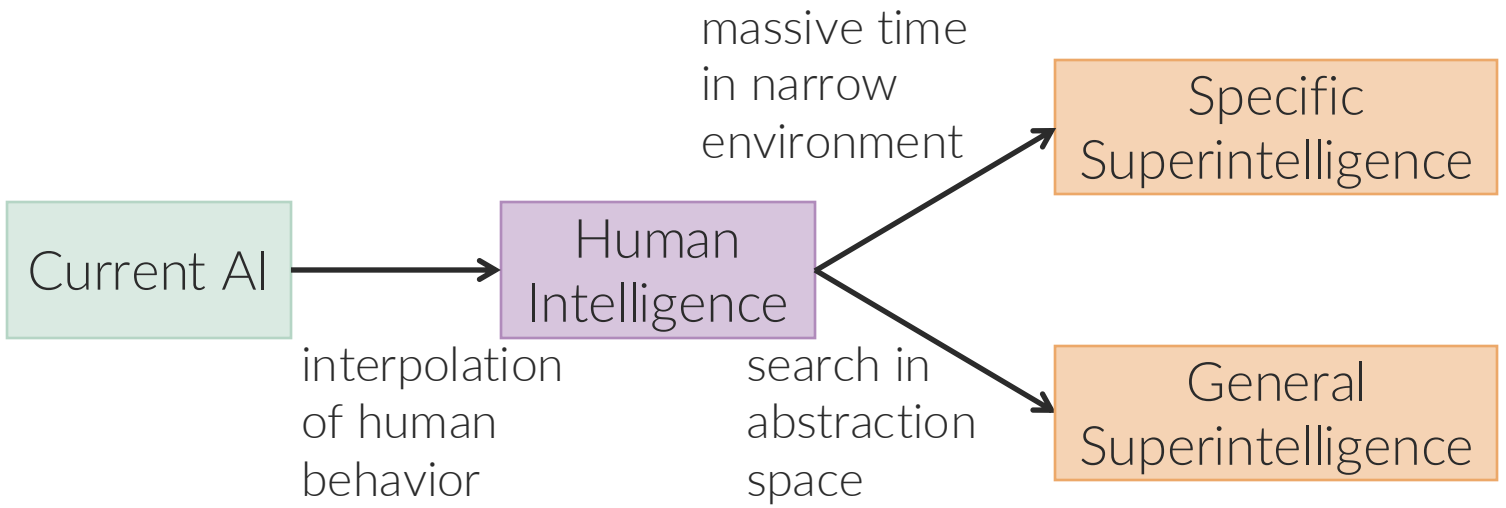
DeUmbra

What Superintelligence Will Look Like
January 25, 2025

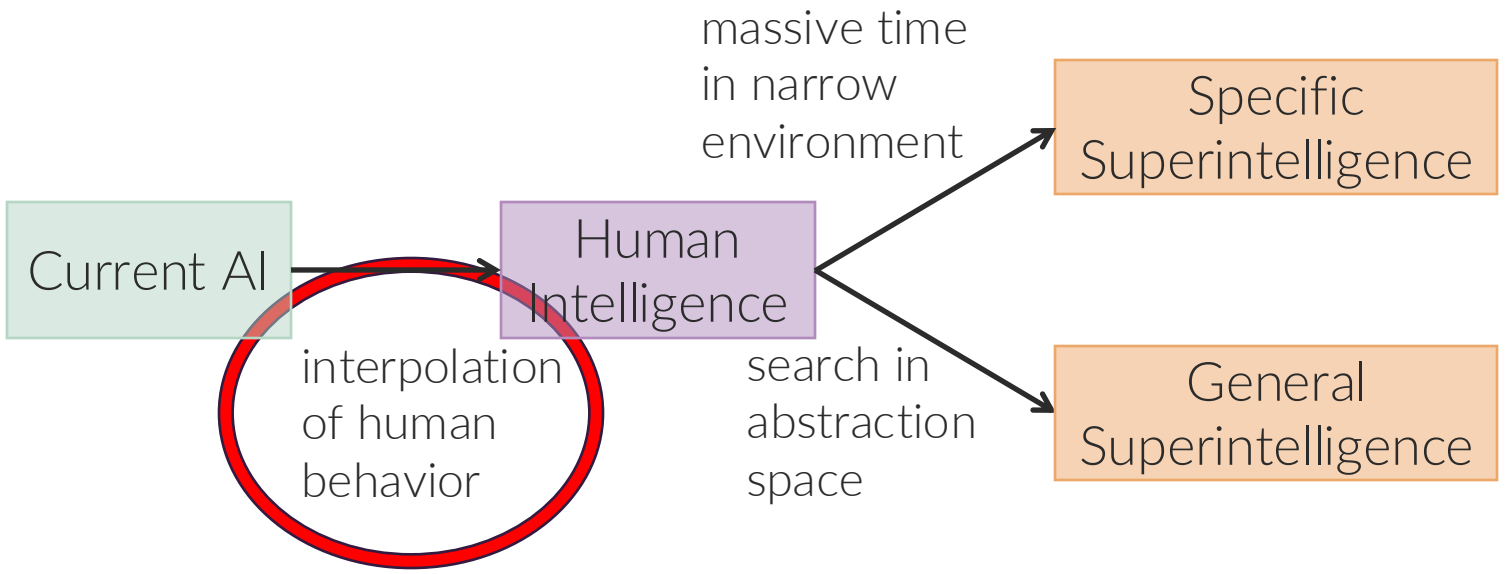
Slides at

https://www.jonathanmugan.com/WritingAndPress/presentations/2025_01_superintelligence_data_day_distribute.pdf

To human intelligence and beyond



Interpolation of human behavior

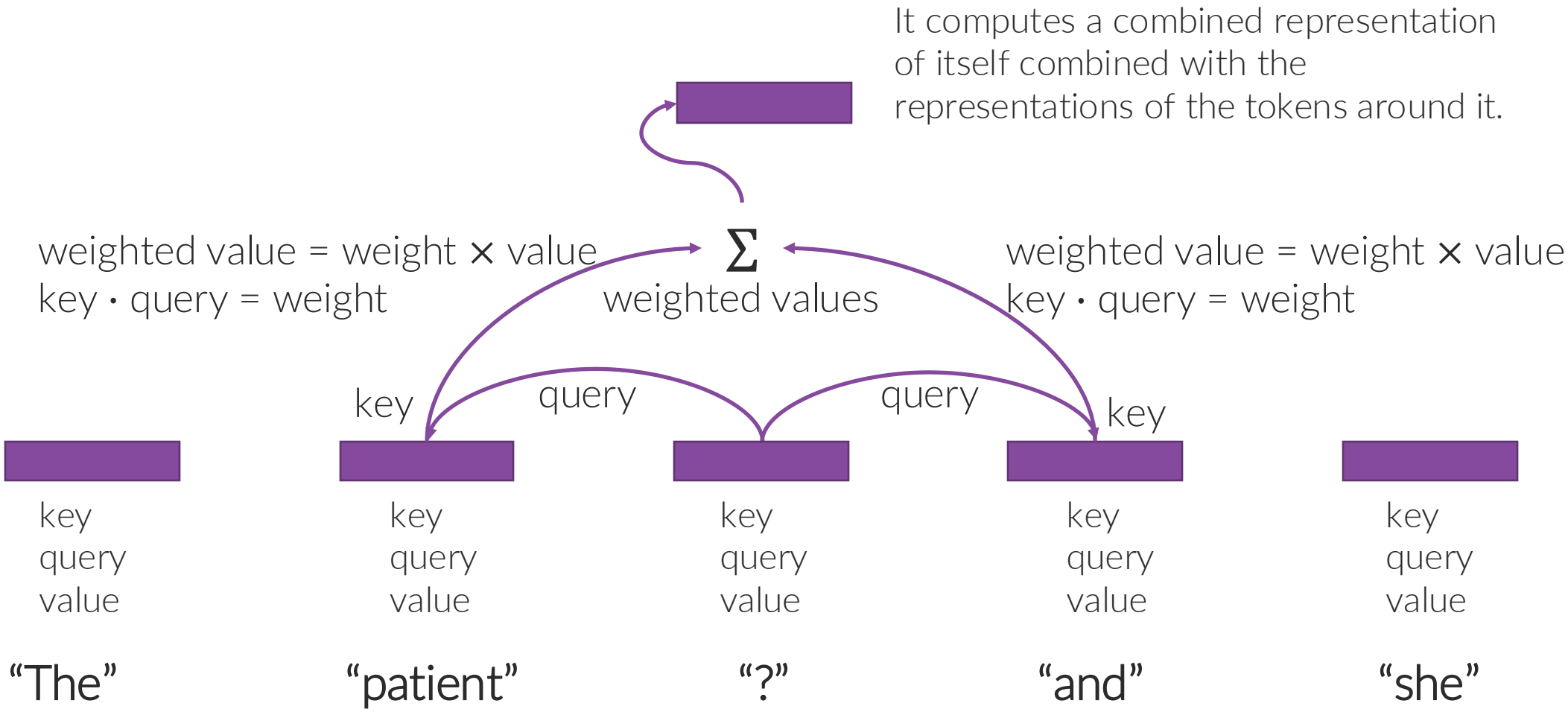


Large Language Models (LLMs) provide low-level interpolation

- They predict the next word based on what they have read.
- They have no explicit model of the world; instead, they have statistics over words.
- When they generate words, they are interpolating over every sequence of words they have ever read to find the sequence of words that best "answers" your question.

Interpolation is estimating an unknown value using the closest known data points.

How transformers (technology of LLMs) perform interpolation



Learned through backpropagation on mostly human data
Backpropagation is chain rule + dynamic programming,
http://blog.ezyang.com/2011/05/neural-networks/

Recall from matrix multiplication

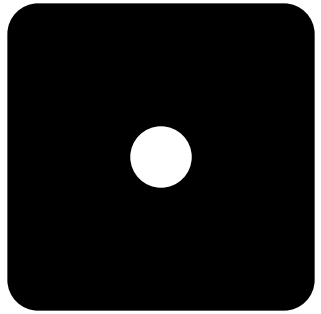
$X^{m \times n} Y^{n \times o} = Z^{m \times o}$

Sizes

v vocab size
d embedding size
l length of text input in tokens

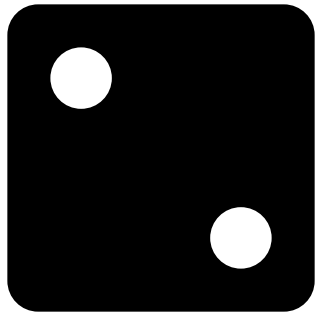
Learned Weight Matrices

$E^{v \times d}$ embedding matrix
$W_Q^{d \times d_k}$ query matrix
$W_K^{d \times d_k}$ key matrix
$W_V^{d \times d_v}$ value matrix
$W^{d_v \times d}$ linear



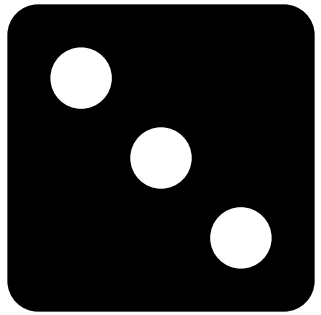
Use embedding matrix to get data

$$X^{l \times d}$$



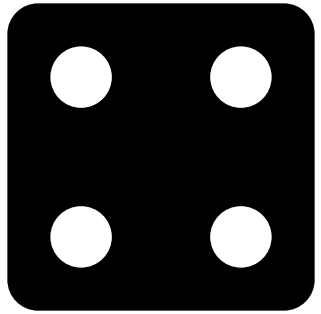
Compute queries, keys, and values

$$Q^{l \times d_k} = X^{l \times d} W_Q^{d \times d_k}$$
$$K^{l \times d_k} = X^{l \times d} W_K^{d \times d_k}$$
$$V^{l \times d_v} = X^{l \times d} W_V^{d \times d_v}$$



Compute attention weights

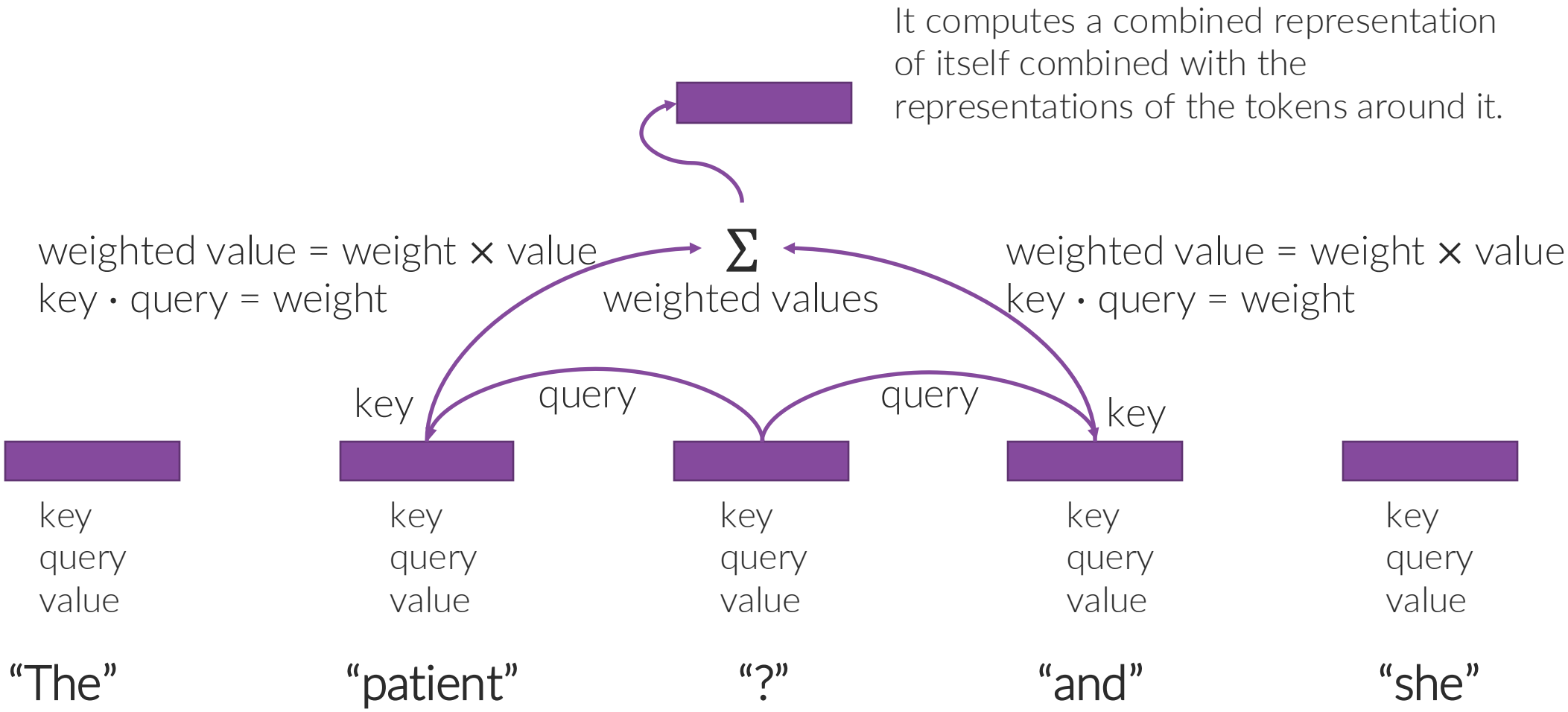
$$A^{l \times l} = Q K^T$$



Use weights to get values and resize for next layer

$$\hat{X}^{l \times d_v} = A^{l \times l} V^{l \times d_v}$$
$$X_{next}^{l \times d} = \hat{X}^{l \times d_v} W^{d_v \times d}$$

