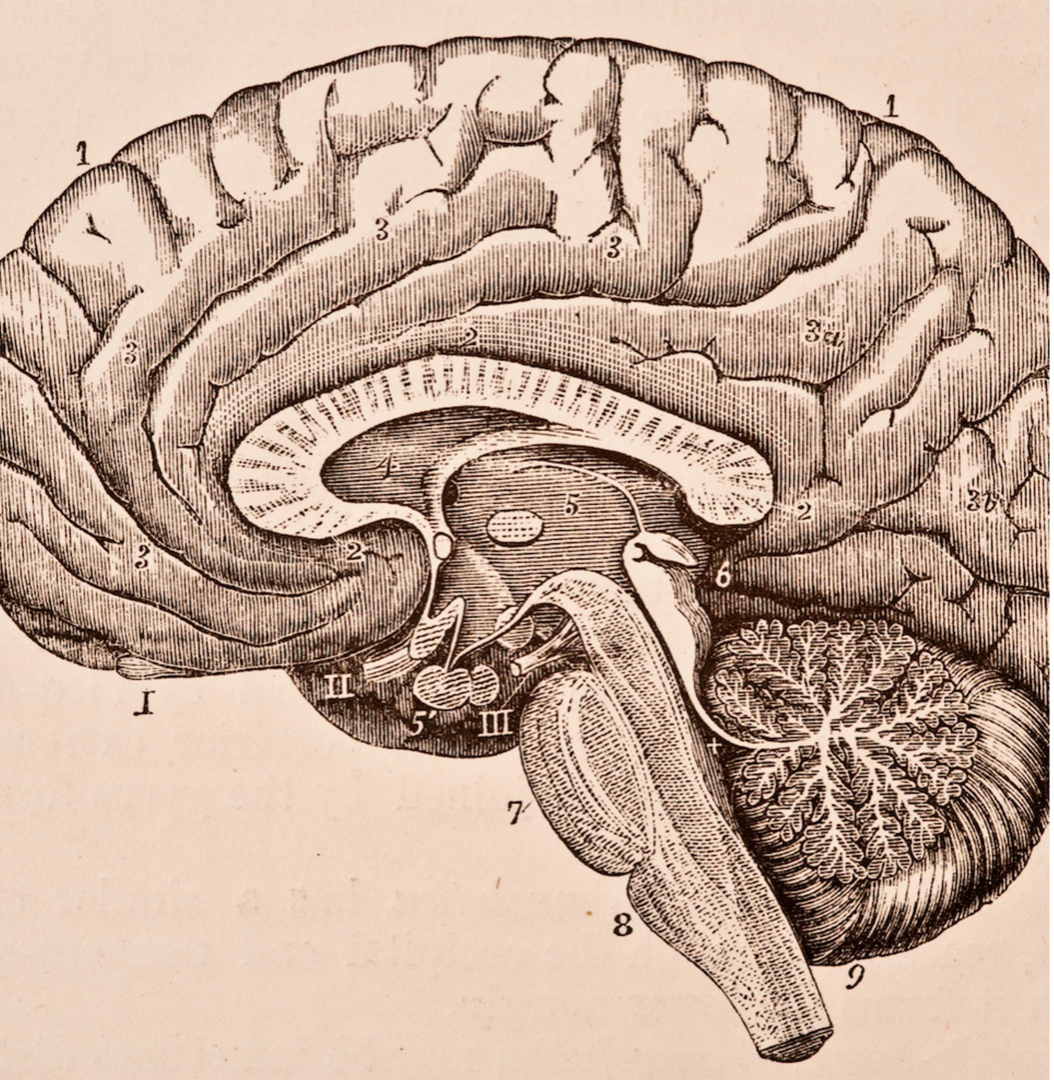


The Evolution of Human Intelligence

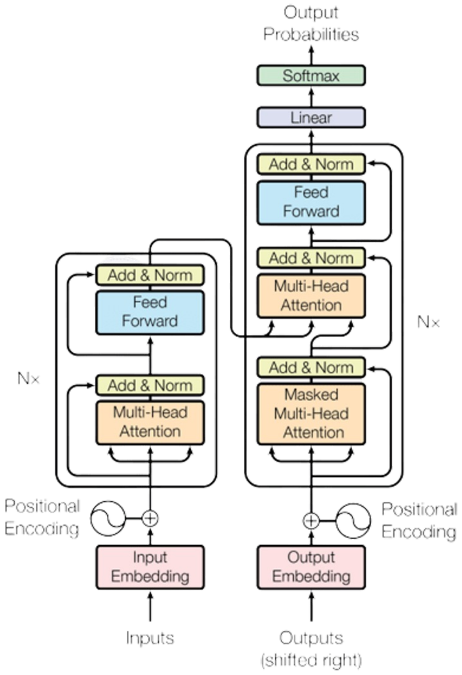
What it Teaches Us and Why it Matters

Max S. Bennett

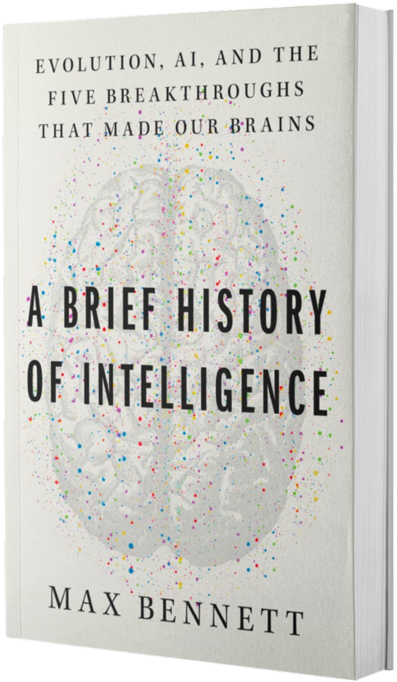
Bluecore

AI platform used by 25% of the largest brands in US,
including **Nike**, **Under Armour**, and **CVS**.

A.I.


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

Brain



Wonderful supporters, thought partners, mentors:

Joseph LeDoux, New York University

Karl Friston, University College London

Dileep George, DeepMind

Jeff Hawkins, Numenta

David Redish, University of Minnesota

Eva Jablonka, Tel Aviv University

Kent Berridge, University of Michigan

..and more..

The goal: Understanding how the brain “works”

My primary focus

Marr’s 3 levels of analysis

Marr’s Level	Definition	Neuroscience example
Computation	The function/goal of the system	Cognition, memory, planning, ...
Algorithm	The abstract mathematical operations being performed	Predictive coding, active inference, ...
Implementation	Physical implementation of the algorithm	Neurons, synapses, neuromodulators, ...

Why does understanding the brain matter to AI?

Reason #1: AI still performs badly in important ways

Continual learning

Common sense

Fine motor skills

Robustness

Active learning / interventional agents

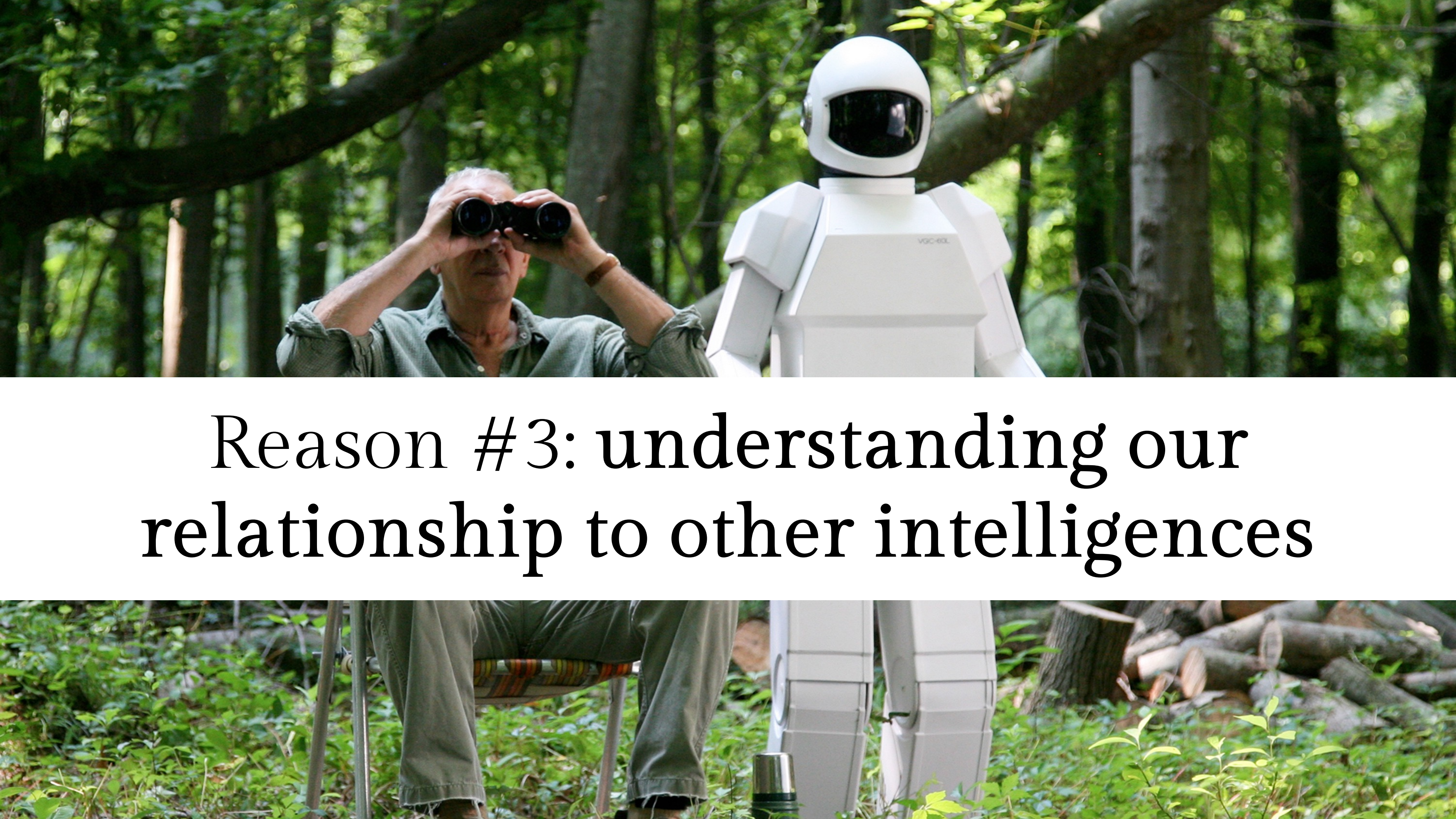
Explainability

Reason #2: AI costs too much energy

Business | Schumpeter

Generative AI has a clean-energy problem

What happens when the AI revolution meets the energy transition


Reason #3: understanding our
relationship to other intelligences

The current paradigm: Functional Decomposition

Computation

Cogn

Perce

Decision

Plan

Mer

Motor

gions:

ortex

ortex

mpus

ala

em

lum

Two challenges with this approach:

1. Functions are distributed

   Examples:

   a. Language (neocortex, thalamus, basal ganglia)
   b. Visual object recognition (distributed in SC, dual neocortical streams, amygdala)
   c. Decision making (some decisions in frontal cortex, others striatum, others SC, etc.)

1. Regions do multiple functions

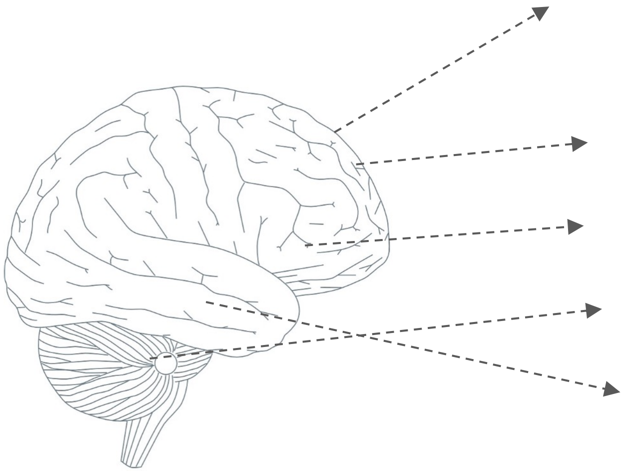
   Examples:

   a. Motor-sensory signals are overlapping in many neocortical regions
   b. Dopamine is both a learning signal and a wanting signal

"The expected mapping between cognitive operations and neural regions has not come to pass"

- Michael Anderson (Anderson, Kinnison, & Pessoa, 2013, as quoted in Cisek 2024)

Focusing on algorithms over functions can help (a bit)...



Neocortical microcircuit? - predictive coding? Generative models? Auto-association?

Neocortical-thalamic circuit - blackboard? Relay?

Basal ganglia - actor-critic system?

Cerebellum - Adaptive filter?

Hippocampus - pattern separation & completion?

but...

Directly reverse engineering these algorithms is hard. Our technological tools are currently limited

An underutilized tool we need in our toolbox: **evolution**

The human brain was not designed from scratch, it *evolved* over a long sequence of steps.

↓

Evolutionary steps are constrained and path dependent

↓

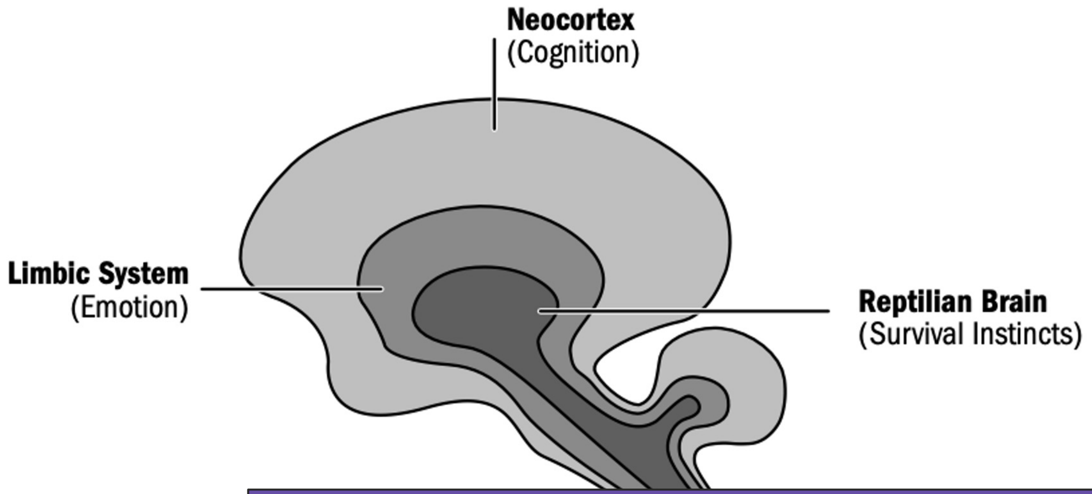
Thus, knowing prior steps helps us understand subsequent steps

↓

Start from first brains, track the modifications and the corresponding intellectual faculties that emerged at each major step.

