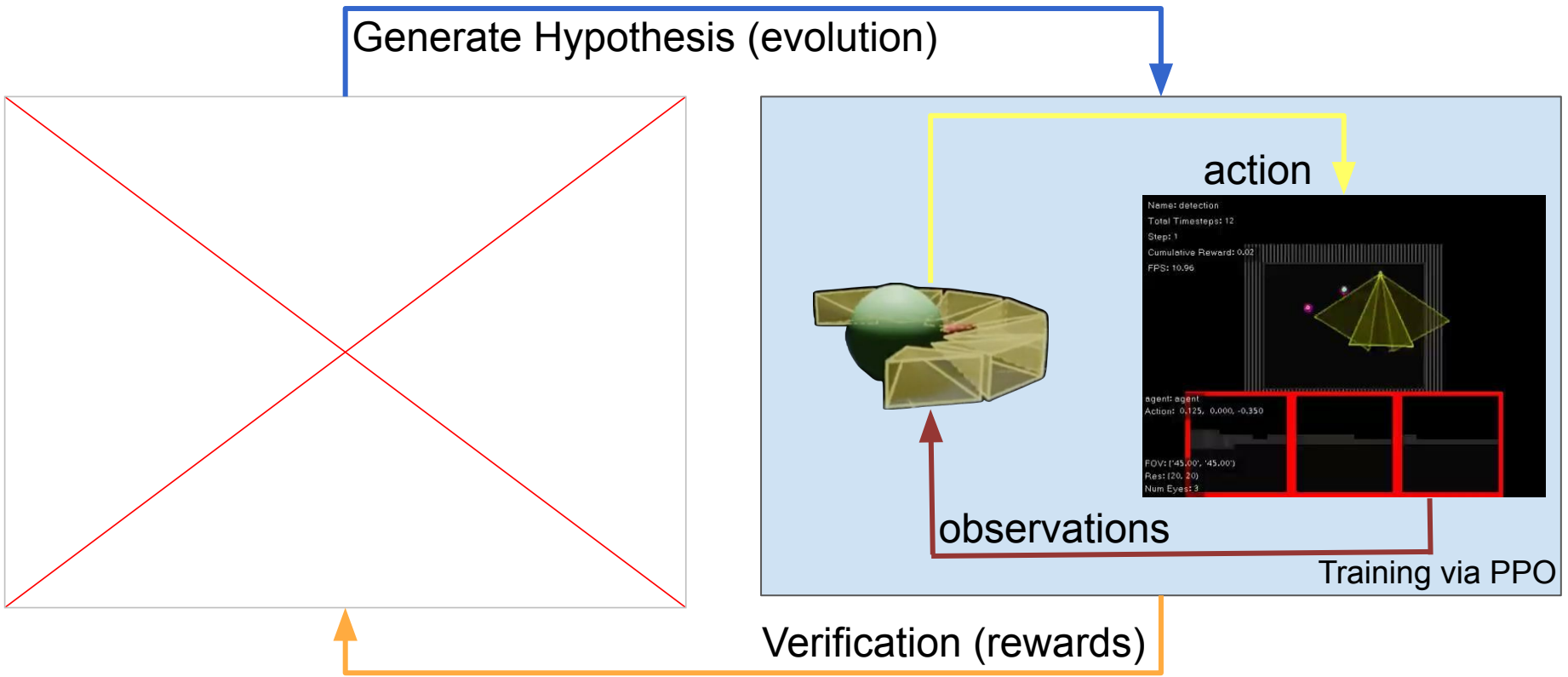


Generative Design of Visual Intelligence



Tiwary et. al, A Roadmap for Generative Design of Visual Intelligence, MIT Press

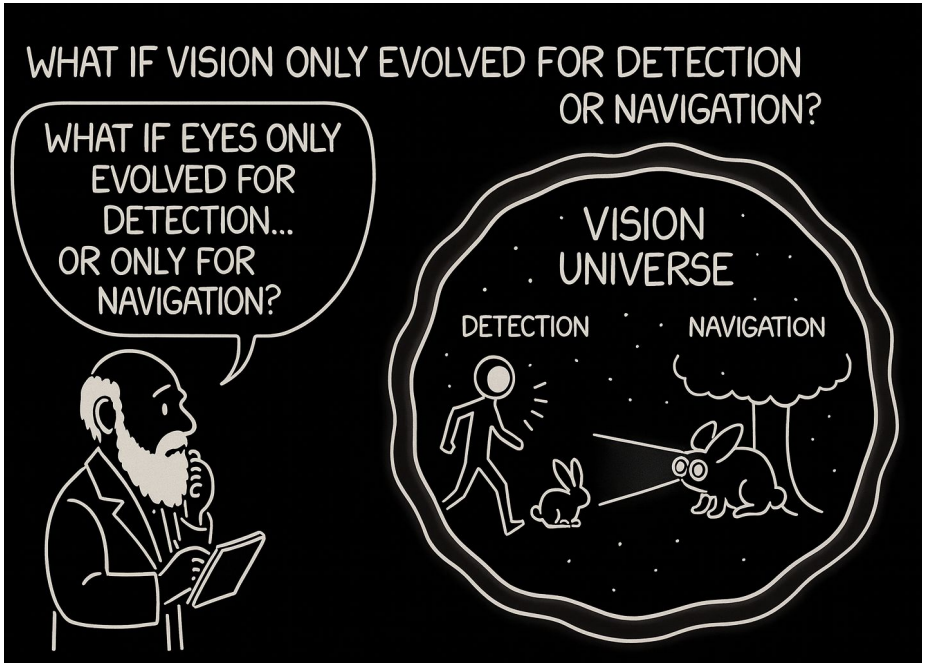
Model perception-action loops with an outer evolutionary loop



Tiwary* et. al, What if Eye? Computationally Recreating Vision Evolution, Science Advances, eyes.mit.edu

What if Eye…?