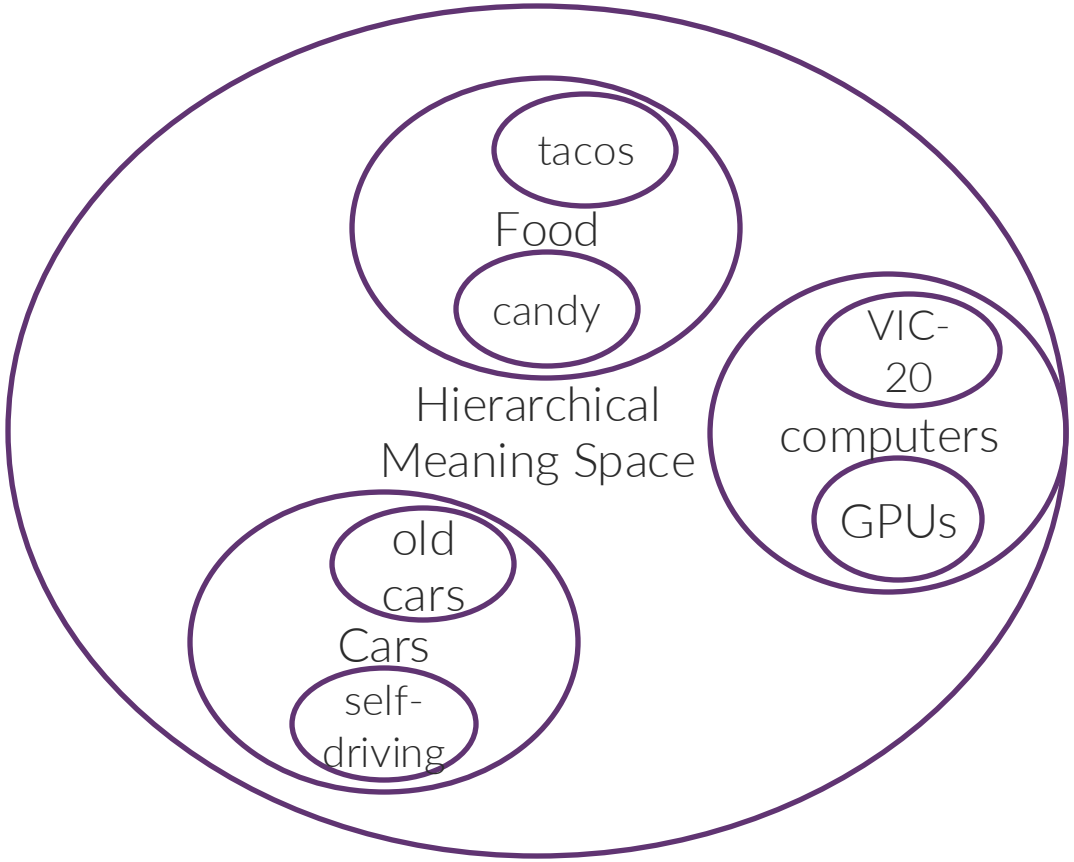
How transformers (technology of LLMs) perform interpolation



GPT-3: 96 heads, 1248 embedding size, 48 layers, plus details like positional encoding

Transformers (technology of LLMs) perform interpolation

That's why they hallucinate, they are just interpolating, and sometimes the interpolation doesn't make sense, but they have no way of knowing that, because there is no explicit model of the world.



Interpolation insufficient for superintelligence

Low-level interpolation is too slow for learning and search

- Their interpolative representation results in a superposition of all possible meanings (See 3Blue1Brown, https://www.youtube.com/watch?v=9-Jl0dxWQs8).
- But to learn quickly, AI must ignore many possible meanings and encode crisp borders.
- Adding examples to the prompt is dropping points in the interpolation space. This only works if the space is well populated in that area. It can't learn new concepts that way.

Low-level interpolation is too power hungry

- Too much power consumption
- Talking about building power stations to expand. https://www.youtube.com/watch?v=58zHJL1dKtw , "Why Amazon, Microsoft, Google And Meta Are Investing In Nuclear Power"
- Bad for the planet. It's already too hot in Texas.

Interpolation can still get you pretty far

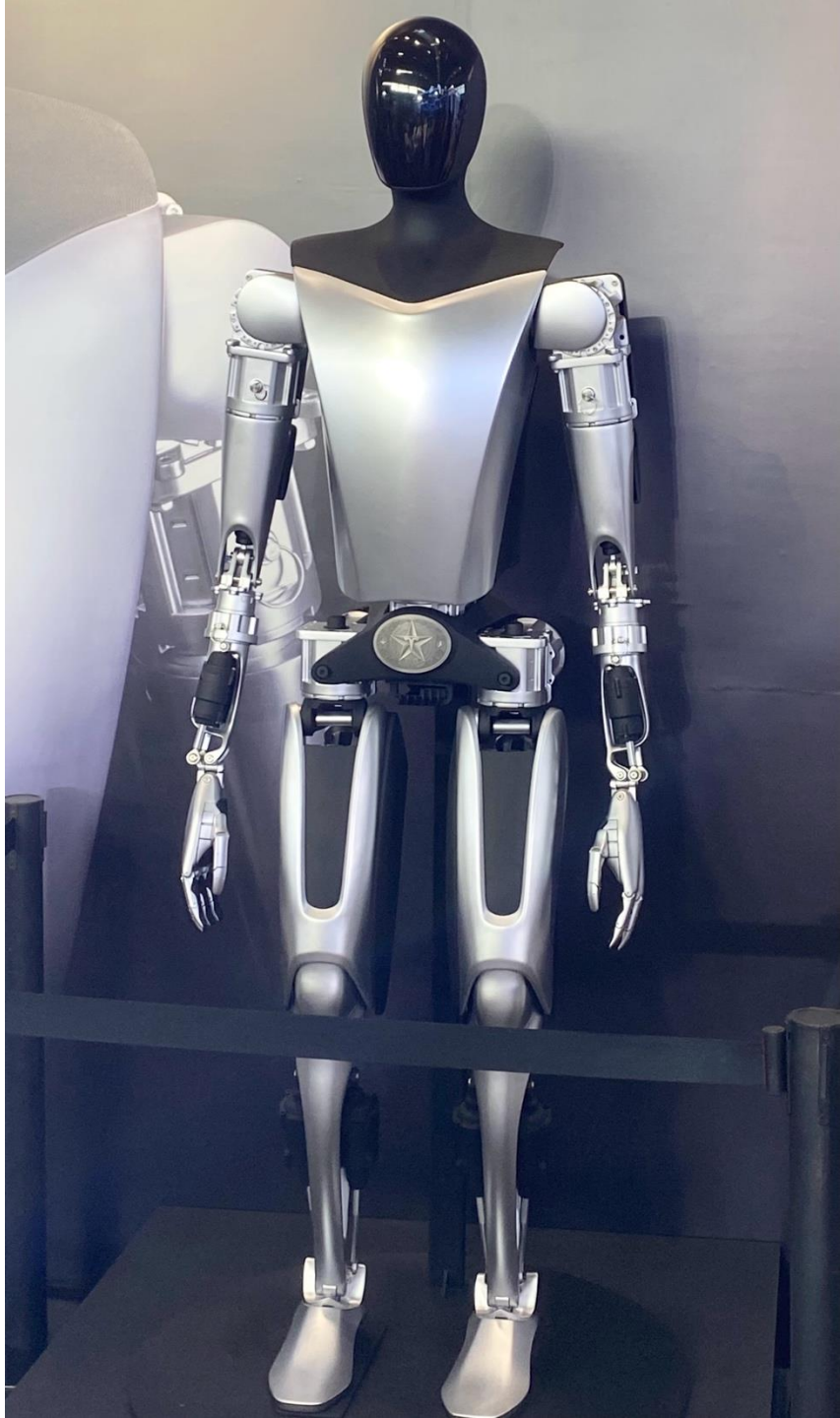
OpenAI o1 and Meta concept models are still in interpolation space

- Chain of thought is good, but still in that low-level space
- Meta: large concept models https://ai.meta.com/research/publications/large-concept-models-language-modeling-in-a-sentence-representation-space/

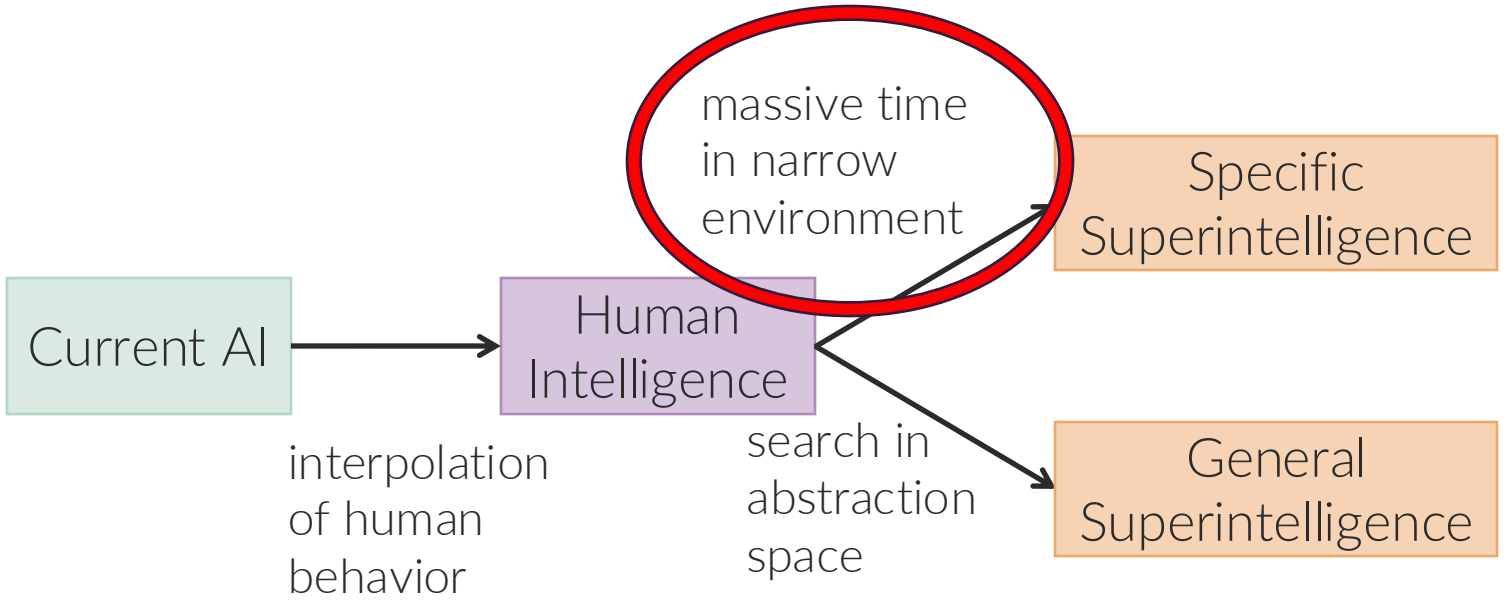
Good enough for household robots

- Basic tasks can be done with interpolation

Photo by Benjamin Ceci, Public domain, via Wikimedia Commons



Massive time in narrow environment



If we have a simulator for an environment, an agent can gain superhuman performance using reinforcement learning.

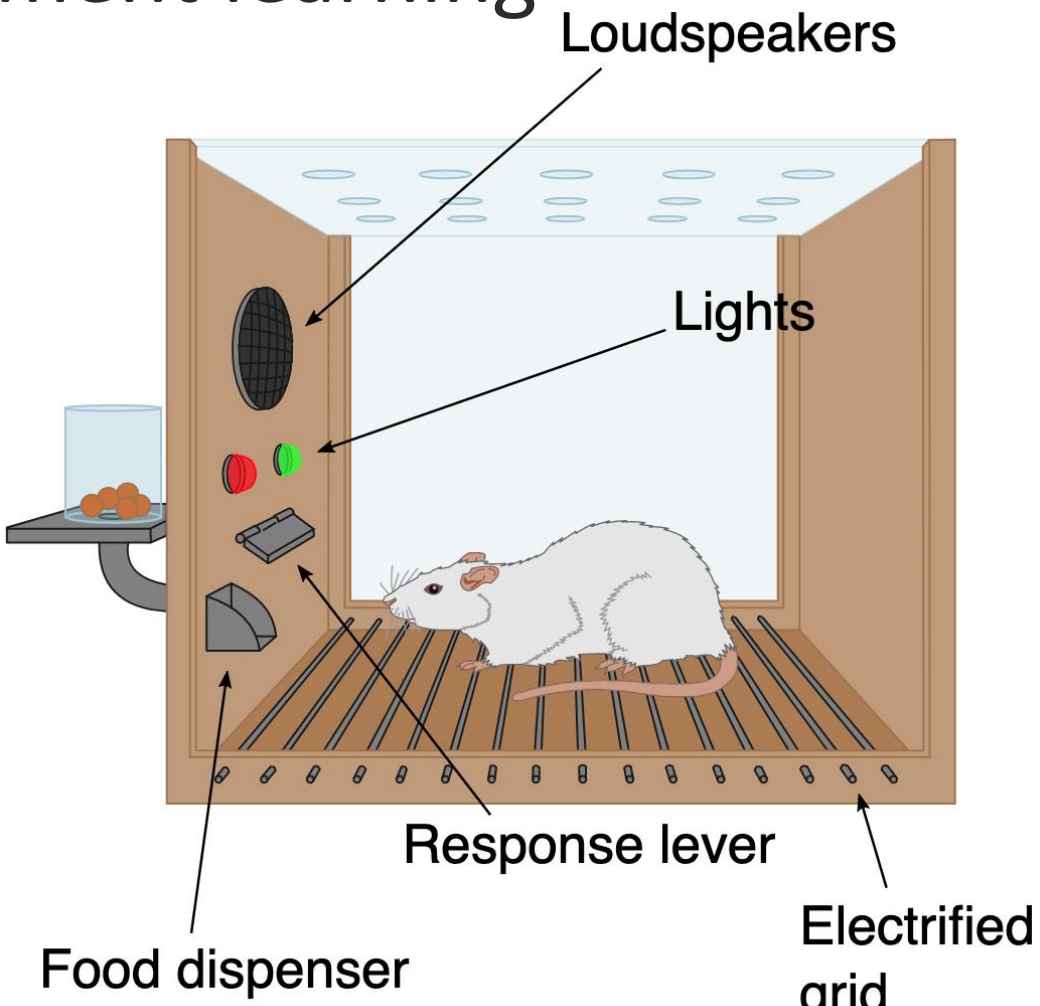
Examples include StarCraft, Go, and Poker

But the set of tasks and the environment must be relatively narrow because learning is so slow.
This also works for narrow domains, such as geometry (AlphaGeometry)
https://deepmind.google/discover/blog/alphageometry-an-olympiad-level-ai-system-for-geometry/

These methods often use reinforcement learning

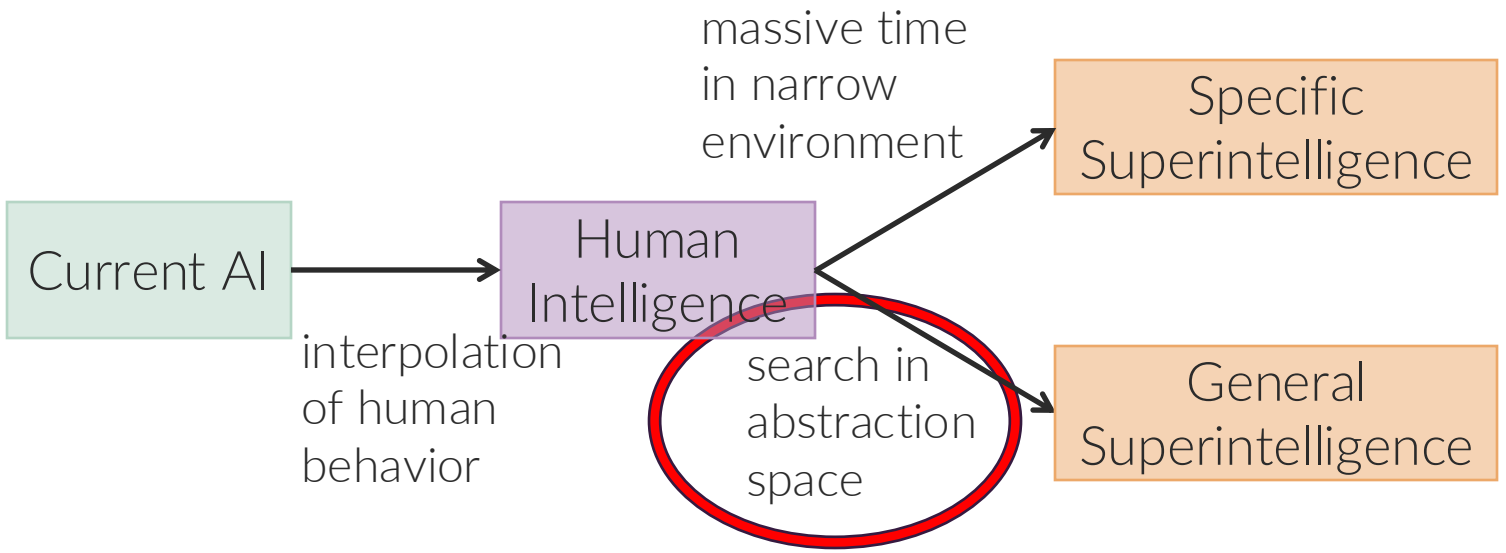
- Some behaviors arise more from a gradual stamping in [Thorndike, 1898].
- Became the study of Behaviorism [Skinner, 1953] (see Skinner box on the right).
- Formulated into artificial intelligence as Reinforcement Learning [Sutton and Barto, 1998].