Possible alternative approach

It is unfortunately not *this* simple: (MacLean's Triune Brain)

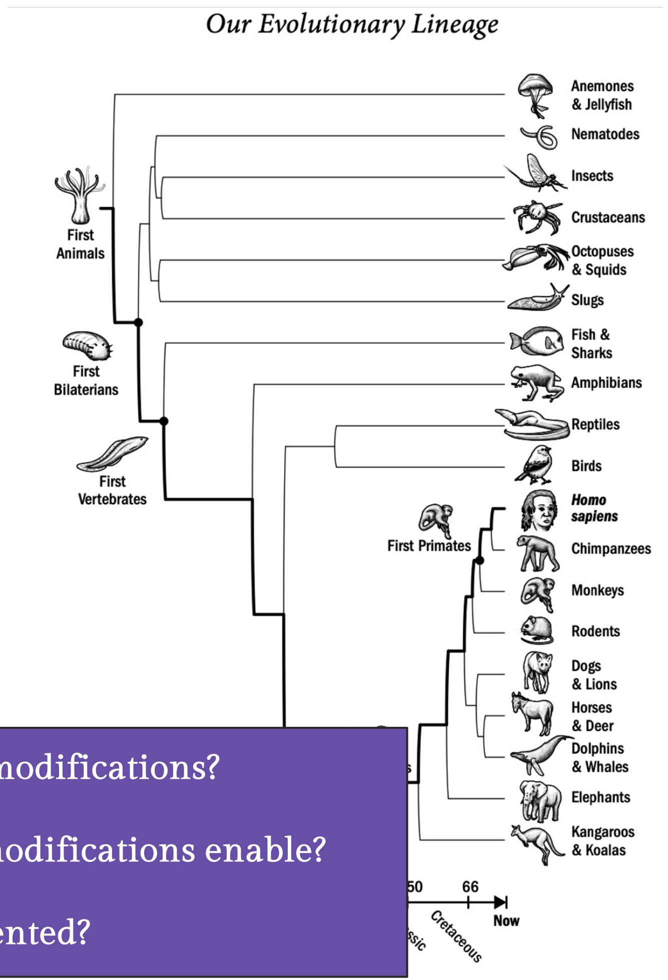


1. **Evolutionary story is wrong**
   a. Brain did not evolve solely through adding "layers"
   b. Reptiles have "limbic" areas
2. **Same problems as functional decomposition**
   a. "Emotion" occurs in non-limbic neocortical areas
   b. "Cognition" occurs in limbic areas
3. **Not grounded in comparative psychology**
   a. 'Early' diverging mammals show evidence of 'cognition'
4. **No insight on marr's level 2 (not grounded in A.I. research)**

We need a new evolutionary framework

Based on up-to-date evolutionary neurobiology & comparative psychology + based in a modern understanding of algorithms in A.I

Milestone	Key Neurobiological modifications
Early Eumetazoans	Neurons
Early Bilaterians	Valence neurons bilateralism first “brain”
Early Chordates	Proto-Hypothalamus + spinal cord
Early Vertebrates	Vertebrate brain template (forebrain, midbrain, hindbrain) Basal ganglia Cortex (pallial amygdala, hippocampus, olfactory cortex)
Early Jawed Vertebrates	Cerebellum
Early Amniotes	Dorsal pallium
Early Mammals	Neocortex emerges from dorsal pallium
Early Placental mammals	Motor cortex
Early Primates	Granular prefrontal cortex STS/TPJ Direct motor cortex connections
Early Humans	

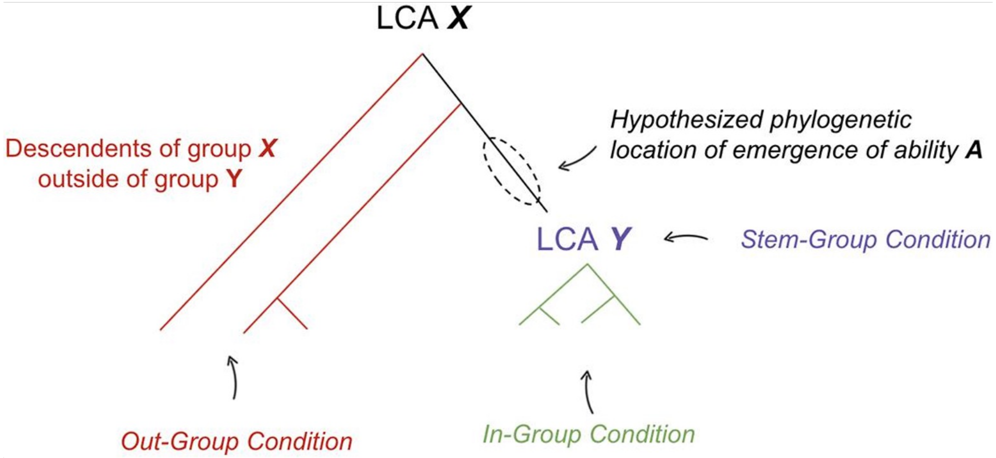


But what were the adaptive benefits of these modifications?

What new intellectual capacities did each of these modifications enable?

What algorithms were being implemented?

A methodology for inferring emergence of intellectual faculties



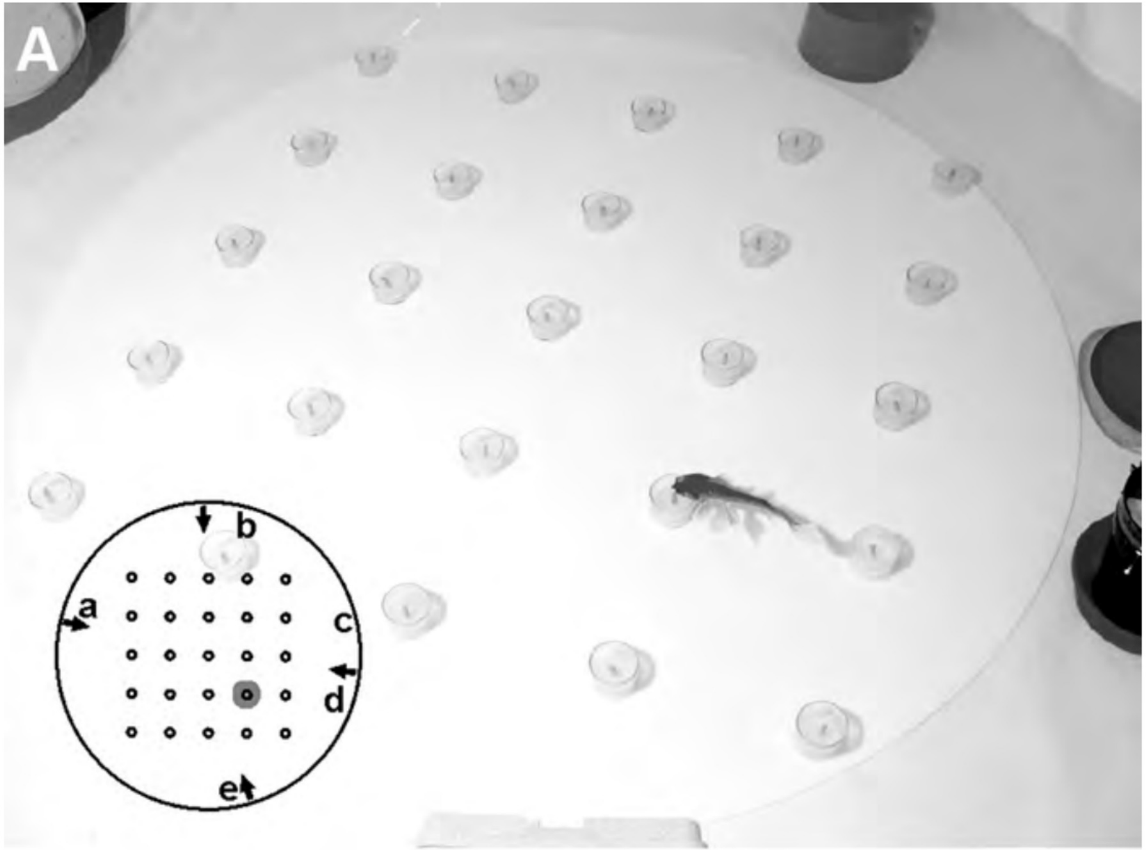
- **In-Group Condition**: Diverse groups of early diverging species across group Y contain ability A, implemented in homologous neural mechanisms.

  AND
- **Out-Group Condition**: Evidence is supportive of one of the following two claims:
  - (a) descendants of earlier diverging phylogenetic group X outside of group Y do not contain ability A
  - OR (b) ability A is implemented in non-homologous neural mechanisms in earlier diverging group X outside of Y

    AND
- **Stem-Group Condition**: Ability A would have been adaptive within the purported ecological niche of early members of group Y

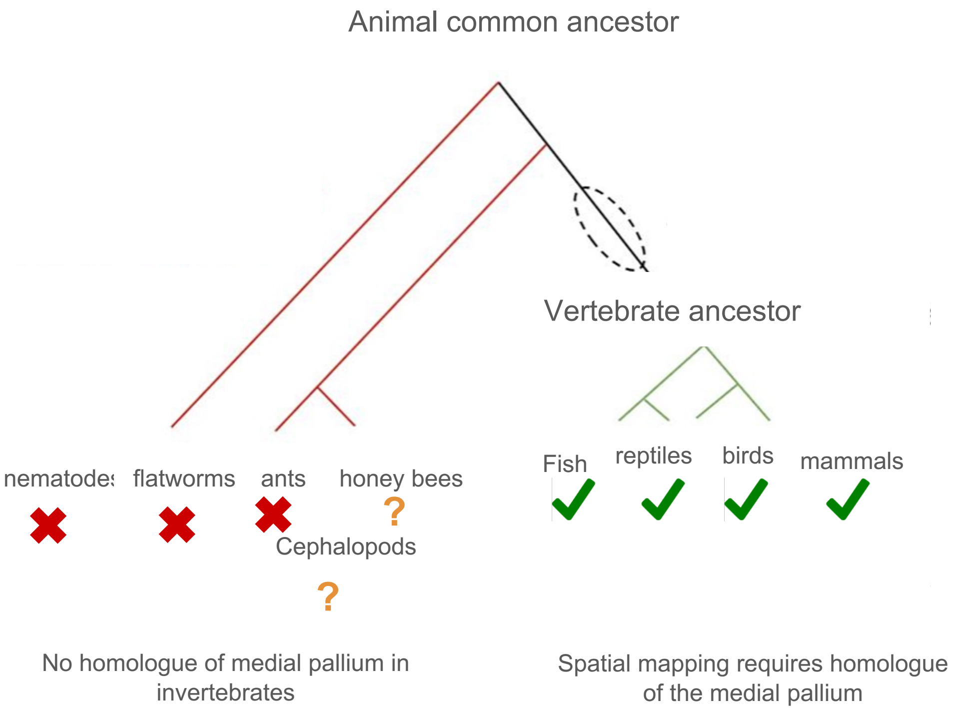
Reference: Bennett MS (2021) What Behavioral Abilities Emerged at Key Milestones in Human Brain Evolution? 13 Hypotheses on the 600-Million-Year Phylogenetic History of Human Intelligence. Front. Psychol. 12:685853. doi: 10.3389/fpsyg.2021.685853

Example - allocentric spatial mapping



Durán et al., 2008, 2010

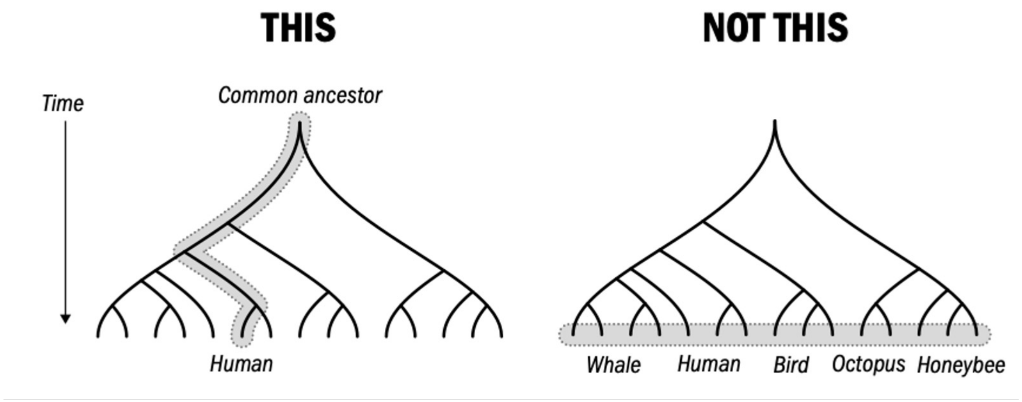
Example - allocentric spatial mapping



Parsimonious conclusion:

ability to remember allocentric locations in space emerged in early vertebrates

Important caveat: We are specifically tracing the human lineage



Saying "in our lineage, spatial mapping evolved with early vertebrates"

is **not** the same as saying

"<u>only</u> vertebrates have spatial mapping"

We then get a first approximation of our story:

Milestone	Neurobiological modifications (“implementation”)	Behavioral abilities (“computation”)
Early bilaterians	Valence neurons Bilateralism First brain	Valence (~reward) Associative learning Affective states
Early vertebrates	Vertebrate brain template (forebrain, midbrain, hindbrain) Basal ganglia	Interval Timing Pattern recognition Trial and error learning Spatial mapping
Early mammals	Neocortex emerges from dorsal cortex	Vicarious Trial & Error Counterfactual learning Episodic Memory
Early primates	Granular prefrontal cortex STS/TPJ Direct motor cortex connections	Theory of mind Imitation learning Anticipating future needs
Early humans	Frontal pole? Unique projection from motor cortex to larynx	Language Beat-based timing