Universe

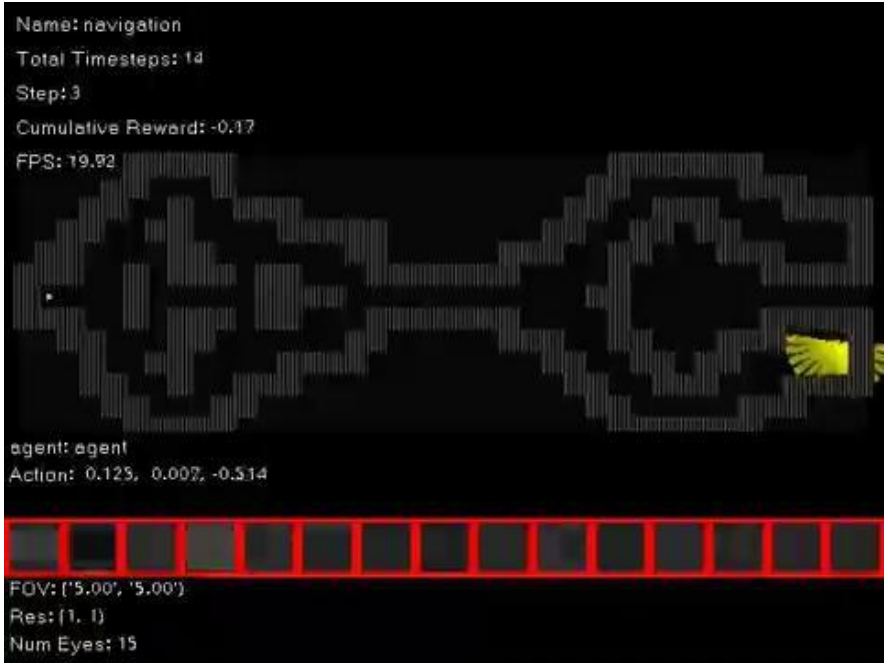


https://eyes.mit.edu/

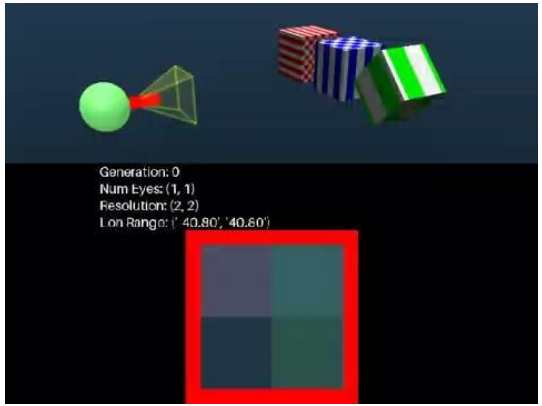
Detection Only Universe

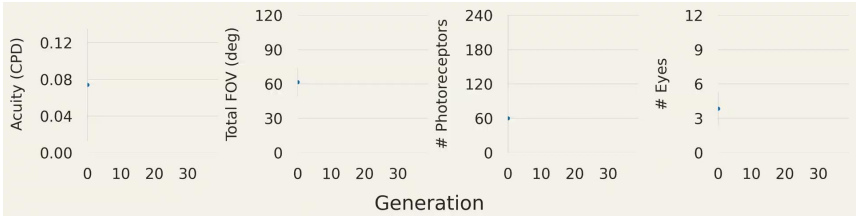


Navigation Only Universe

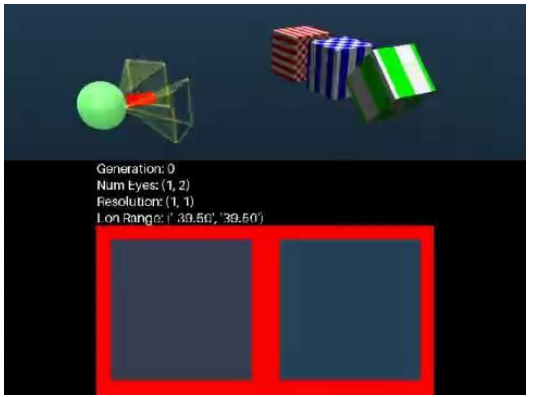


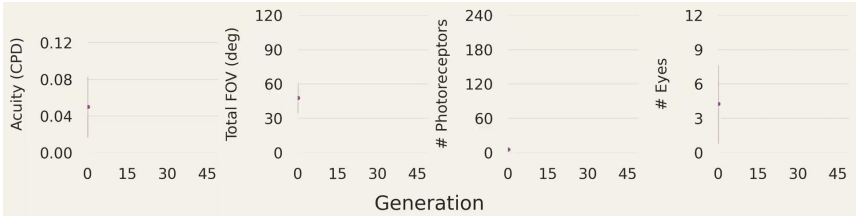
Detection Only Universe



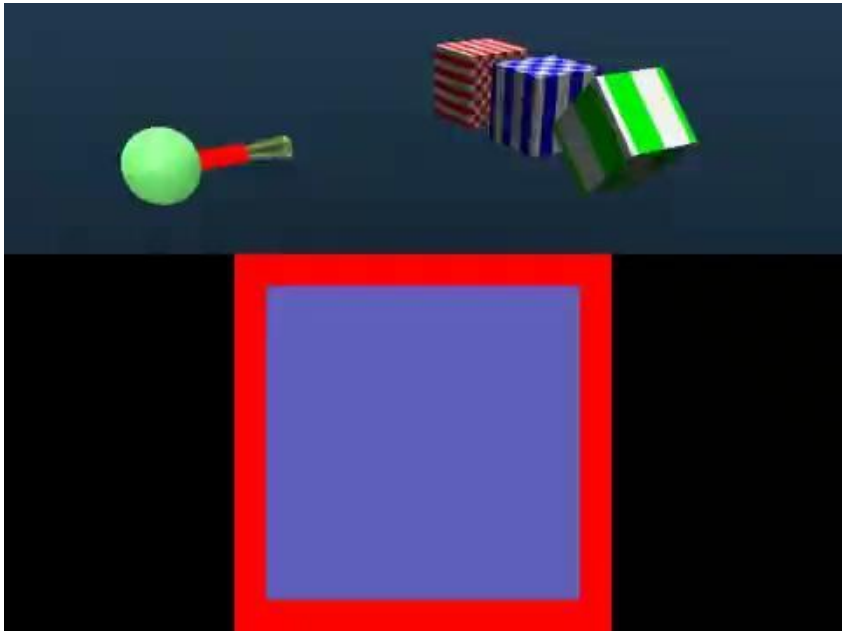


Navigation Only Universe

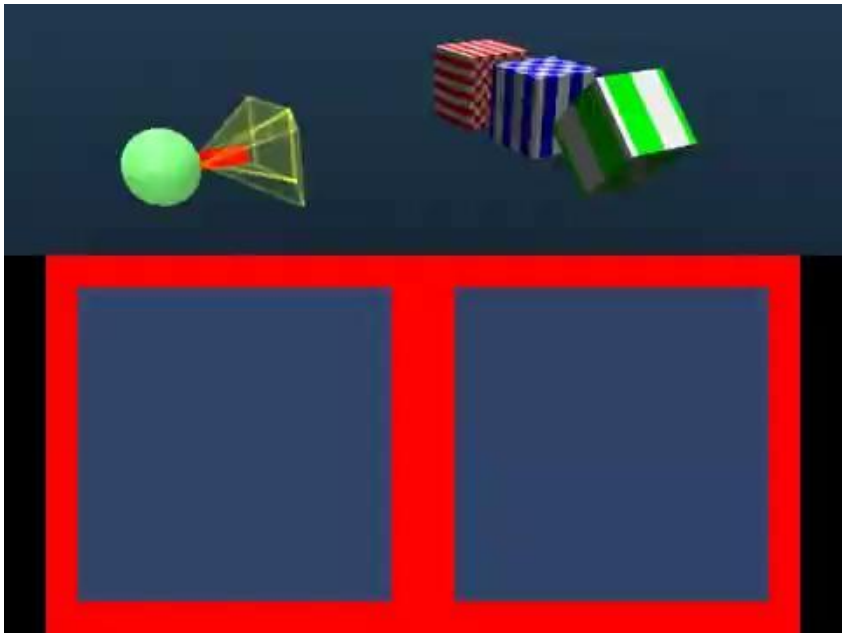




But how do we generate embodied eyeballs?



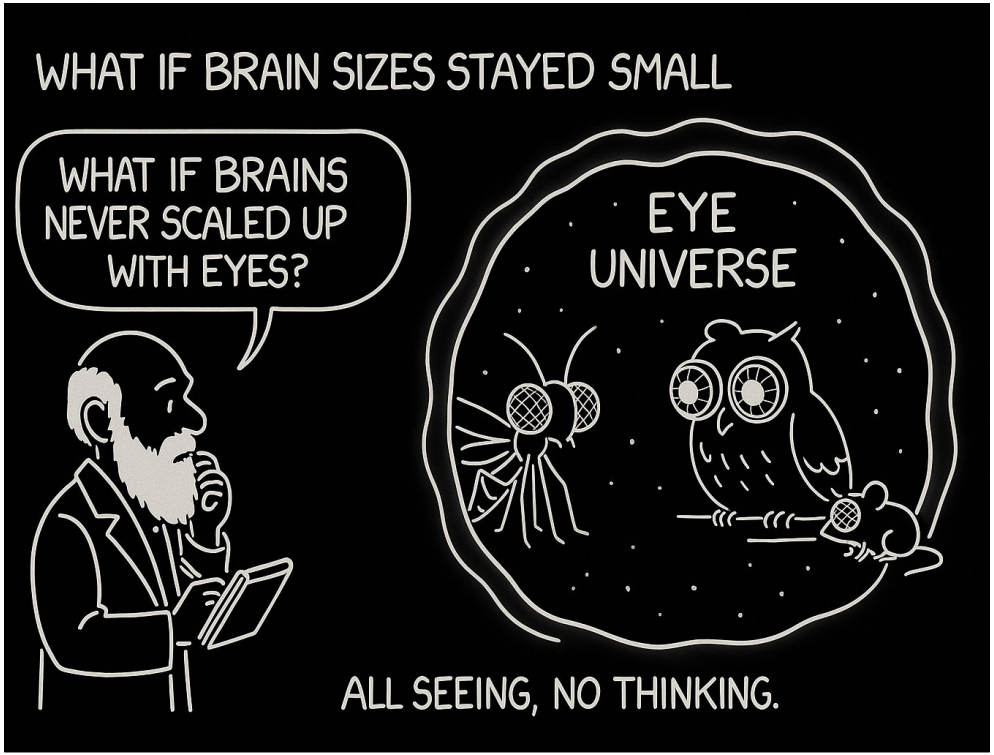
#eyes



#photoreceptors

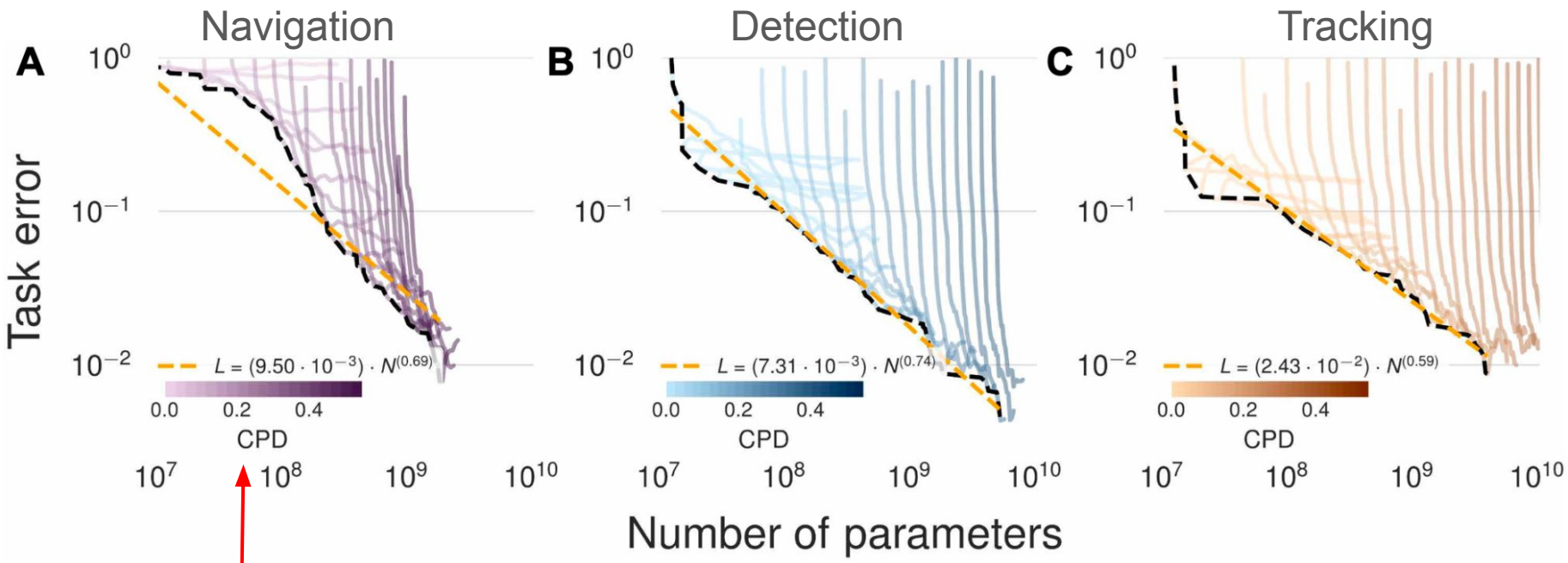
What if Eye…?