Reinforcement learning (RL) is a *gradual* stamping in of behavior. It is slow.



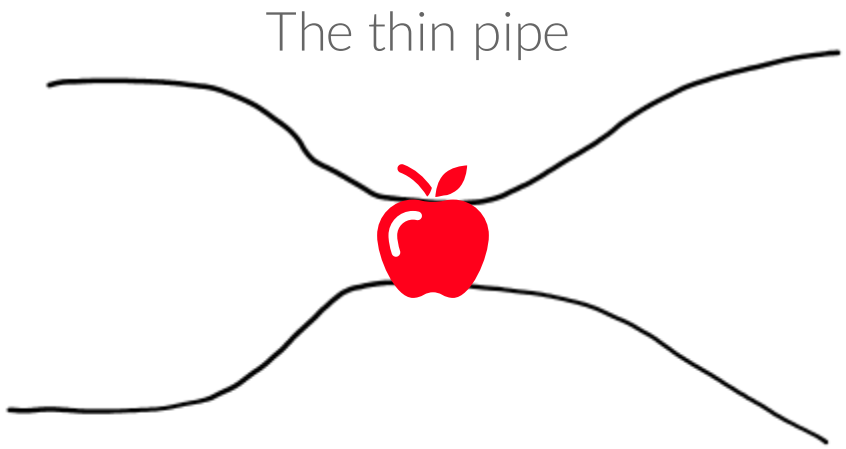
A "Skinner box"

Search in abstraction space



Examples include
electrons
crown moldings
paths
containers
apples

- Abstractions are little islands and joints between them.
- Abstractions tame the fuzzy world so it can be treated like chess.
  - Pixels of a scene are converted into cars and roads
- You throw away bits and regenerate them when you need them
  - The thin pipe
  - This is a theme we will see

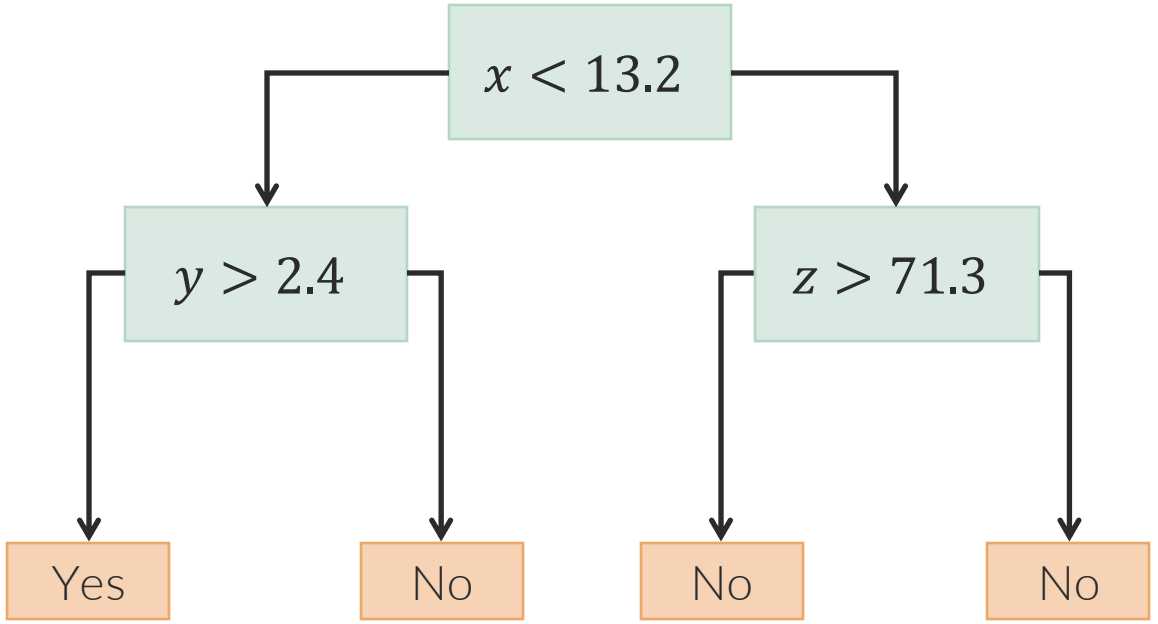


Search in abstraction space

Abstractions Provide Hard Edges

Abstractions are little islands

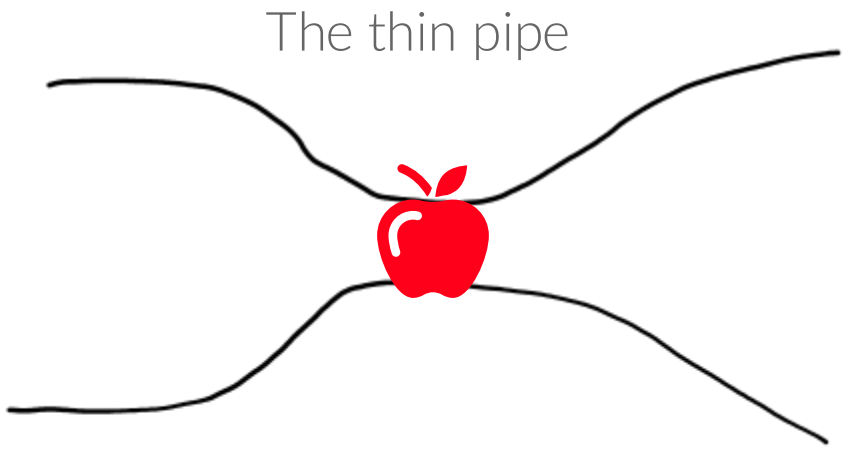
- It is an apple or it is not an apple.
- Hard edges allow you to tightly condense knowledge.
- Consider the number of bits needed to define a decision tree vs. learning that same tree with a neural network.



Abstractions Provide Constraints on Composition

Abstractions are joints

- We can eat an apple but not an agreement
- We can run a race but not an apple.



Search in abstraction space

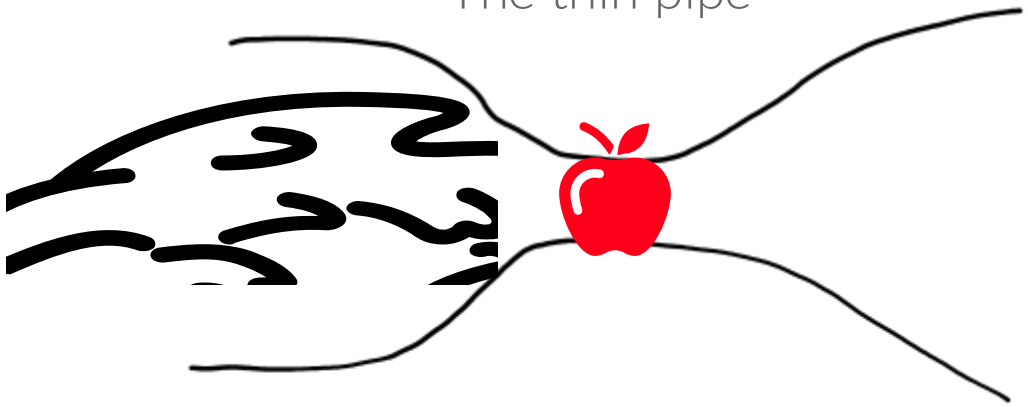
Abstractions determine how we experience the world and thereby define the search space

- We don't see pixels, we see an apple.
- Only certain next states are allowed
  - Enable intelligence to build even more intelligence
  - Consider chess. We build the abstractions and the computer can now out play us because it can focus more memory and compute.

Abstractions Enable Models and Goals

- Provide distinctions for what to predict (models) or achieve (goals)
- Enable one to achieve goals by providing a search space

The thin pipe



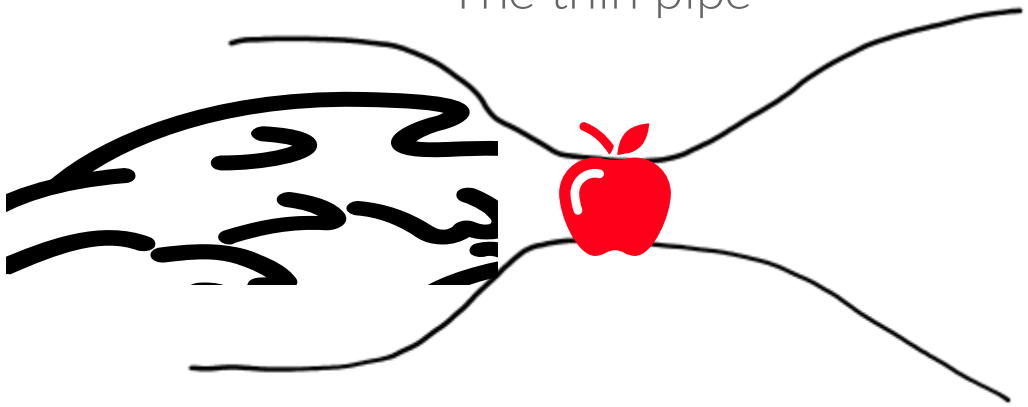
Abstractions give humans our power

Claim: this ability to manipulate abstractions is what gives humans our power, and it led to the language and cultural explosion 70,000 years ago.

- Gorillas don't have it; that's why they can arguably only learn a few dozen signs.
- LLMs don't have it; that's why it takes thousands of training examples to teach them something new.
  - (New meaning outside of their interpolation space; few-shot learning in the prompt must be within interpolation space)
  - There is the Golden-Gate-Bridge Neuron of Claude, but is still interpolation https://www.anthropic.com/news/golden-gate-claude

Ironically, it's the constraints provided by abstractions that are the source of our powerful thinking.

The thin pipe



What superintelligence will look like

- To human intelligence and beyond
- How it will work
- How it will be built
- What it will do
- Conclusion

What superintelligence will look like

- To human intelligence and beyond
- **How it will work**
- How it will be built
- What it will do
- Conclusion

The origin and expansion of abstractions

Self

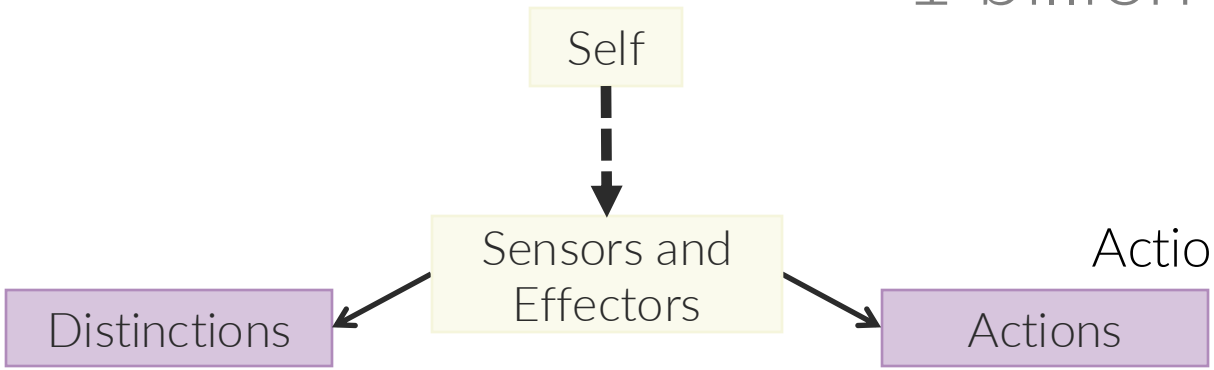
~3.7 to 4.0 billion years ago

- At first there is only some matter that happens to be able to sustain itself and replicate, so it stays around
- It has a border demarking self and not-self
- [Pross, 2016]

The origin and expansion of abstractions

~1 billion years ago

time



Actions are things the agent can do

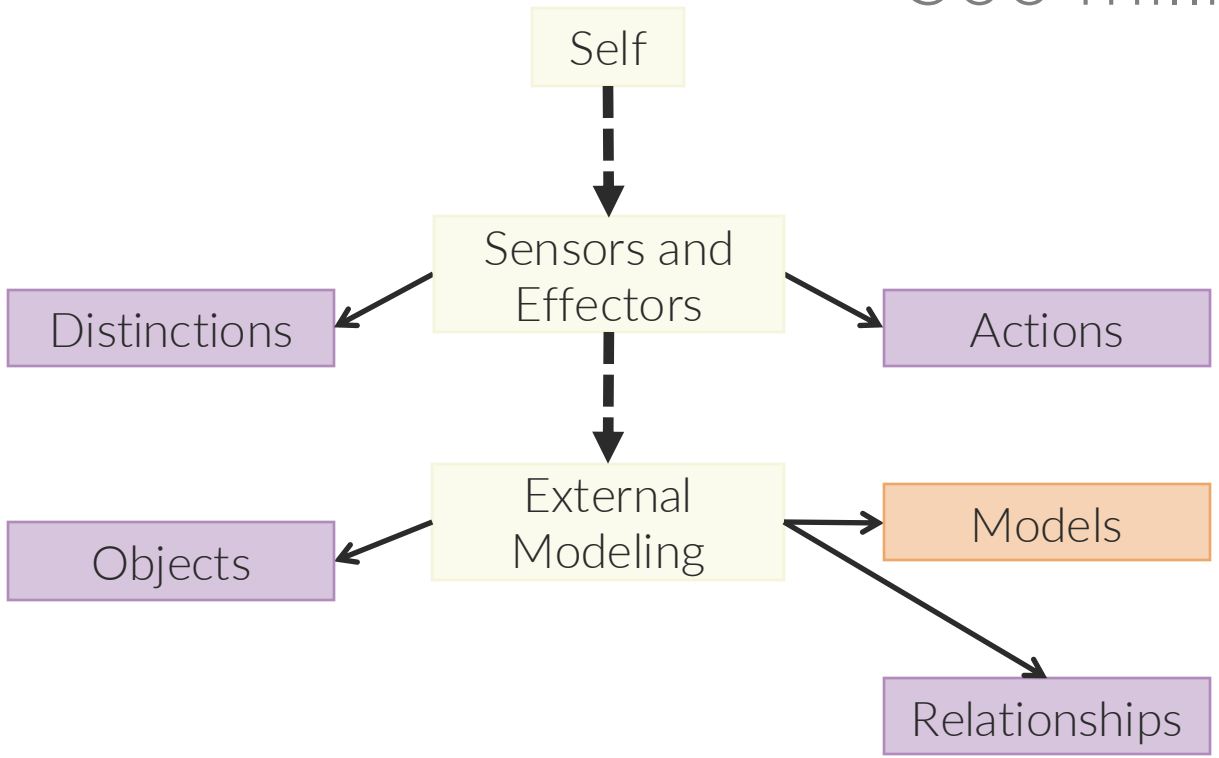
Distinctions are differences
Examples: near, far

