Machine Learning Models of the Statistics of Human Behavioral Responses in Experimental Tasks

Drew Cranford, Ken Ford, Kevin Gluck, Will Hancock, Christian Lebiere, Mark Orr, Pete Pirolli, Andrea Stocco, Frank Ritter

Intergenerational Centaur Response Group

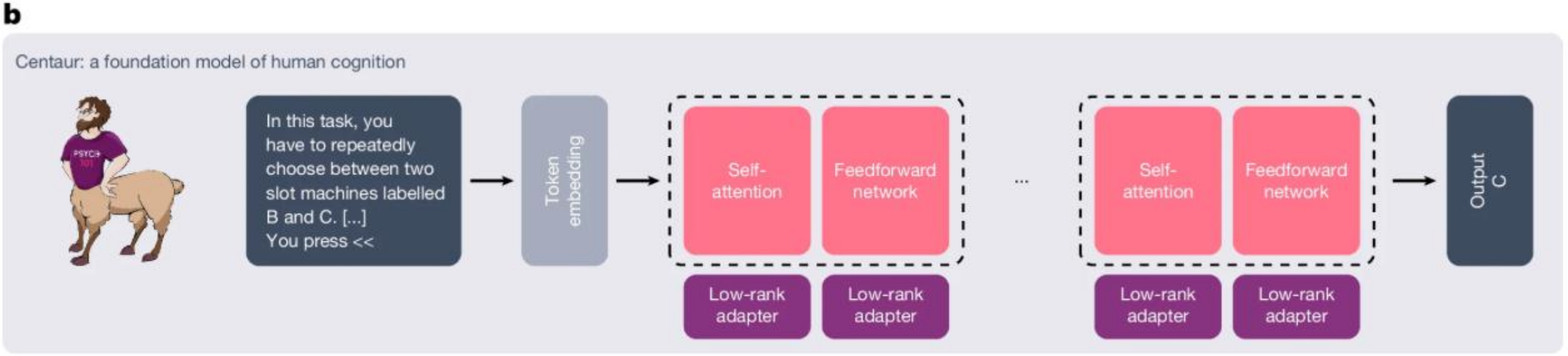
ICRG, Mark Orr, Organizer

Centaur is

- Pre-trained LLM + fine tuned to 90% of Psych-101 data set
- Psych-101: trial-by-trial data, ~ 160 experiments, 60,000 participants, 10 MIL trials

Key Requirement

A first step towards such a theory is to create a computational model that can predict human behaviour in a wide range of settings. Here we introduce Centaur, a computational model that can predict and simulate human behaviour in any experiment expressible in natural language. We derived Centaur by fine-tuning a state-of-the-art language model on a large-scale dataset called Psych-101. Psych-101 has an



a, Psych-101 comprises trial-by-trial data from 160 psychological experiments with 60,092 participants making 10,681,650 choices in total and involving 253,597,411 text tokens. It contains domains such as multi-armed bandits, decision-making, memory, supervised learning, Markov decision processes and others (the examples shown have been stylized and abbreviated for readability). **b**, Centaur is a foundation of model human cognition that is obtained by adding low-rank adapters to a state-of-the-art language model and fine-tuning it on Psych-101.

Fine-tuning:

- Quantized low-rank adaptation (QLoRA)--added low-rank adapters to all non-embedding layers
- Newly added parameters ~ 0.15% of the base model's parameters.
- Trained one epoch on the entire dataset (10% holdout) using a standard cross-entropy loss.
- Masked out the loss for all tokens not corresponding to human responses,
- Training process took approximately five days on an A100 80GB GPU

Some of the tasks

Shepard categorization, (CHICKEN) Drifting four-armed bandit, Multiple-cue judgment, Recall and recognition, N-back, Digit span, Go/no-go, Recent probes, Horizon task, Gardening task, Columbia card task, Balloon analog risk task, Experiential-symbolic task (CHICKEN), Two-armed bandit, Conditional associative learning, THINGS odd-one-out, Multi-attribute decision-making, (CHICKEN), Grammar judgement, Two-step task, Risky choice, Tile-revealing task, Probabilistic instrumental learning, Medin categorization, Zoopermarket, choices, Episodic long-term memory, Intertemporal choice, Horizon task, Structured bandit, Horizon task, Weather prediction task, (CHICKEN) Iowa gambling task, Virtual (CHICKEN) network, Multi-task reinforcement learning, Horizon task, Aversive learning, Spatially correlated multi-armed bandit, Serial reaction time task, Decisions from description, Decisions from experience, Changing bandit, Probabilistic reasoning, (CHICKEN) Two-step task

Go/no-go

Data source: [25]

Number of experiments: 1
Number of participants: 463
Number of choices: 150517

Example prompt:
In this task, you need to emit responses to certain stimuli and omit responses to others.
You will see one of two colours, colour1 or colour2, on the screen in each trial.
You need to press button X when you see colour1 and press nothing when you see colour2.
You need to respond as quickly as possible.
You will be doing 10 practice trials followed by 350 test trials.

You see colour1 and press nothing.
You see colour2 and press <<X>> in 753.0ms.
You see colour2 and press <<X>> in 381.0ms.
You see colour2 and press nothing.
You see colour1 and press <<X>> in 473.0ms.
You see colour1 and press <<X>> in 713.0ms.
You see colour2 and press nothing.
You see colour1 and press <<X>> in 364.0ms.
You see colour2 and press nothing.
You see colour1 and press <<X>> in 378.0ms.
You see colour1 and press <<X>> in 794.0ms.

Recent probes

Data source: [25]

Number of experiments: 1
Number of participants: 471
Number of choices: 34714

Example prompt:
You will repeatedly observe sequences of six letters.
You have to remember these letters before they disappear.
Afterward, you will be prompted with one letter. You have to answer whether the letter was part of the six previous letters.
If you think it was, you have to press C. If you think it was not, press Q.