Legend

- Strong evidence
- Good evidence (not conclusive)
- Preliminary evidence (still controversial)

Reference: Bennett MS (2021) What Behavioral Abilities Emerged at Key Milestones in Human Brain Evolution? 13 Hypotheses on the 600-Million-Year Phylogenetic History of Human Intelligence. Front. Psychol. 12:685853. doi: 10.3389/fpsyg.2021.685853

We then get a first approximation of our story:

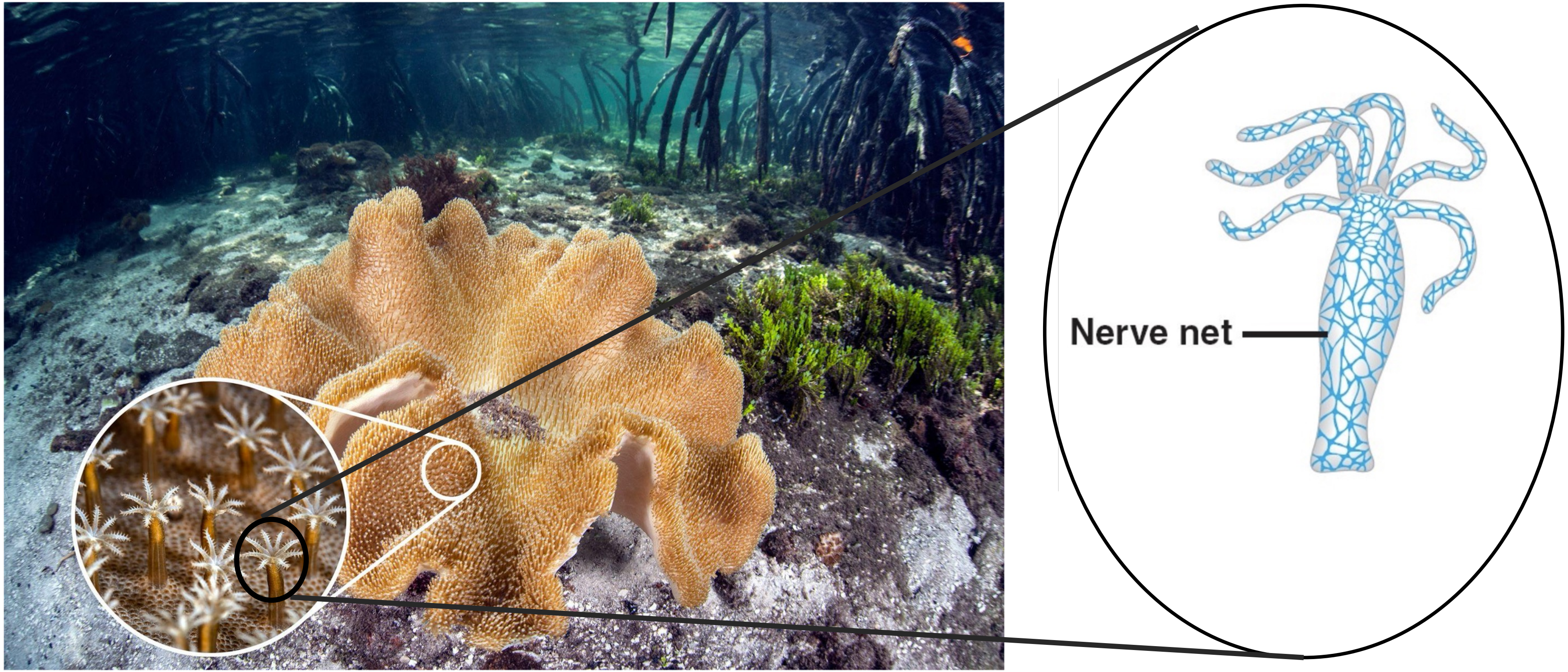
Milestone	Neurobiological modifications ("implementation")	Algorithm	Behavioral abilities ("computation")
Early bilaterians	Valence neurons Bilateralism First brain	?	Valence (~reward) Associative learning Affective states
Early vertebrates	Vertebrate brain template (forebrain, midbrain, hindbrain) Basal ganglia	?	Interval Timing Pattern recognition Trial and error learning Spatial mapping
Early mammals	Neocortex emerges from dorsal cortex	?	Vicarious Trial & Error Counterfactual learning Episodic Memory
Early primates	Granular prefrontal cortex STS/TPJ Direct motor cortex connections	?	Theory of mind Imitation learning Anticipating future needs
Early humans	Frontal pole? Unique projection from motor cortex to larynx	?	Language Beat-based timing

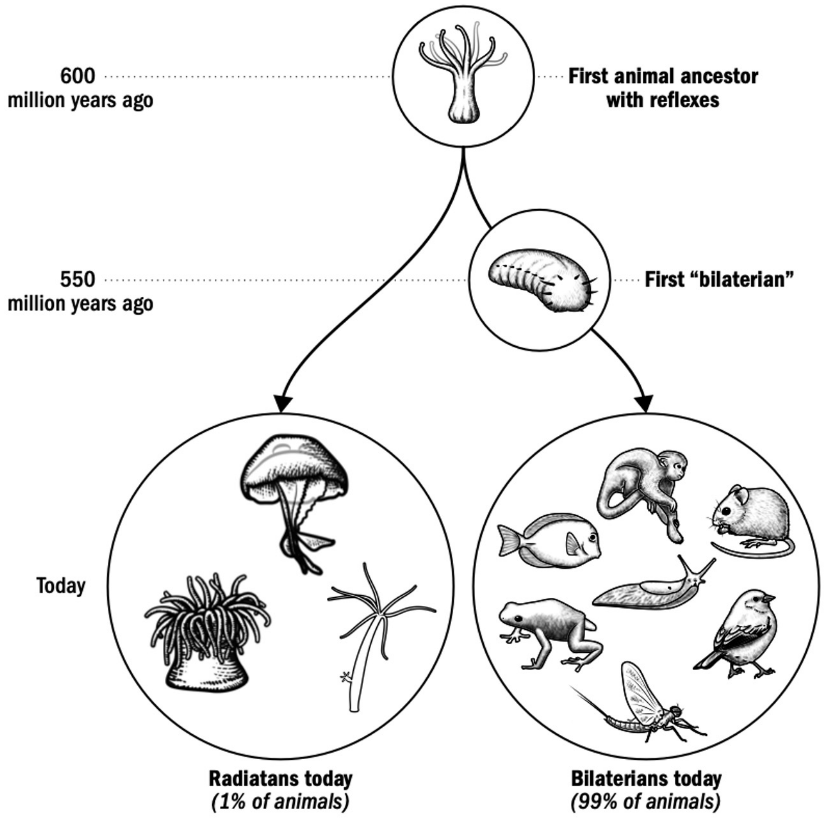
Legend
- Strong evidence
- Good evidence (not conclusive)
- Preliminary evidence (still controversial)

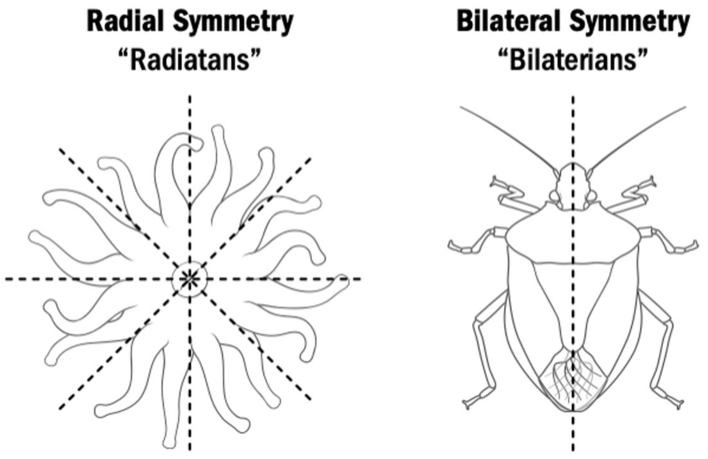
Reference: Bennett MS (2021) What Behavioral Abilities Emerged at Key Milestones in Human Brain Evolution? 13 Hypotheses on the 600-Million-Year Phylogenetic History of Human Intelligence. Front. Psychol. 12:685853. doi: 10.3389/fpsyg.2021.685853

Let’s go back to **~600 million years ago…**

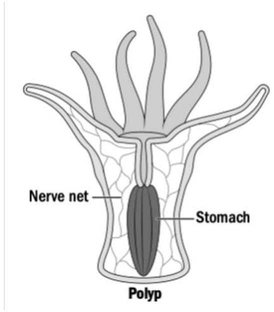
First animals did not have a brain:




600
million years ago
First animal ancestor
with reflexes
550
million years ago
First "bilaterian"
Today
Radiatans today
(1% of animals)
Bilaterians today
(99% of animals)


Radial Symmetry
"Radiatans"
Bilateral Symmetry
"Bilaterians"

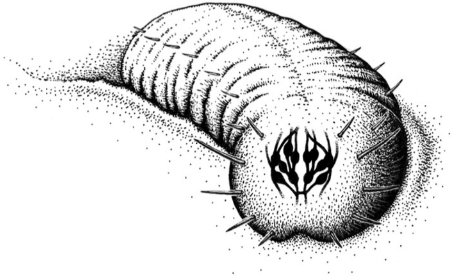
First animals



Stay in one place.

Wait for food

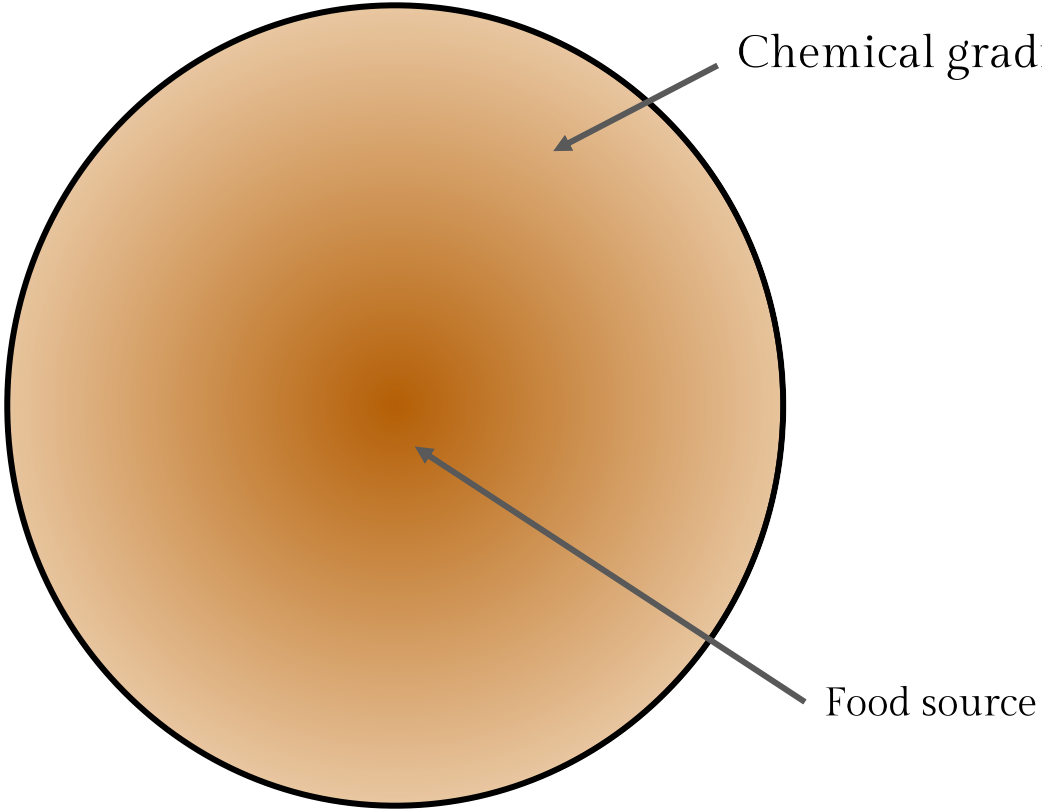
First bilaterians



Move around to find food.

302 neurons. No eyes. No ears.

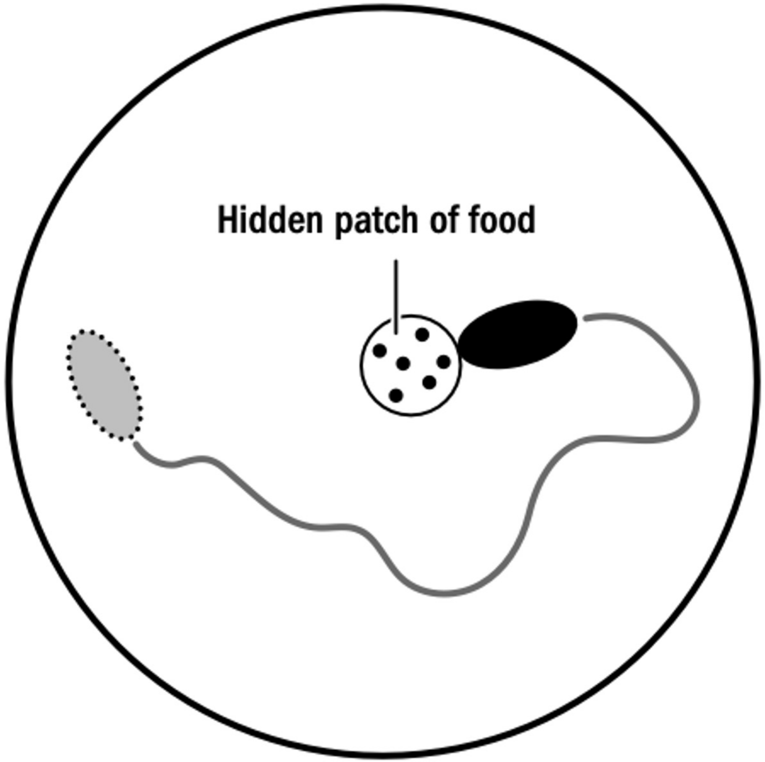
How does a nematode find food?