Universe



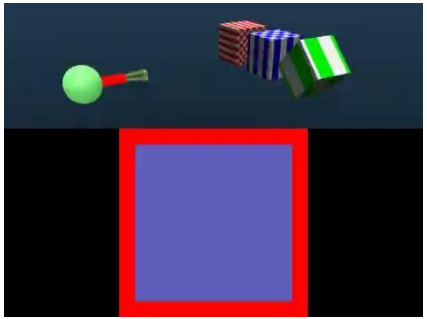
https://eyes.mit.edu/

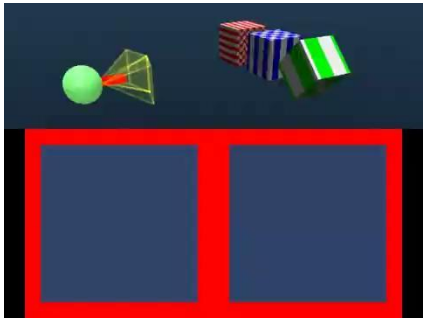
Framework can also discover scaling laws

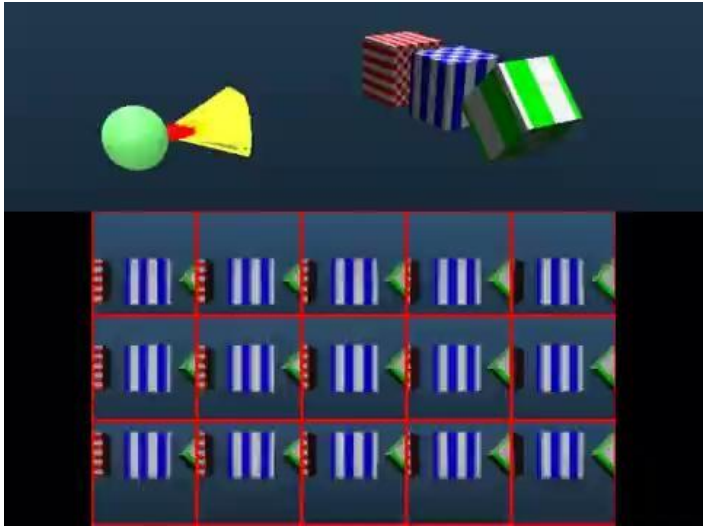


Measure of visual acuity

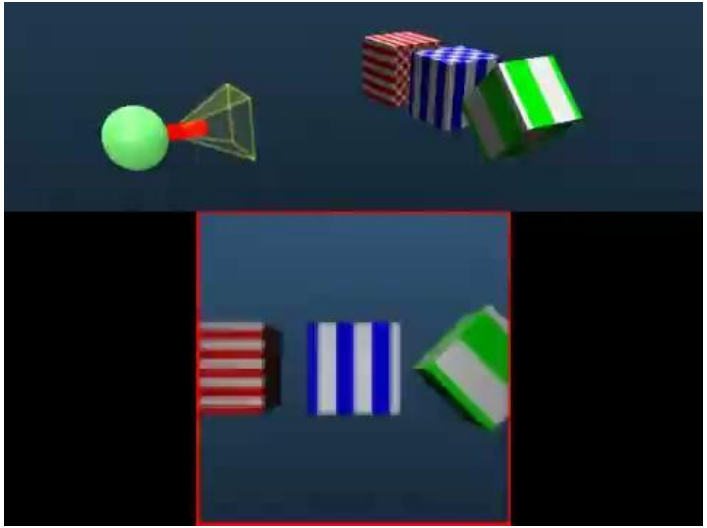
But how do we generate embodied eyeballs?







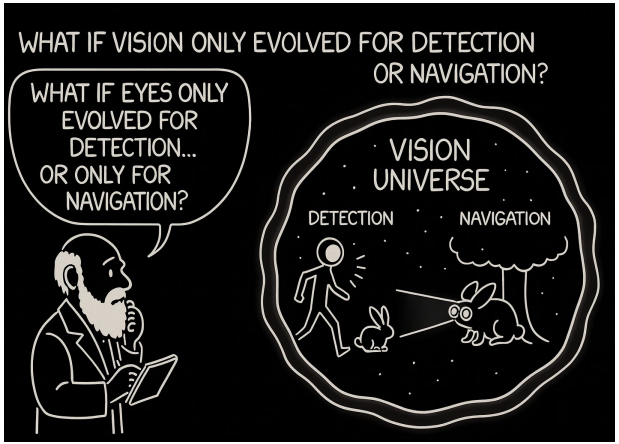
position

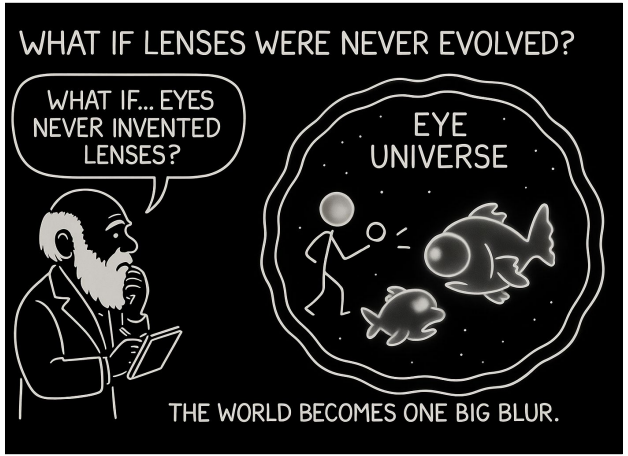


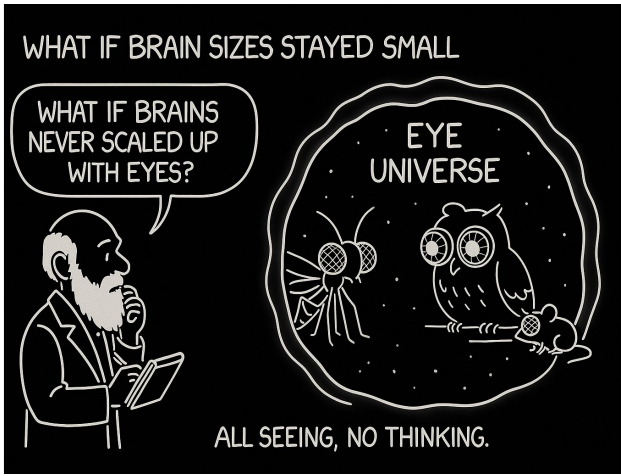
pupil size

What if Eye…?

Universe







https://eyes.mit.edu/

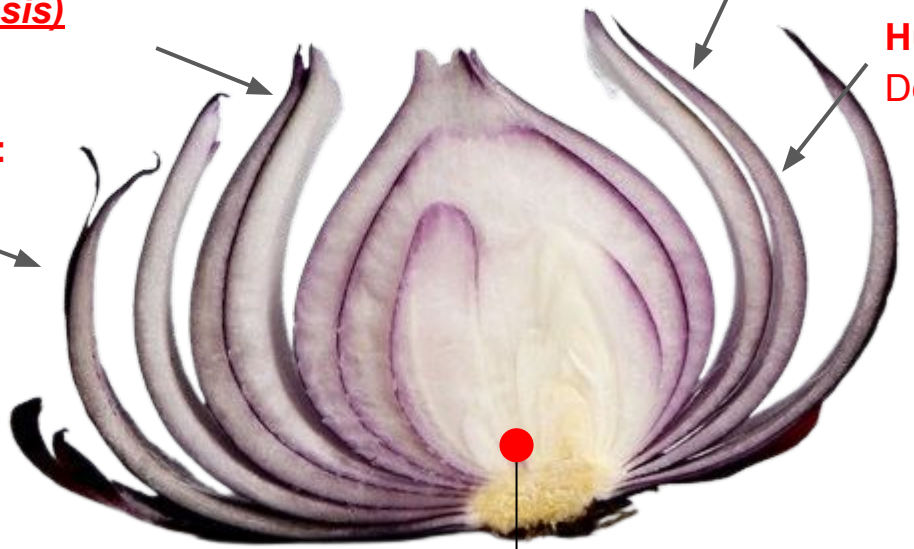
But are we at the center of the onion? (remember goal #2 from earlier)

Human-Abstraction 3: physics can provide structure to the learning problem

Human-Abstraction 4: Perception-Action loops ***(my hypothesis)***

Human-Abstraction 2: Object Detection, 3D, Point Clouds

Human-Abstraction 1: Geometric CV, filters

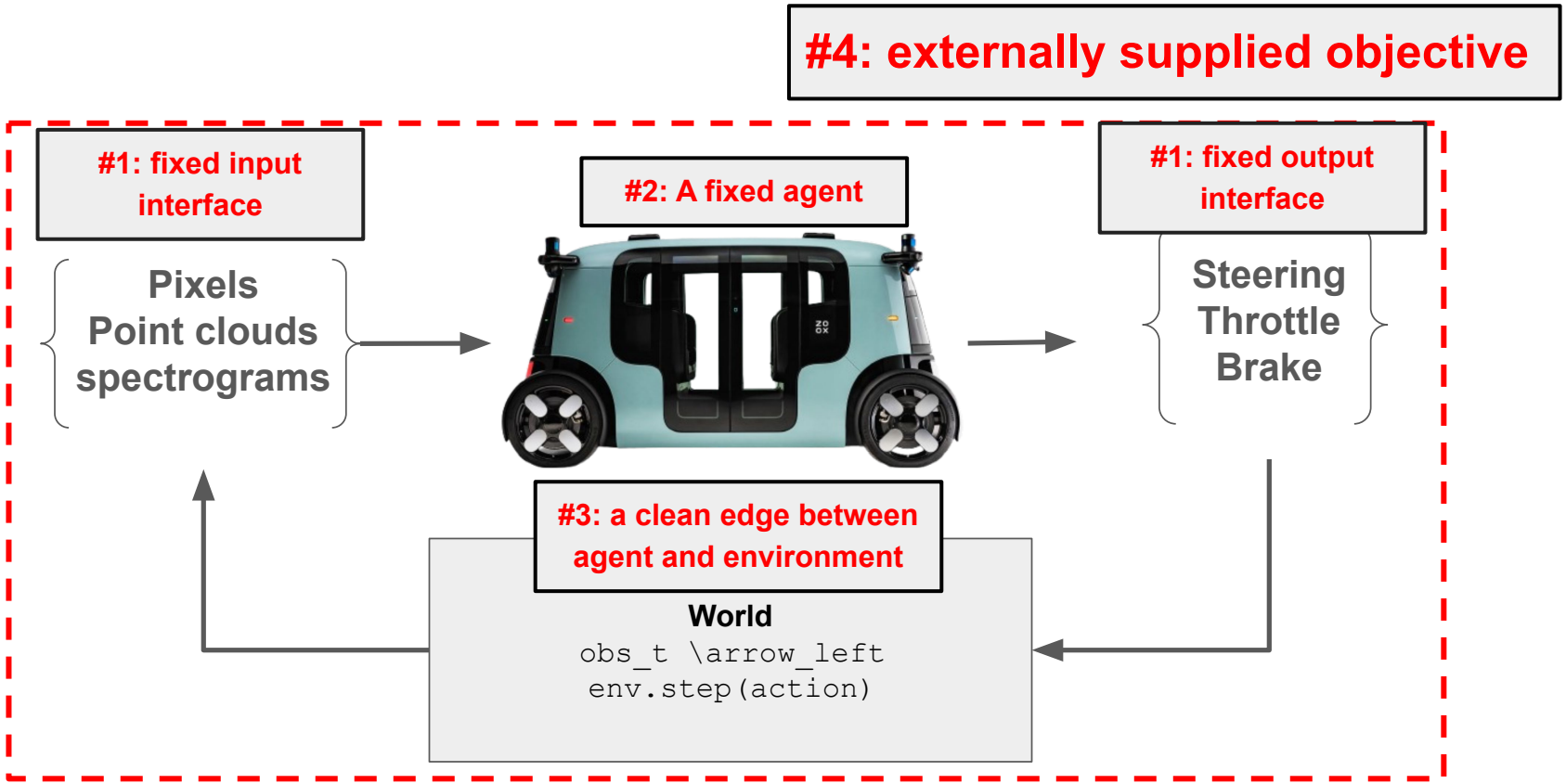


This talk: What is the *"generation-process"* that can generate all of natural and artificial intelligence?