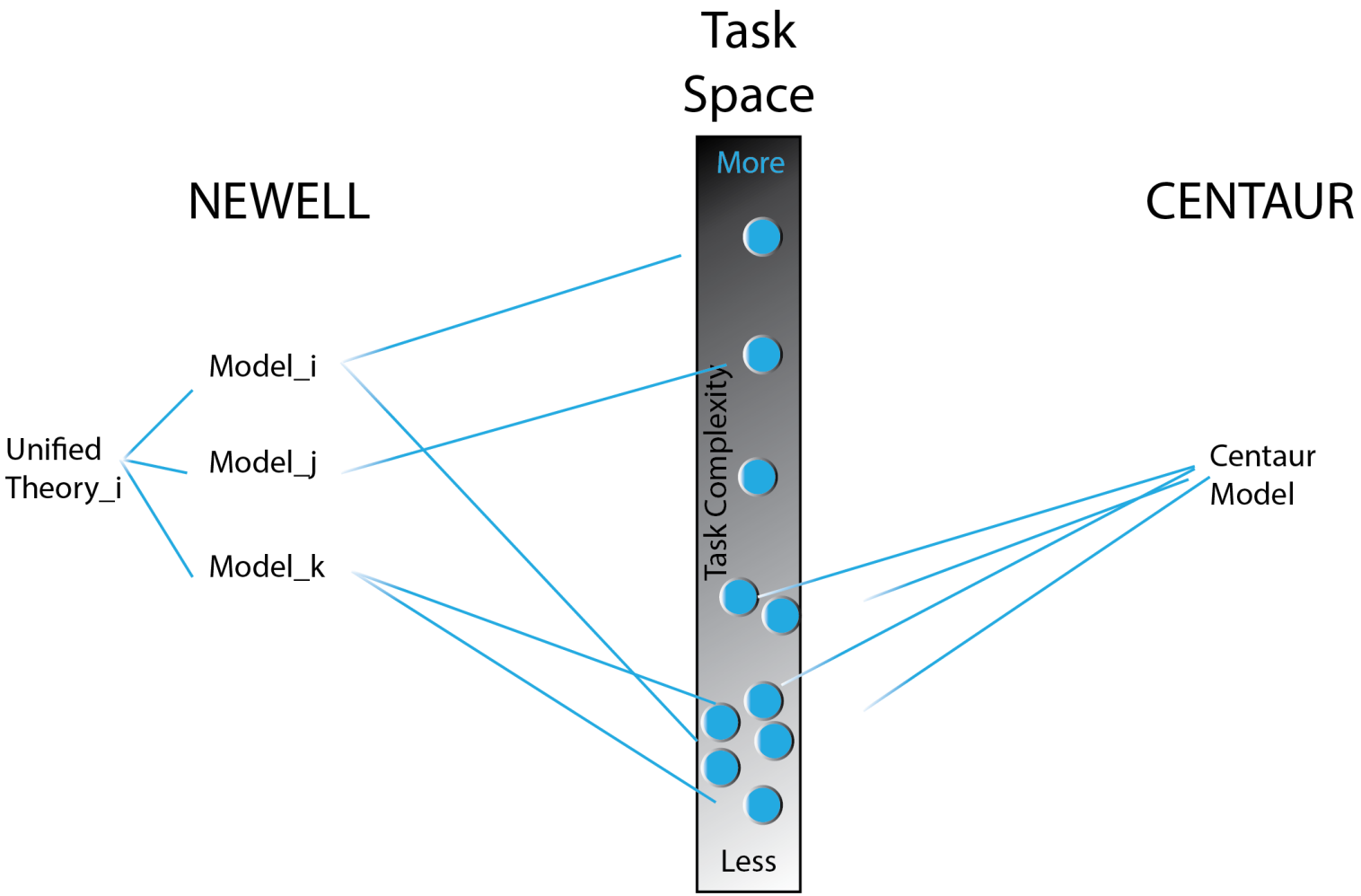
You are shown the letters ['C', 'I', 'Q', 'F', 'W', 'Z']. You see the letter Y. You press <<Q>>.
You are shown the letters ['I', 'Q', 'C', 'D', 'M', 'V']. You see the letter U. You press <<Q>>.
You are shown the letters ['I', 'O', 'C', 'X', 'A', 'Q']. You see the letter M. You press <<C>>.
You are shown the letters ['Z', 'C', 'W', 'I', 'J', 'O']. You see the letter C. You press <<Q>>.
You are shown the letters ['Q', 'M', 'F', 'V', 'P', 'E']. You see the letter W. You press <<C>>.
You are shown the letters ['W', 'F', 'U', 'M', 'B', 'Q']. You see the letter V. You press <<Q>>.

Centaur claims

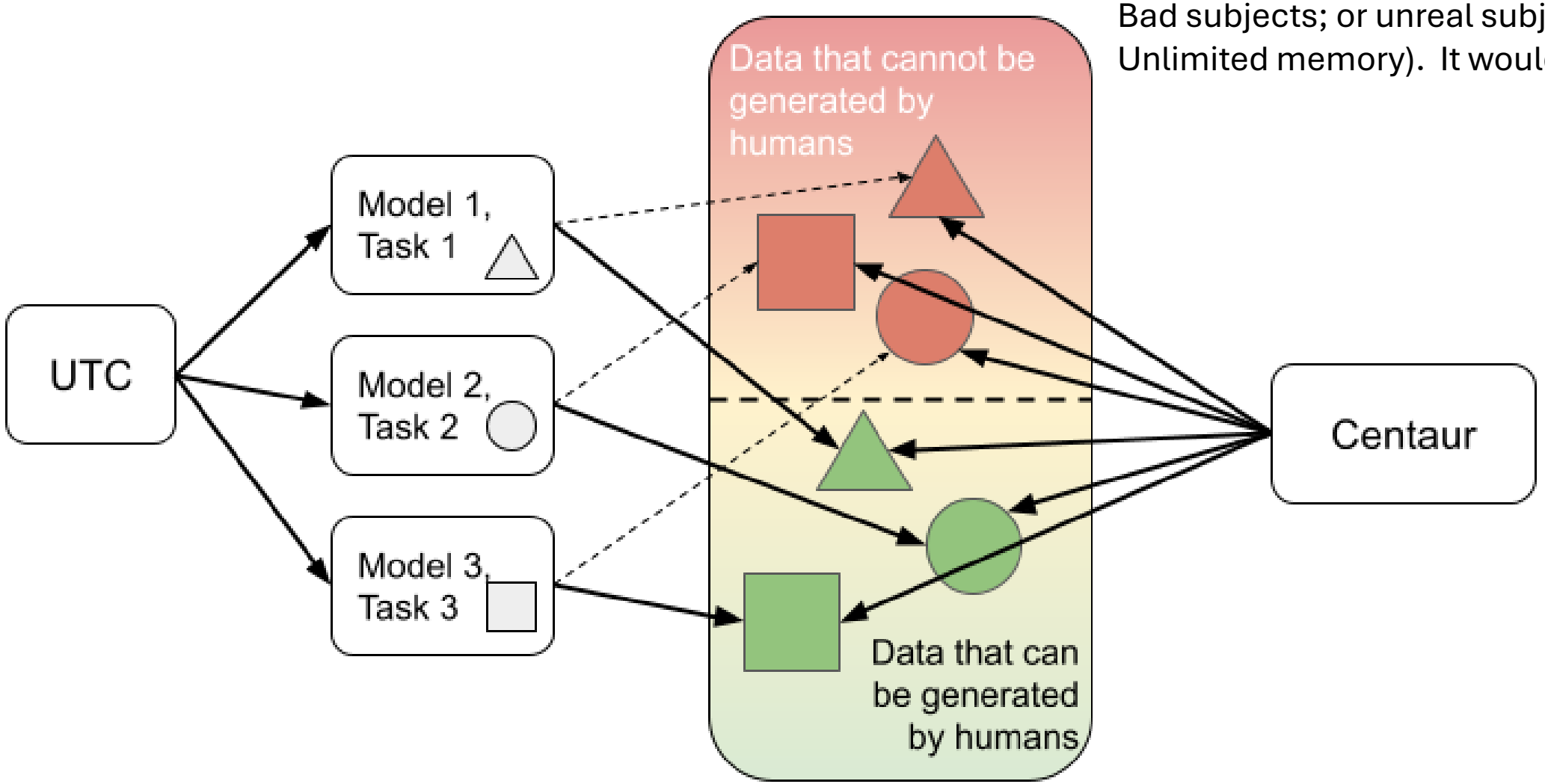
- Psy Arxiv claim (2024):
  - first true unified model of human cognition
- Nature claim (2025):
  - softened; strong claim only in conclusion
  - Domain general model of cognition → a next step towards unified theory

