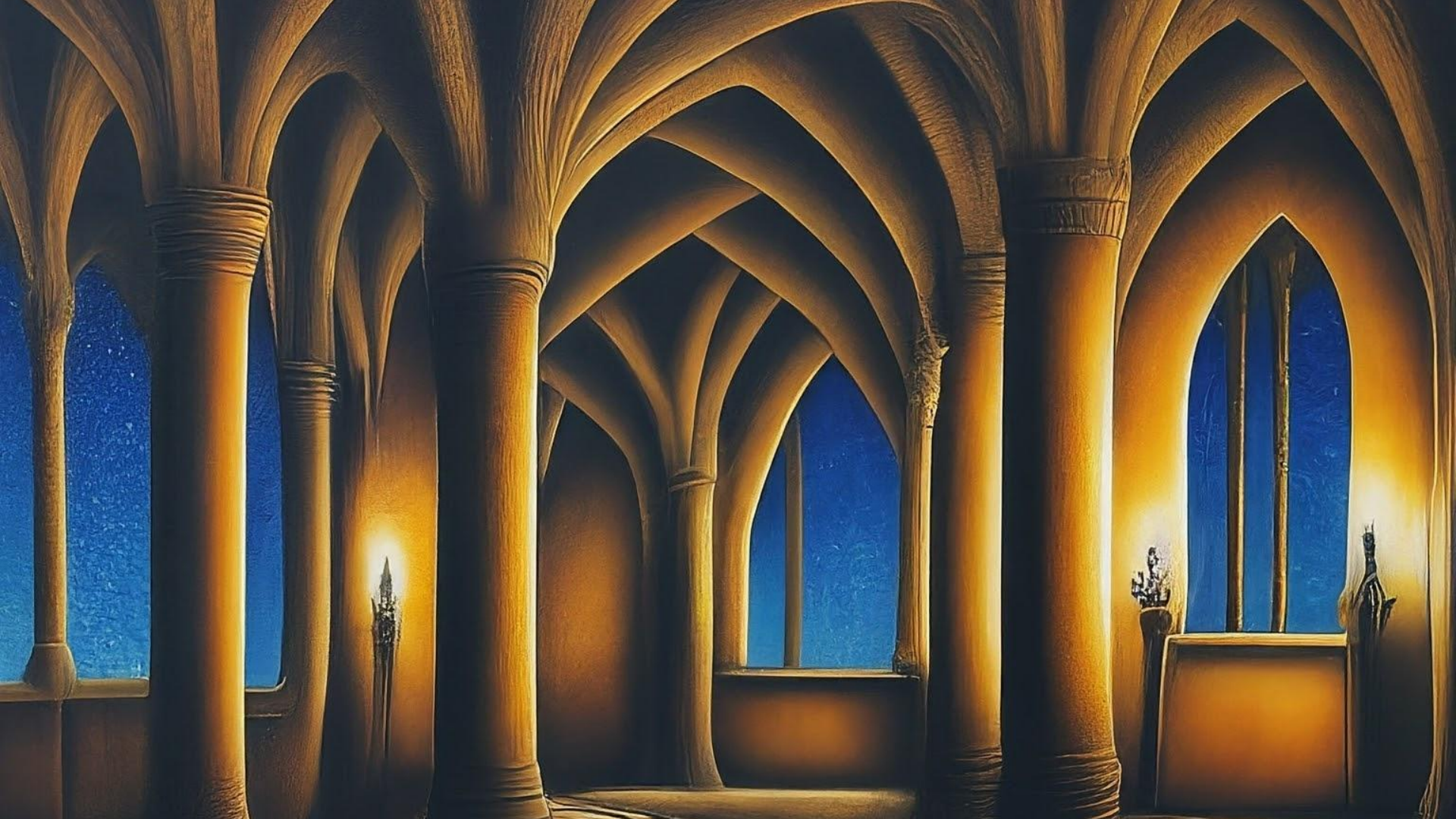
Google DeepMind

Symbolic Behaviour in AI

Some possible lessons for whales?

Andrew Lampinen

Simons Institute for the Theory of Computing
June 29th, 2024



DeepMind

https://arxiv.org/pdf/2102.03406.pdf

Symbolic Behaviour in Artificial Intelligence

Adam Santoro[*,1], **Andrew Lampinen**[*,1], **Kory Mathewson**[1], **Timothy Lillicrap**[1] **and David Raposo**[1]

[*]Equal contributions, [1]DeepMind

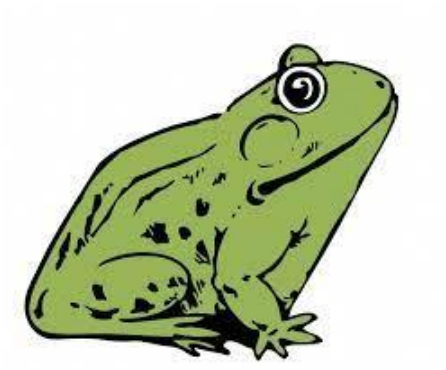
DeepMind

1 Symbols

What is a symbol?

```python
def loss_fn(images, labels):
  mlp = hk.Sequential([
      hk.Linear(300), jax.nn.relu,
      hk.Linear(100), jax.nn.relu,
      hk.Linear(10),
  ])
  logits = mlp(images)
  return jnp.mean(softmax_cross_entropy(logits, labels))
```

And to whom?



Symbols

- Newell and Simon define symbols as a set of interrelated “physical patterns” that could “designate any expression whatsoever”

- But designate to whom?

Symbols

Reconstructing Physical Symbol Systems

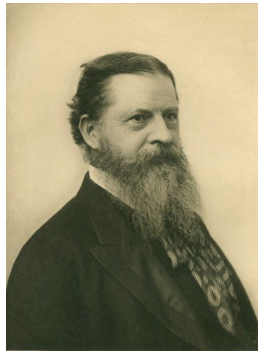
DAVID S. TOURETZKY AND DEAN A. POMERLEAU
Carnegie Mellon University

1. INTRODUCTION

In attempting to force ALVINN[1] into their already bulging symbolist tent, Vera and Simon (1993a, 1993b, 1993c) have burst the seams of the physical symbol system hypothesis (PSSH; Newell, 1980a; Newell & Simon, 1976).

The primary error in Vera and Simon's (1993a) argument is to mistake *signal* for *symbol*: "We call patterns symbols when they can designate or denote" (p. 9). This leaves no possibility of nonsymbolic representations.

Charles Sanders Peirce on Symbols



- A symbol is a representation that has its meaning established by *convention (“agreed upon link”)*.
- Not by superficial similarity (an icon), or by physical/temporal correlation/analogy (an index).

Icon

Indicator/Index

Symbol

Why does definition of symbols matter?