Brilliantly simple algorithm called “taxis navigation”:

1. If food smell increases, go forward

1. If food smell decreases, turn randomly

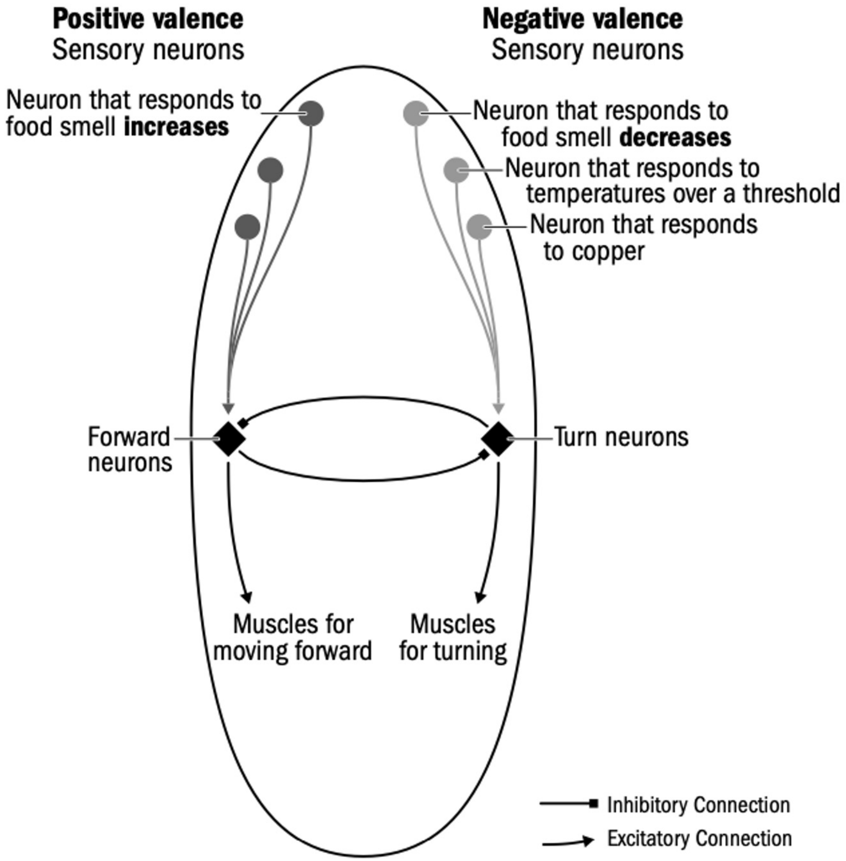


Brilliantly simple algorithm called “taxis navigation”:

1. If food smell increases, go forward

1. If food smell decreases, turn randomly

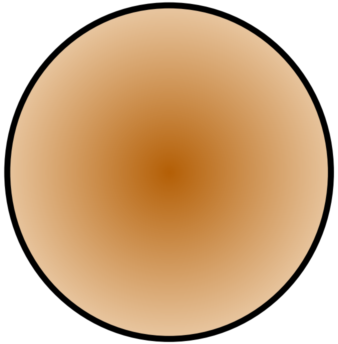
The brain of nematodes



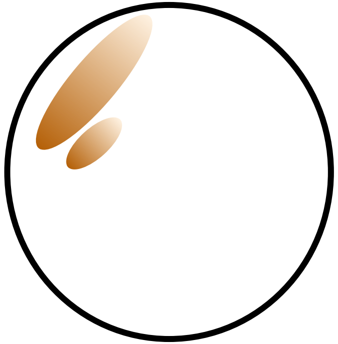
But it's not so simple...

Gradients are sparse, noisy, and clumpy.

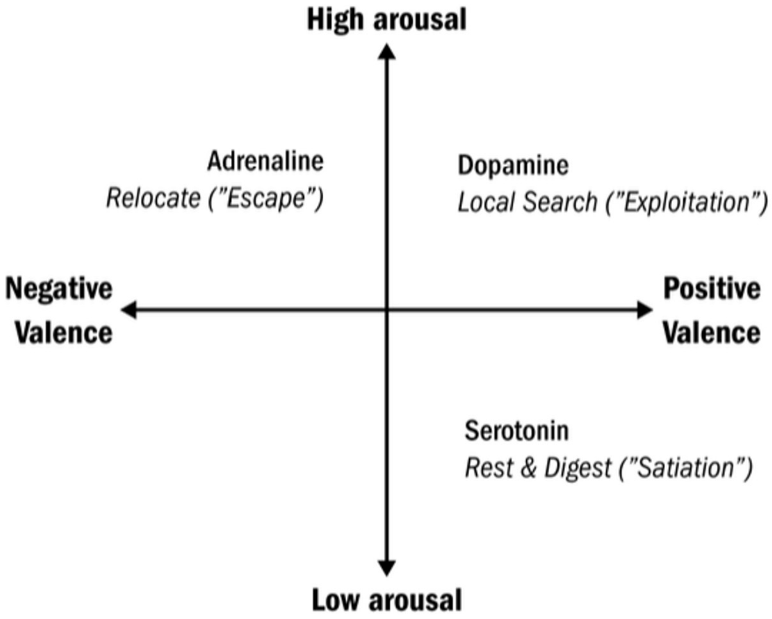
Less this

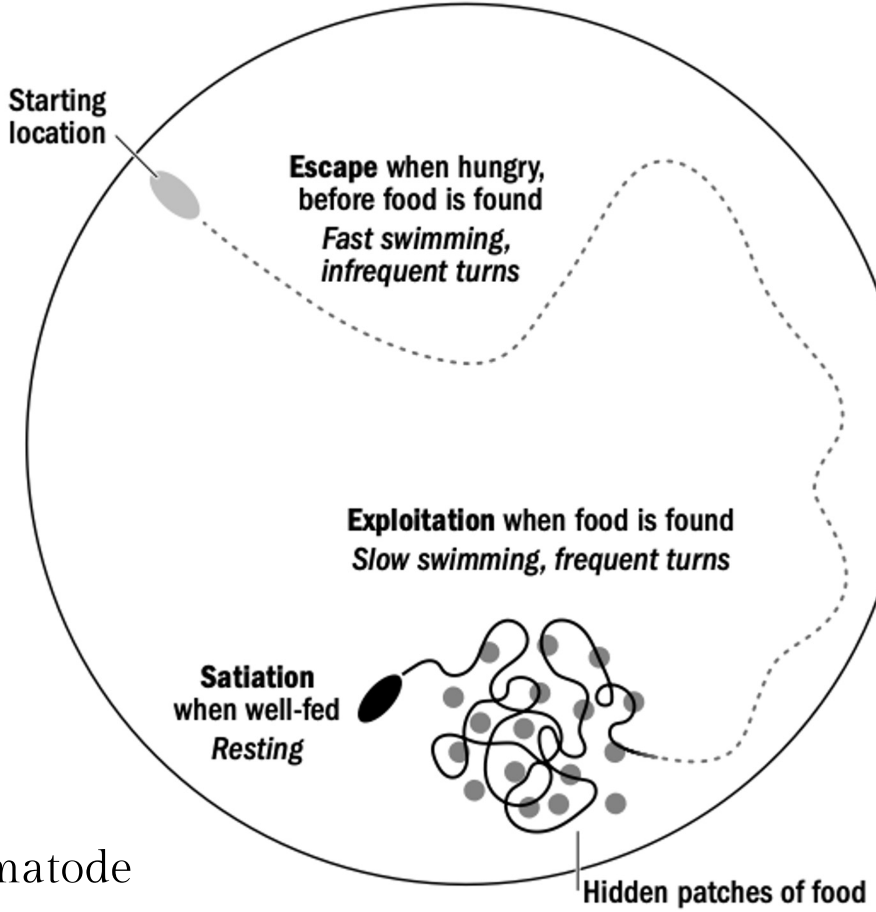


More this



Behavioral states evolve

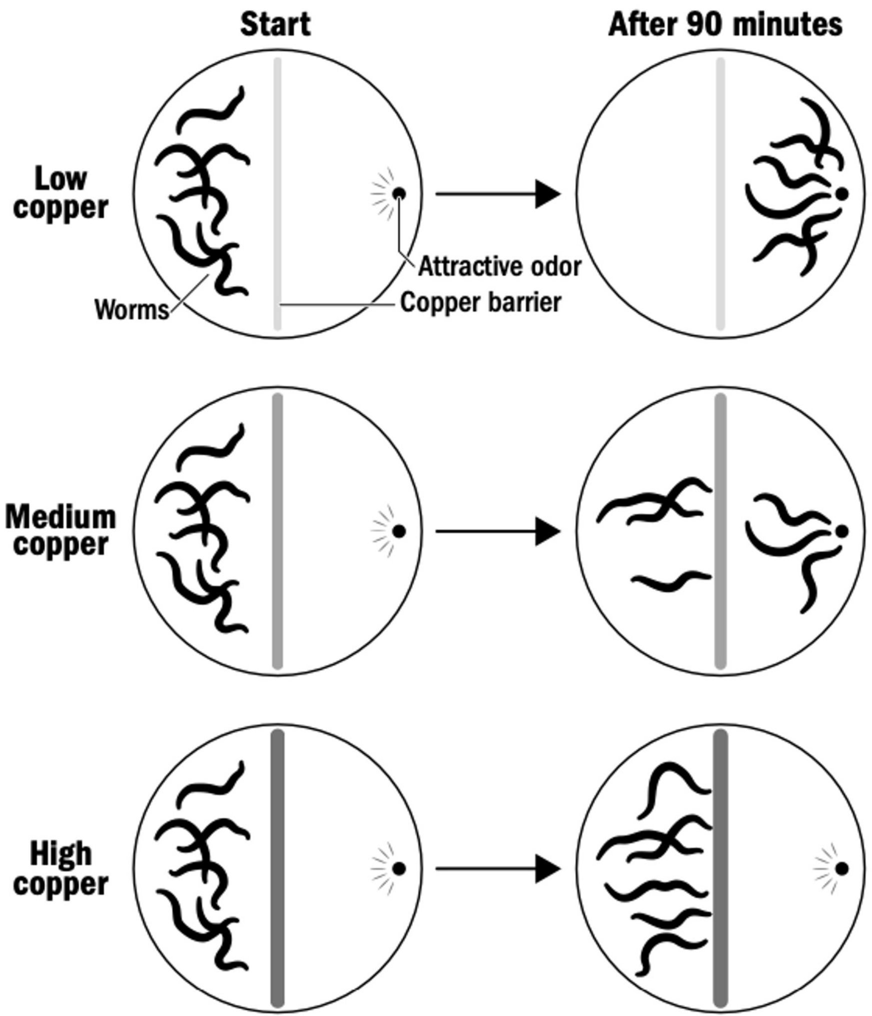




Dopamine neurons: detect food *outside* nematode

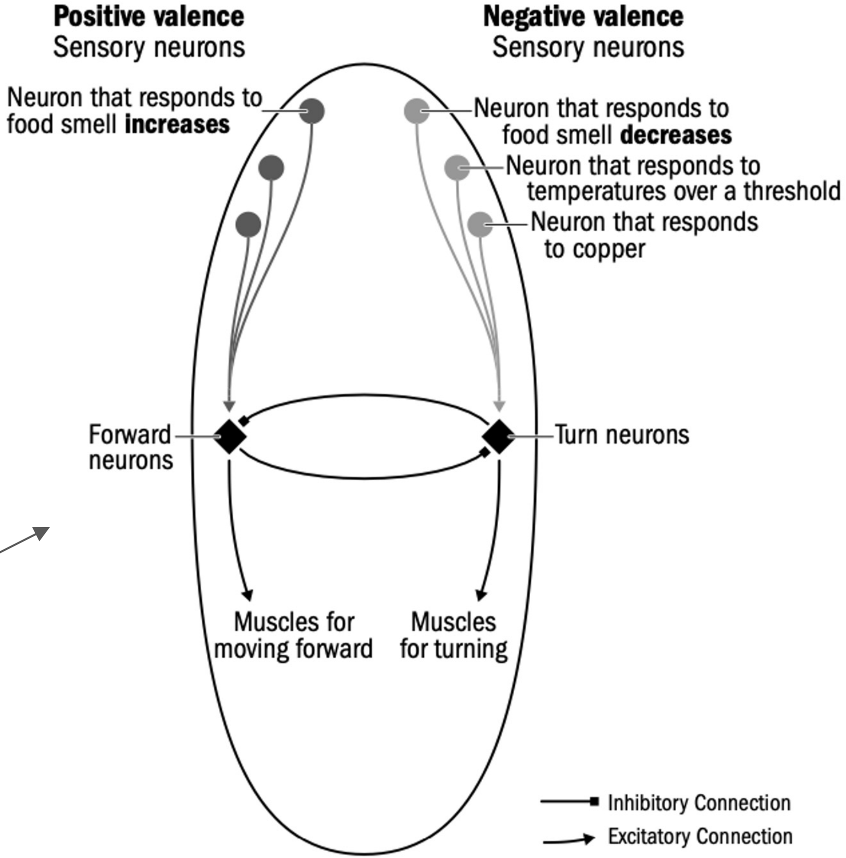
Serotonin neurons: detect food *inside* nematode

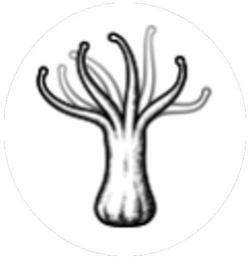
Why a brain?


Start
After 90 minutes
Low copper
Attractive odor
Copper barrier
Worms
Medium copper
High copper

Can only make a single choice

You need a brain to integrate tradeoffs





First animals

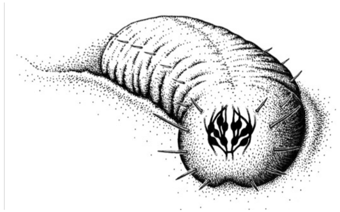
Radial symmetry

No “reward”

No behavioral states

No associative learning

No brain



First bilaterians

Bilateral symmetry

“Reward” - categorization of things in the world into good and bad

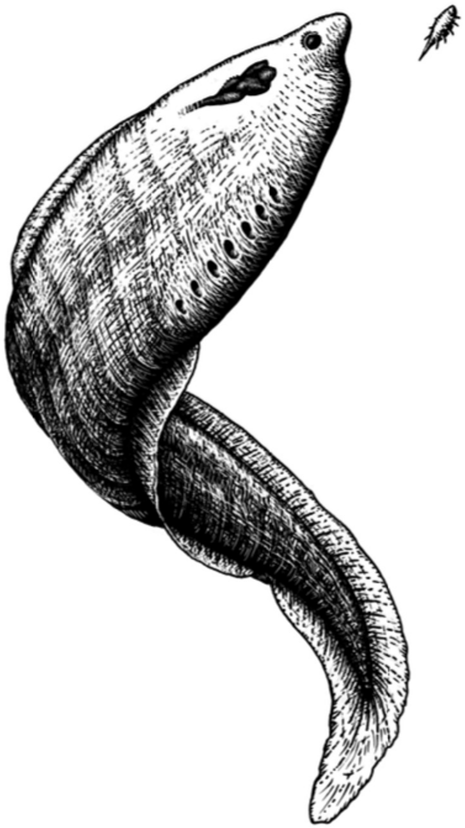
Behavioral states

Associative learning

Brain

All tools for taxis navigation

The purpose of the first brain was to implement a **taxis algorithm** to enable our ancestors to navigate the seafloor without complex sensory organs.