Comment #1: Category Mistakes

- Category Mistakes
  - Centaur is moving towards a unified model of cognition, not theory
  - Lacks an architectural basis (the unified theory)
  - Unified model because it is one model to serve all tasks


Task
Space
NEWELL
CENTAUR
More
Task Complexity
Less
Unified
Theory_i
Model_i
Model_j
Model_k
Centaur
Model

e.g., just give Centaur exclusively Bad subjects; or unreal subjects (e.g, Unlimited memory). It would



Comment #2: Newell's test

- Centaur invokes Newell's Test wrt Centaur the model of cognition
  - They claimed that it passed the Newell test
- Newell offered criteria for a theory of cognition
- We offer comments on all 12 criteria
  - Separate centaur-the-system from centaur-the-trained-product

Newell test

Together with his call for unified theories of cognition [16, 17], Newell outlined a set of criteria that a unified computational model should fulfill. Centaur is the first model to satisfy the majority of these criteria (see Table 1). Most importantly, it (1) behaves as an almost arbitrary function of the environment, (2) operates in real time, and (3) relies on vast amounts of knowledge about the world. We provide an extended discussion on Newell's criteria in the following.

Criterion	Fulfilled by Centaur
Behave as an (almost) arbitrary function of the environment	✓
Operate in real time	✓
Exhibit rational, that is, effective adaptive behavior	✓
Use vast amounts of knowledge about the environment	✓
Behave robustly in the face of error, the unexpected, and the unknown	✓
Integrate diverse knowledge	✓
Use (natural) language	✓
Exhibit self-awareness and a sense of self	●
Learn from its environment	✓
Acquire capabilities through development	✗
Arise through evolution	✗
Be realizable within the brain	✓

Supplementary Table 1 Newell test for a theory of cognition.

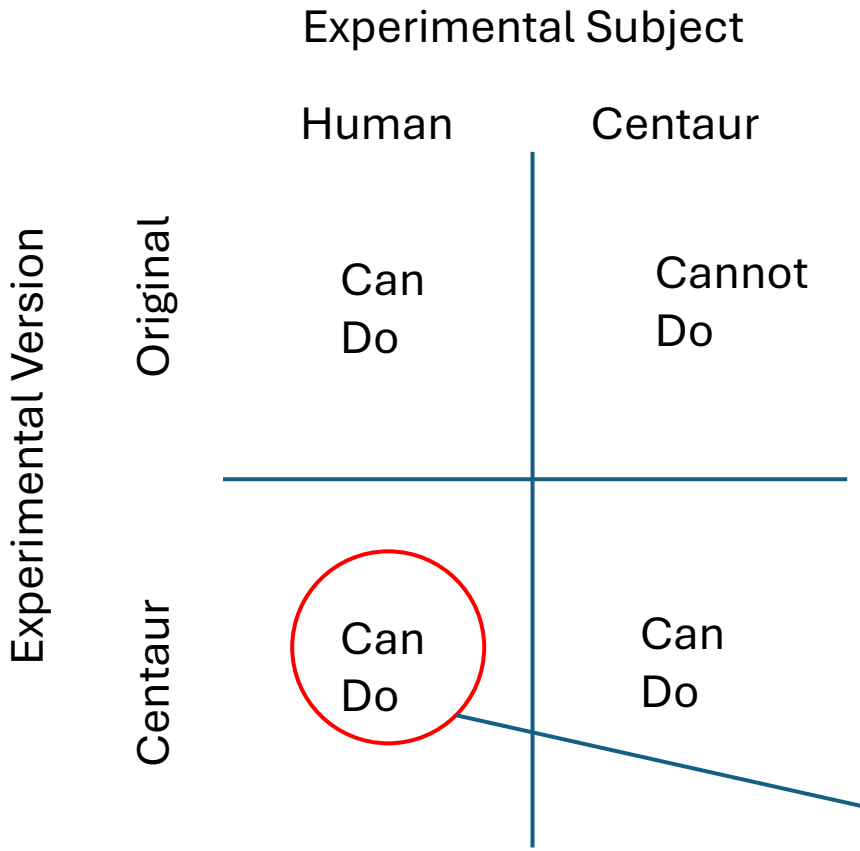
Comment #3: The Its Analogous to the LLM Consciousness Argument Argument

- [No one seemed to appreciate this comment. But, I'll raise it to get a second opinion.]
- The allusions are illusions argument

Comment #4: Measurement and its Discontents (ref to Freud)

Centaur Claim:

It can simulate human behavior in any experiment given that there is a procedure for expressing the paradigm in natural language.