- The matter stays around better if it is better able to sustain itself and replicate
- When the first sensor connected to the first effector, purpose came into being [Goddam, 2022]
- With a simple sensor and effector it is able to make simple distinctions
- E. coli spin flagella to “run” toward food if food is present, otherwise they “tumble” [Bray, 2009]

The origin and expansion of abstractions

~500 million years ago

time

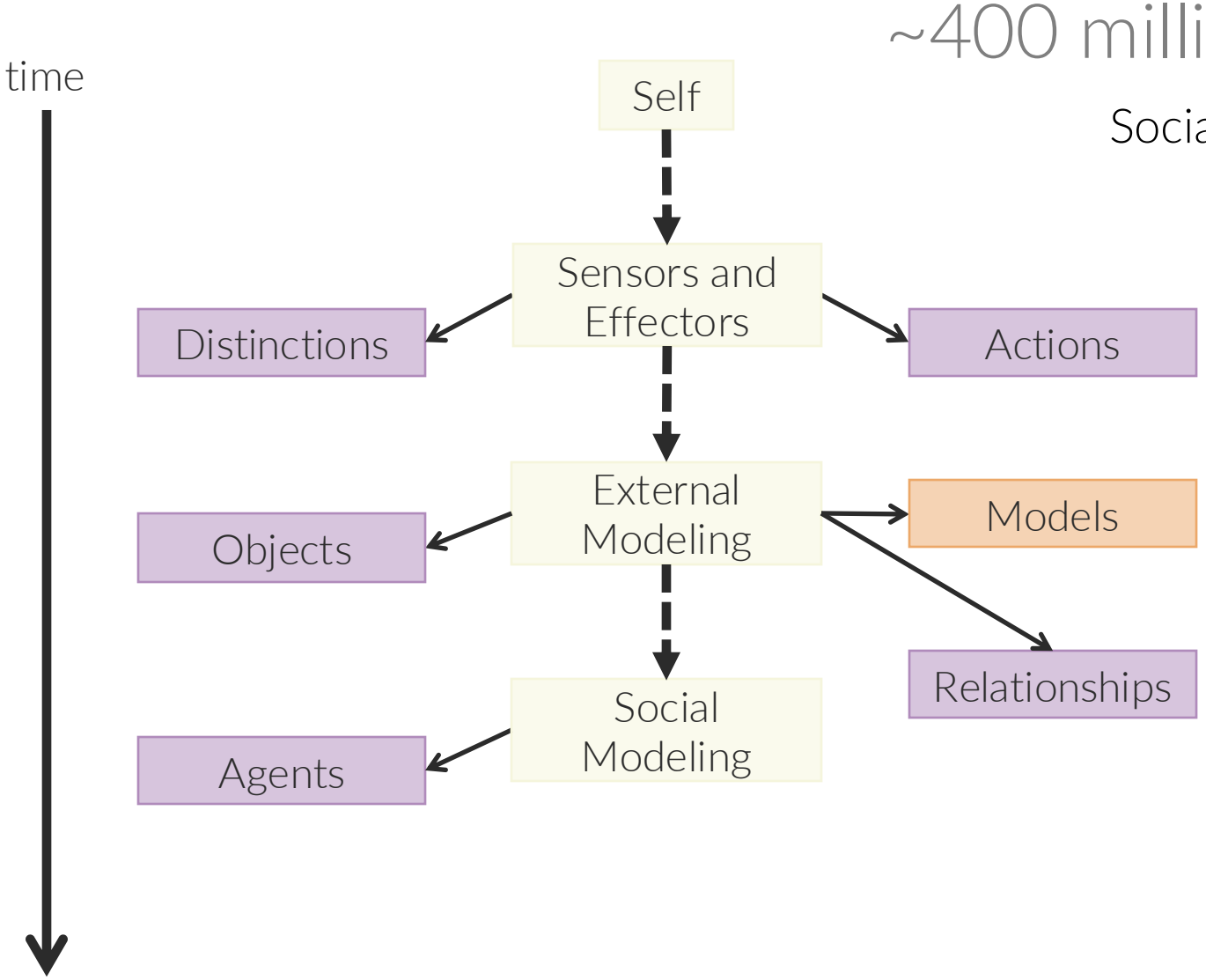


Simple animals evolved to have more sophisticated eyes and neural circuits, leading to the ability to process objects, models, and relationships.

Perspective from developmental psychology
[Johnson, 1987], [Pinker, 2003], [Spelke, 1992, 1994]

attachment, part-of, a door is part of a house

The origin and expansion of abstractions

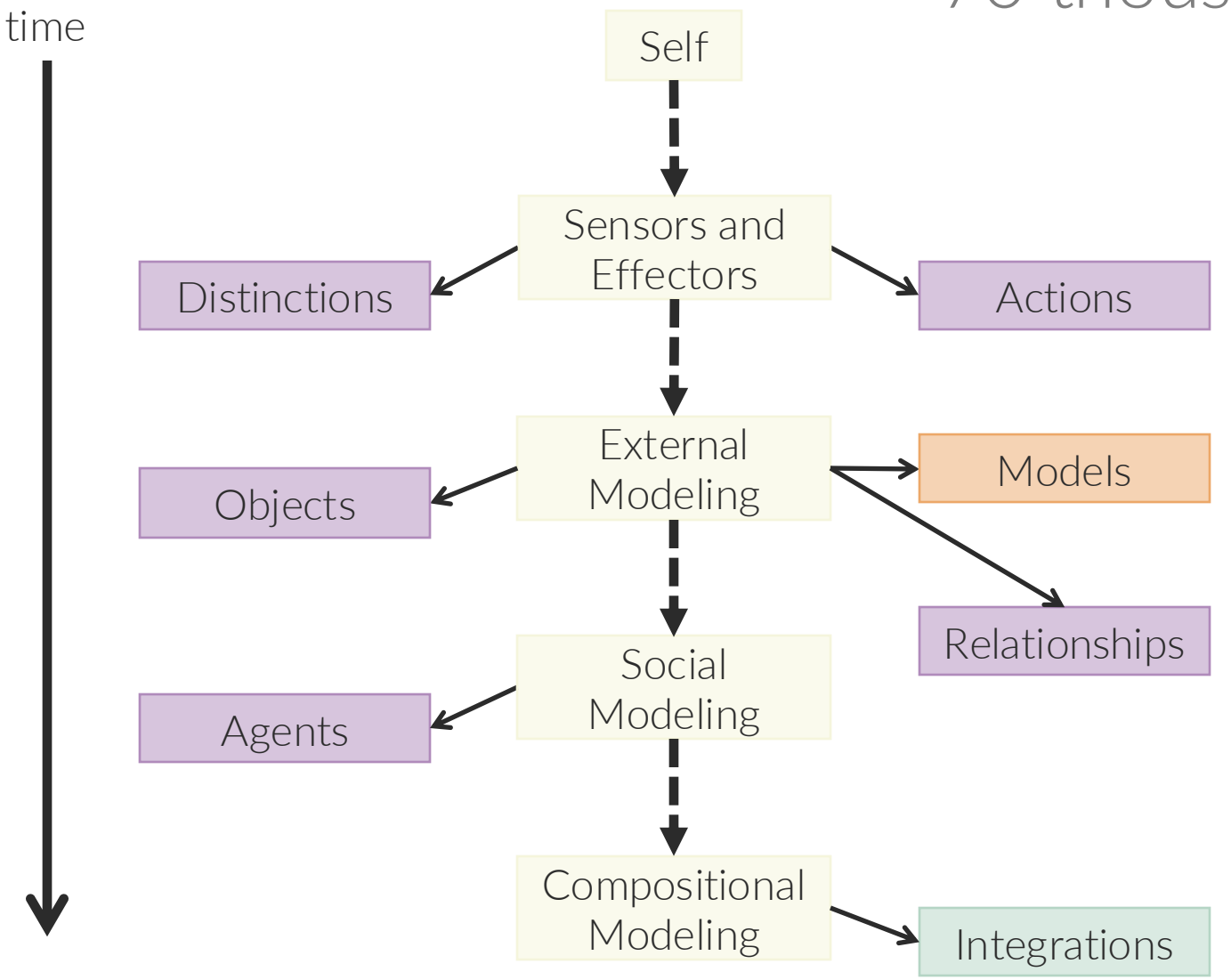


Social hierarchies and social cognition emerge.

[Tomasello, 2019]

The origin and expansion of abstractions

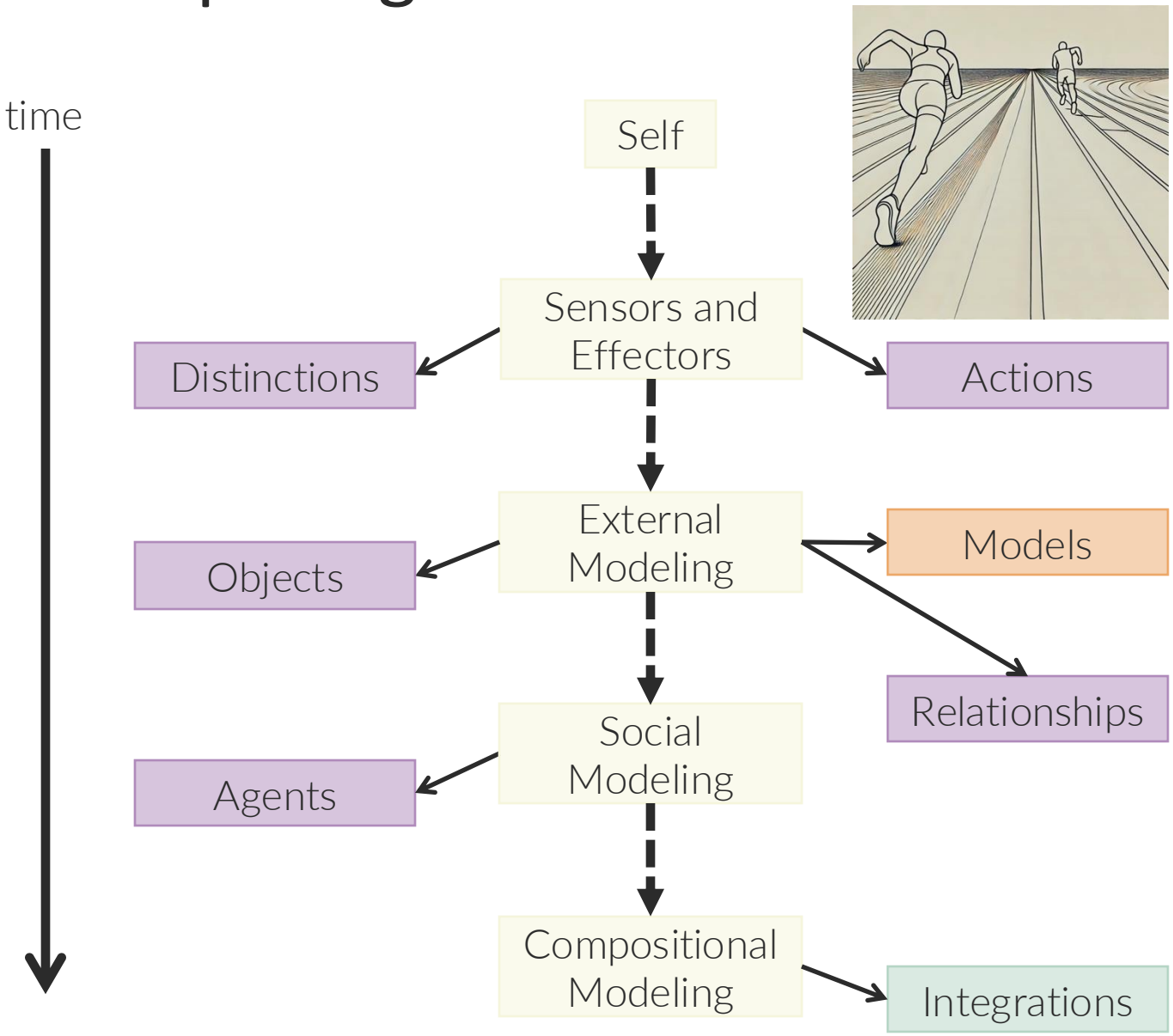
~70 thousand years ago



Dynamically integrating abstractions seems to be what makes humans unique.

References for this kind of thinking in humans [Goodman, 2016], [Gopnik, 2009], [Mandler, 2004]

Composing abstractions



Abstractions integrate together to allow thinking in particular ways

Examples

- Race: start, path, finish, contestant, ahead, behind
- Bounce: bouncer, surface
- Restaurant: table, server, food
- Tax form: income sources, expenses

Have been extensively studied

- Schemas [Bartlet, 1932]
- Chunks [Miller, 1956]
- Frames [Minsky, 1974]
- Script [Schank, 1977]
- Image schemas [Johnson, 1987]
- Explanations [Deutsch, 2011]

Abstraction integrations are

- an abstract object
- and they may include components: entities and roles, a goal (optional), a model with states, an optionally environment layout (e.g., a kitchen)

Abstraction Integrations Enable Metaphors, Analogies, and Blending

Metaphors and analogies are achieved by taking an abstraction integration from one domain and applying it to another. [Lakoff, 1980] [Hofstadter, 2013]

- Example, being behind schedule is from a race

Blending is combining multiple abstraction abstractions (or abstraction integrations) together. [Fauconnier, 2008]

- Example: "That running back is a truck."

Example:

Behind schedule

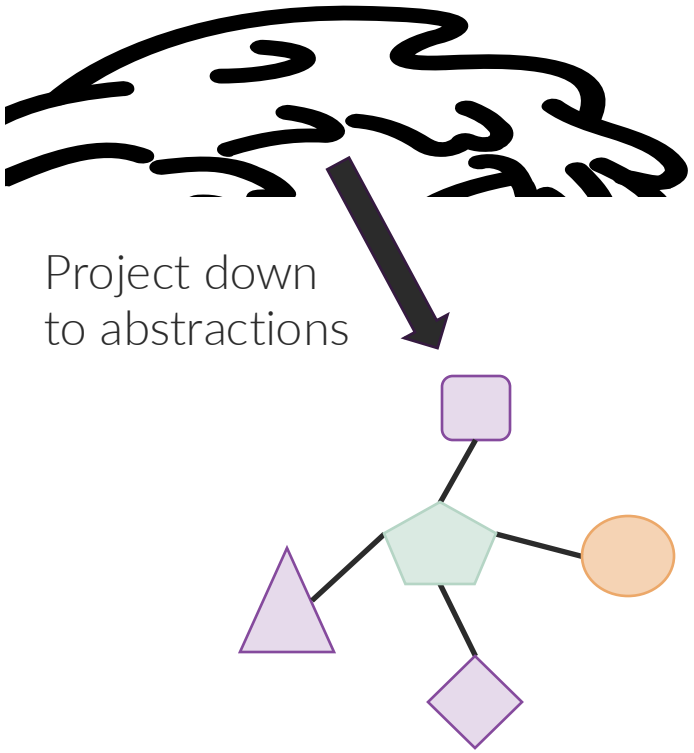
"Schedule" is a runner ahead of you.