- At the time we wrote this paper, there was a fair amount of hand-wringing about the need for “symbolic” inductive biases to allow deep learning to “really” do symbol manipulation. Which architectures are/can be symbolic depends on the definition of symbols!
- But maybe the focus on architecture is itself misleading…

BEHAVIORAL AND BRAIN SCIENCES (2017), Page 1 of 72
doi:10.1017/S0140525X16001837, e253

Building machines that learn and think like people

Brenden M. Lake
Department of Psychology and Center for Data Science, New York University, New York, NY 10011
brenden@nyu.edu
http://cims.nyu.edu/~brenden/

Tomer D. Ullman
Department of Brain and Cognitive Sciences and The Center for Brains, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139
tomeru@mit.edu
http://www.mit.edu/~tomeru/

Joshua B. Tenenbaum
Department of Brain and Cognitive Sciences and The Center for Brains, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139
jbt@mit.edu
http://web.mit.edu/cocosci/josh.html

Samuel J. Gershman
Department of Psychology and Center for Brain Science, Harvard University, Cambridge, MA 02138, and The Center for Brains, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139
gershman@fas.harvard.edu
http://gershmanlab.webfactional.com/index.html

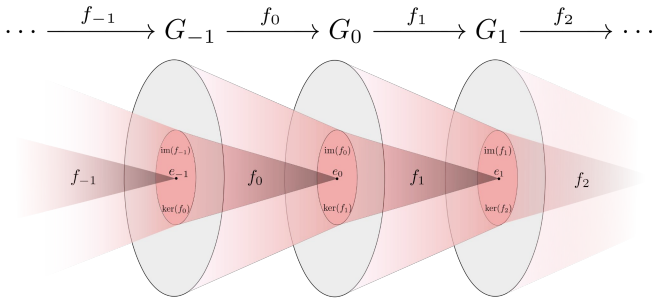
Neurosymbolic AI: The 3^{rd} Wave

Artur d’Avila Garcez[1] and Luís C. Lamb[2]
[1] City, University of London, UK
a.garcez@city.ac.uk
[2] Federal University of Rio Grande do Sul, Brazil
luislamb@acm.org

December, 2020

Instead, should we focus on behaviour

- What we ultimately care about are the behavioural consequences of symbols—what do symbol users do?
- Taking the perspective that symbols are conventional affords different types of **behaviour** than classical perspectives.
- And we suggest that these behaviours are actually more aligned with human capabilities.



What behaviours demonstrate understanding of symbols?

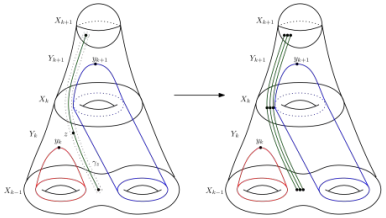
DeepMind

2 Symbolic Behaviour

Receptive

Constructive

Meaningful



Symbolic Behaviour

Malleable

Embedded

Graded

Symbolic Behaviour is *Receptive*

The ability to appreciate existing conventions, and to receive new ones.

- For example, learning a word from a definition.

Examples are common in AI:

- We often impart our conventions onto models: classify with human labels, or imitate human language.
- Some capacity for rapid receptiveness. E.g. meta-learning in LMs or RL.
- Many animals exhibit receptiveness too with enough training; e.g. dogs learn commands.

Being receptive (or any other criterion alone) is not sufficient for symbolic behavior!

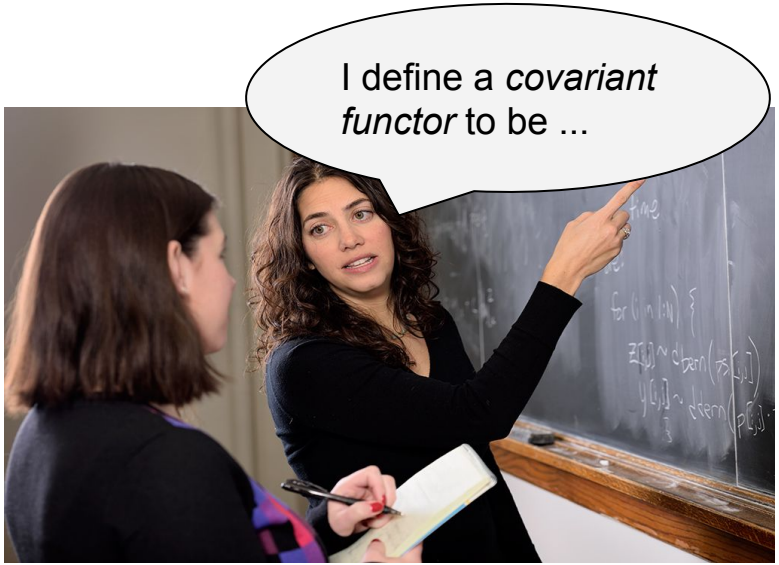
Symbolic Behaviour is *Constructive*

The ability to create new conventions, imposing new meaning on an arbitrary substrate.

- For example, define a new mathematical concept in order to prove a theorem more easily.
- This convention could be with one's self; one benefit of symbol use may be this ability to define symbols which reduce mental burden.

This capacity is much rarer in AI. Examples could include a model:

- Inventing a mathematical concept (or reinventing one we held out) to prove a theorem.
- Defining a new word to help explain something.
- Agents on a team inventing coded language that allows them to communicate secretly.



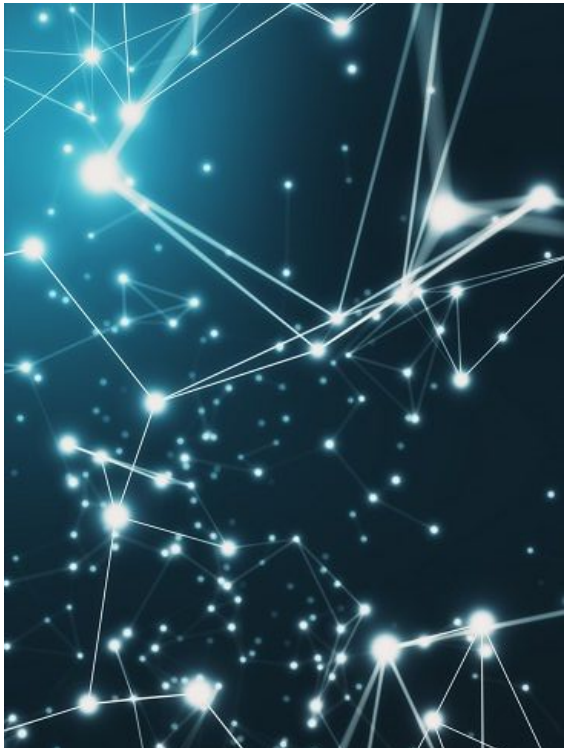
Symbolic Behaviour is *Embedded*

Symbols are part of a larger knowledge system; they cannot be understood outside of it.