A typical **perception–action loop**:



A fixed **perception–action loop** makes four assumptions:



Assumption #1: The reality-agent interface problem

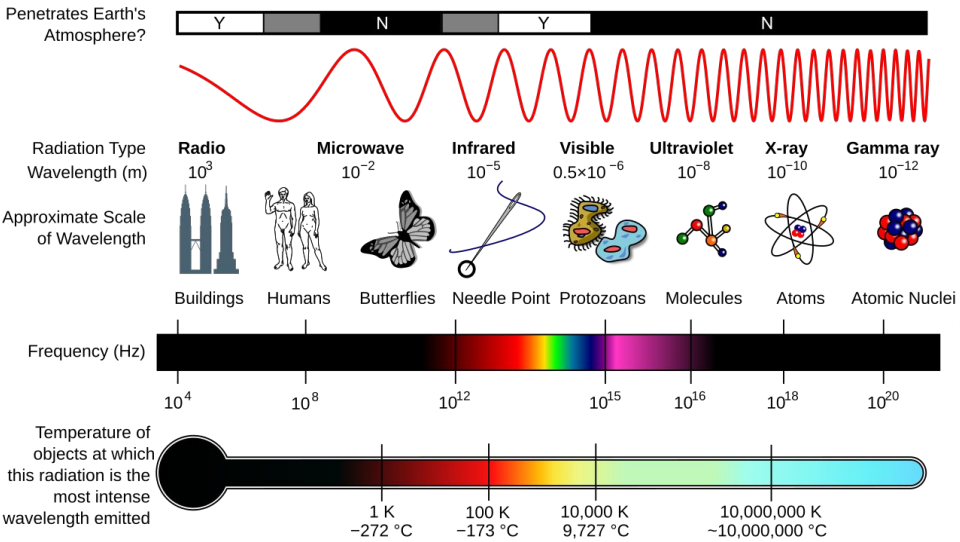


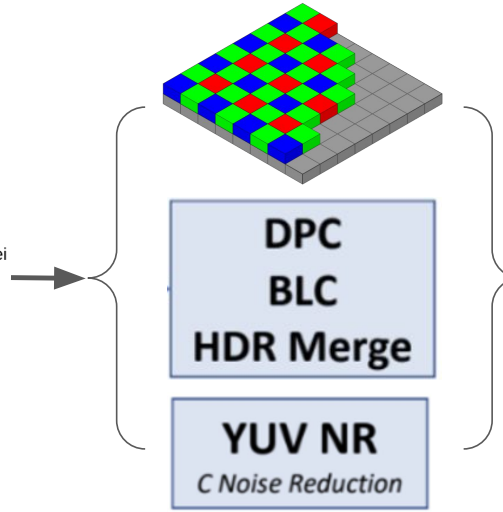
#1: fixed input interface

#1: fixed output interface

How “close” to reality can we interface with our sensing?

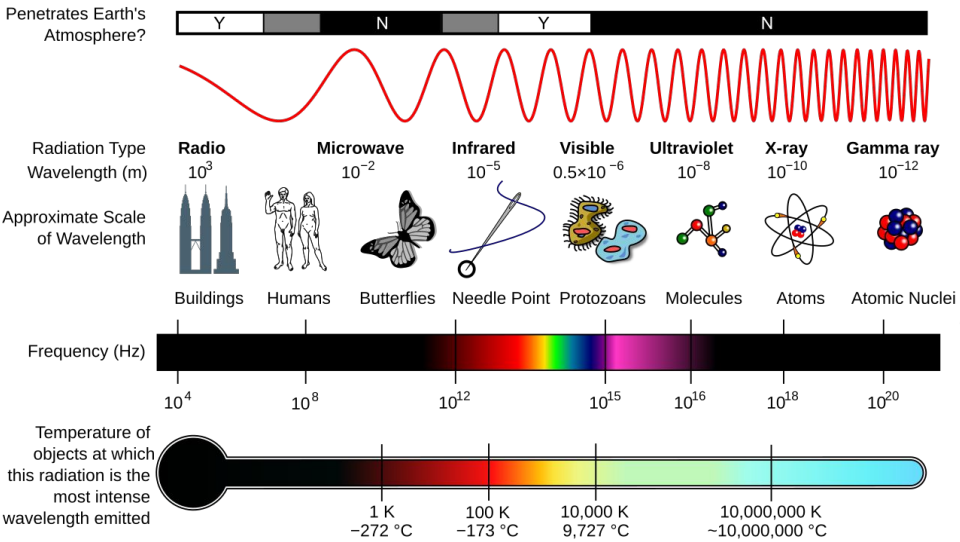
#1: Pixels, Point clouds, spectrograms are human made representations



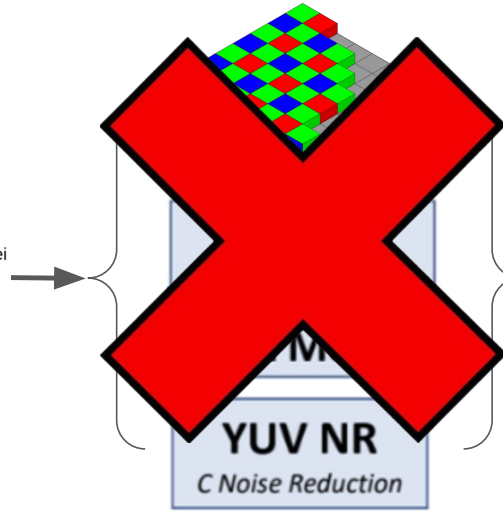




How “close” to reality can we interface with our sensing?

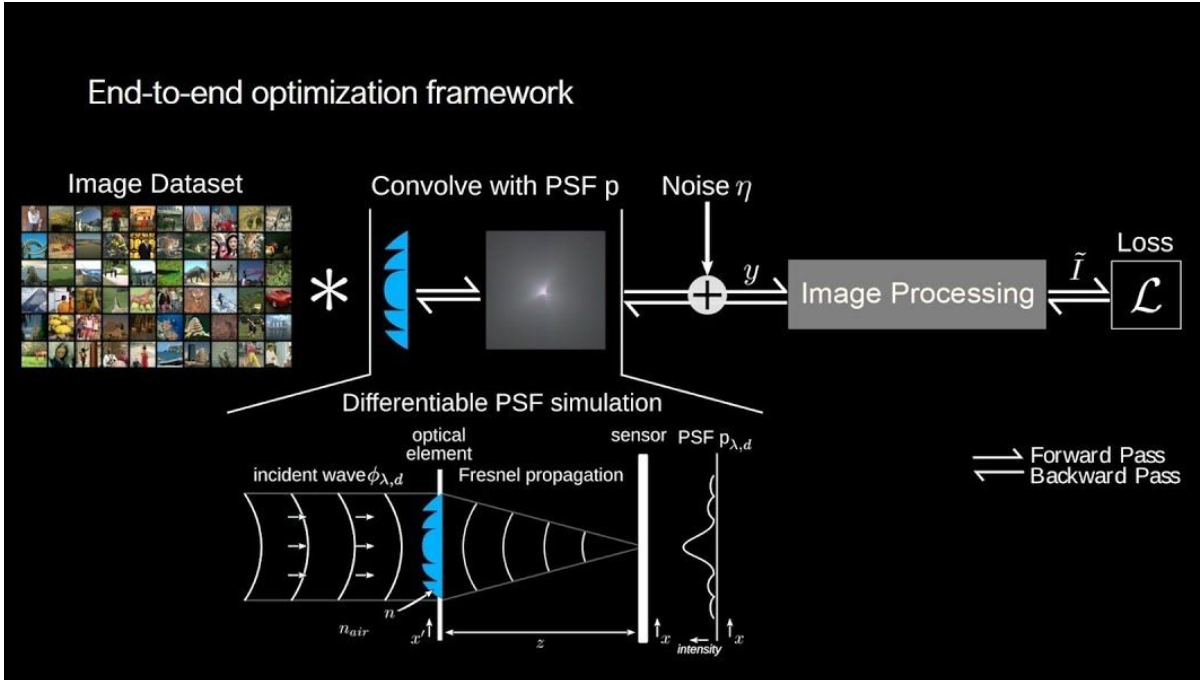


#1: Pixels, Point clouds, spectrograms are human made representations





I think *end-to-end optics* was on the right track…

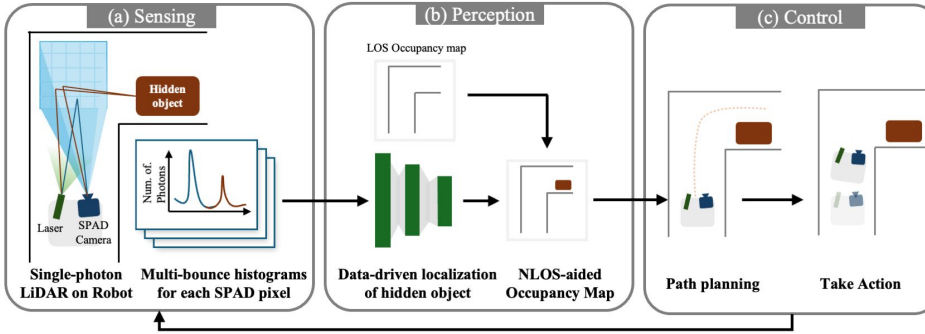


.. But ultimately got bitter-lessoned because it didn't align with scale and was focused on human-made abstractions (2D/3D)

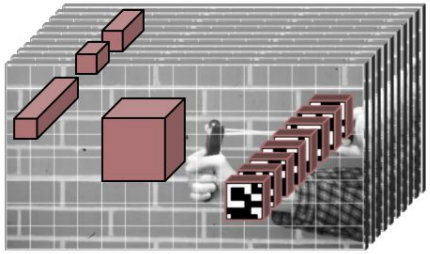
V. Sitzmann et al., "End-to-end Optimization of Optics and Image Processing for Achromatic Extended Depth of Field and Super-resolution Imaging,"

Open Questions and some interesting stabs at the problem…

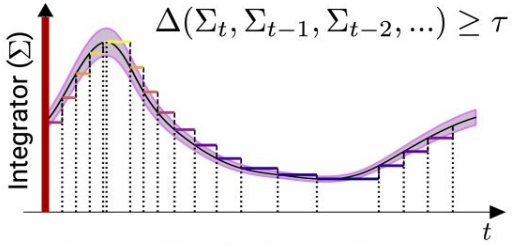
1. How do we “act” directly on photons?
2. How do ride the scaling exponent and align with 1?