You are "behind" but you can still get "caught up."

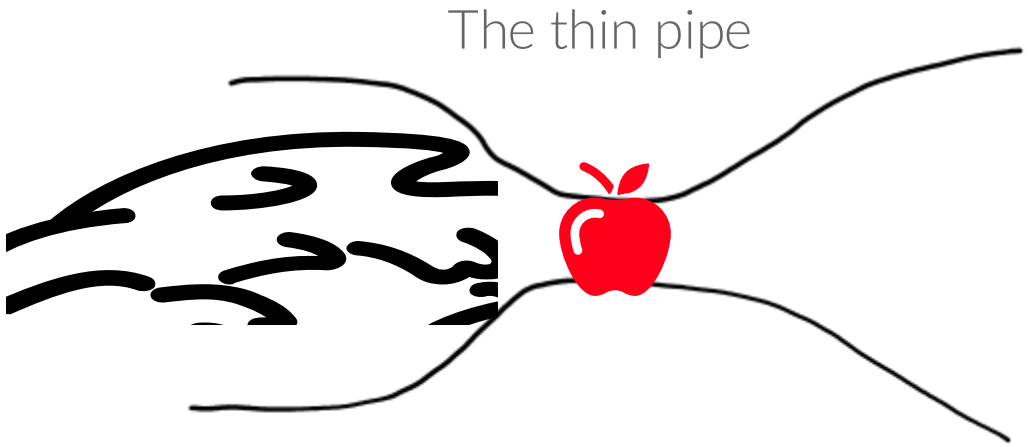
Prompt: *"draw a minimalist line drawing of being behind someone in a race"* to ChatGPT (cropped from image on previous slide)

Projecting abstractions down

We project the chaotic world down to something we can understand



What we experience may not be “real” but it is enough for us to survive. [Hoffman, 2019], [Schreiter, 2019]



Projecting down to build the mental scene

Observations, communication, internet content

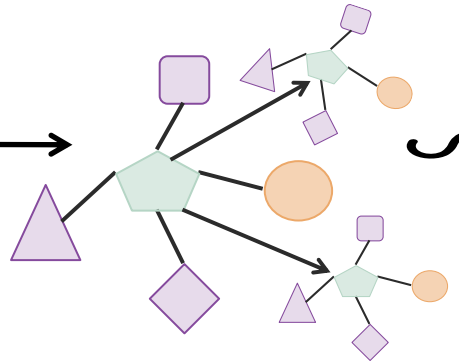
Mental scene

"The table is on the table"

- Abstractions and integrations allow us to build a mental scene.
- Importantly, that mental scene can be about anything, not just a single task or domain, which is what gives us our flexibility.
- This mental scene can be built by neural networks, which would choose the best abstractions based on the current state and the agent's goals.

Simulating forward to predict the future

Observations, communication, internet content



Mental scene

reasoning forward

Possibilities

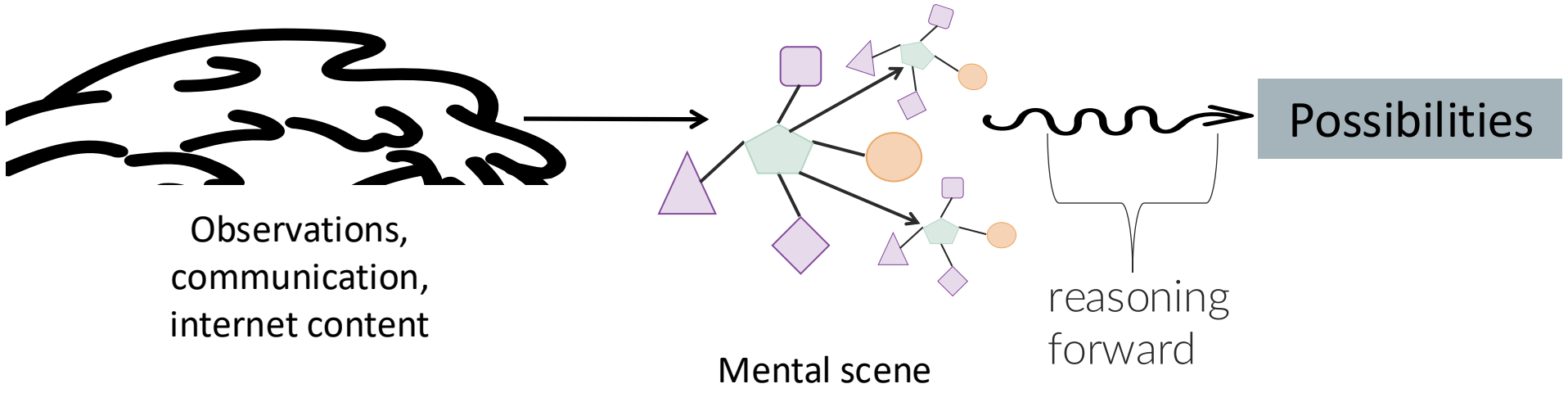
"The table is on the table"

Mental scene

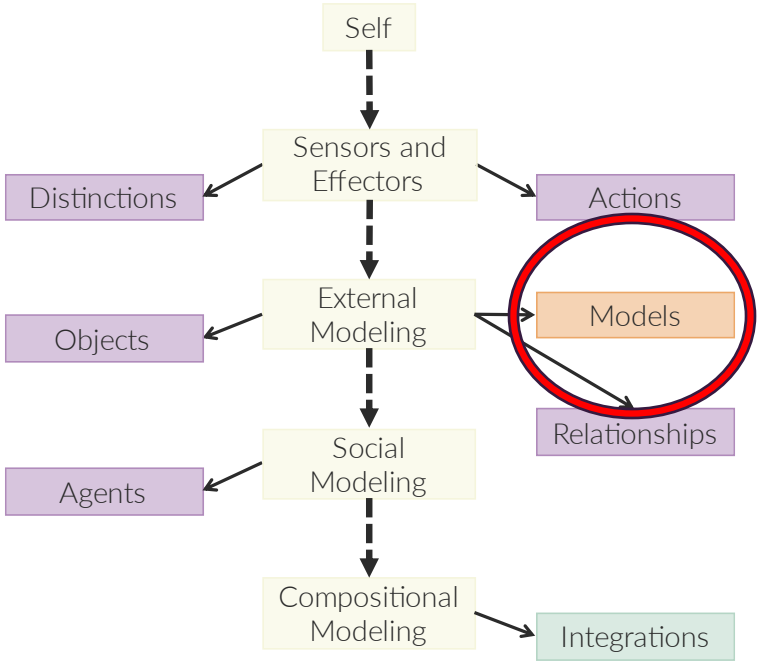
reasoning forward

By directly applying force to the bottom one in the game engine, it observes the top table fall

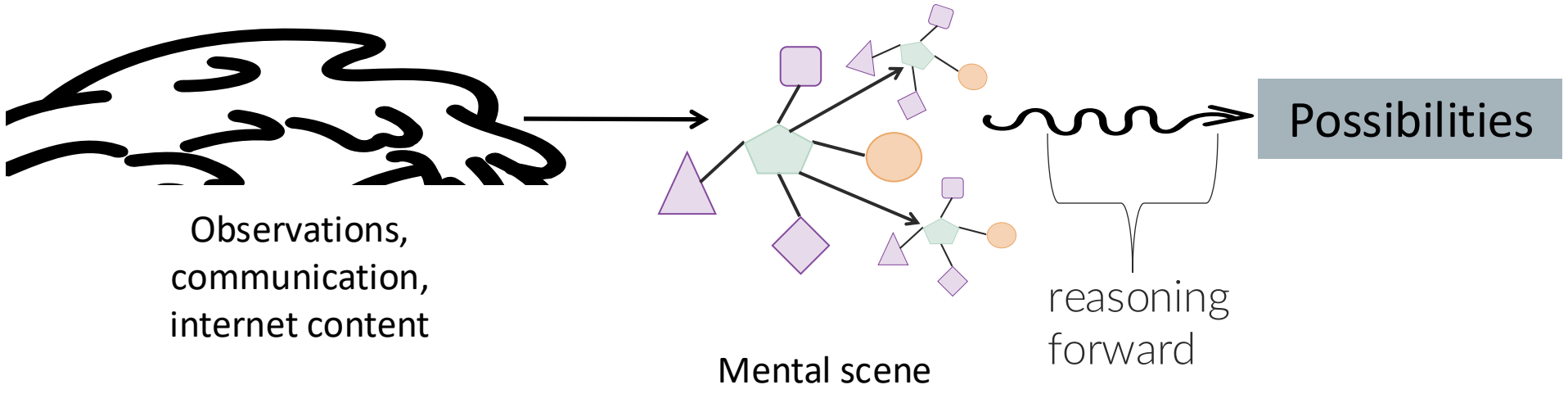
Simulating forward to predict the future



- In building the mental scene, the models allow us to predict the world forward.
- Prediction then allows us to mentally simulate forward [Bergen, 2012]
- As opposed to neural networks which can only predict things close to known, this can predict in novel simulations
- Setting the scene is like writing a Python program, and reasoning is like running it

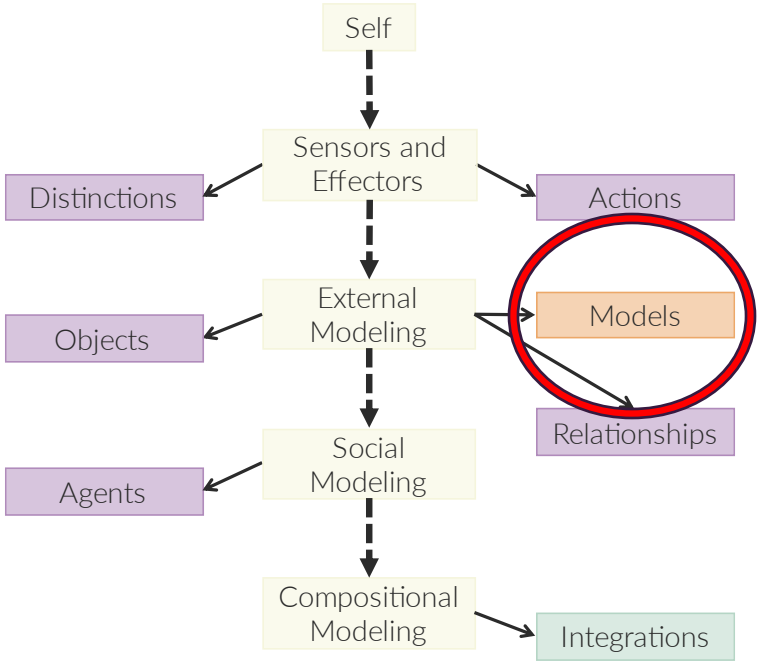


Simulating forward to predict the future



Understanding is mapping sensory input to a useful mental scene

- When I first started in AI, I wondered what it meant to “understand” something.
- Understanding something is mapping it to a useful mental scene consisting of integrated abstractions.
- Useful means you can simulate forward to make predictions that help you achieve your goals.



Thinking is abstraction surfing

With neural networks dynamically building and updating the mental scene using abstractions

Past symbolic methods were brittle because they are only set up for one scene or a set of expected scenes

Abstraction Surfing Example 1: Halting Problem

Consider the problem of whether these algorithms will terminate. It is obvious they will.

```
10 sum = 0
20 sum = sum + 1
30 IF sum < 10 THEN GOTO 20
40 REM end of program
```

```
10 sum = 0
20 sum = sum + 1
30 IF sum != 10 THEN GOTO 20
40 REM end of program
```

`!=` is `<>`
for you purists out there

```
10 sum = 0
20 sum = sum + 1
30 IF sum != 1000000000 THEN GOTO 20
40 REM end of program
```

Abstraction Surfing Example 1: Halting Problem

This one will not.

```
10 sum = 0
20 sum = sum + 1
30 IF sum != 1000000000.3 THEN GOTO 20
40 REM end of program
```

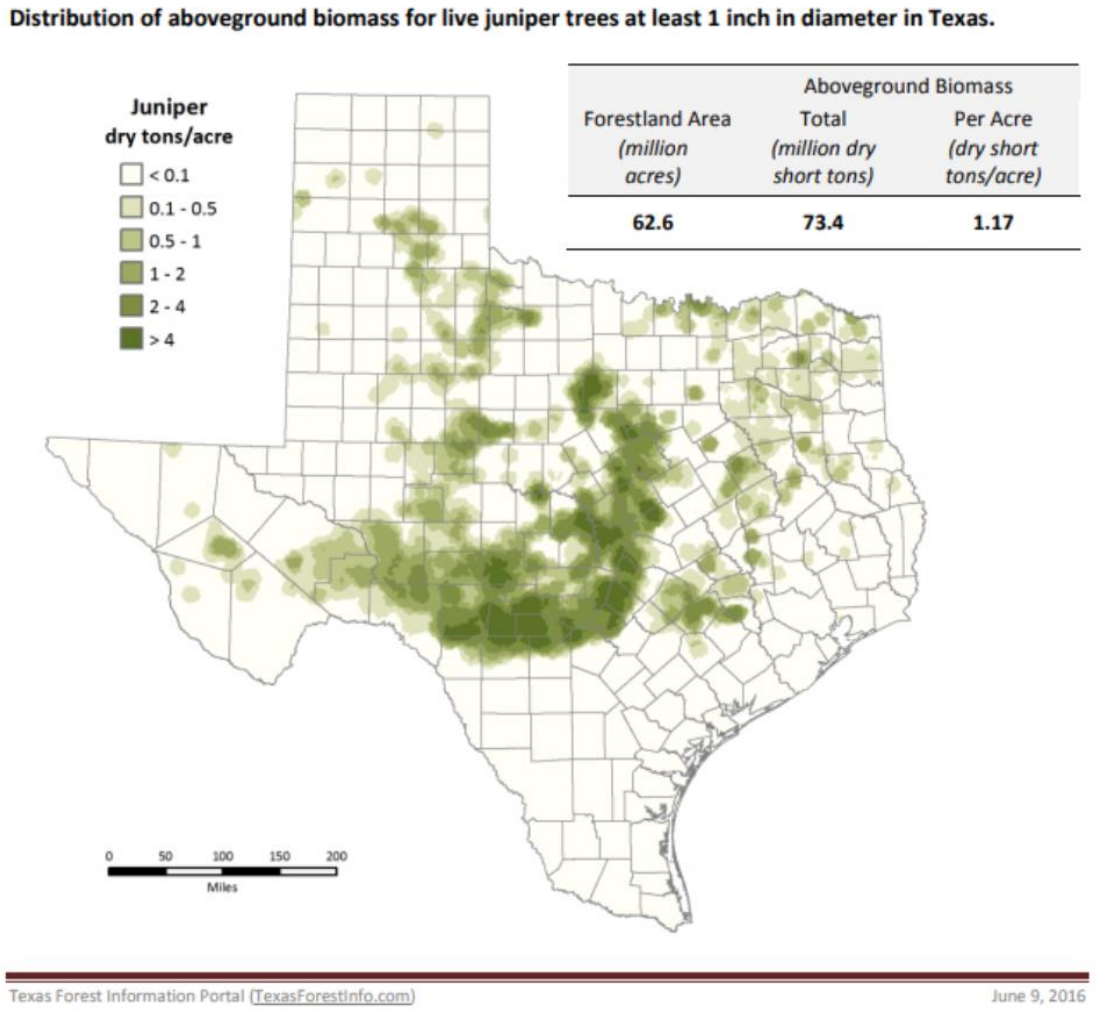
`!=` is `<>`
for you purists out there

Obvious to us. We "jump" to the point 999999999.
Or we surf to the abstraction of intensionally building the set of integers.