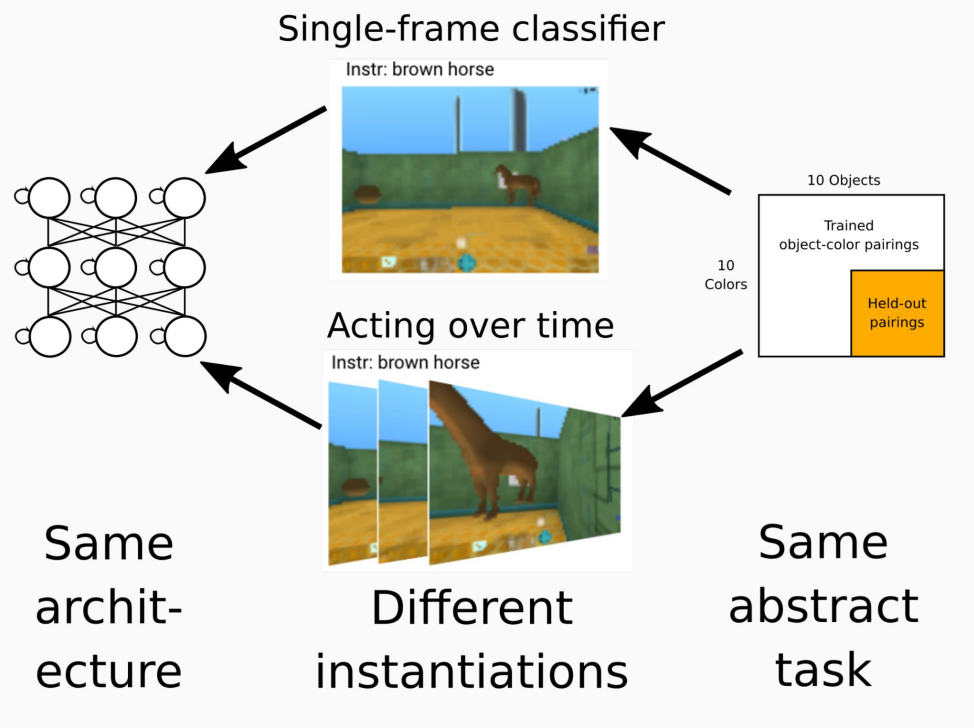
- Meaning is determined in part through interactions.
- Theoretical developments (category theory in mathematics, or genes in biology) can fundamentally change the way we think about a field.
- Humans rely on embodied understanding, such as gestures, to help us understand abstract concepts from math or science.

Examples:

- LMs and other NNs are strongly biased towards embedded understanding.
- Language models may need situated, multimodal experience to understand some of language.

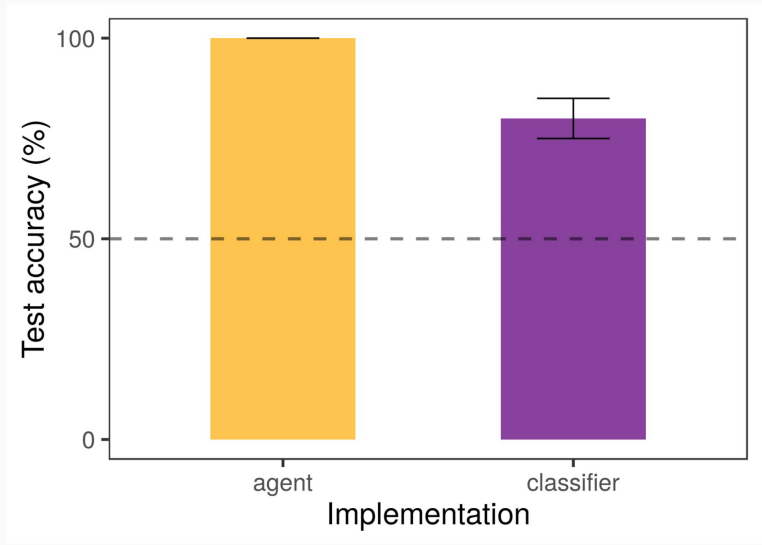


A digression on embodiment & language:



ENVIRONMENTAL DRIVERS OF SYSTEMATICITY AND GENERALISATION IN A SITUATED AGENT

Felix Hill[1], Andrew Lampinen[3]*, Rosalia Schneider[1], Stephen Clark[1]
Matthew Botvinick[1,2], James L. McClelland[1,3] & Adam Santoro[1]



Embedding symbols in a rich environment can be essential to learning their meaning!

Symbolic Behaviour is *Malleable*

Because symbol meaning is conventional, it can be contextual, or can even require a re-definition.

- Meaning is situational and pragmatic in human communication.
- More fundamentally, our models should have the *epistemic humility* to consider that meaning could be otherwise. Human progress has often required redefining symbols.

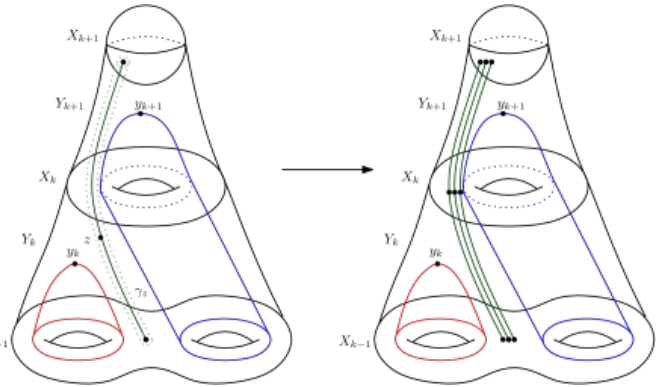
Examples:

- LMs exhibit **some** pragmatics.
- Learning implies (passive) malleability—if you change the data distribution, the model will eventually learn.
- But humans adapt with purpose. Could a model redefine a symbol **because** the old one was limiting?

Symbolic Behaviour is *Meaningful*

Meaning is essential to symbol use, even in formal domains like math.