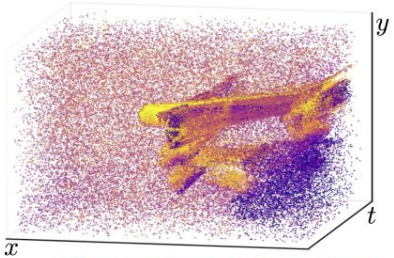
Generalized event camera



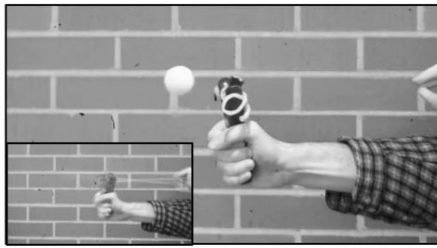
Motion adaptive | Spatial patches | Coded exposures



Enhanced spatio-temporal support
Noise-aware thresholding



Stream of integrator values
$(x, y, t, \Sigma(x, y, t))$



Reconstruction
3025 FPS at 30 FPS bitrate

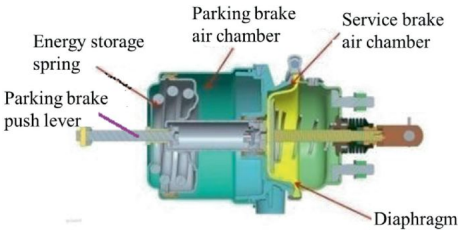
Sundar*, Dutson* et. al, Generalized Event Cameras

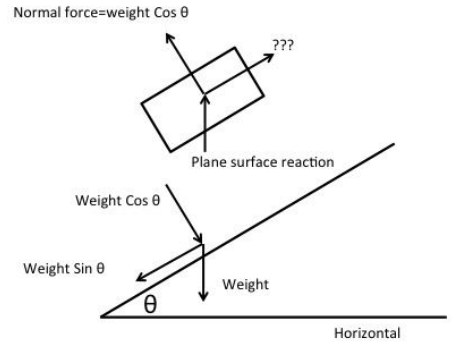
Young*, Batagoda* et. al, Enhancing Autonomous Navigation by Imaging Hidden Objects using Single-Photon LiDAR

How “close” to reality can we interface with our actuation?



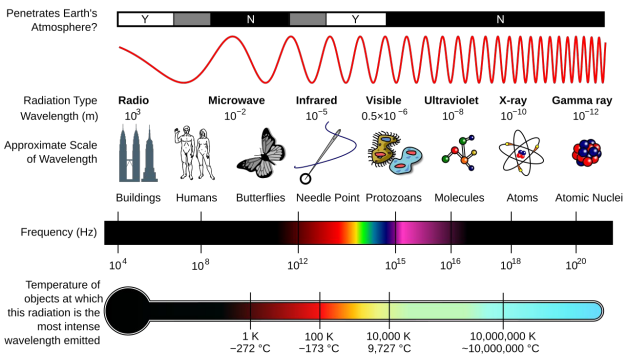
Steering
Throttle
Brake

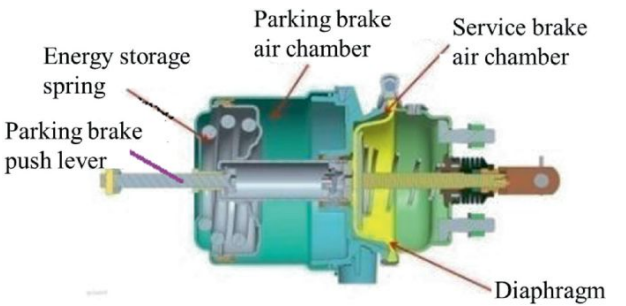




Get rid of abstractions,
directly model force!

Takeaway #1: Remove all abstractions between you and reality





- Be mindful of your abstraction over your sensing
- Be mindful of your abstraction over your actuation

Models are too powerful now and will pick up (or will be limited by) human-made abstractions

Interface with reality?

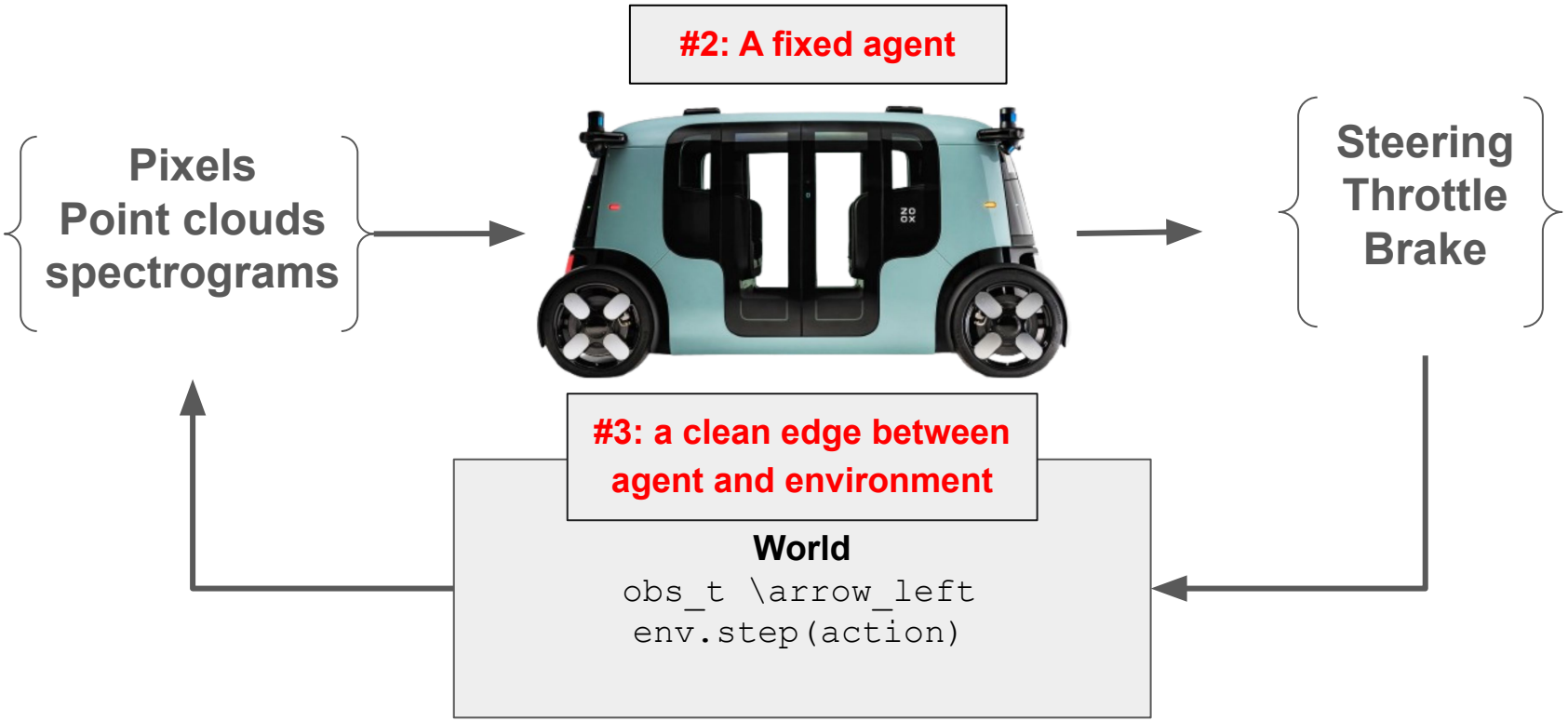
Human-Abstraction 3:
Perception-Action loops *(current)*

Human-Abstraction 2: Object Detection, 3D, Point Clouds

Human-Abstraction 1:
Geometric CV, filters



Assumption #2: The agent-environment assumption



Where's ~~Waldo~~ the Agent?



Stove?

The cook?

The knife?

Cook's hand?

The painter?

Rice & Beans I, Oil on canvas, Ricky Vasan, 2023,
https://www.rickyvasan.com/

Collective view on agents:

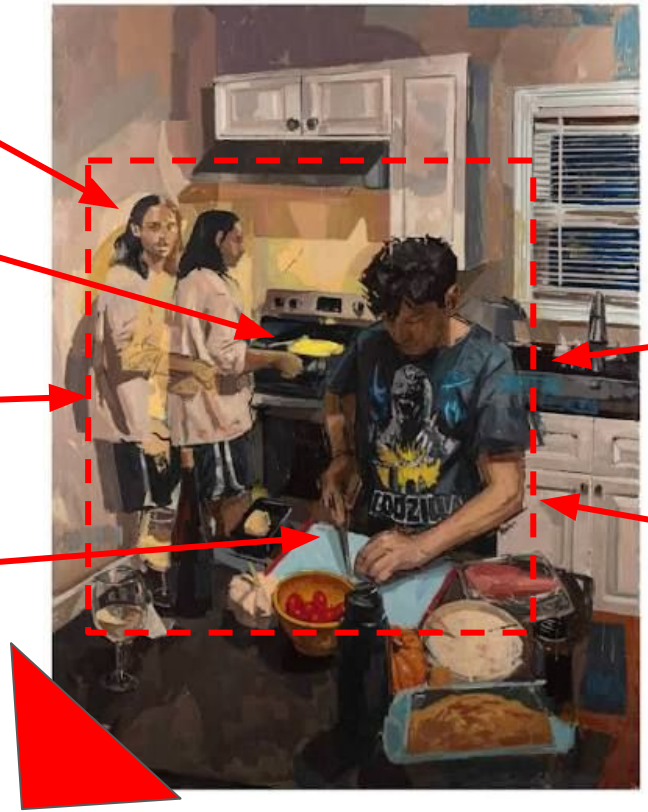
The other cook?