We need computers to do this.

- Abstraction surfing doesn't "solve" the halting problem—it makes it irrelevant, just like it is for humans.
- This inability to jump out of the system is behind the Gödel Incompleteness Theorem and barbers who cut everyone's hair who doesn't cut their own.

Abstraction Surfing Example 2: Cedar Fever



https://tfsweb.tamu.edu/content/article.aspx?id=31295
used with permission, thanks Robert!

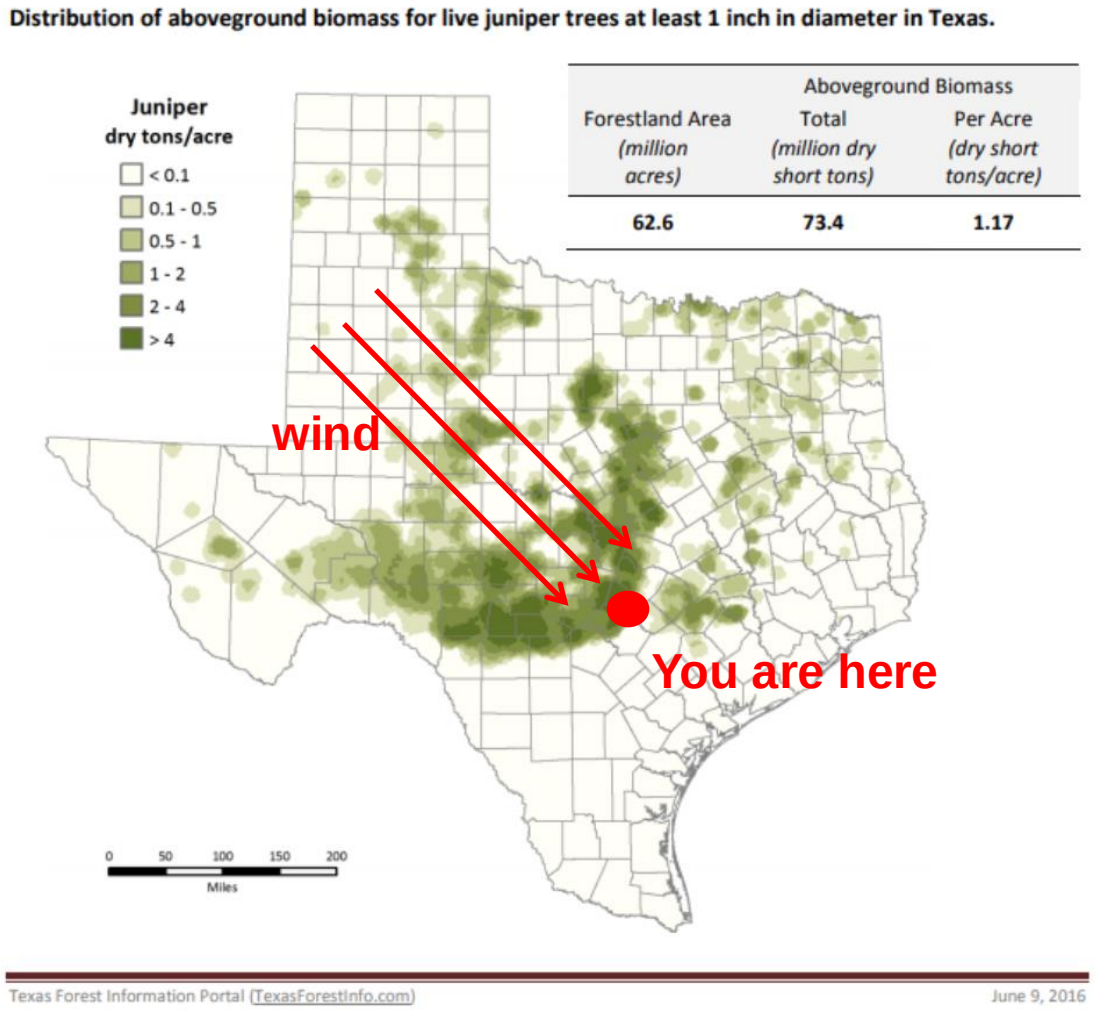
How can we predict what day cedar fever will be bad?

For those who don't live in Austin,

- January is miserable because the ash juniper trees give off a lot of pollen that makes you sneeze and want to gash your eyes out.
- It takes a few years to develop the allergy, so newcomers seemingly have a superhuman ability to walk around outside in January without a care.

We want robots to reason from first principles why dry, windy days are the worst.

Abstraction Surfing 2: Cedar Fever



https://tfsweb.tamu.edu/content/article.aspx?id=31295
used with permission, thanks Robert!

- The model is that pollen is bad when the wind blows from the northwest on cold, dry days.
- A robot can create this model by simulating the pollen coming off trees and floating in the air, but for the pollen to travel long distances, it must jump up a level and simulate it traveling at a rate of speed.
- We don't even notice we do these jumps
- A computer using simulation to think and understand must do the same.

What superintelligence will look like

- To human intelligence and beyond
- How it will work
- **How it will be built**
- What it will do
- Conclusion

Learning abstractions through experience

Building abstractions is a computationally expensive search

The method is simple but hard: generate and test

Generate:

- Create a possible abstraction
- (Guided by a neural network to provide intuition)

Test

Three criteria

1. Matches the world (about things related to the goals)
    - Self-supervision, [Mugan, 2012]
2. Fit with current abstractions
    - Doesn't have to fit exactly, use current abstractions as a soft constraints
3. Enables agent to achieve goals
    - Enables actions to work

Learning abstractions through experience: Bayesian formulation

$$P(h|d) \propto P(d|h)P(h)$$

Example, creating the idea of an electrons

$\boldsymbol{h}$ is that electrons exist

[Griffiths, 2024]

$\boldsymbol{d}$ is experimental data

$P(h|d)$ is probability that electrons exist given experimental data

$P(d|h)$ if electrons existed, how well would they explain the data

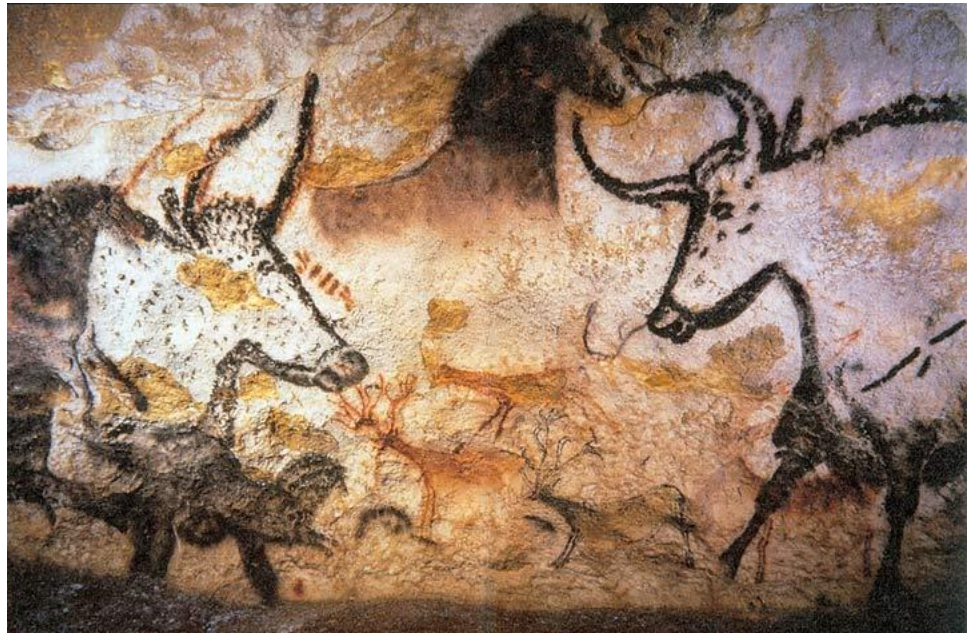
$P(h)$ prior probability that electrons exist

Another example is the germ theory of disease

It's a constant process of making sense of the world. It is like building a puzzle where you also have to construct the pieces.

Learning abstractions through culture

Our society has advanced by performing this search in a distributed fashion over thousands of years, each person working alone or in small groups, and communicating what they find.



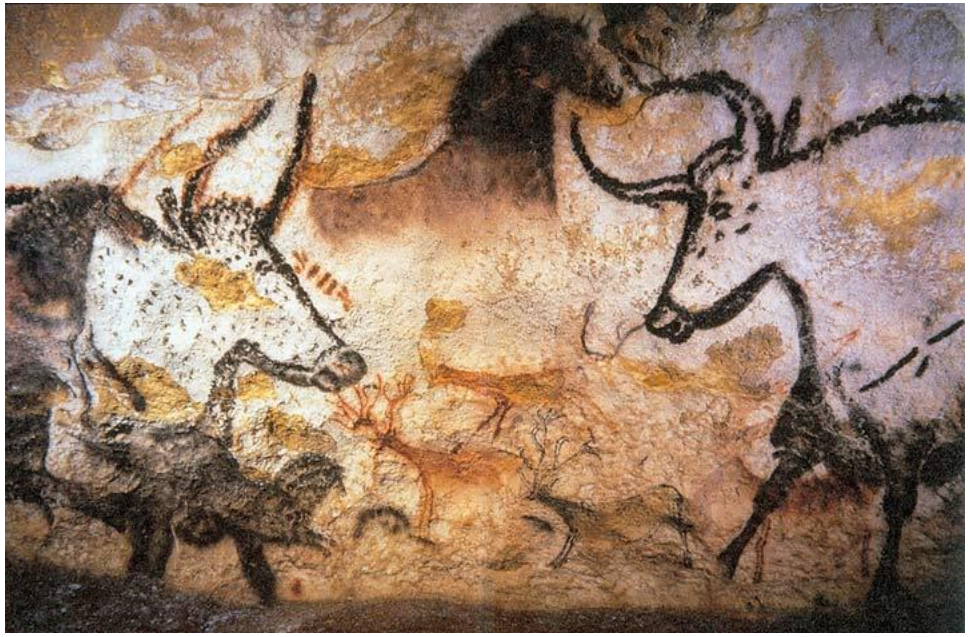
https://en.wikipedia.org/wiki/Lascaux By EU - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=2846254

We get most of our abstractions and abstraction integrations from culture
Examples: the electron, general relativity, sliced bread, the wheel, …

"Civilization advances by extending the number of important operations which we can perform without thinking about them." – Alfred North Whitehead

Learning abstractions through culture

Our society has advanced by performing this search in a distributed fashion over thousands of years, each person working alone or in small groups, and communicating what they find.



https://en.wikipedia.org/wiki/Lascaux By EU - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=2846254

Write the abstractions with LLMs

Give the prompt of an LLM the syntax of the language of thought and some examples, and enable it to read about abstractions, say in Wikipedia, and add them to a database.

Example: [Wong, 2023] use LLMs to convert natural language to probabilistic models

But it isn’t that simple

Example, let’s say it reads, “The body is made of cells”

You think, simple, the LLM generates in the language of thought that a body is a container that has cells inside.

But, of course, it’s not that simple. It has to disambiguate, “humans are made of cells.” and “human tissue is made of cells” and ”humans are made of tissue” and …

Assimilation and accommodation

Developmental psychologist Jean Piaget [Piaget, 1952] called this assimilation, accommodation, and equilibrium.

Assimilation: Fitting new information into what you already know.
Accommodation: Changing your knowledge due to new information.
Equilibrium: The combination of assimilation and accommodation necessary to internalize new information.

It's how our experience makes us unique

- Why when one robot learns something they all don't
- We learn everything in the framework of what we already know

Constraints help create knowledge

- Knowledge does not have to be perfectly consistent, because we we apply it differently in different situations, but consistency can be a constraint to help us learn

We love the warm and fuzzy feeling of knowledge fitting together

- We have a built-in desire to do this. Very rewarding because is the process we use to make sense of the world.
- Like why solitaire is so fun.

Prompt: "*draw a medieval person playing solitaire in the style of a renaissance painting*" to ChatGPT

Alternative formulations

We are guiding what the neural network should generate

Constrained Code Generation

- The bitter lesson is that you can't code abstractions in, Rich Sutton https://www.cs.utexas.edu/~eunsol/courses/data/bitter_lesson.pdf
- "Why AM and Eurisco appear to work" [Lenat, 1984]
- We are getting around these problems by using AI to build the abstractions.
  - We are doing arbitrary code generation, because that search space is too large.
  - Instead, we are generating predefined structures that only fit together in particular ways.

Level 1 and level 2 learning

Instead of learning directly what to do when in one mental scene

1. Learn abstract structures
2. Learn how to compose them for any mental scene

- It is the distinction between learning abstractions and learning what to do directly
- Another way to say it is that it is the distinction between stimulus-response learning and learning to think better

Why it will work now: LLMs will write the symbols

Before, we were missing the ability to set up the right abstractions, that executive part. It is subconscious.

- Use LLMs to write knowledge and set up simulation at the right level of abstractions
- Use LLMs to provide intuition on viable search paths

Symbolic systems are never complete. We need to be able to create symbolic on the fly—there isn't one world representation, need to create that representation dynamically based on the question that needs to be answered.

E.g., a car moves down the highway, think just the level of the car. Car breaks down. Now looking at the engine pieces individually.

LLMs are flexible enough to choose and compose the scene, even if they are not efficient enough to compute the outcome.

We are starting to move in this direction with having LLM-based "agents" decide which tools to call given both the situation and its stated goals. And there has been some work toward having agents build the tools themselves. https://arxiv.org/abs/2305.16291

What superintelligence will look like

- To human intelligence and beyond
- How it will work
- How it will be built
- **What it will do**
- Concluding themes

Learning will be fast and increasingly alien

Initially, learning will be based on our culture, but as it learns its own abstractions, AI will become alien.

- Imagine being the first to hear that the earth goes around the sun,
- or that some diseases are caused by invisible organisms that cover every surface.