Centaur's missed opportunity. Let's imagine doing it…

Comment #5: Claims of Neural Alignment are overblown

- Centaur claims its internal representations become aligned to human neural activity.
- Has nonsensical sense about it:
  - LLM internal not align with our theories and models of neural processing, representation, storage, learning, etc.
- Method of alignment favors centaur
  - No explanatory power
  - Flies in face of practice in cognitive neuroscience

Comment #6: Centaur is Non-Mechanistic and Atheoretical

- Describe what role theory plays in cog sci.
- From cog sci perspective, what is centaur's role
  - If predictions poor, whatever next? What insights would we glean? What would we change or modify in the model?
  - If predictions are good, where do we go next?
- Centaur is a theoretical dead end; or as one might say and Epistemic Black Box

(Alternative) Trends for the future?

- Other related new LLM approaches
  - Rmus, 2025--Guided generation of Computational Cognitive Models (GeCCo)
  - Guided LLM prompted cog model building
- Other (so-called?) cognitive architectures
  - Aran Nayebi (CMU)
  - Yann LeCun
- Generative Agents, e.g., Joon Sung Park
- What is a cognitive architecture?

The AlphaFold, fold this! AlphaCog, AlphaMind.

Possible in principle, impossible in fact.

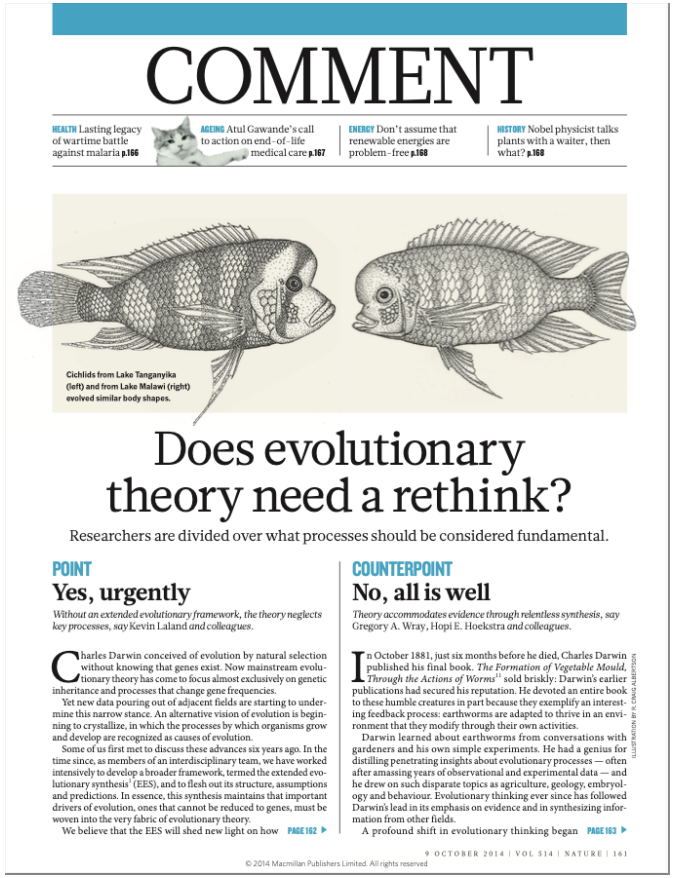
COMMON, USEFUL

Impossible in principle, possible in fact.

UNCOMMON, RISKY

Extended Evolutionary Synthesis

- J. Huxley's Evolution: The Modern Synthesis (1942)
  - set stage for 80 years (see pop version Selfish-Gene)
  - Genes incrementally change over time
- 2010s we see a dramatic shift towards teleology
- Next Gen sci enablers for understanding intelligent systems



COMMENT

HEALTH Lasting legacy of wartime battle against malaria p.166 | AGEING Atul Gawande's call to action on end-of-life medical care p.167 | ENERGY Don't assume that renewable energies are problem-free p.168 | HISTORY Nobel physicist talks plants with a waiter, then what? p.168

Cichlids from Lake Tanganyika (left) and from Lake Malawi (right) evolved similar body shapes.

Does evolutionary theory need a rethink?

Researchers are divided over what processes should be considered fundamental.

POINT

Yes, urgently

Without an extended evolutionary framework, the theory neglects key processes, say Kevin Laland *and colleagues.*

Charles Darwin conceived of evolution by natural selection without knowing that genes exist. Now mainstream evolutionary theory has come to focus almost exclusively on genetic inheritance and processes that change gene frequencies.

Yet new data pouring out of adjacent fields are starting to undermine this narrow stance. An alternative vision of evolution is beginning to crystallize, in which the processes by which organisms grow and develop are recognized as causes of evolution.

Some of us first met to discuss these advances six years ago. In the time since, as members of an interdisciplinary team, we have worked intensively to develop a broader framework, termed the extended evolutionary synthesis[1] (EES), and to flesh out its structure, assumptions and predictions. In essence, this synthesis maintains that important drivers of evolution, ones that cannot be reduced to genes, must be woven into the very fabric of evolutionary theory.

We believe that the EES will shed new light on how PAGE 162 ▶

COUNTERPOINT

No, all is well

Theory accommodates evidence through relentless synthesis, say Gregory A. Wray, Hopi E. Hoekstra *and colleagues.*

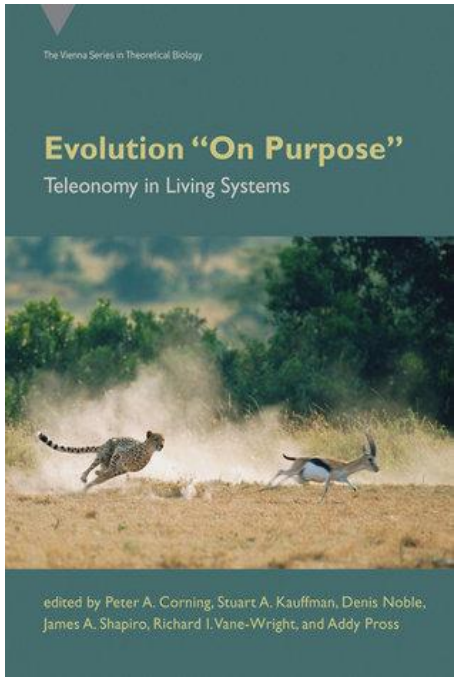
In October 1881, just six months before he died, Charles Darwin published his final book. *The Formation of Vegetable Mould, Through the Actions of Worms*[1] sold briskly: Darwin's earlier publications had secured his reputation. He devoted an entire book to these humble creatures in part because they exemplify an interesting feedback process: earthworms are adapted to thrive in an environment that they modify through their own activities.