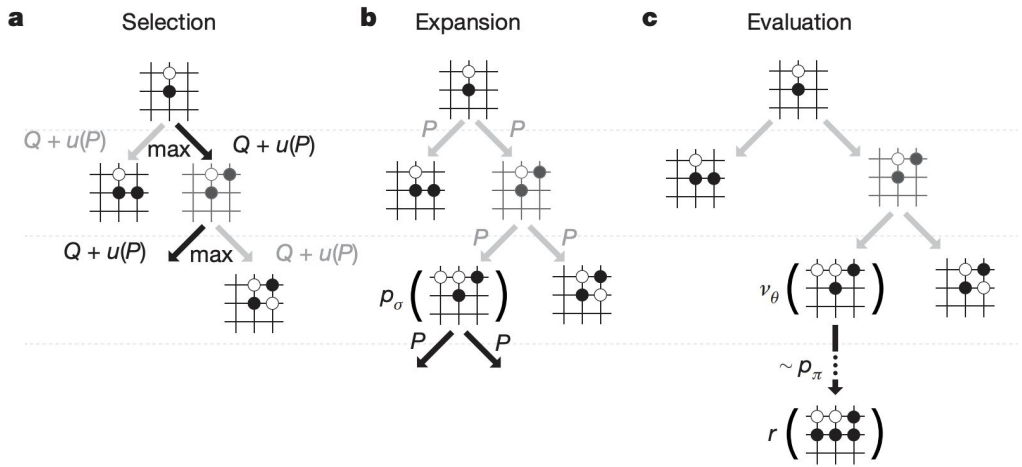
- Syntactic manipulations are only useful insofar as they are meaningful. “I never was able to successfully analyze proofs as a combinatorial ‘game’ played with symbols on paper. [To reason productively] one must essentially forget that all proofs are eventually transcribed in this formal language.” - Paul Cohen
- It’s meaning that solves the frame problem. “Strict formalism can’t explain which of many formulas matter [...] the choice is determined by ideas and experience.” - Saunders Mac Lane
- We want deep proofs that convey **why** a theorem is true.
- Models need to understand their own reasoning processes.



Symbolic Behaviour is *Meaningful*

Examples:

- AlphaZero uses learned meaning as a search heuristic, but cannot understand its hand-engineered MCTS reasoning, nor share knowledge among different branches of the search tree, etc.



Symbolic Behaviour is *Graded*

Symbol use is not a binary capacity. Instead, our capacities are graded.

- Children are receptive before constructive, and receptive to repeated meanings before they can reliably learn one-shot.
- Although we highlighted malleability, relatively few humans might be capable of finding a better meaning for any particular symbol.
- Symbol use will always be limited by cognitive, conceptual, or cultural factors.
- We should expect each of these abilities to be graded in our models, as they are in humans.

We see symbolic behaviour as a constellation of graded capacities for engaging with and creating meaning.

Symbolic Behaviour is Not Necessarily …

- Symbolic behaviour is not equivalent to certain syntactic manipulations. Systems need to **interpret** the entities they are reasoning over as symbols.
- GOFAI cannot, nor can contemporary neurosymbolic models.
- Symbolic behaviour is not necessarily rule-based.

Case study: Ethics

These issues are salient when considering ethics.

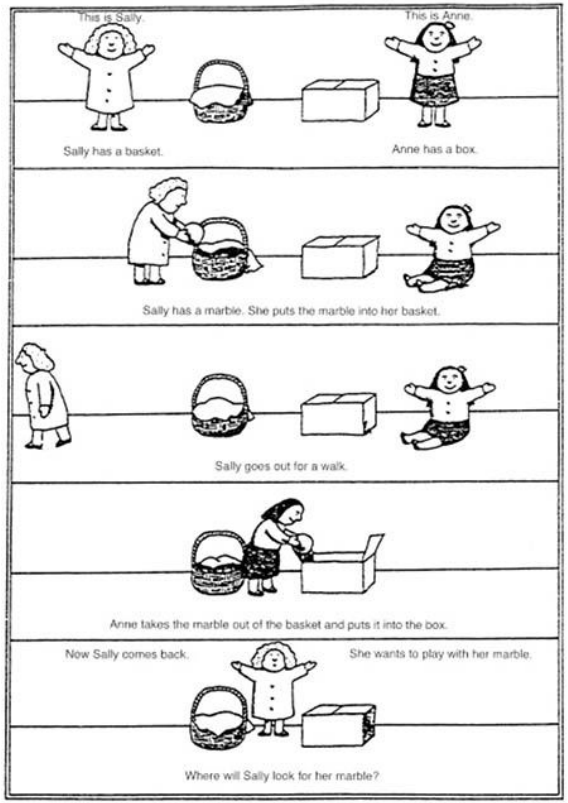
- Philosophers still can't agree on a rule basis for ethics, let alone how to impart that basis to a machine.
- Human ethics is contextual rather than fixed and rule-like.
- For example "don't hurt a human"—many situations this rule should be broken, e.g. re-break an arm to set it better.
- Rules are too easy to circumvent, e.g. "don't discriminate on the basis of race" could just result in the use of proxy variables (c.f. redlining).
- AI should understand ethics as a holistic, meaningful framework.

DeepMind

3 Origins of human symbolic behaviour

How do humans develop symbolic behaviour?

Perspectives, beliefs, and alignment



How do humans develop symbolic behaviour?