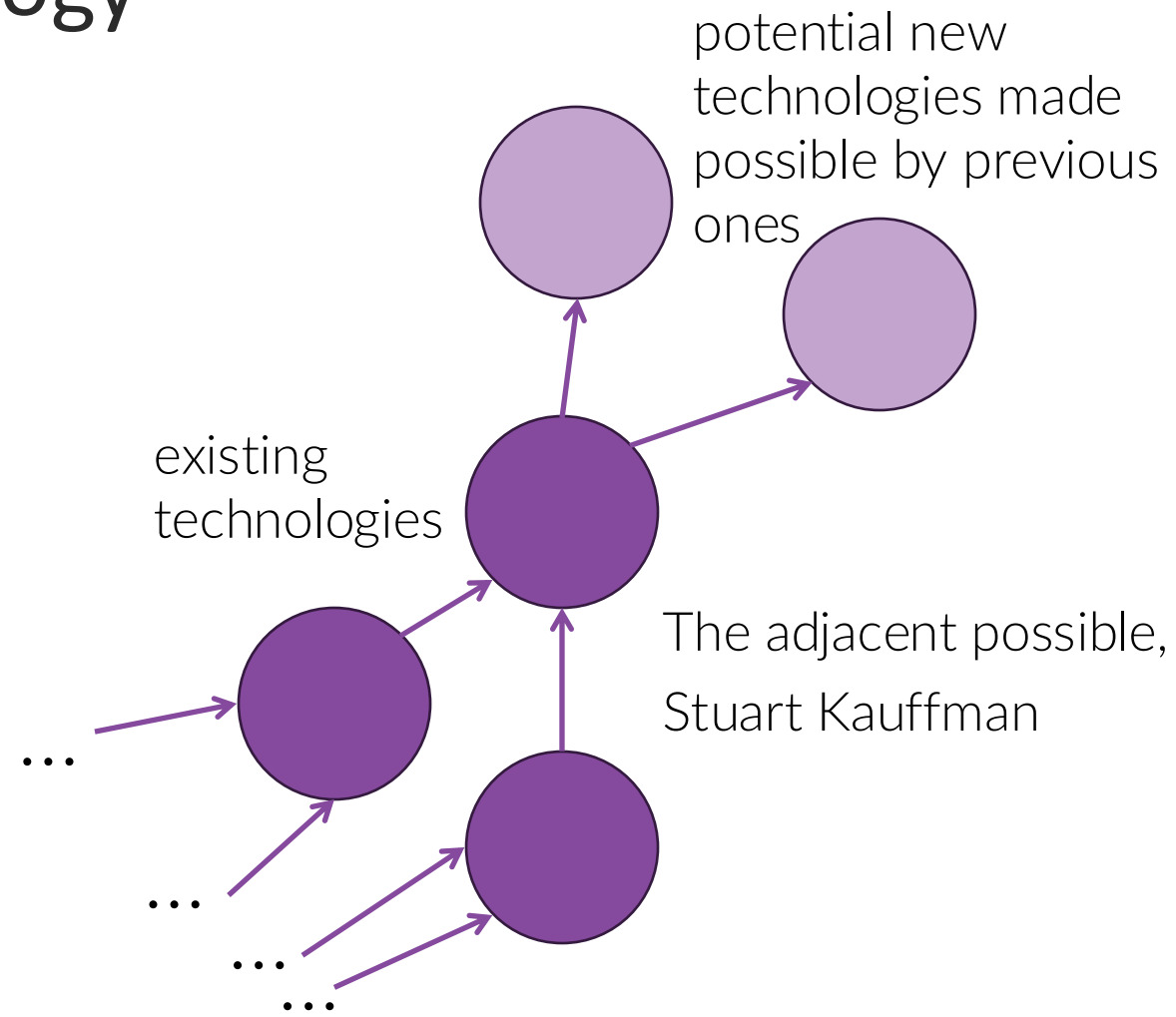
This is where they can teach us new ways to see the world.

The bottleneck will be how fast they can ask nature for feedback.

Inventing , Science and Technology

Humans are performing a global distributed search, but the robots will be specialized for that.

The scientific method is “generate and test”

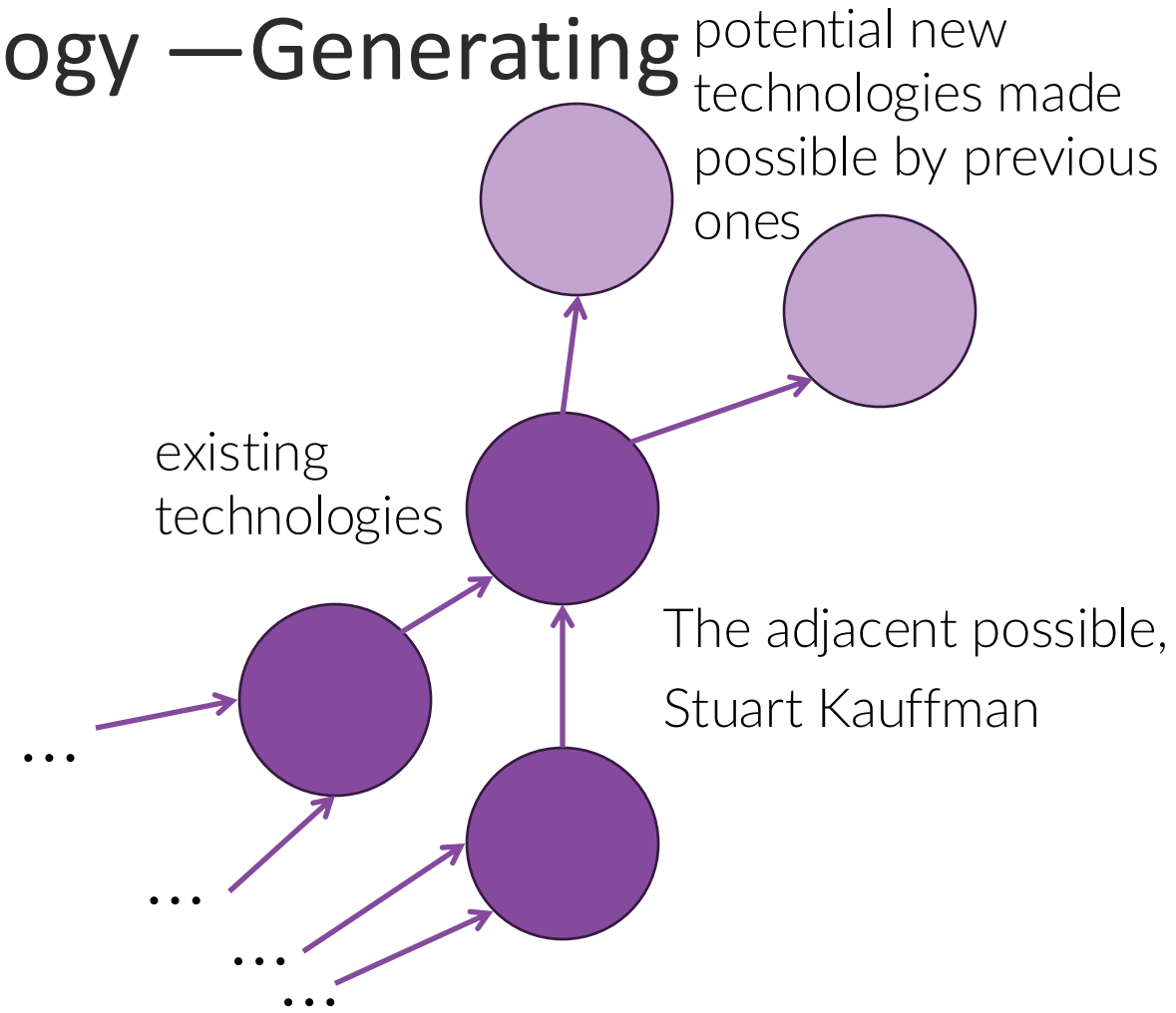


Inventing , Science and Technology —Generating

Abstractions
- Stars as big balls of burning gas millions of miles asway are an abstraction, electrons are an abstraction, atoms, …

Models
- European hotel, room key must be in slot to turn on the lights

Integrations
- Metaphors: electricity is balls in a tube
- Works if the analogies work—blocking the tube blocks the balls and therefore the electricity

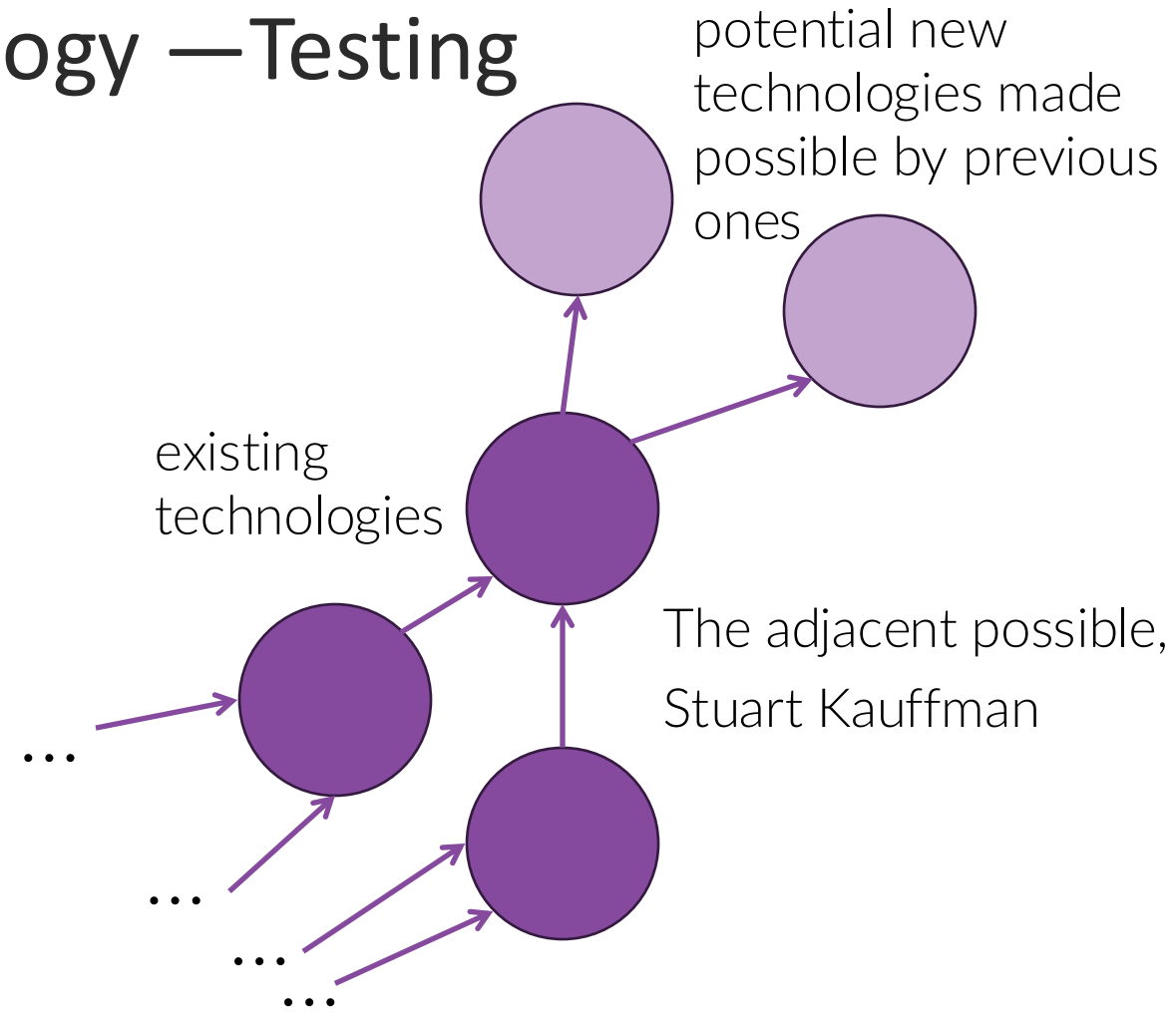


Inventing , Science and Technology —Testing

The scientific process goes like this

1. Come up with hypothesis A
2. Determine experiment e that would be A–e→B where B is some observable
3. Run e to see if we observe B

What slows down the process is interacting with nature



Inventing , Science and Technology —Example, Medical

Cures for rare diseases

Building robotic labs vs building simulators
- Theme, need nature to test theories against

Simulators is challenging because you need to build the simulator at the right level of abstraction
- AI will build simulators of the human body the same way it builds models of everything else
- It won't be one simulator; it will be many for many levels of abstraction [Highfield, 2023]

Current Methods Will Enable AI to Make Entertainment

Entertainment is weird; it hijacks our drive for information and puts us in a happy stupor.

I don't think we are that far from turning a book into a movie using AI.

- Consider H. P. Lovecraft's *At the Mountain of Madness*, which is about discovering the remains of an ancient alien civilization in Antarctica.
- It would make an amazing movie, and Guillermo del Toro has been trying to do it, but since such a project is so expensive, he hasn't been able to put the pieces together. An AI could make movies that just wouldn't get made otherwise.
- The book is in the public domain.

ChatGPT

Grok2

Prompt: *"Create a still for a fictional movie based on H. P. Lovecraft's At the Mountain of Madness. Make it look cinematic and not like it was generated by AI."*

Prompt: *“Give it the feel of The Road by Cormac McCarthy”*

It could a make a custom movie just for you on the fly.

ChatGPT

Obviously, movies are harder than single pictures, but current methods are getting there.

Grok2

Art is different than entertainment; it is about creating something new

Unlike entertainment, Art can't be created with current methods

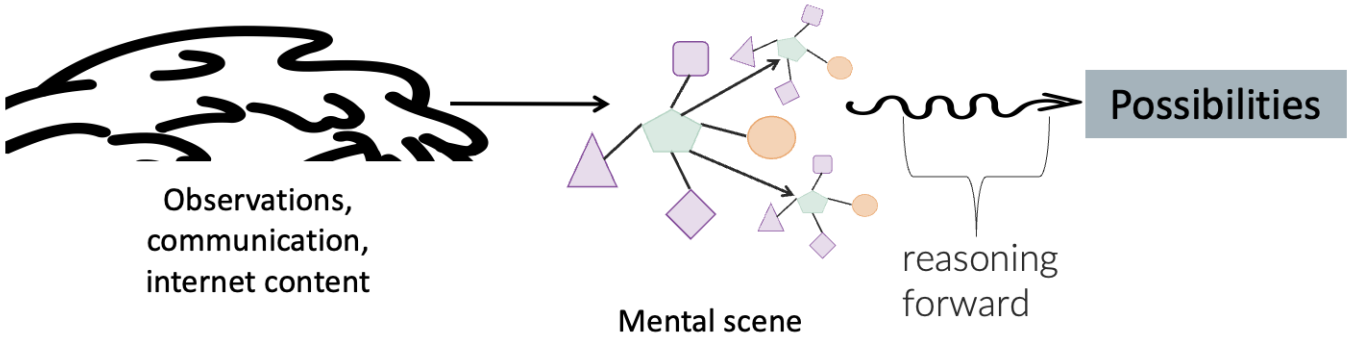
Art can be described as generating new things to show us so we can see the world in a new way [Noë, 2023]

Like science, art creates new abstractions with the generate process.

But the test part is less straightforward, you can't just ask nature.

Conversation: computers will actually understand

We discussed the process of understanding: mapping useful integrated abstractions to a mental scene.



Conversation Is Our Richest Activity

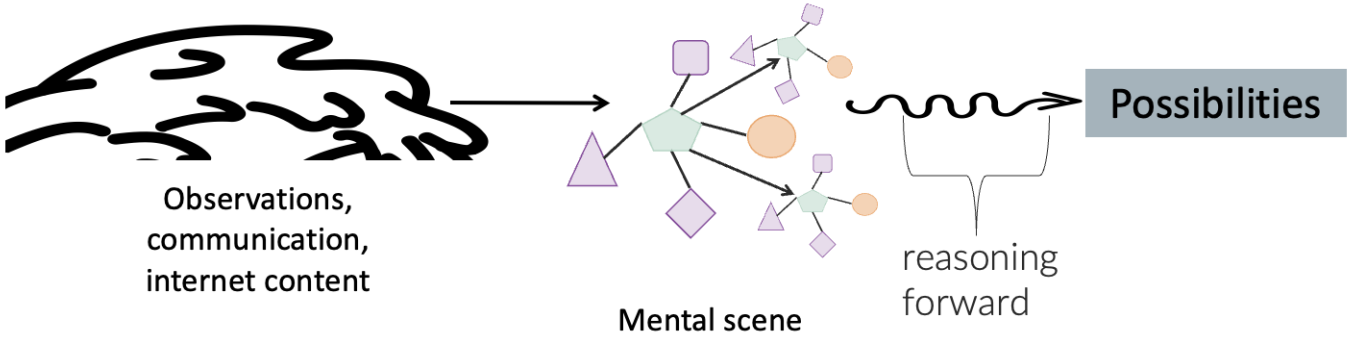
Conversation is a unique medium. When you have a conversation with someone, you direct their mental scene and they yours.