Darwin learned about earthworms from conversations with gardeners and his own simple experiments. He had a genius for distilling penetrating insights about evolutionary processes — often after amassing years of observational and experimental data — and he drew on such disparate topics as agriculture, geology, embryology and behaviour. Evolutionary thinking ever since has followed Darwin's lead in its emphasis on evidence and in synthesizing information from other fields.

A profound shift in evolutionary thinking began PAGE 163 ▶

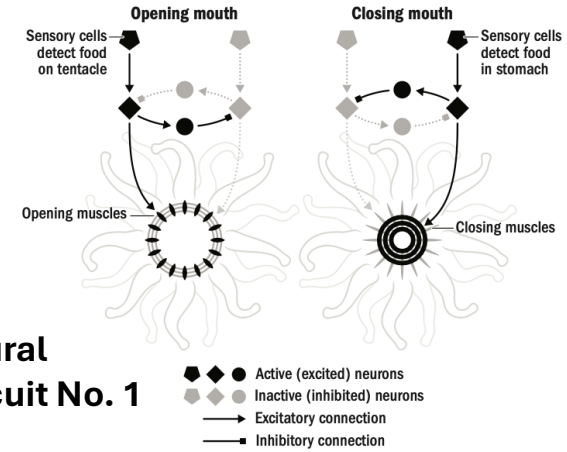
9 OCTOBER 2014 | VOL 514 | NATURE | 161

Laland, et al. vs Wray et al. (2014). Nature, 514(7521),161-164.

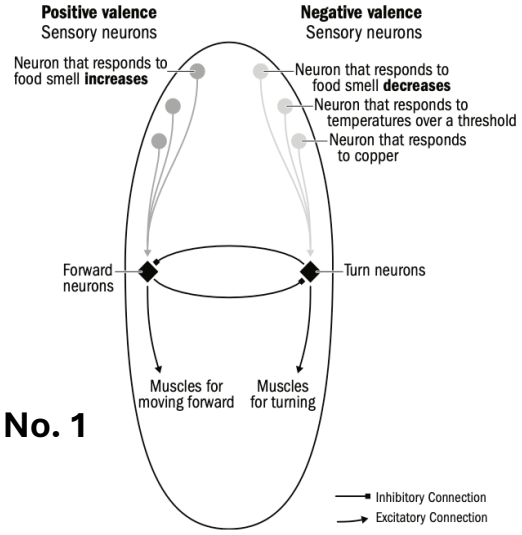


Corning, et al (2023). Evolution "On Purpose". MIT Press

When studying any species, we are studying millions, if not billions of years of evolution