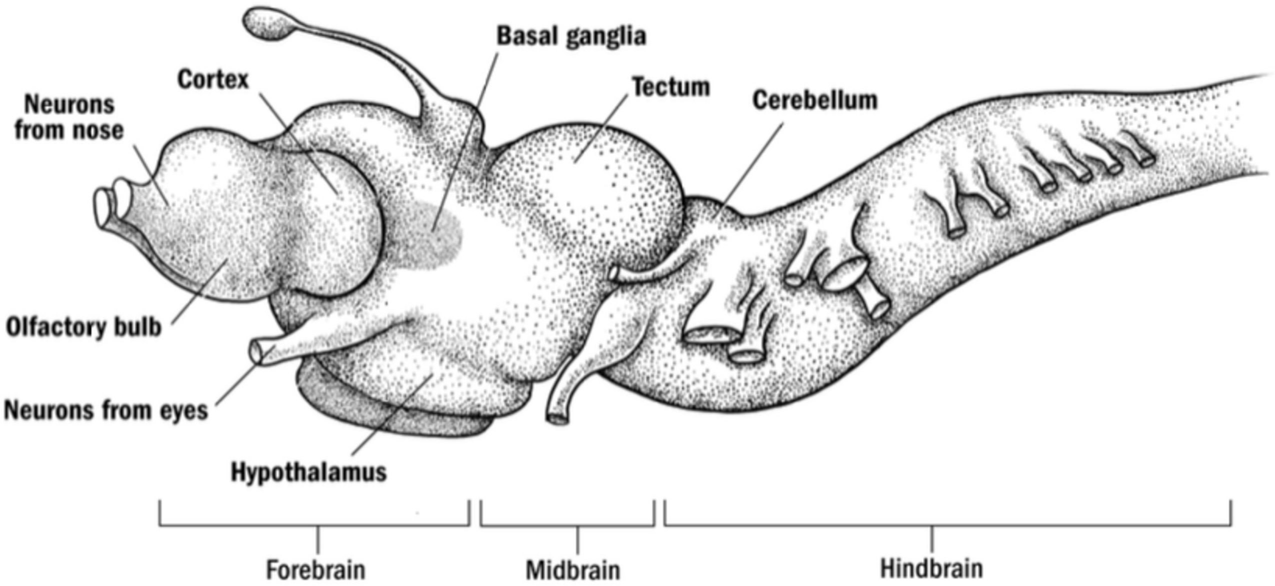


Your brain 500 million years ago

The First Bilaterians

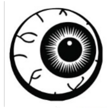
The First Vertebrates

The brain of the first vertebrates



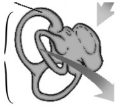
Lamprey fish

Sensory organs of vertebrates



Lens shaped eyes

Ears



Vestibular system

Olfactory neurons

Taste cells

Sensory processing of vertebrates

- *Identify objects despite rotations*
- *Smell pattern recognition*
- *Spatial memory*
- *3D locations*
- *Interval timing*

Fish can..

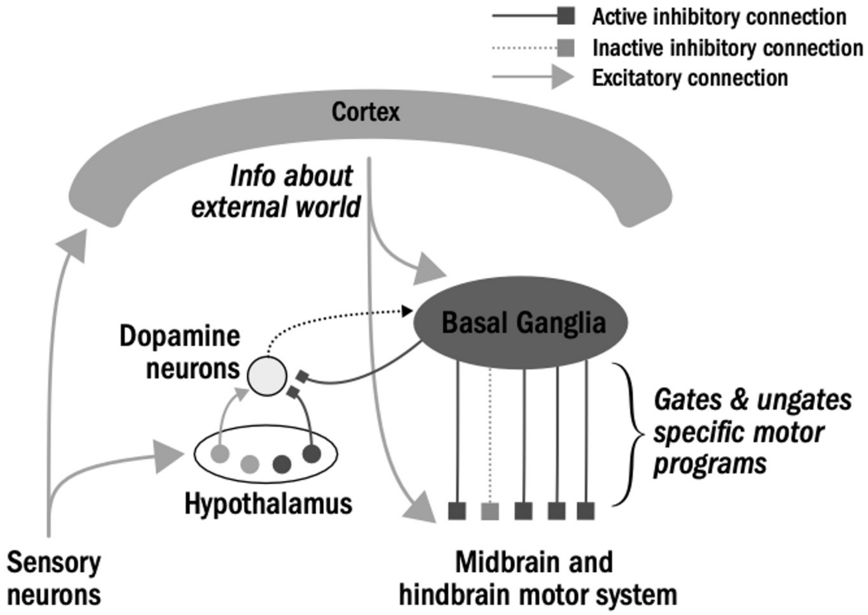
Learn to swim through mazes for rewards, remember 1 year later

Learn to jump through hoops to get rewards

Learn to find and push buttons to get rewards

**Nematodes & flatworms can't do any of these*

Cortex + basal ganglia enable **temporal difference learning algorithm**



Dopamine was repurposed from a general average of nearby food, to a precise predicted future reward signal. (TD learning signal)

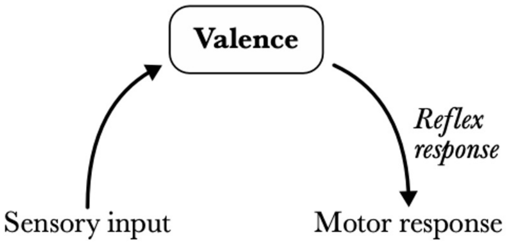
*see sten grillner for more great work on vertebrate brain evolution and conserved forebrain systems

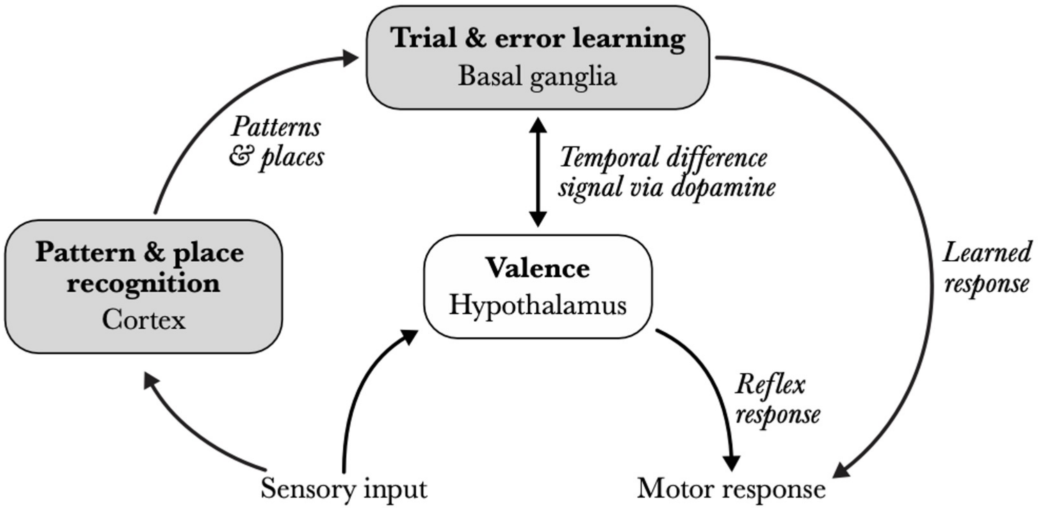
Early Bilaterian Brain

Steering

Early Vertebrate Brain

Reinforcement learning





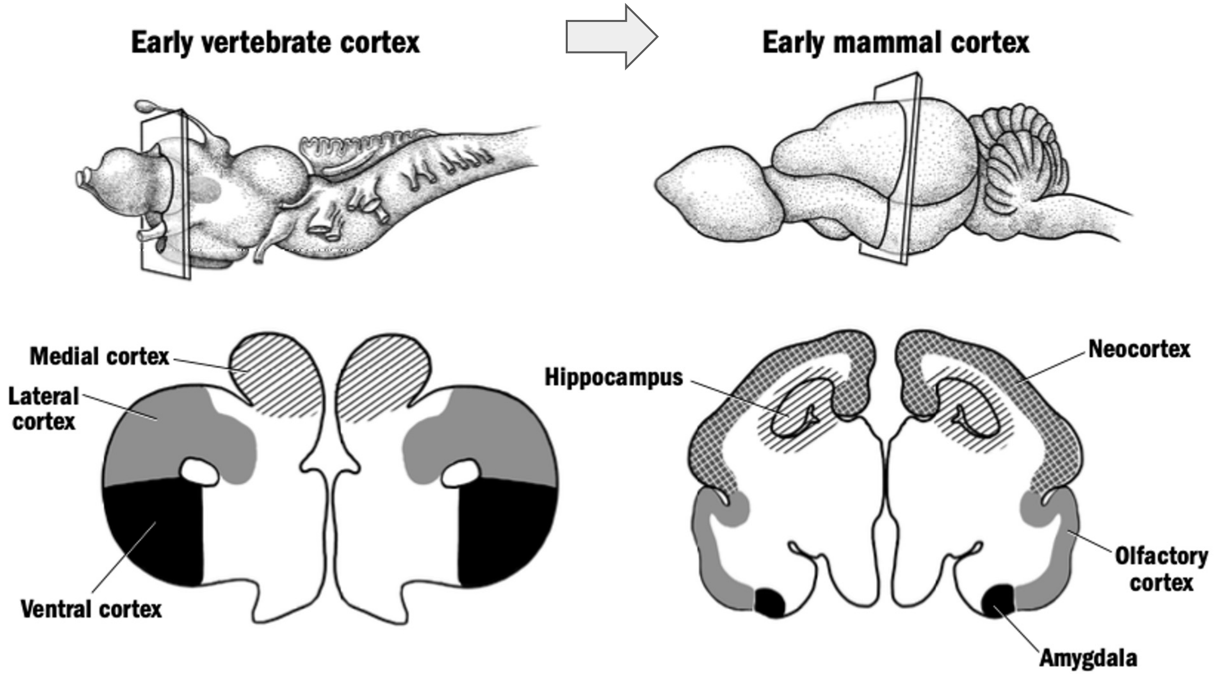
Your brain 200 million years ago

The First Bilaterians

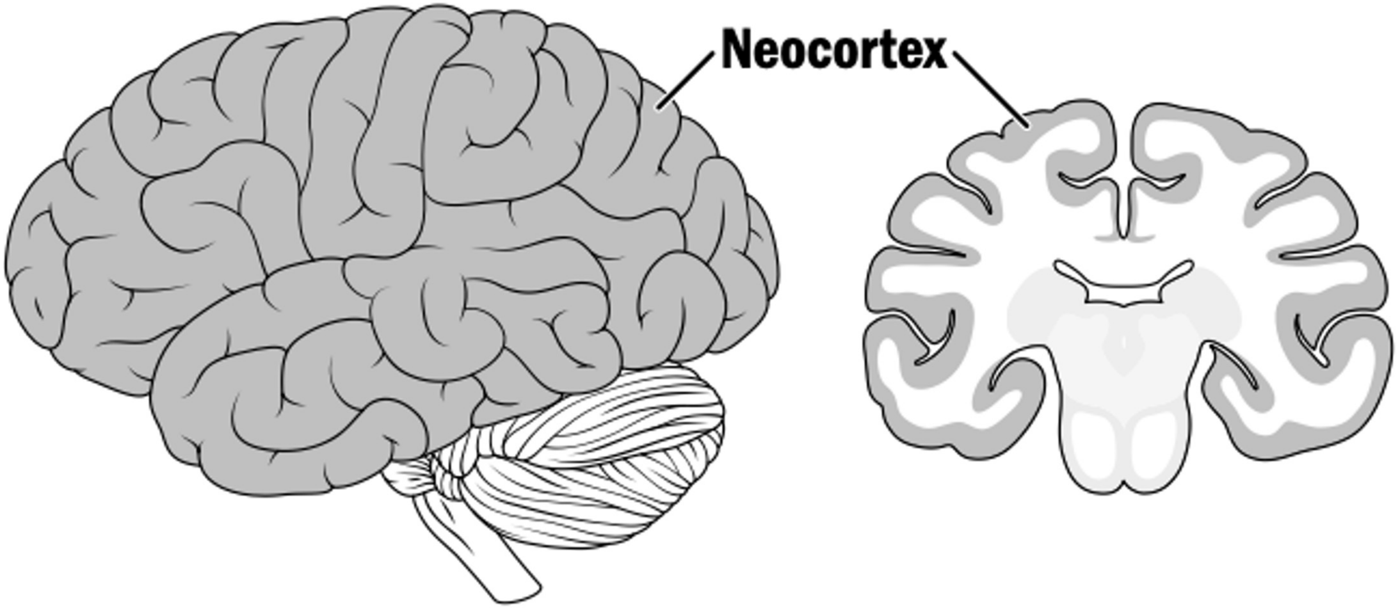
The First Vertebrates

The First Mammals

Main brain modification was emergence of the **neocortex**

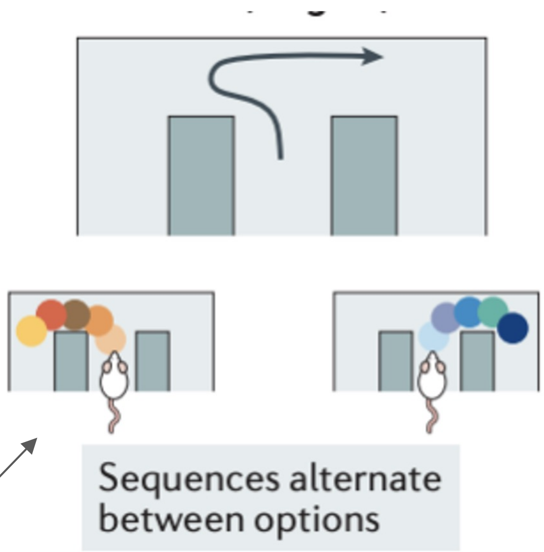


Main brain modification was emergence of the **neocortex**



What was the adaptive value of the neocortex?