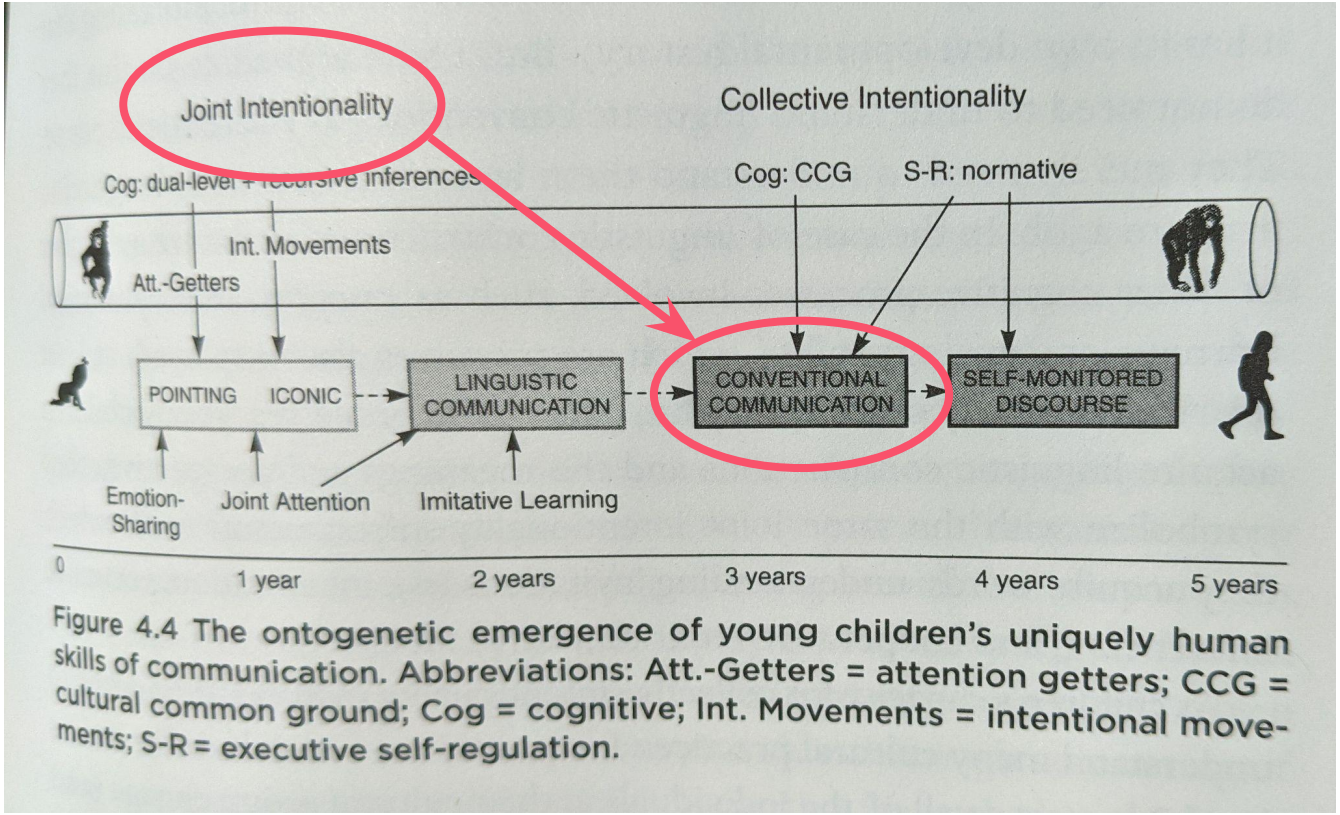
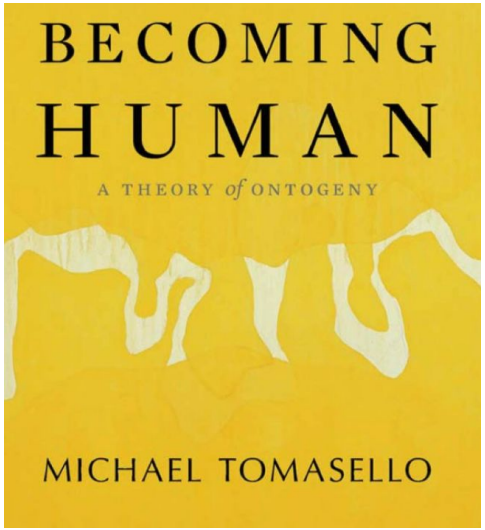


Figure 4.4 The ontogenetic emergence of young children's uniquely human skills of communication. Abbreviations: Att.-Getters = attention getters; CCG = cultural common ground; Cog = cognitive; Int. Movements = intentional movements; S-R = executive self-regulation.



Interactions and shared or unique perspectives

Developing symbolic capabilities

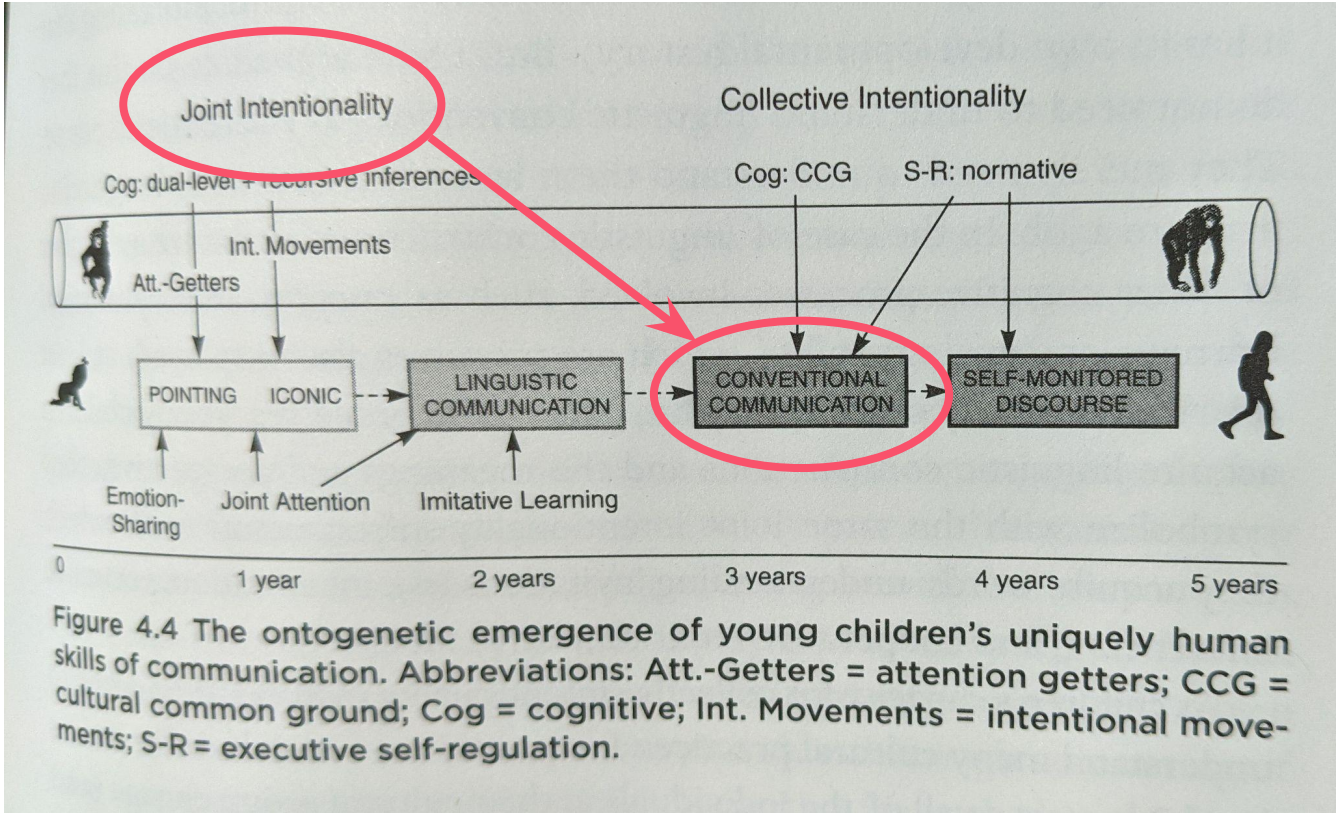
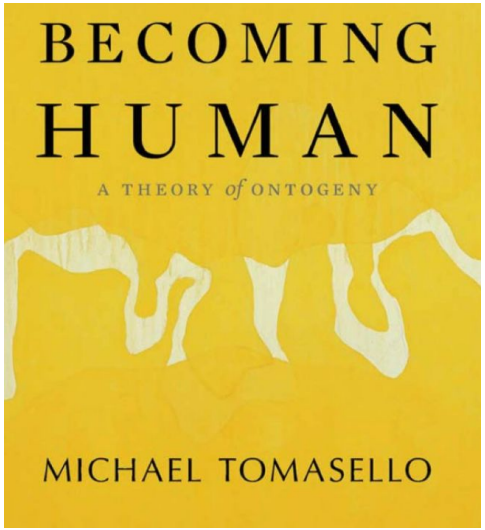


Figure 4.4 The ontogenetic emergence of young children's uniquely human skills of communication. Abbreviations: Att.-Getters = attention getters; CCG = cultural common ground; Cog = cognitive; Int. Movements = intentional movements; S-R = executive self-regulation.