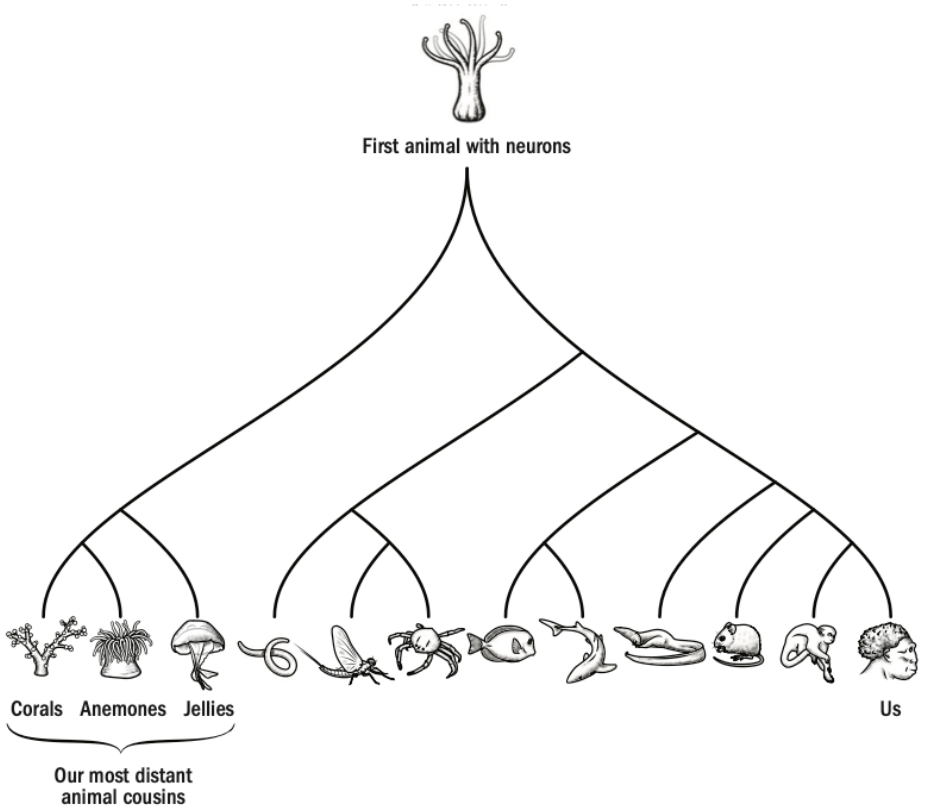


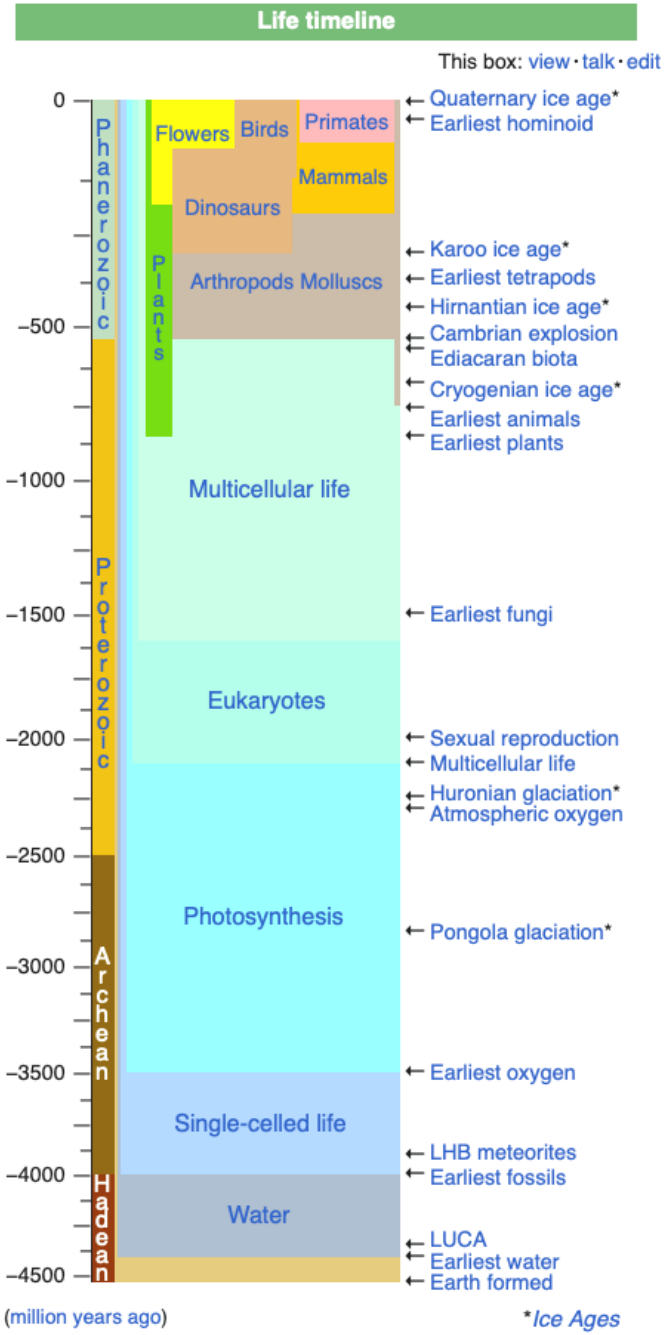
Neural Circuit No. 1



Brain No. 1

"Let's start with Artificial Rat-level intelligence (ARI), then move onto Artificial Cat-level intelligence (ACI), and so on to Artificial Human-level Intelligence (AHI)." -Yann Lecun





Rodney Brooks (1990):

”It is instructive to reflect on the way in which earth-based biological evolution spent its time. … This suggests that problem solving behavior, language, …are all rather simple once the essence of being and reacting are available.”

Elephants Don't Play Chess

Rodney A. Brooks
MIT Artificial Intelligence Laboratory, Cambridge, MA 02139, USA

Robotics and Autonomous Systems 6 (1990) 3-15

Keywords: Situated activity; Mobile robots; Planning; Subsumption architecture; Artificial Intelligence.

Rodney A. Brooks was born in Adelaide, Australia. He studied Mathematics at the Flinders University of South Australia and received a Ph.D. from Stanford in Computer Science in 1981. Since then he has held research associate positions at Carnegie Mellon University and the Massachusetts Institute of Technology and faculty positions at Stanford and M.I.T. He is currently an Associate Professor of Electrical Engineering and Computer Science at M.I.T. and a member of the Artificial Intelligence Laboratory where he leads the mobile robot group. He has authored two books, numerous scientific papers, and is the editor of the *International Journal of Computer Vision*.

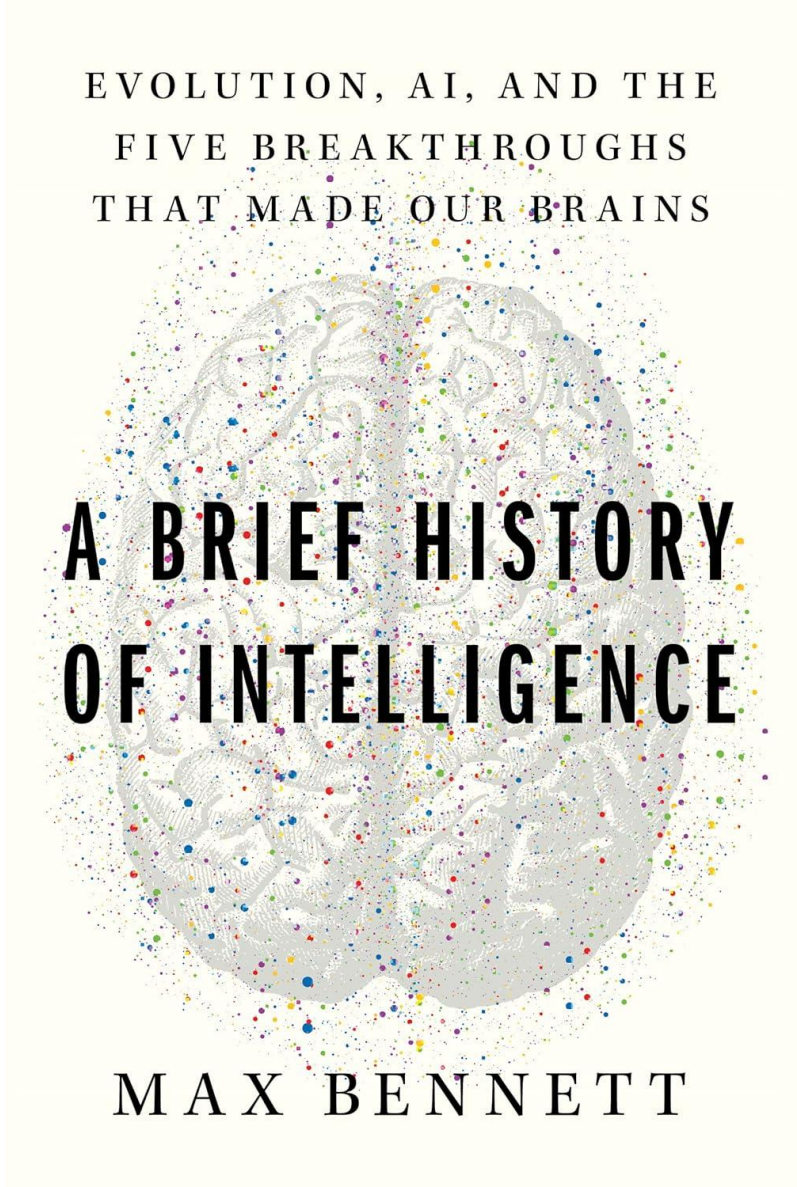
There is an alternative route to Artificial Intelligence that diverges from the directions pursued under that banner for the last thirty some years. The traditional approach has emphasized the abstract manipulation of symbols, whose grounding, in physical reality has . rarely been achieved. We explore a research methodology which emphasizes ongoing physical interaction with the environment as the primary source of constraint on the design of intelligent systems. We show how this methodology has recently had significant successes on a par with the most successful classical efforts. We outline plausible future work along these lines which can lead to vastly more ambitious systems.