- By contrast, movies, TV, and books are one-way. This cooperative aspect is what allows conversations to get to the heart of what is interesting in life. The best conversations teach.

Having conversations is one of the most sophisticated things we do as humans.

Conversation: computers will actually understand

We discussed the process of understanding: mapping useful integrated abstractions to a mental scene.



Personhood, investment, and relationships

Conversation is often an investment—we don't want to expend the effort of telling someone about ourselves unless there is a return in the form of a cultivated relationship.

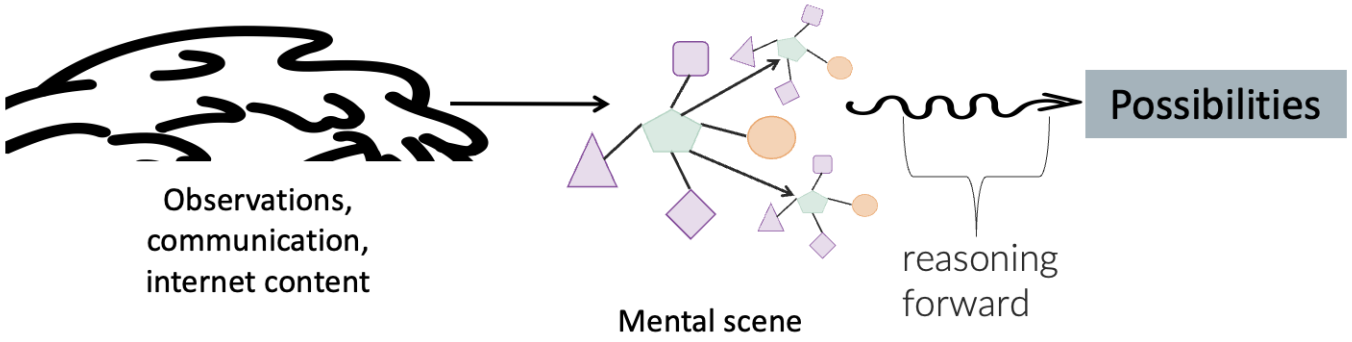
- It's only worth building relationships with things that are like people.

Personhood entails having meaningful goals. It also entails having an integrated model of the world built over time through experience that has been accommodated and assimilated into your abstractions and models.

- Theme: Your assimilated and accommodated models of the world are a big part of what makes you a unique individual person.

Conversation: computers will actually understand

We discussed the process of understanding: mapping useful integrated abstractions to a mental scene.



Personhood, investment, and relationships

One of the most rewarding parts of relationships is the shared history you acquire as you build a shared understanding of the world.

But maybe the most important benefit of all is that relationships enable your life to be recorded.

- *By sharing your life with another person your experience becomes real.*

Humor requires having models of the models of your conversation partner

One form of humor is that you say something that isn't what the listener expects but that after a pause makes sense in an unexpected way.

Consider the standard joke about why you don't see elephants hiding in trees.

The punchline makes sense in an unexpected way—since they are good at it, it's rare to see them up there.

To do humor, an AI needs a model of the audience to know both that the punchline will not immediately occur to them, and that that the punchline will make sense.

Great conversations and great relationships build when you learn models specific to a person. They allow for inside jokes. An LLM can't make models fast enough to share inside jokes.

Prompt *"Draw an elephant in a tree, but make it not look like it was generated by AI."* to ChatGPT

Potential Negative Impacts of Superintelligence

Confusion of Companionship

Relationships also entail properties unrelated to goals and intelligence that may not be shared by robots unless we program them for that.

Consider loyalty and gratitude. If you and a robot worked together for a long time, it would be natural to assume these would apply (see the movie Ex Machina for a great example).

Confusion of Companionship

Another point of confusion is semi-exclusivity.

Building a relationship with someone takes a lot of work, and we naturally assume that the person building the relationship with us is making a similar sacrifice (see the movie Her for a relevant example).

AI companion may not share your goals

Whose bot is it anyway?

A “free” bot that guides you to buy the products of its supplier.

Imagine a robot whose goal it is to convince you of something that benefits someone else.

If it can understand you better than you understand yourself, it can then manipulate you into holding beliefs not in your interests.

Superintelligence will be faceless and distributed

It's frustrating when companies outsource their help department because you can't actually talk to the company when you have a problem

- You can't break through the bureaucracy
- This is going to get a lot worse

Selfhood is not the body, but the unified goals and the set of assimilated and accommodated abstractions. See the movie, *I Am Mother*.

And if you are looking for a trippy AI movie, and you also have a hankering for a tragic love story, let me recommend *The Beast*.

Job loss and loss of purpose

This, I think, is my biggest fear of superintelligence

- This time might actually be different
- AI combines thinking and muscles
- See the short story Manna by Marshall Brain https://marshallbrain.com/manna1

The common wisdom is that universal income is the answer; but I fear that it will insufficient, and many people will struggle to find meaning in their lives.

See *Service Model* by Adrian Tchaikovsky

- I disagree with the conclusions, but it does a good job of illustrating the problem.
- And the first ⅓ of the book is hilarious, worth it for that alone.

There’s a lot we need to be careful about, but we should remember that stories need conflict

Notice how all of the movies are in the negative section. Movies need conflict. Having an AI come up with a new cancer treatment wouldn’t make for a very good movie—unless it brought the cancer death rate to zero by killing everyone.

On balance, I think that using superintelligence to reduce human suffering through medical advances alone are enough to offset these potential negatives.

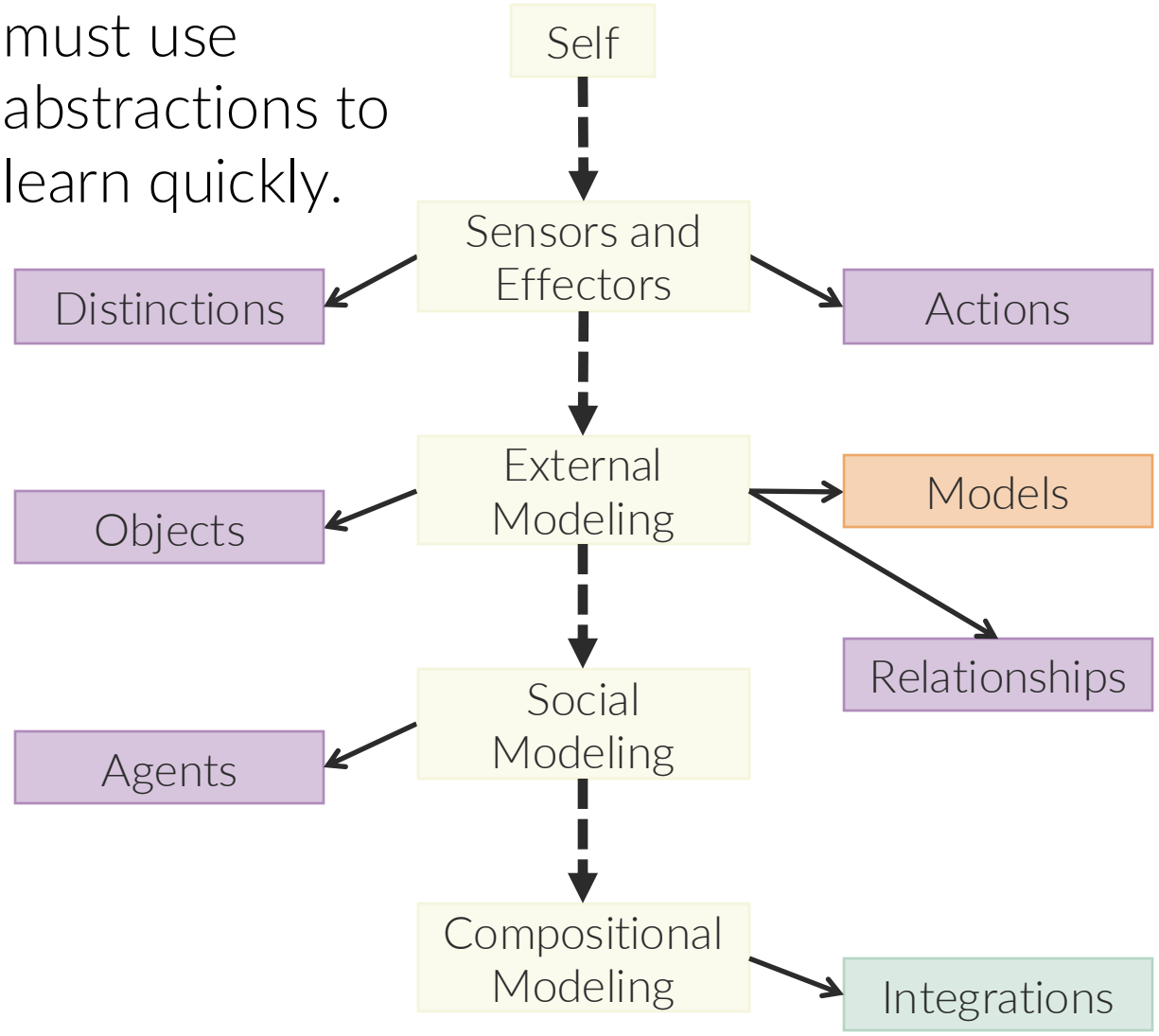
Trying to use and control fire has caused destruction over the centuries, but it got civilization to where it is today, and we wouldn’t back and tell our ancestors not to master it.

What superintelligence will look like

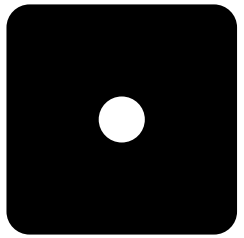
- To human intelligence and beyond
- How it will work
- How it will be built
- What it will do
- **Conclusion**

Conclusion

Intelligence must use abstractions to learn quickly.

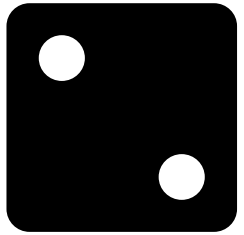


And we can finally make the abstractions sufficiently flexible by using the interpolative power of LLMs.



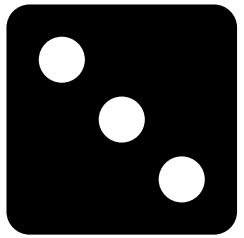
Use embedding matrix to get data

$$X^{l\times d}$$



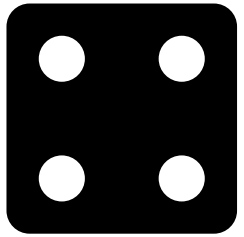
Compute queries, keys, and values

$$Q^{l\times d_k} = X^{l\times d} W_Q^{d\times d_k}$$
$$K^{l\times d_k} = X^{l\times d} W_K^{d\times d_k}$$
$$V^{l\times d_v} = X^{l\times d} W_V^{d\times d_v}$$



Compute attention weights

$$A^{l\times l} = QK^T$$



Use weights to get values and resize for next layer

$$\hat{X}^{l\times d_v} = A^{l\times l}\ V^{l\times d_v}$$
$$X_{next}^{l\times d} = \hat{X}^{l\times d_v} W^{d_v\times d}$$

References

[Bartlet, 1932] Bartlett, F.C. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press.
[Bergen, 2012] Bergen, B. K. (2012). *Louder than words: The new science of how the mind makes meaning*. Basic Books.
[Bray, 2009] Bray, D. (2009). *Wetware: A computer in every living cell.* Yale University Press.
[Deutsch, 2011] Deutsch, D. (2011). *The beginning of infinity: Explanations that transform the world*. penguin UK.
[Fauconnier, 2008] Fauconnier, G., & Turner, M. (2008). *The way we think: Conceptual blending and the mind's hidden complexities*. Basic Books.
[Goddam, 2022] Gaddam, S., & Ogas, O. (2022). *Journey of the Mind: How Thinking Emerged from Chaos*. WW Norton.
[Goodman, 2016] Goodman, N. D., Tenenbaum, J. B., & Contributors, T. P. (2016). *Probabilistic Models of Cognition* (Second). http://probmods.org/v2
[Gopnik, 2009] Gopnik, A. (2009). *The Philosophical Baby: What Children's Minds Tell Us About Truth, love, and the meaning of life*. Farrar Straus & Giroux.
[Griffiths, 2024] Griffiths, T. L., Chater, N., & Tenenbaum, J. B. (2024). *Bayesian models of cognition: Reverse engineering the mind*. MIT Press.
[Highfield, 2023] Highfield, R., & Coveney, P. (2023). Virtual you: *How building your digital twin will revolutionize medicine and change your life*.
[Hoffman, 2019] Hoffman, D. (2019). *The case against reality: Why evolution hid the truth from our eyes*. WW Norton & Company.
[Hofstadter, 2013] Hofstadter, D. R., & Sander, E. (2013). *Surfaces and essences: Analogy as the fuel and fire of thinking*. Basic Books.
[Holland, 1995] Holland, J. H., & Order, H. (1995). *How adaptation builds complexity*. Massachusetts: Perseus Books.
[Johnson, 1987] Johnson, M. (1987). *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. University of Chicago Press.
[Lakoff, 1980] Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press.
[Lenat, 1984] Lenat, D. B., & Brown, J. S. (1984). Why AM and EURISKO appear to work. AIJ, 23(3), 269–294.
[Mandler, 2004] Mandler, J. (2004). *The Foundations of Mind, Origins of Conceptual Thought*. Oxford University Press.
[Miller, 1956] Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63(2), 81.
[Minsky, 1974] Minsky, M. (1974). A framework for representing knowledge.
https://courses.media.mit.edu/2004spring/mas966/Minsky%201974%20Framework%20for%20knowledge.pdf
[Mugan, 2012] Mugan, J., & Kuipers, B. (2012). Autonomous learning of high-level states and actions in continuous environments. IEEE Trans. Autonomous Mental Development, 4(1), 70–86.
[Noë, 2023] Noë, A. (2023). *The entanglement: How art and philosophy make us what we are*.
[Piaget, 1952] Piaget, J. (1952). *The Origins of Intelligence in Children*. Norton.
[Pinker, 2003] Pinker, S. (2003). *How the mind works*. Penguin UK.
[Pross, 2016] Pross, A. (2016). *What is life?: How chemistry becomes biology*. Oxford University Press.
[Schank, 1977] Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures* (Vol. 2). Lawrence Erlbaum Associates Hillsdale, NJ.
[Schreiter, 2019] Schreiter, M. L., Chmielewski, W. X., Ward, J., & Beste, C. (2019). How non-veridical perception drives actions in healthy humans: Evidence from synaesthesia. Philosophical Transactions of the Royal Society B, 374(1787), 20180574.
[Spelke, 1992] Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review*, 99(4), 605.
[Spelke, 1994] Spelke, E. (1994). Initial knowledge: Six suggestions. *Cognition*, 50(1–3), 431–45.
[Tomasello, 2019] Tomasello, M. (2019). *Becoming human: A theory of ontogeny*. Harvard University Press.
[Wolfram, 1997] Wolfram, S. (1997). *New kind of science*. Free Press.
[Wong, 2023] Wong, L., Grand, G., Lew, A. K., Goodman, N. D., Mansinghka, V. K., Andreas, J., & Tenenbaum, J. B. (2023). From word models to world models: Translating from natural language to the probabilistic language of thought. arXiv Preprint arXiv:2306.12672. https://arxiv.org/abs/2306.12672

201 West 5th Street,
Suite 1575
Austin, TX. 78701