We suggest AI should develop symbolic behaviour as humans do — learning through social interactions and culture.

Some steps in that direction

Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam
Girish Sastry Amanda Askell Sandhini Agarwal Ariel Herbert-Voss
Gretchen Krueger Tom Henighan Rewon Child Aditya Ramesh
Daniel M. Ziegler Jeffrey Wu Clemens Winter
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray
Benjamin Chess Jack Clark Christopher Berner
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

Imitating Interactive Intelligence

Interactive Agents Group*
DeepMind

GPT-3

DeepMind

4 Reconciling with other perspectives on symbols

Reconciling: behaviour

Other views on symbols are motivated by behaviour, starting from Newell & Simon. What behaviour?

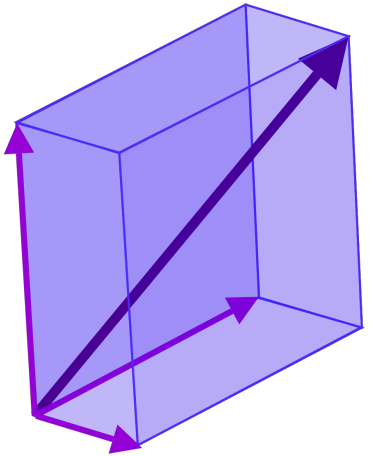
- Often compositional generalization, but how systematic are humans actually?
- When humans achieve compositional generalization 80-90% of the time, it is cited as evidence of our compositional skill; yet when a transformer achieves 98.4% accuracy on difficult calculus problems it's cited as a failure of the model class.
- Systematic behaviour may instead be a **graded** competency afforded by environment and education.
- Importantly, this means that mechanisms that guarantee systematicity may not be the right direction for achieving symbolically-fluent AI.
- Especially if they interfere with the aspects of symbolic behaviour that we highlight.

Reconciling: discrete?

Many perspectives on symbols and symbolic AI assume discrete tokens. We think this is misleading.

- The sounds that make up speech or the image of a Canadian flag are not discrete, yet they are substrates for symbols.
- Even within classical symbols, there is an important distinction between an entity being discrete and that entity serving as a discrete unit in a larger symbolic framework.
- For example, a **continuous** vector can be a **discrete** element of a set of basis vectors.

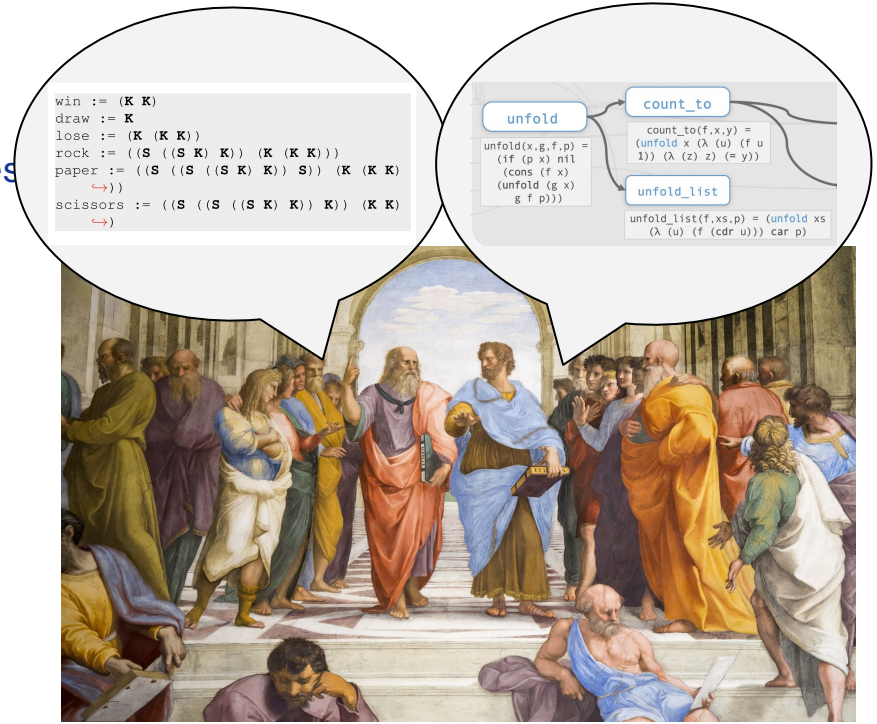
Our definition of a symbol does not place any restriction on the substrate.



Reconciling: neo-classical

Some new works attempt to allow aspects like construction of new symbols via classical mechanisms like combinators or “neo-classical” ones like probabilistic program induction.

- These approaches may suffer from the same weaknesses as GOFAI, at least as presently implemented.
  - Hence demonstrated in toy domains.
- If these challenges are overcome, we suggest that these approaches would also benefit from our perspective:
  - Focusing on symbolic behaviour.
  - Using social forces to encourage its emergence.



Interim summary

- We take symbols to have meaning by convention.
- This perspective highlights certain behaviours as evidence of symbolic, conventional understanding.
- We suggest interactive social development as a route to achieving these capabilities.
- We think this will provide a fruitful path for AI research to achieve more human-like capabilities.