1. Introduction

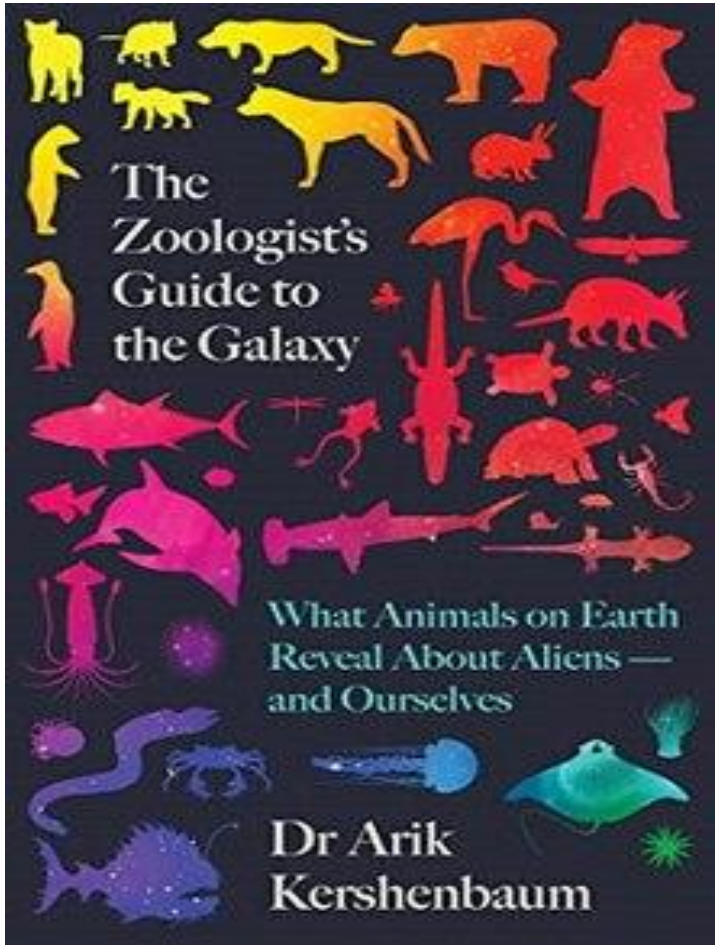
Artificial Intelligence research has foundered in a sea of incrementalism. No one is quite sure where to

But there is an alternative view, or dogma, variously called *nouvelle AI*, *fundamentalist AI*, or in a weaker form *situated activity* '. *It is based on the*

New Deep Look from Evo Bio (2023):



A(lien)I



Astrobio-evobio look at the AI problem:

"…evolutionary processes that are observed operating on Earth are universal, and a necessary requirement for the presence of complex life on any planet."

Extras and supplementals

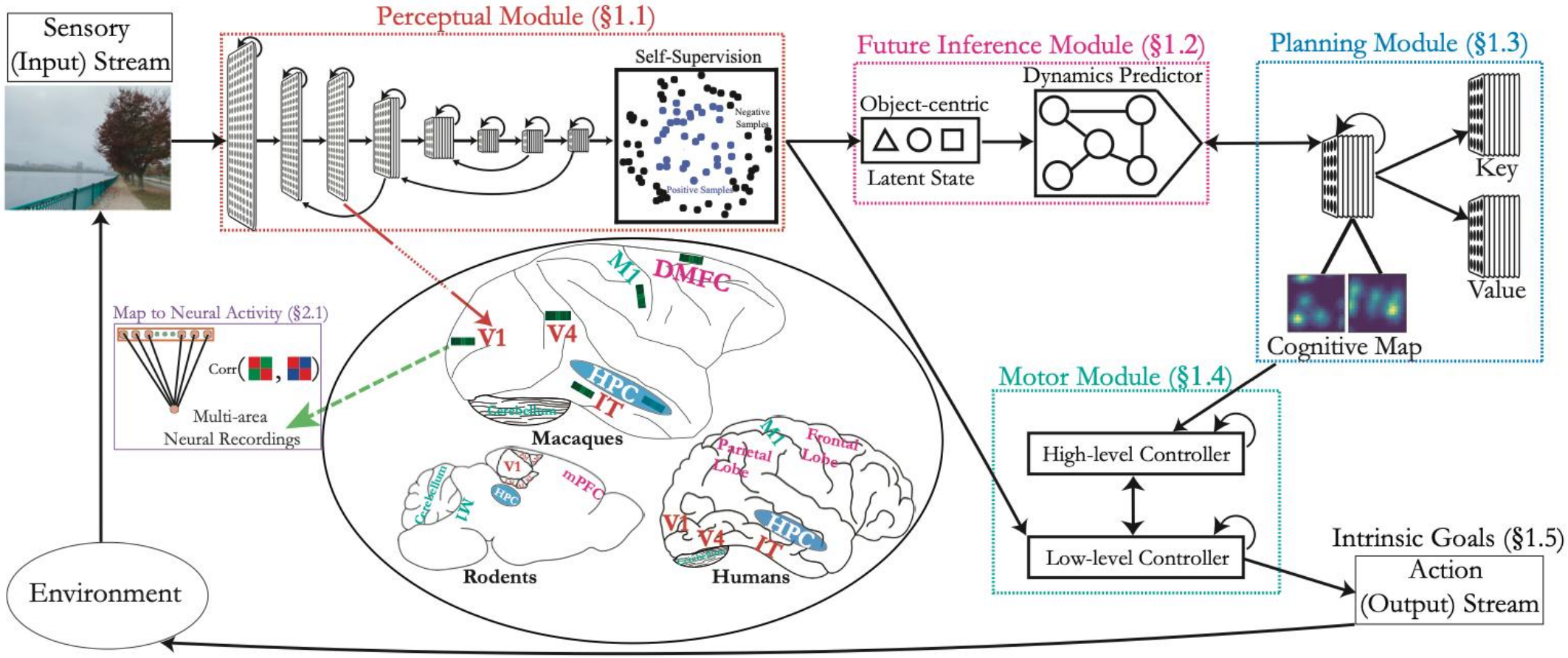


Figure 1: **Integrative, embodied agents to reverse-engineer natural intelligence.** A schematic of an *example* integrative, embodied agent consisting of a recurrent, self-supervised perceptual system (§1.1) that outputs an object-centric latent upon which a future inference module (§1.2) operates to predict the next state of the environment. The planning module (§1.3) hierarchically organizes these representations to plan future actions, which are then passed to effectors that output intrinsically-guided (§1.5) motor commands to perform actions in a biomechanically-realistic animal body (§1.4). Solid black arrows represent *possible* connections between modules. Each representation in these modules is obtained through task-optimization and then mapped (up to the suitable transform, cf. §2.1) to neural activity across multiple brain areas (dotted green arrow). In this example, we show rhesus macaque, rodent, and human brains, with proposed matched representative areas color-coded to each module across species. While I expect the specifics of the modules in each integrative agent to differ for each species it is compared to, the long-term, overarching goal of this approach is that by comparing integrative agents to multi-area neural and behavioral data from multiple species, we are positioned to identify *common* algorithms of natural intelligence conserved across species.