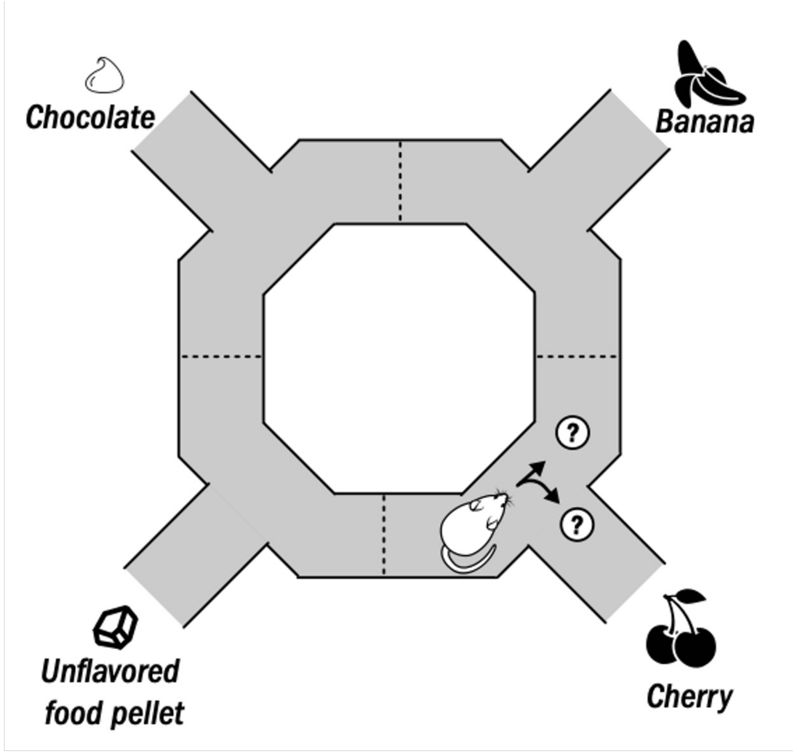
Mammals engage in "Vicarious Trial & Error"



Rats imagine possible outcomes before deciding

Johnson & Redish 2007

Mammals engage in “Counterfactual learning”

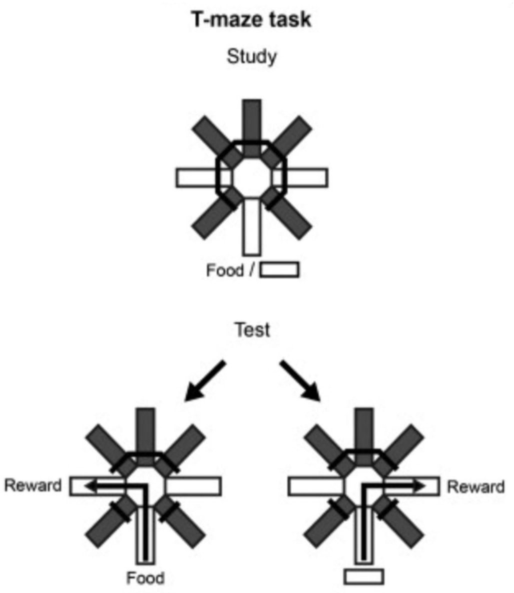


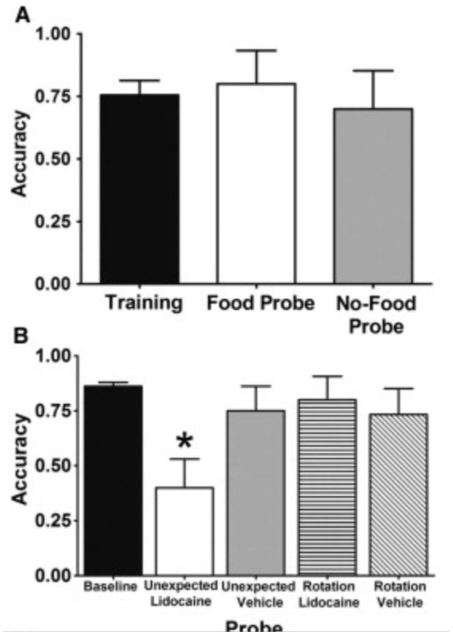
When choosing to skip OK reward NOW for possibility of GOOD reward, but then finding out there is a long delay, rats regret their choice:

- Rats look back
- Rats re-activate representation of foregone choice in neocortex
- Rats change future choices

Steiner and Redish, 2014

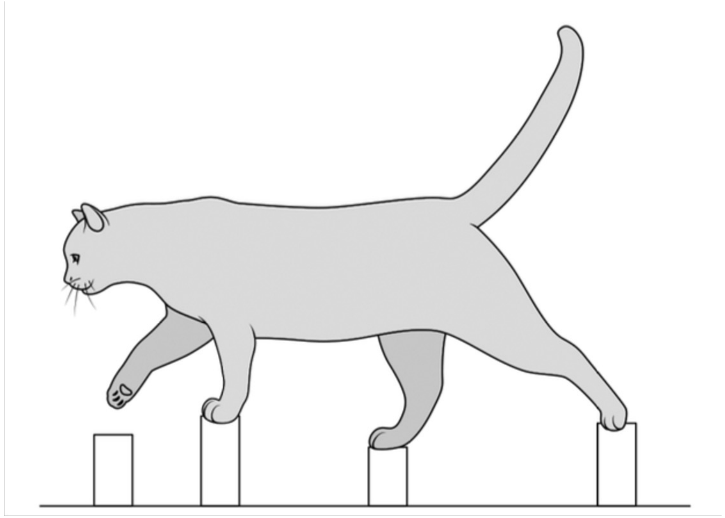
Mammals engage in “Episodic memory”





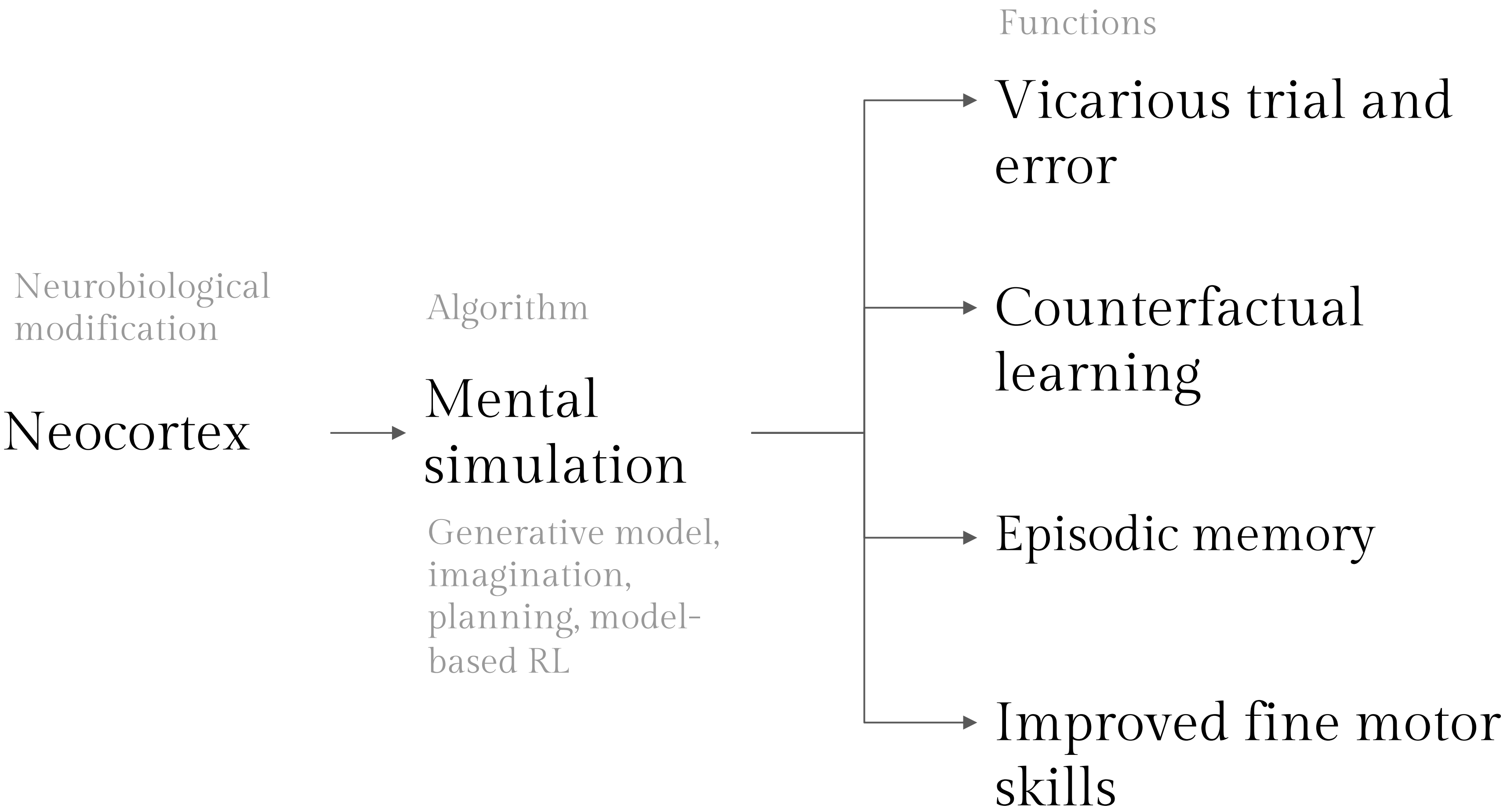
Zhou et. al., 2012

Mammals have uniquely good fine motor skills relative to ancestral amniote

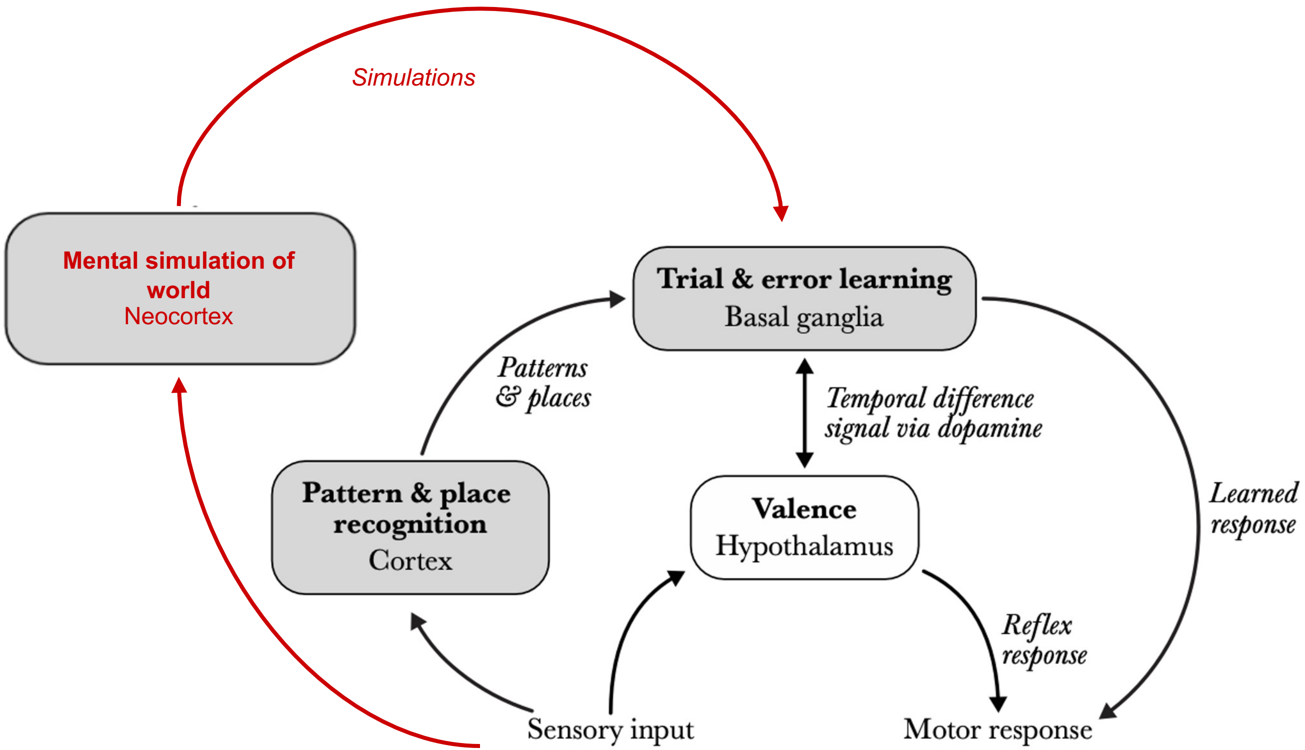


Lizard feet movement do not anticipate obstacles

(Kohlsdorf and Navas, 2007; Olberding et al., 2012; Parker and Mc-Brayer, 2016; Tucker and McBrayer, 2012.)