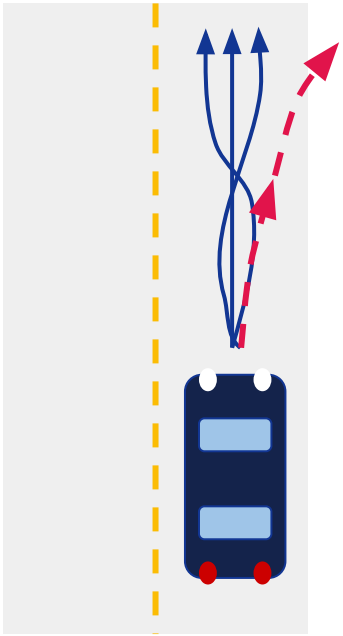
DeepMind

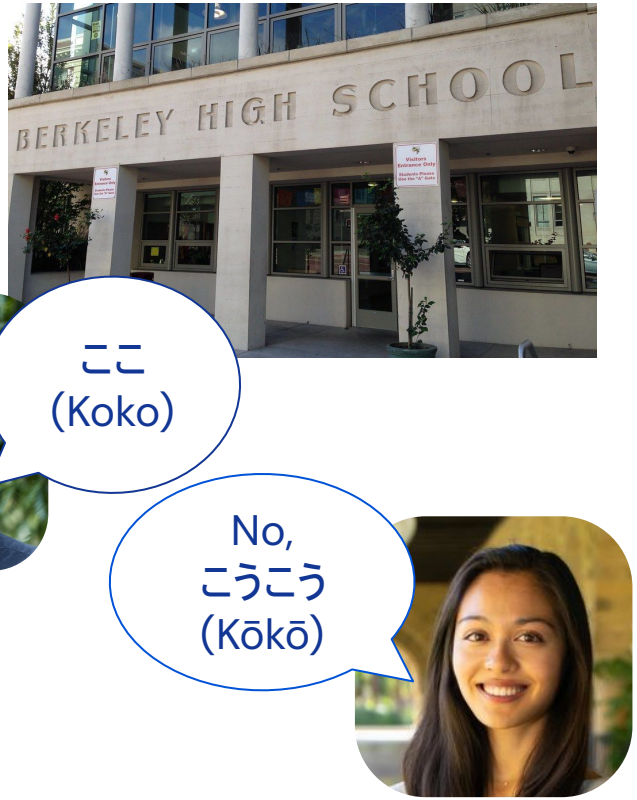
5 Whales?

Participating in interactions or observing + imitating?

GPT-4

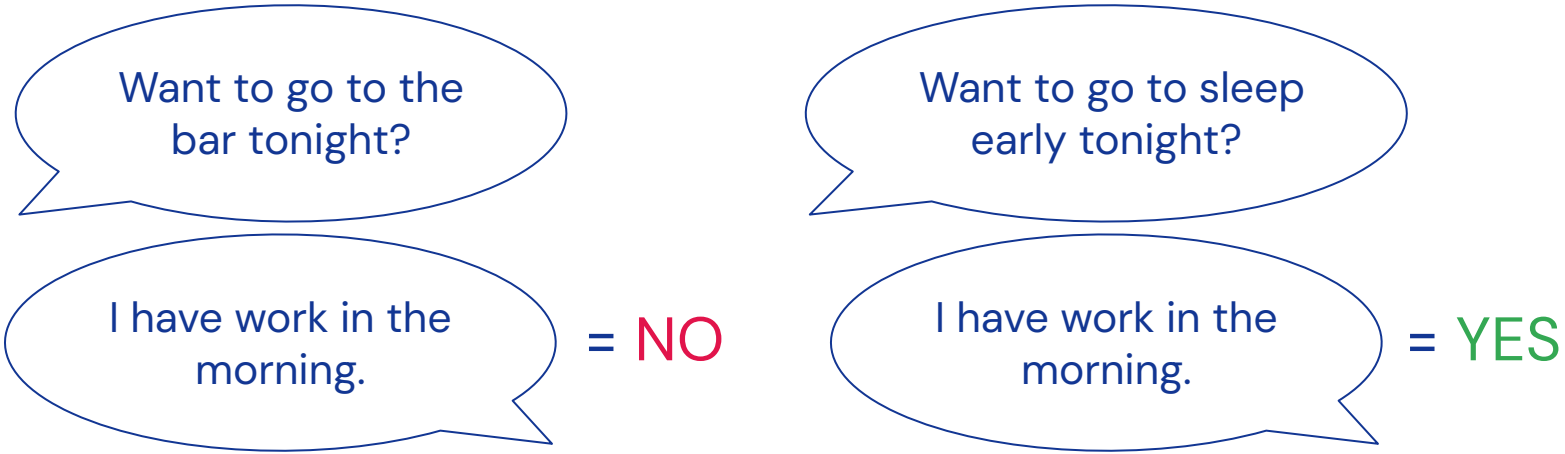
When does interaction matter?





Interaction is important for actively correcting a system's mistakes that others haven't made in the training data

Are the pragmatics of implicature hard for LMs to learn?



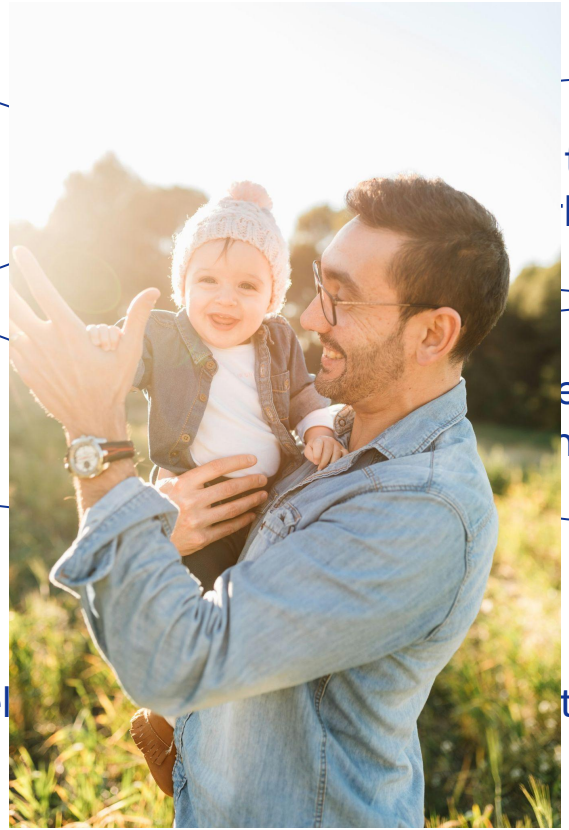
Not learned well by pure LMs! Only with certain kinds of tuning.

(Ruis et al., 2023)

Are the pragmatics of implicature hard for LMs to learn?

Want to go to the bar tonight?

to go to sleep ly tonight?



I have work in the morning.

e work in the morning.

= YES

Not learned wel tain kinds of tuning.