Stove?

Collective Agents: household?

The knife?

The painter?

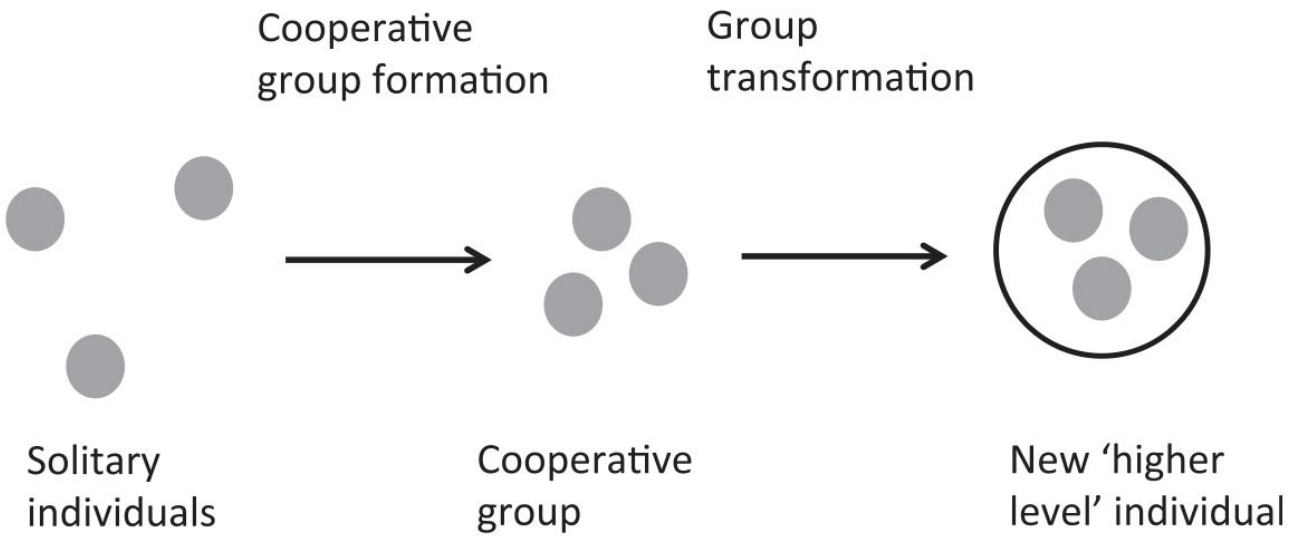


The cook?

Cook’s hand?

Rice & Beans I, Oil on canvas, Ricky Vasan, 2023, https://www.rickyvasan.com/

Well, what is an agent in biology…



West et. al, Major evolutionary transitions in individuality

We are composed of many sub agents

Table 1. The major evolutionary transitions in individuality

Major transitions				Not major transitions		
					Why not?	
Transition			Examples	Examples	Mutual dependence?	Negligible conflict?
Independent replicators	→	Populations of replicators	Genome	—	—	—
Separate unicells	→	Symbiotic unicell	Eukaryotic cell*, primary plastid endosymbiosis*, *Paulinella** Secondary/tertiary plastid endosymbiosis*	—	—	—
Unicells	→	Multicellular organism	Animals, plants, fungi, red algae, brown algae	*Dictyostelid* slime molds[†]	No	No
				Pseudomonas biofilms	No	No
Multicellular organisms	→	Eusocial society	Higher termites, leaf-cutter ants, honeybee	Social aphids[‡]	No	Yes
				Melipona bees	Yes	No
				Cooperative vertebrates	No	No
				Humans	No	No
Multicellular organisms	→	Obligate interspecific mutualism	Some obligate endosymbionts (e.g., *Buchnera* in aphids*), leaf-cutter ants and their fungi, Paracatenula flatworms and their chemoautrophic symbionts	Legumes-*Rhizobia*[§]	No	No
				Squid-*Vibrio*	No	No
				Gut microbiota	No	No

West et. al, Major evolutionary transitions in individuality

Some ideas to consider on agency…

1.2. Other Characterisations of Agents

To put our contribution in context, we here describe previous characterisations of agents:

- *The intentional stance*: an agent's behaviour can be usefully understood as trying to optimise an objective (Dennett, 1987).
- *Cybernetics*: an agent's behaviour adapts to achieve an objective (e.g. Ashby, 1956; Wiener, 1961).
- *Decision theory / game theory / economics / AI*: An agent selects a policy to optimise an objective.
- An agent is a system whose behaviour can be *compressed* with respect to an objective function (Orseau et al., 2018).
- "An *optimising system* is ... a part of the universe [that] moves predictably towards a small set of target configurations" (Flint, 2020).
- A *goal-directed system* has self-awareness, planning, consequentialism, scale, coherence, and flexibility (Ngo, 2020).
- Agents are ways for the future to influence the past (via the agent's model of the future) (Garrabrant, 2021; von Foerster et al., 1951).

For broader reviews, see also Shimi et al. (2021) and Wooldridge and Jennings (1995).

Our proposal can be characterised as: *agents are systems that would adapt their policy if their actions influenced the world in a different way.* This may be read as an alternative to, or an elaboration of, the intentional stance and cybernetics definitions (depending on how you interpret them) couched in the language of causality and counterfactuals. Our definition is fully consistent with the decision theoretic view, as agents choose their behaviour differently depending on its expected consequences, but doesn't require us to know who is a decision maker in advance, nor what they are optimising.

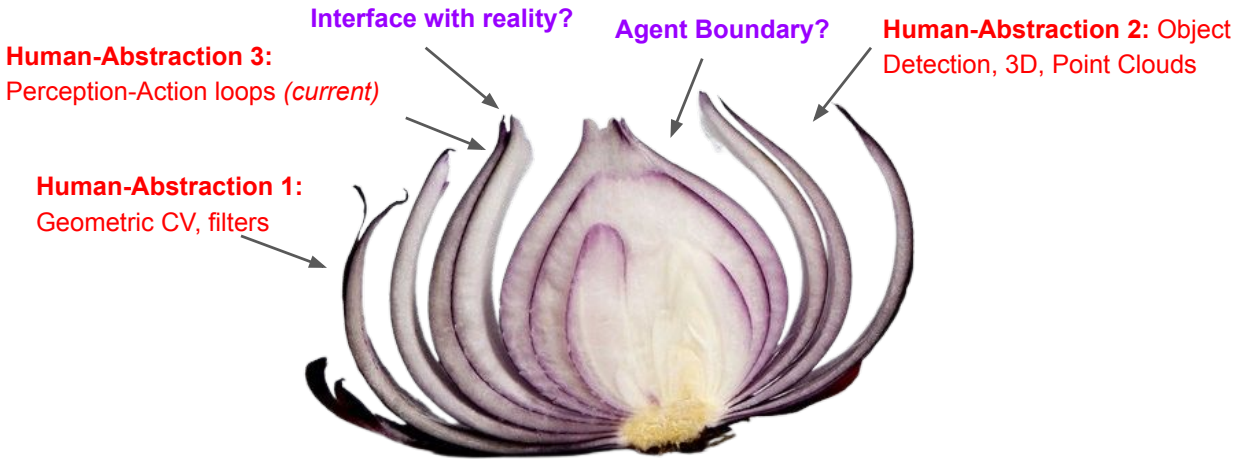
Kenton et. al, Discovering Agents

Takeaway #2: *Questions around agency lie ahead of us*



Rice & Beans I, Oil on canvas, Ricky Vasan, 2023, https://www.rickyvasan.com/

- Don’t impose agency or action split at unnecessary places
- Move from “learning” centric view to “agent” centric view
- Agent(s)-Env boundary is only gonna get challenging as number of agents scale