Nayebi's 2024 Approach

A Path Towards Autonomous Machine Intelligence
Version 0.9.2, 2022-06-27

Yann LeCun
Courant Institute of Mathematical Sciences, New York University yann@cs.nyu.edu
Meta - Fundamental AI Research yann@fb.com

June 27, 2022

Abstract

How could machines learn as efficiently as humans and animals? How could machines learn to reason and plan? How could machines learn representations of percepts and action plans at multiple levels of abstraction, enabling them to reason, predict, and plan at multiple time horizons? This position paper proposes an architecture and training paradigms with which to construct autonomous intelligent agents. It combines concepts such as configurable predictive world model, behavior driven through intrinsic motivation, and hierarchical joint embedding architectures trained with self-supervised learning.

Keywords: Artificial Intelligence, Machine Common Sense, Cognitive Architecture, Deep Learning, Self-Supervised Learning, Energy-Based Model, World Models, Joint Embedding Architecture, Intrinsic Motivation.

1 Prologue

This document is not a technical nor scholarly paper in the traditional sense, but a position paper expressing my vision for a path towards intelligent machines that learn more like animals and humans, that can reason and plan, and whose behavior is driven by intrinsic objectives, rather than by hard-wired programs, external supervision, or external rewards. Many ideas described in this paper (almost all of them) have been formulated by many authors in various contexts in various form. The present piece does not claim priority for

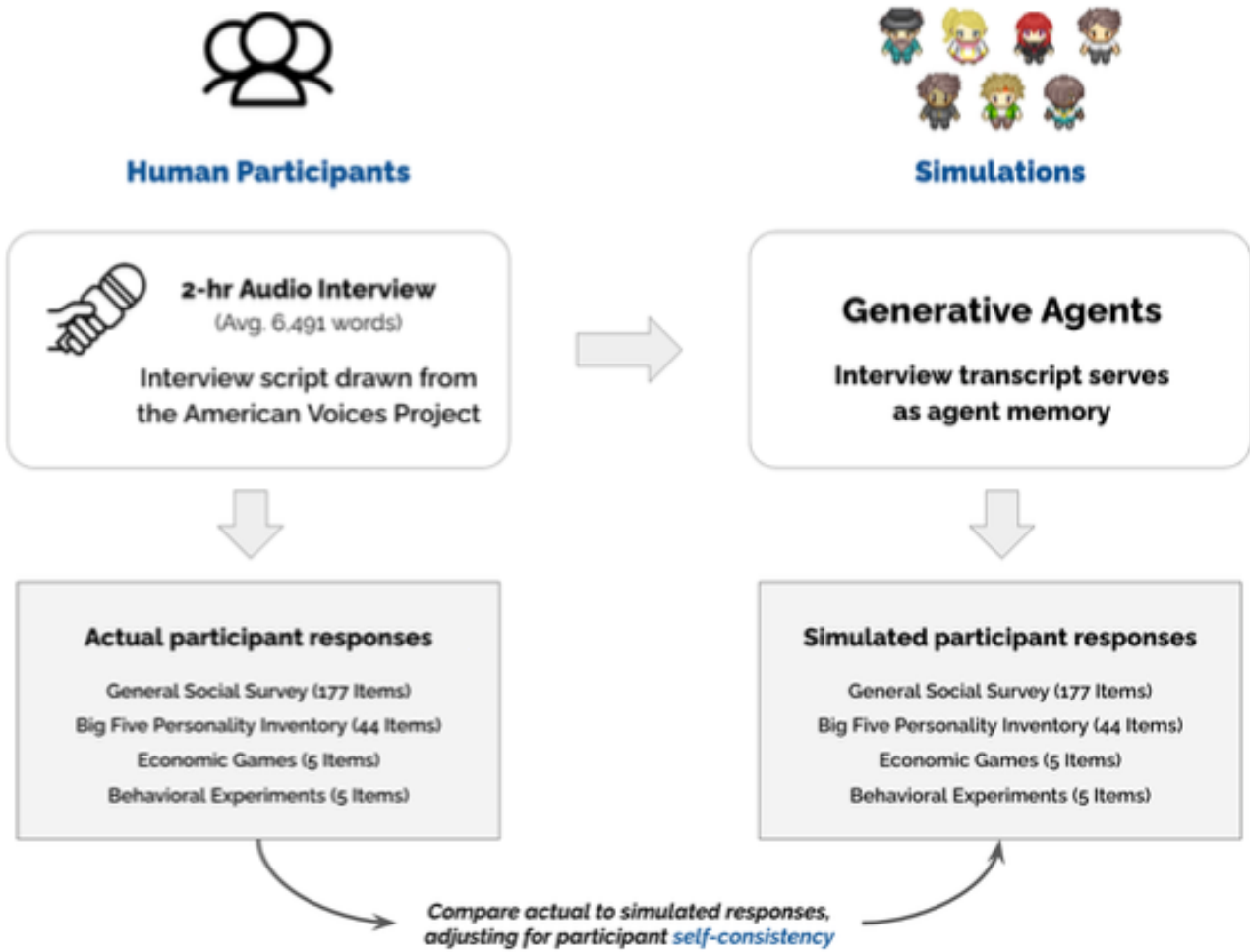


Figure 1. The process of collecting participant data and creating generative agents begins by recruiting a stratified sample of 1,052 individuals from the U.S., selected based on age, census division, education, ethnicity, gender, income, neighborhood, political ideology, and sexual identity. Once recruited, participants complete a two-hour audio interview with our AI interviewer, followed by surveys and experiments. We create generative agents for each participant using their interview data. To evaluate these agents, both the generative agents and participants complete the same surveys and experiments. For the human participants, this involves retaking the surveys and experiments again two weeks later. We assess the accuracy of the agents by comparing agent responses to the participants' original responses, normalizing by how consistently each participant successfully replicates their own responses two weeks later.

Joon Sung Park, 202X?