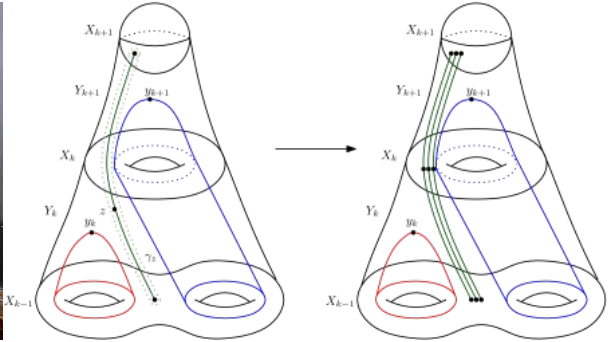
(Ruis et al., 2023)

Symbolic behaviour is learned through interacting and developing contextual use of conventional behaviour

Receptive

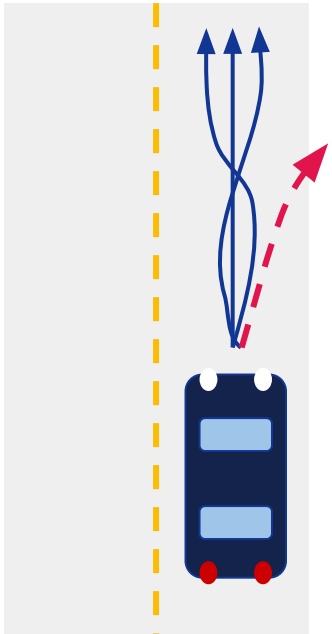
Constructive

Meaningful

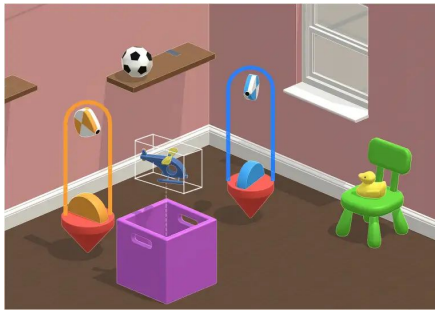


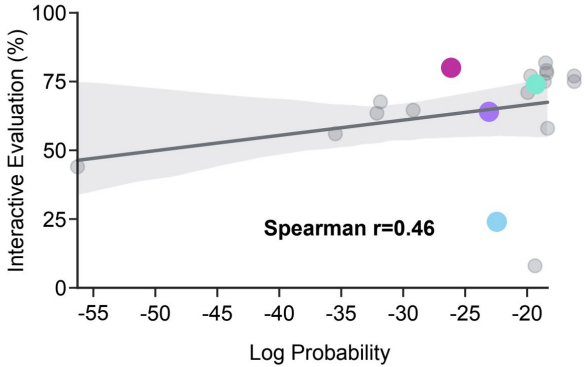
… and we don’t fully know how much of that learning can be from offline interaction, and how much needs to be online

Interaction is also important for assessment



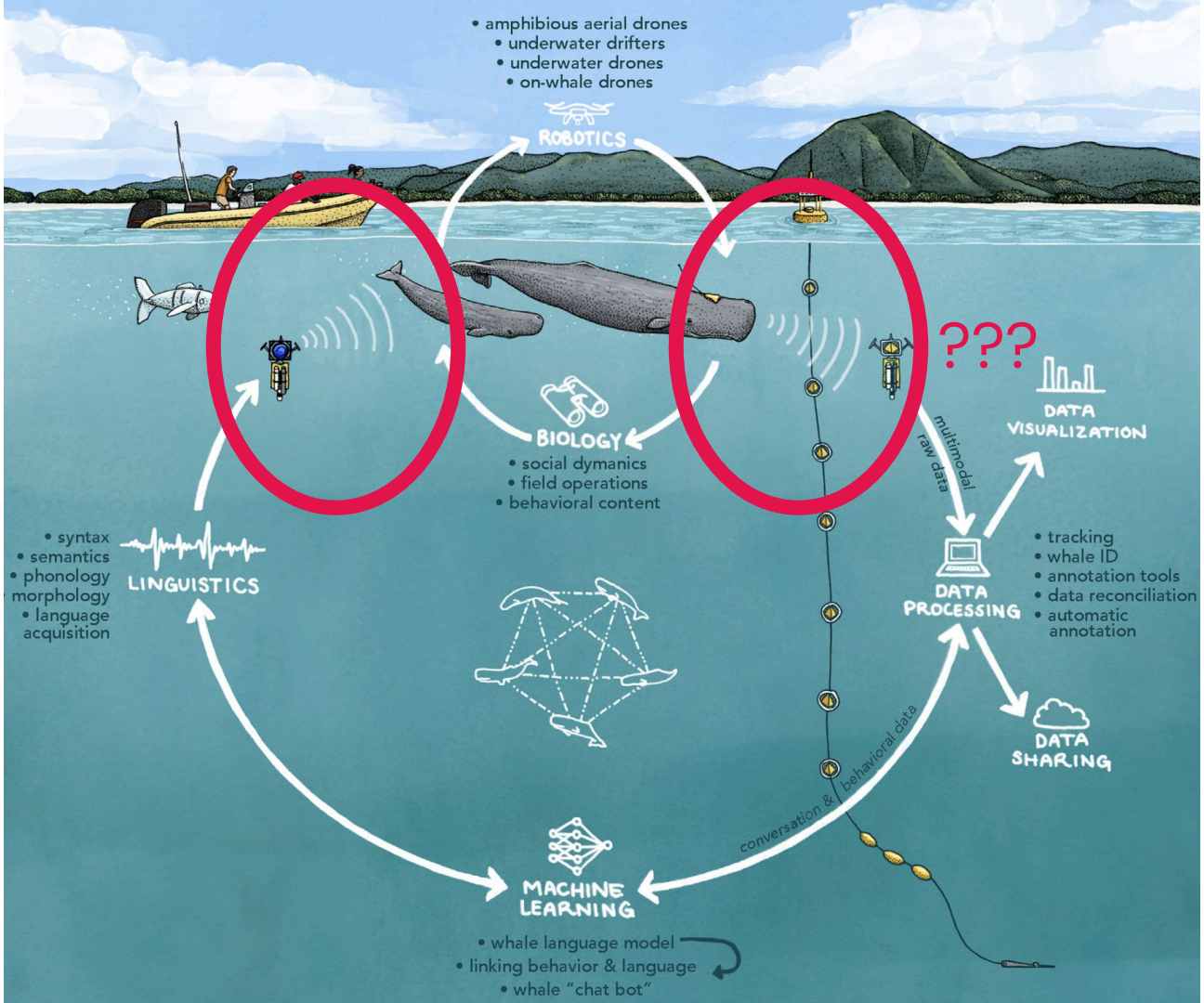
Grounded Agents





Language Models

(Interactive Agents Team, 2022)


• amphibious aerial drones
• underwater drifters
• underwater drones
• on-whale drones
ROBOTICS
???
DATA VISUALIZATION
BIOLOGY
• social dymanics
• field operations
• behavioral content
multimodal raw data
• syntax
• semantics
• phonology
• morphology
• language acquisition
LINGUISTICS
DATA PROCESSING
• tracking
• whale ID
• annotation tools
• data reconciliation
• automatic annotation
DATA SHARING
conversation & behavioral data
MACHINE LEARNING
• whale language model
• linking behavior & language
• whale "chat bot"

Thanks!