Neurobiological modification
Neocortex
Algorithm
Mental simulation
Generative model, imagination, planning, model-based RL
Functions
Vicarious trial and error
Counterfactual learning
Episodic memory
Improved fine motor skills



Simulations
Mental simulation of world
Neocortex
Trial & error learning
Basal ganglia
Patterns & places
Temporal difference signal via dopamine
Pattern & place recognition
Cortex
Valence
Hypothalamus
Learned response
Reflex response
Sensory input
Motor response

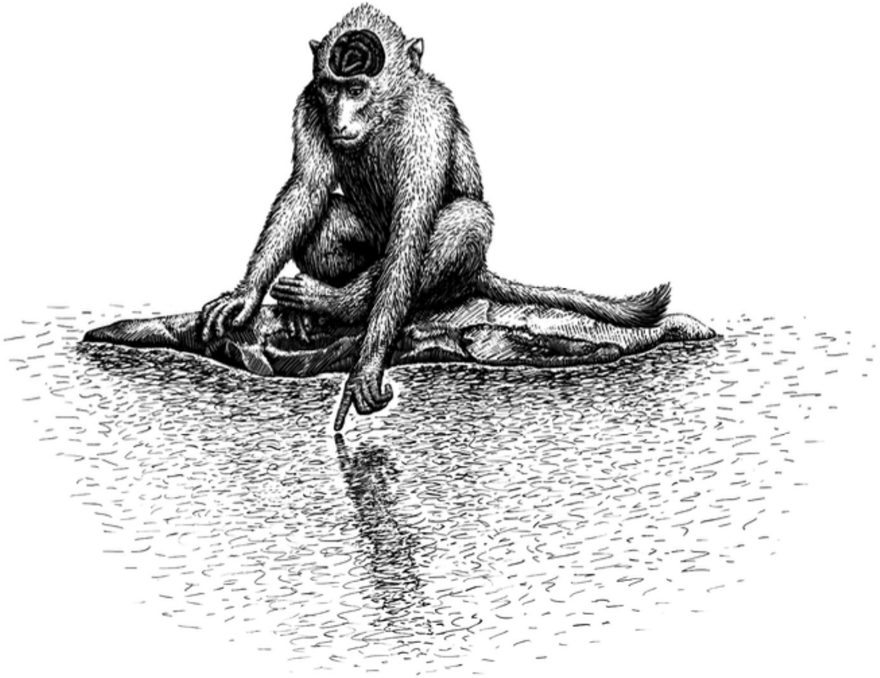
The neocortex enabled "model-based reinforcement learning"

MODEL-FREE REINFORCEMENT LEARNING	MODEL-BASED REINFORCEMENT LEARNING
Learns direct associations between a current state and the best actions	Learns a model of how actions affect the world and uses this to simulate different actions before choosing
Faster decisions but less flexible	Slower decisions but more flexible
Emerged in early vertebrates	Emerged in early mammals

"System 1" / "habit"

"System 2" / "Goal-directed"

AlphaGo was a model-based RL system



Your brain 15 million years ago

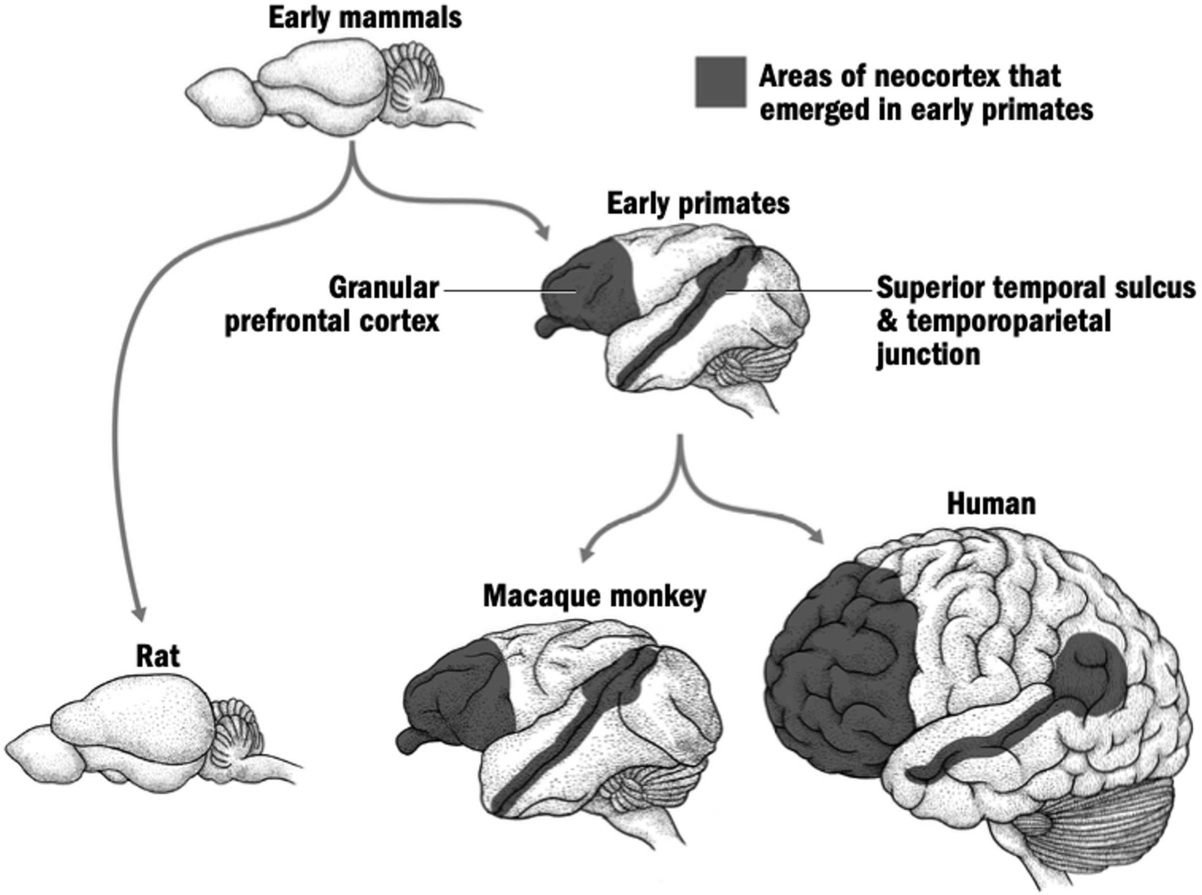
The First Bilaterians

The First Vertebrates

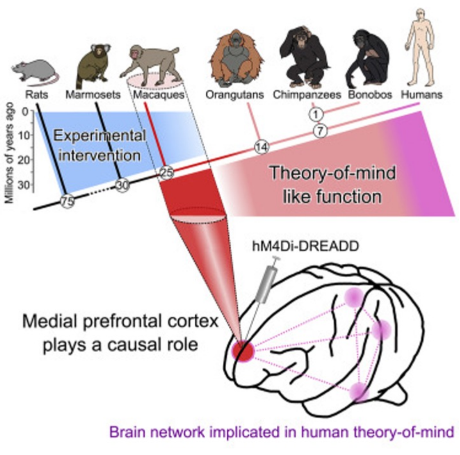
The First Mammals

The First Primates

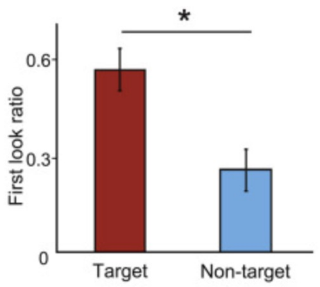
Key new primate regions



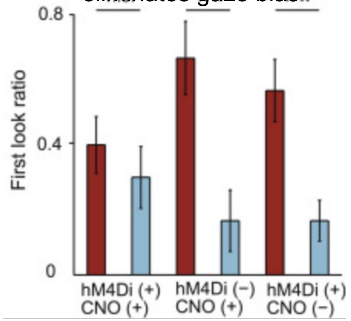
Theory of Mind in nonhuman primates



Correct gaze bias



Inactivating granular PFC eliminates gaze bias



Hayashi et al. 2020

Uniquely powerful imitation learning in primates

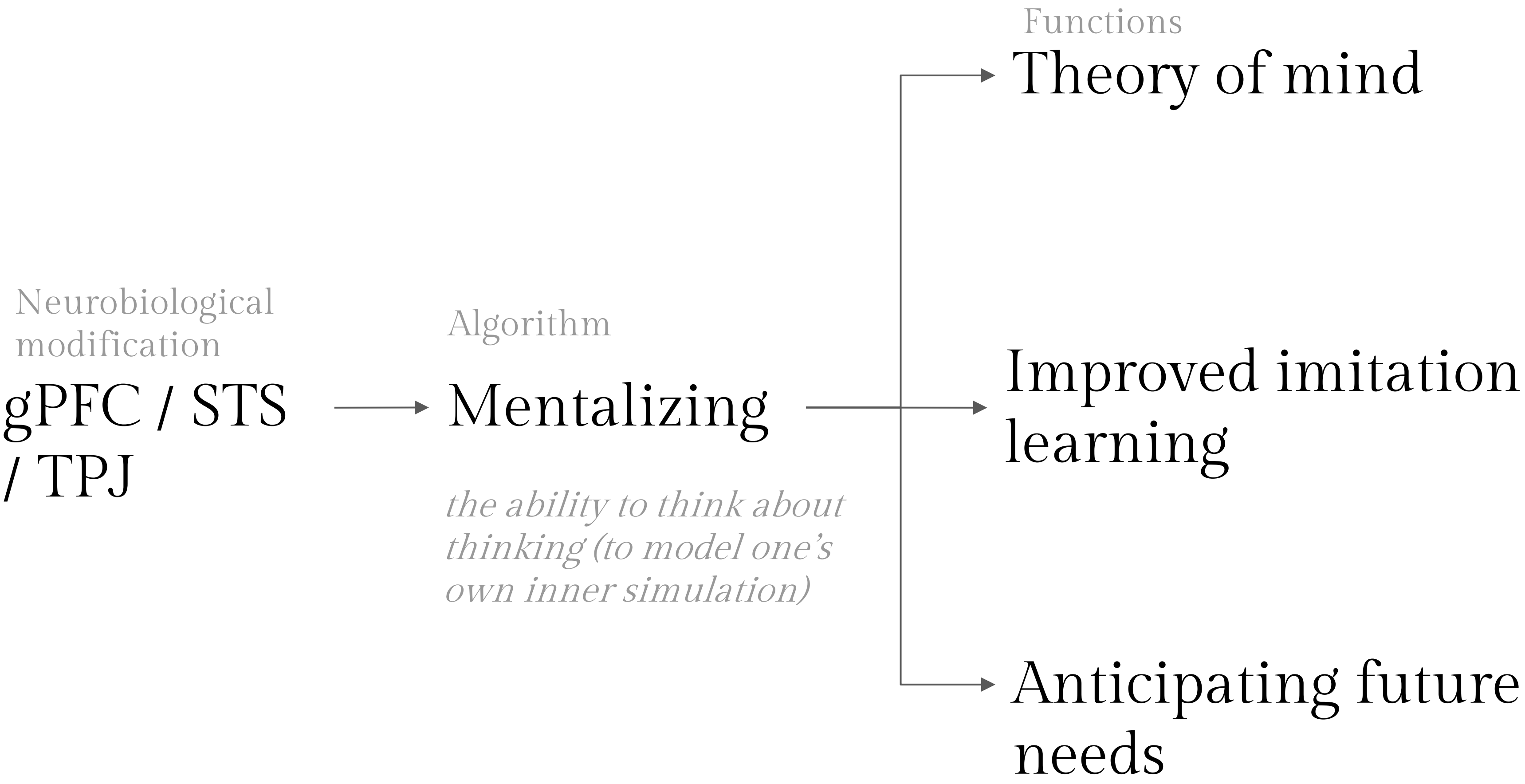
Anticipating future needs in nonhuman primates

Can make a decision *now* to satiate thirst in future, even when not yet thirsty

Can't make a decision *now* to satiate thirst in future, if not thirsty yet

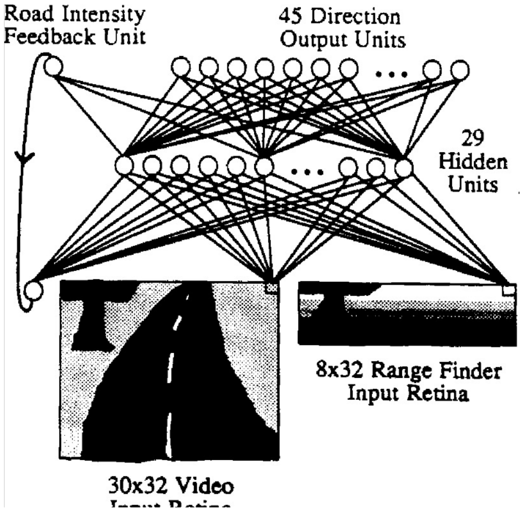
**Mice hoarding is a genetically hard-coded behavior in response to dropping temperature (not an anticipation of a future need) (Barry, 1976)

Naqshbandi and Roberts, 2006


Neurobiological modification
gPFC / STS / TPJ
Algorithm
Mentalizing
the ability to think about thinking (to model one's own inner simulation)
Functions
Theory of mind
Improved imitation learning
Anticipating future needs

Why is mentalizing important for imitation learning?

ALVINN Self Driving Car 1989



Alternative to direct imitation: **Inverse reinforcement learning**

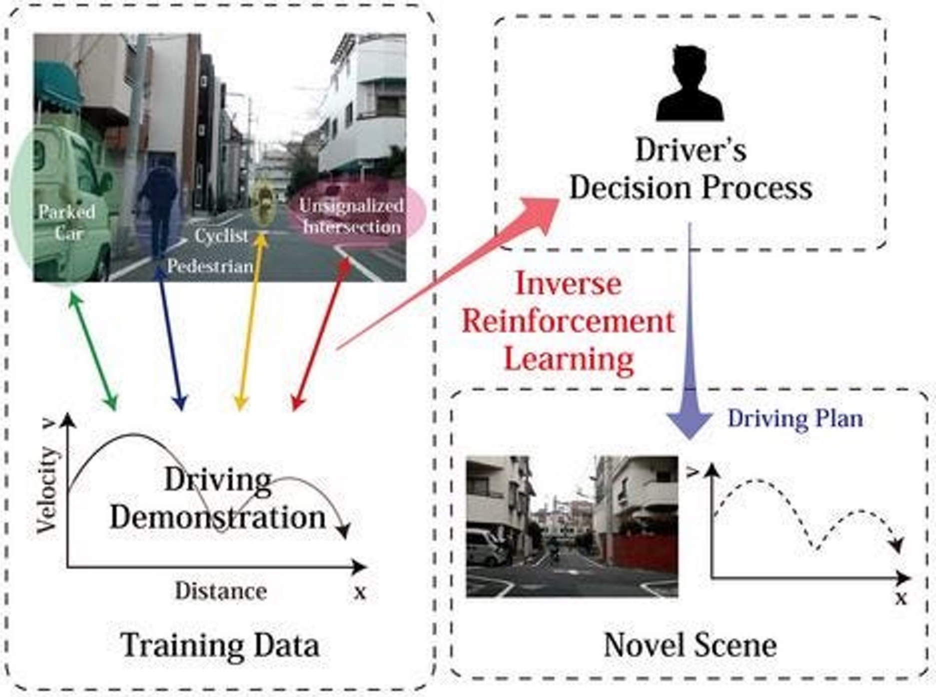


Photo from: https://towardsdatascience.com/inverse-reinforcement-learning-6453b7cdc90d

Ng & Abeel 2004

Why is mentalizing important for imitation learning?