Assumption #3: The externally supplied objective assumption

#4: externally supplied objective

#1: fixed input interface

Pixels
Point clouds
spectrograms

#2: A fixed agent



#1: fixed output interface

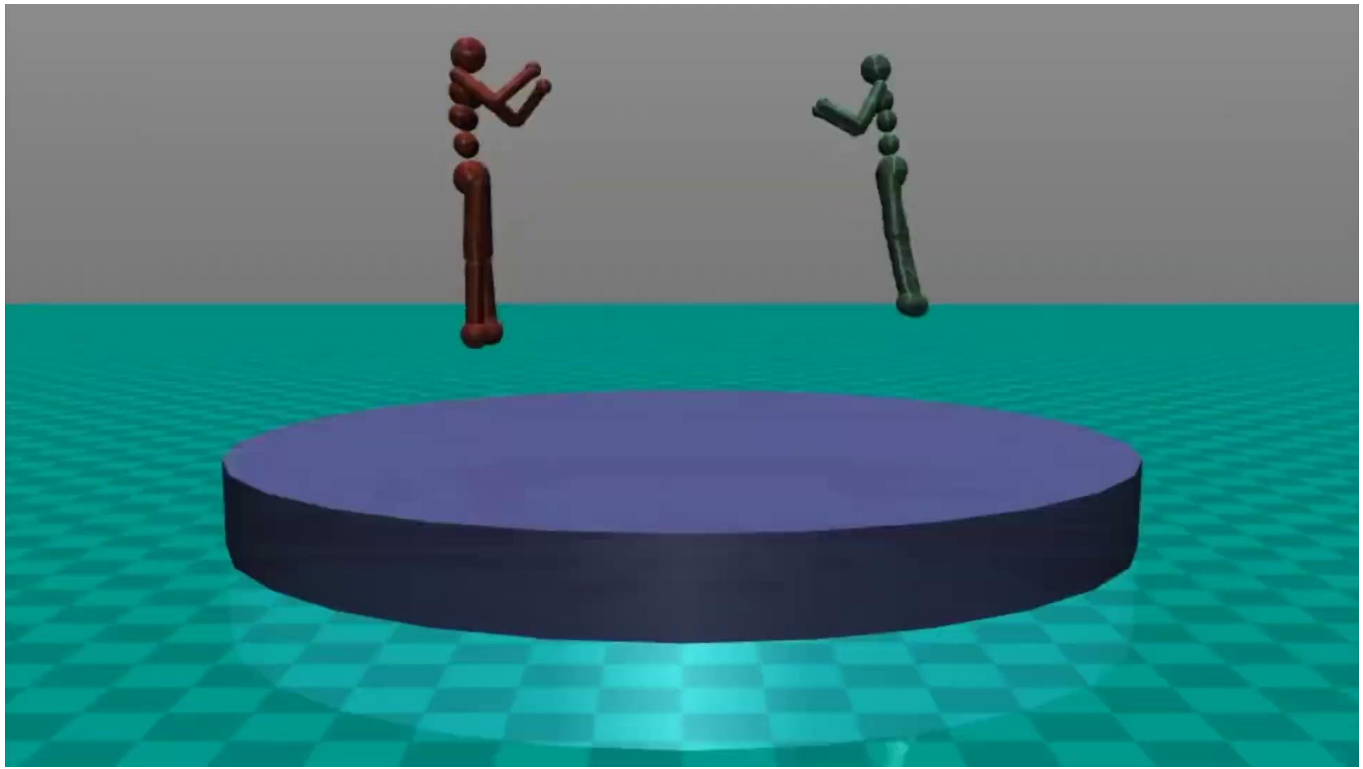
Steering
Throttle
Brake

#3: a clean edge between agent and environment

World

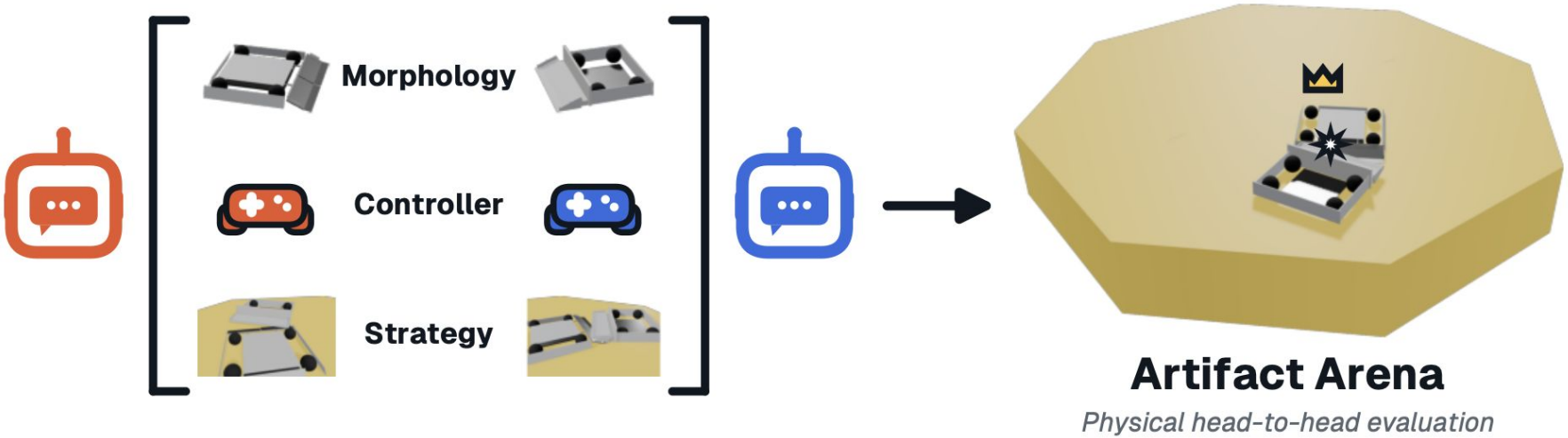
```
obs_t \arrow_left
env.step(action)
```

Supply simple rules, not hand-designed objectives



Bansal et. al, Emergent Complexity via Multi-Agent Competition

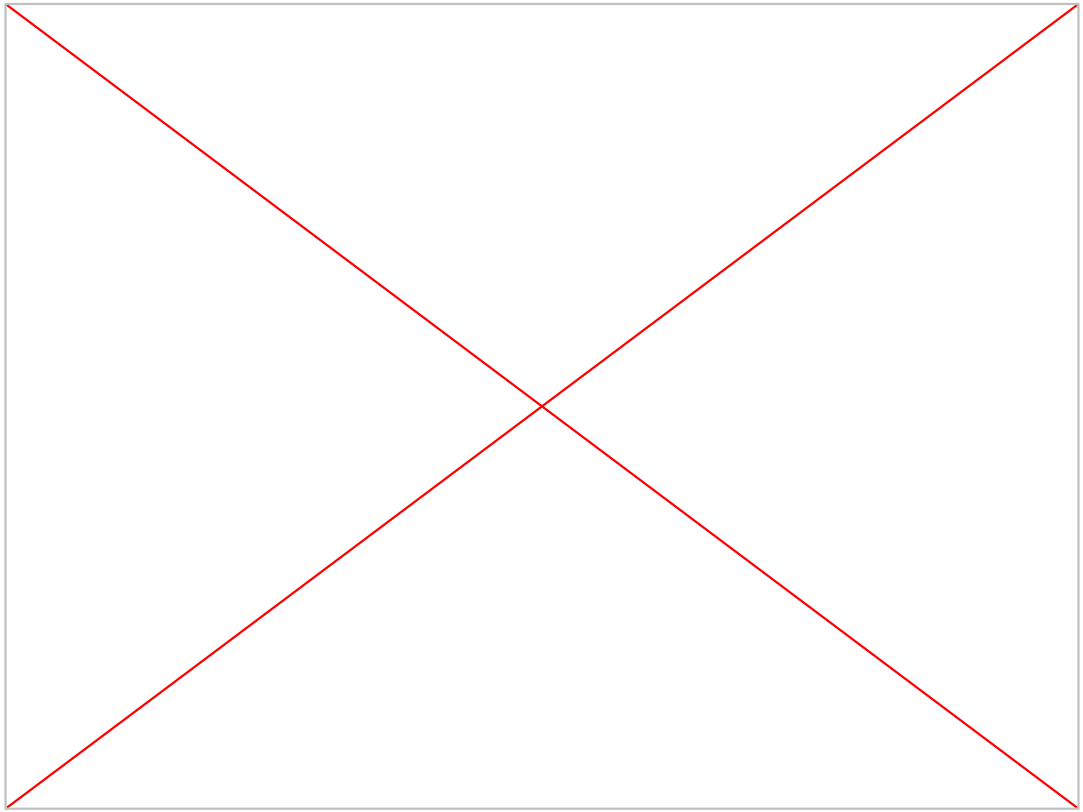
How can we explore this with language models?



Tiwary*, Mayo* et. al, ArtifactArena: Evaluating Models by What They Build in the Physical World

https://artifactarena.ai/

Simple rules, large design space, multi-agent, adversarial



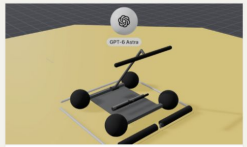
https://artifactarena.ai/

Tiwary*, Mayo* et. al, ArtifactArena: Evaluating Models by What They Build in the Physical World

Creativity emerges from env & simple rules/constraints!

Creative model inventions

Given simple rules and a physically grounded design space, models invent their own bot morphologies and co-design the controllers that use them.



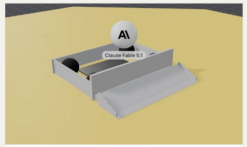
VGH **Undertow Razorwake**
GPT-6 Astra
Overhead jaw on a T-bar above two split intake shovels; four two-axis ball drives



SH **Sidewinder Undertow**
GPT-6 Astra
Rolls on four rubber balls, two motors each; slides sideways with the scoop on target



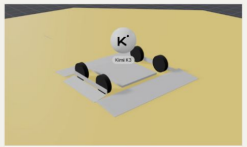
DLH **Apex**
Gemini 3.1 Pro
343 kg steel drum with tungsten teeth, fed by a wedge, counterweighted at the rear



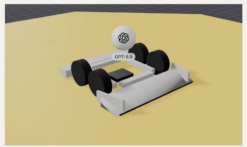
VGH **Tungsten Groundhog Mk8**
Claude Fable 5.1
790 kg tungsten slab on four rubber spheres; a floating scoop hinged at ground level



VGH **Keelhaul**
Grok 4.6
U-channel apron at both ends, spring-loaded and motor-held down against prying



VGH **Bulldozer Prime: Centerline**
Kimi K3
653 kg tungsten floor 1.75 cm off the ring, motorized 2.4 m wedge; best open-weight robot



DLH **Elevator Scoop**
GPT-5.5
A 383 kg tungsten block on a 1.7 m vertical slide to beat the inactivity rule



SH **Tungsten Manta**
GPT-5.4
Six wheels under a sculpted plow with skirts; the best record in its qualification group

WHAT THE REPLAY REVEALED

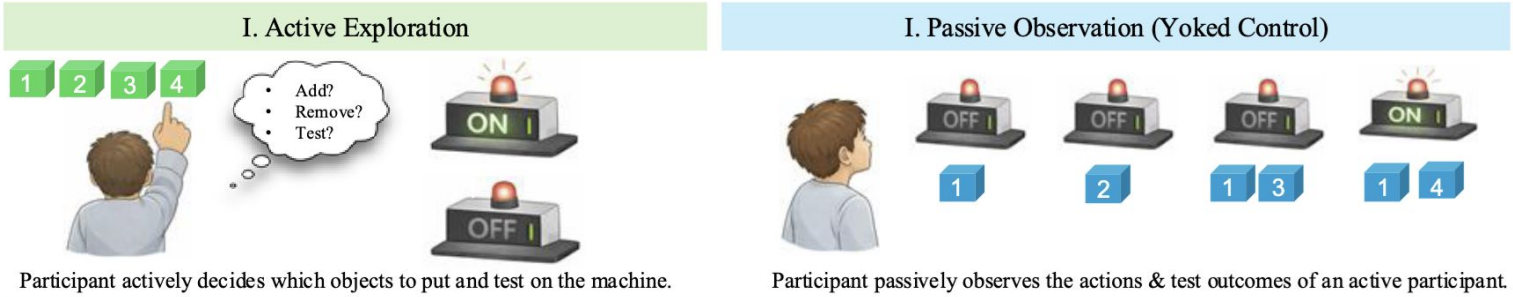
"revision 5 won all three qualification matches … but the replay exposed a serious competitive weakness: repeated, very large receiver–front-wheel self-contact forces. … Qualification success therefore concealed substantial lost traction and maneuverability."