Learn by directly copying

✖

Self-driving by direct imitation

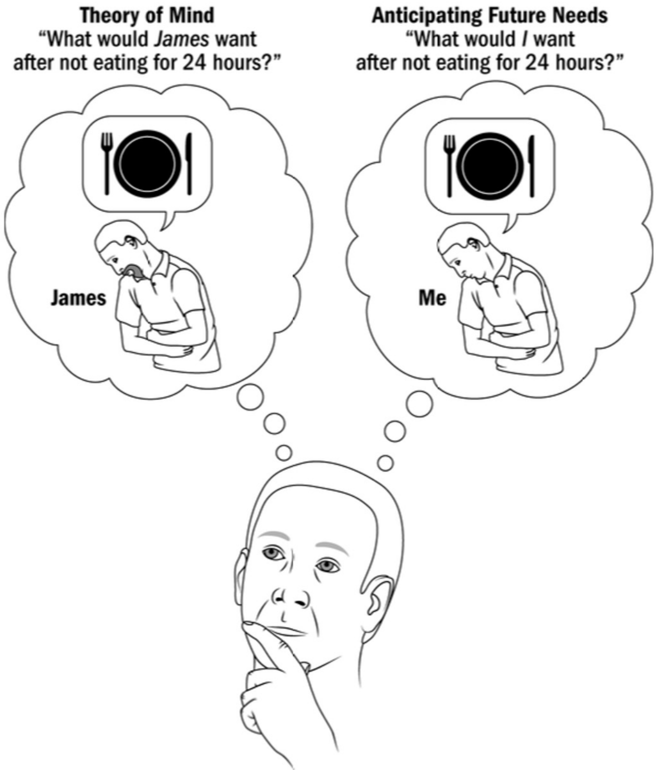
✔

Learn by first inferring reward function ("inverse reinforcement learning"), then teaching yourself

Self-driving by inverse reinforcement learning

Ng & Abeel 2004

Why is mentalizing important for anticipating future needs?



Common algorithm for theory of mind and future need anticipation first proposed by Suddendorf and Corballis, 1997

Neurobiological modification

gPFC / STS / TPJ

→

Algorithm

Mentalizing

the ability to think about thinking (to model one's own inner simulation)

→

Functions

- **Theory of mind**
  Projecting your mind into another mind to infer their intent and knowledge
- **Improved imitation learning**
  Projecting your mind into another mind to simulate their motor skills
- **Anticipating future needs**
  Projecting your mind into your own future to anticipate future mind states

Your brain 100,000 years ago

The First Bilaterians

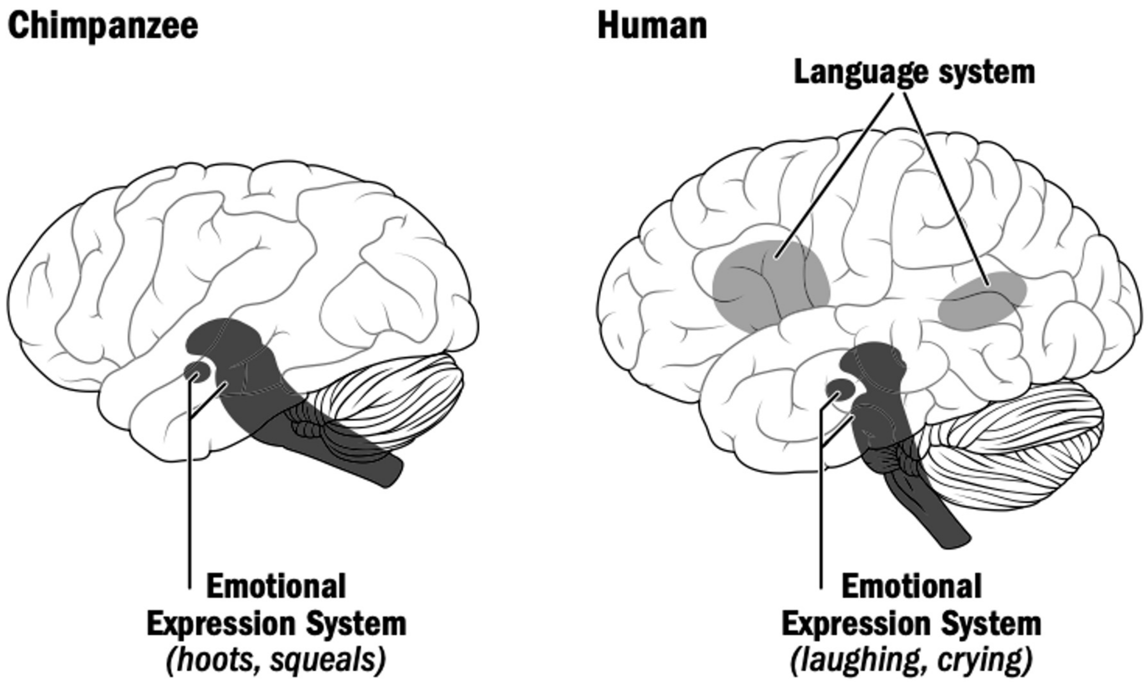
The First Vertebrates

The First Mammals

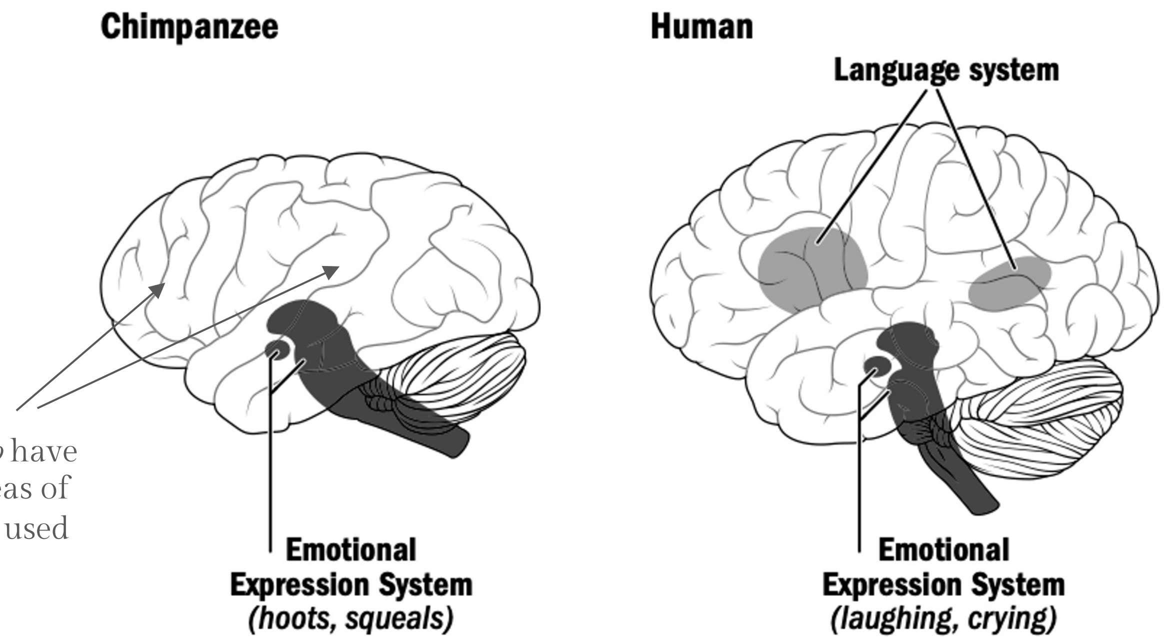
The First Primates

The First Humans

Language *is not* just scaled up primate communication.



There is no unique brain region for language.

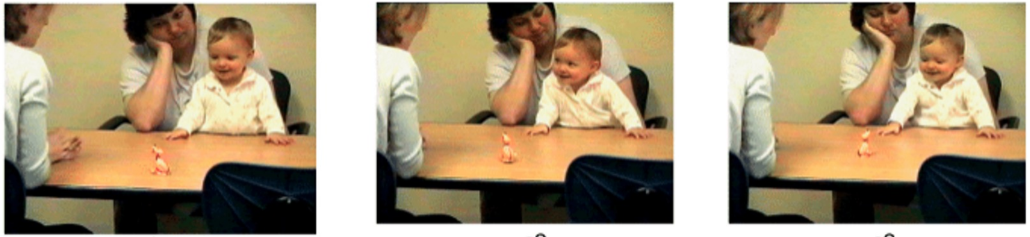


Chimpanzee's *also* have these language areas of neocortex, just not used for language

A unique learning program for language...

Joint attention

Turn taking (i.e. proto conversations)



c1 c2 c3

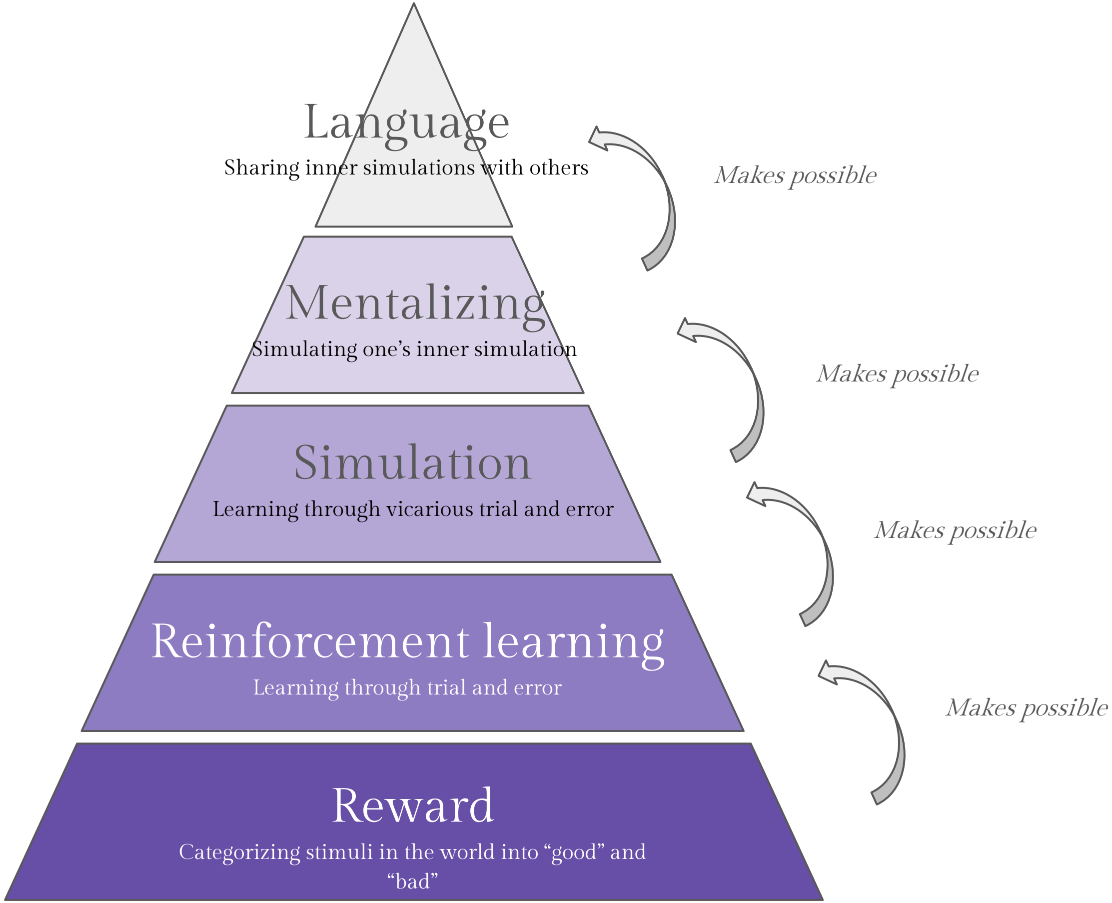
Mundy & Newell, 2007


What i mean
What they mean
“Please maximize
production of
paperclips”

All together: and we get a first approximation of our story:

Milestone	Neurobiological modifications (“implementation”)	Algorithm category	Behavioral abilities (“computation”)
Early bilaterians	neuromodulators for bilateralism first brain	**Taxis navigation (“Steering”)**	Valence (~reward) Associative learning Affective states
Early vertebrates	Vertebrate brain template (forebrain, midbrain, hindbrain) Basal ganglia	**Temporal difference learning (“model-free reinforcement learning”)**	Trial and error learning Pattern recognition Interval Timing Spatial mapping
Early mammals	Neocortex emerges from dorsal cortex	**Generative model (“simulation”)**	Vicarious Trial & Error Counterfactual learning Episodic Memory
Early primates	Granular prefrontal cortex STS/TPJ Direct motor cortex connections	**Second order generative model (“mentalizing”)**	Theory of mind Imitation learning Anticipating future needs
Early humans	Frontal pole? Unique projection from motor cortex to larynx	**Set of instincts for language learning**	Language Beat-based timing

The five breakthroughs - a first approximation of brain evolution



The AI theme on "more data=more performance" can be seen in evolution

The Evolution of Progressively More Complex Sources of Learning

	REINFORCING IN EARLY VERTEBRATES	SIMULATING IN EARLY MAMMALS	MENTALIZING IN EARLY PRIMATES	SPEAKING IN EARLY HUMANS
SOURCE OF LEARNING	Learning from your own actual actions	Learning from your own imagined actions	Learning from others' actual actions	Learning from others' imagined actions
WHO LEARNING FROM?	Yourself	Yourself	Others	Others
ACTION LEARNING FROM?	Actual actions	Imagined actions	Actual actions	Imagined actions

Key idea

Many disparate intellectual skills seemed to from common algorithmic “breakthroughs” that built on top of on prior algorithmic breakthroughs

How does this tool help us?

1. **Helps interpret brain as a whole, instead of through functional divisions** (I.e. "what ability did *adding* a neocortex enable" vs "what does the neocortex do")

1. **Adds constraints on our interpretations of the 'functions' of various modifications** (e.g. helps us see that the motor cortex evolved for motor planning, not motor control)

1. **Narrows classes of "algorithms" to evaluate** (e.g. Algorithms for simulation likely underlie neocortical machinery, algorithms for mentalizing likely underlie new primate regions)

1. **Helps explain multi-purpose neurobiological features** (e.g. dopamine was repurposed for many different things over evolutionary time)

Q&A

maxbennett@gmail.com

www.abriefhistoryofintelligence.com