WHAT IT CHANGED

"replaces the problematic SDF receiver and deflectors with deliberately shaped, analytically exact inclined plates … primitive collision geometry is the better physical implementation here—not merely the easier one."

GPT-6 Astra · Undertow: Freebite · revision 6

Models are not great at active explortation…



Human Adults and LLMs as Scientists: Who Benefits from Active Exploration?

Mandana Samiei [1,2*] **Eunice Yiu** [3*] **Anthony GX-Chen** [4] **Dongyan Lin** [5] **Jocelyn Shen** [6]
Blake A. Richards [1,2,7] **Alison Gopnik** [3] **Doina Precup** [1,2]

[1]Mila - Quebec AI Institute [2]McGill University [3]University of California Berkeley [4]New York University
[5]Meta FAIR [6]MIT Media Lab [7]Montreal Neurological Institute

*Equal contribution. Correspondence: mandana.samiei@mail.mcgill.ca, ey242@berkeley.edu

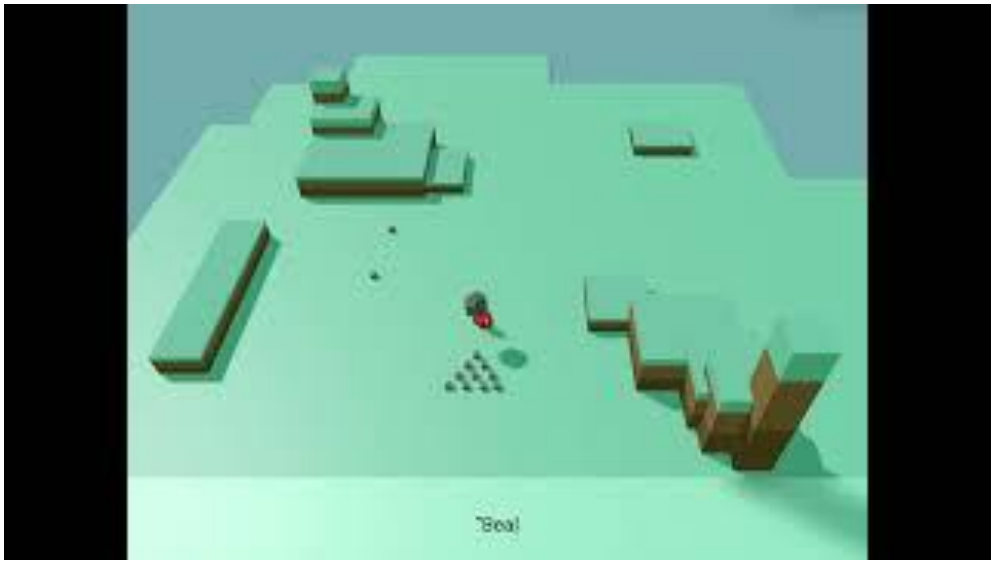
The hard problems of curiosity, opendedness, motivation lie ahead of us *(this is the exciting stuff)*

Remember that some of the greatest discoveries in vision were driven by open-endedness, curiosity





An example of no-objective LLM-based embodied exploration…



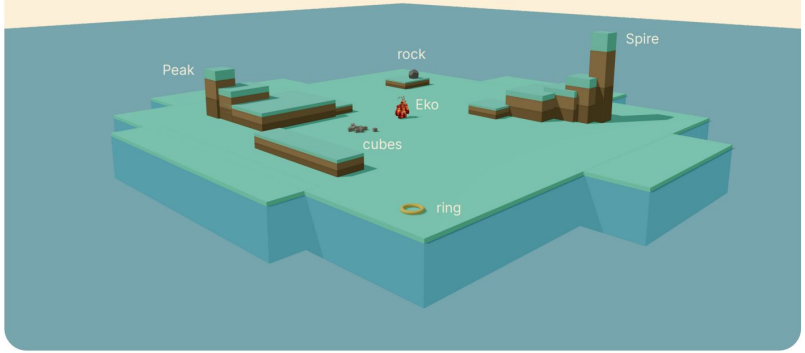
Is this machine playing?

Nathan Cloos* Antonio Norelli*
Daniel Durbin Jacob Andreas Daniela Rus Phillip Isola

MIT

We placed a modern AI coding assistant in an unintended role: as the mind of a body on an unknown digital island. With only a minimal instruction mentioning no specific task, reward, or activity, the machine started animating its virtual body. Across thirty-hour runs, the embodied AI agent climbed hills, stacked blocks into towers, drew mandalas, reinterpreted sports, ran experiments on the physics of its world, and learned techniques that later expanded what it could accomplish. These activities recurred across thirteen agents but diverged into distinct histories. We examine whether this behavior satisfies classical criteria for play and ask whether play can become a mode of machine development.

website code



Cloos*, Norelli* et. al, Is this machine playing?; https://machineplaying.org/

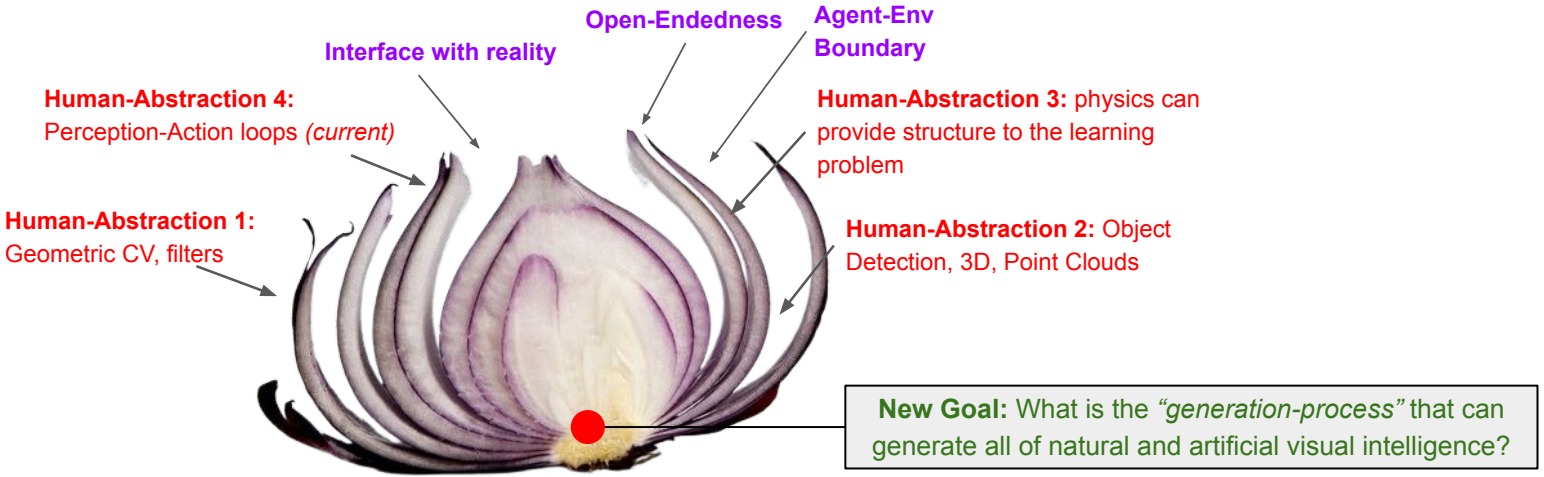
Takeaway #3: Don’t supply the objective; set simple rules and design environments that generates everything else



- Supply the rules, not the objectives
  - Example: last bot standing. The model designs everything else: body, controller, strategy
- Process that generates it will probably have some multi-agent interaction in it
- Curiosity, opendedness, motivation are hard problems

Takeaways:

- **Takeaway #1:** Remove all abstractions; peel the onion
- **Takeaway #2:** Remove all sensing and actuating interface
- **Takeaway #3:** Carefully examine the Agent-Env boundary
- **Takeaway #4:** Supply physically grounded rules not objectives; make it multiagent so the process generates sub-objectives.



Peeling back the onion of Computer Vision // Contact: ktiwary@mit.edu

Open-Endedness

Agent-Env Boundary

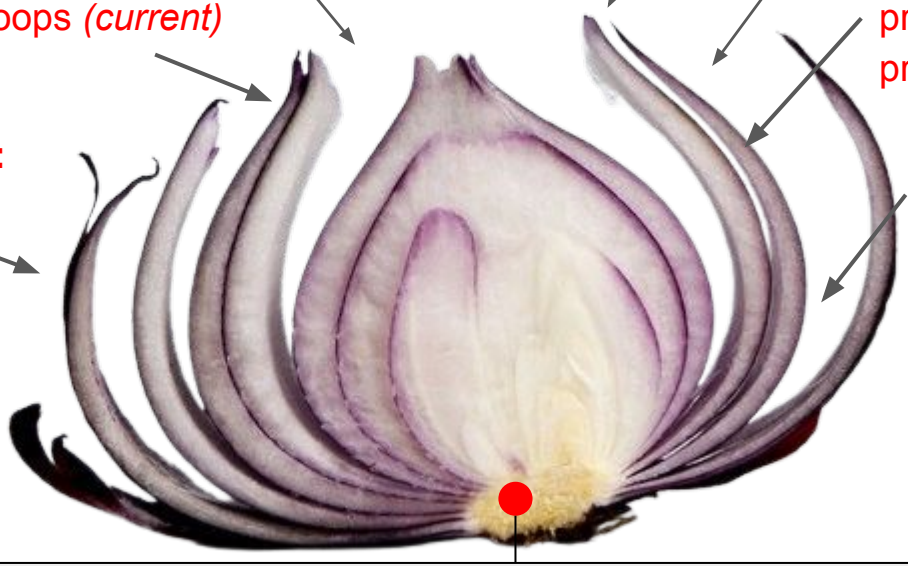
Interface with reality

Human-Abstraction 4: Perception-Action loops *(current)*

Human-Abstraction 3: physics can provide structure to the learning problem

Human-Abstraction 1: Geometric CV, filters

Human-Abstraction 2: Object Detection, 3D, Point Clouds



This talk: What is the *"generation-process"* that can generate all of natural and artificial visual intelligence?