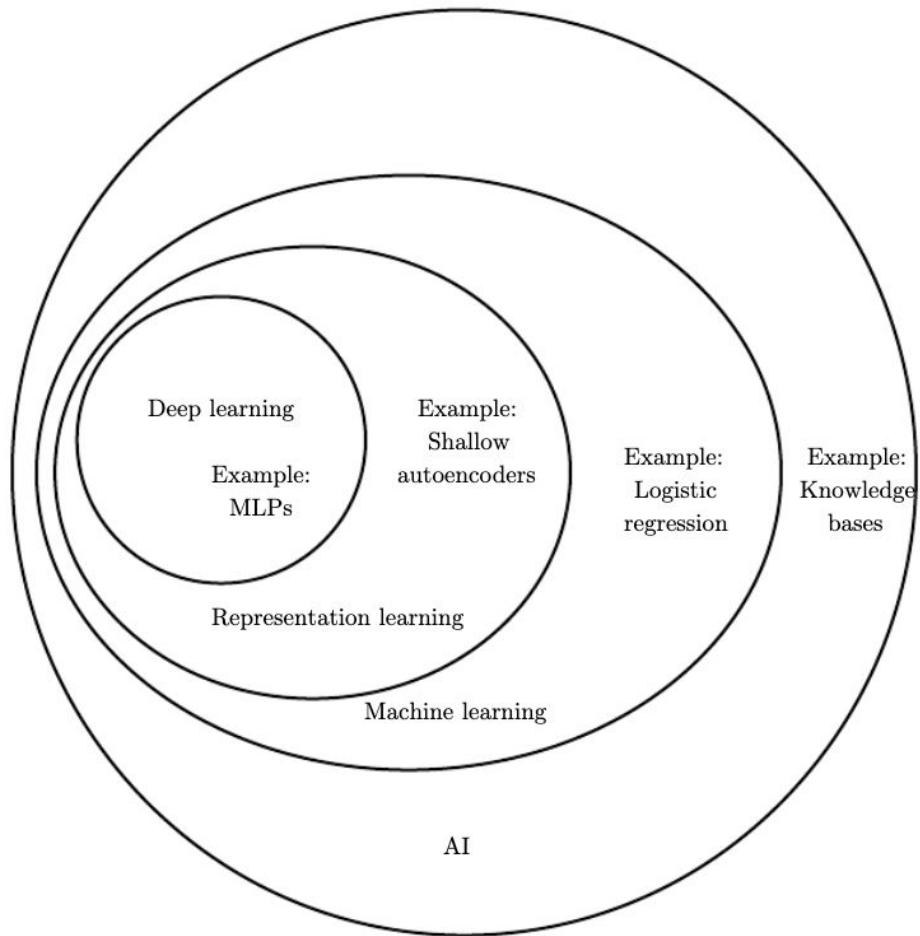


# Consciousness & AI

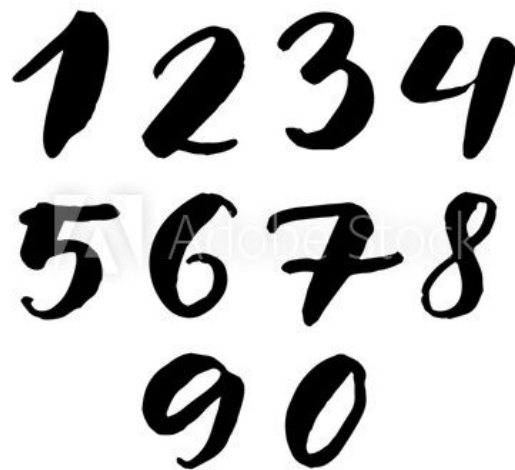
Final Class!

What is AI?



# The Classic Problem

- Imagine coding every possible configuration
- GOFAI



1 2 3 4  
5 6 7 8  
9 0

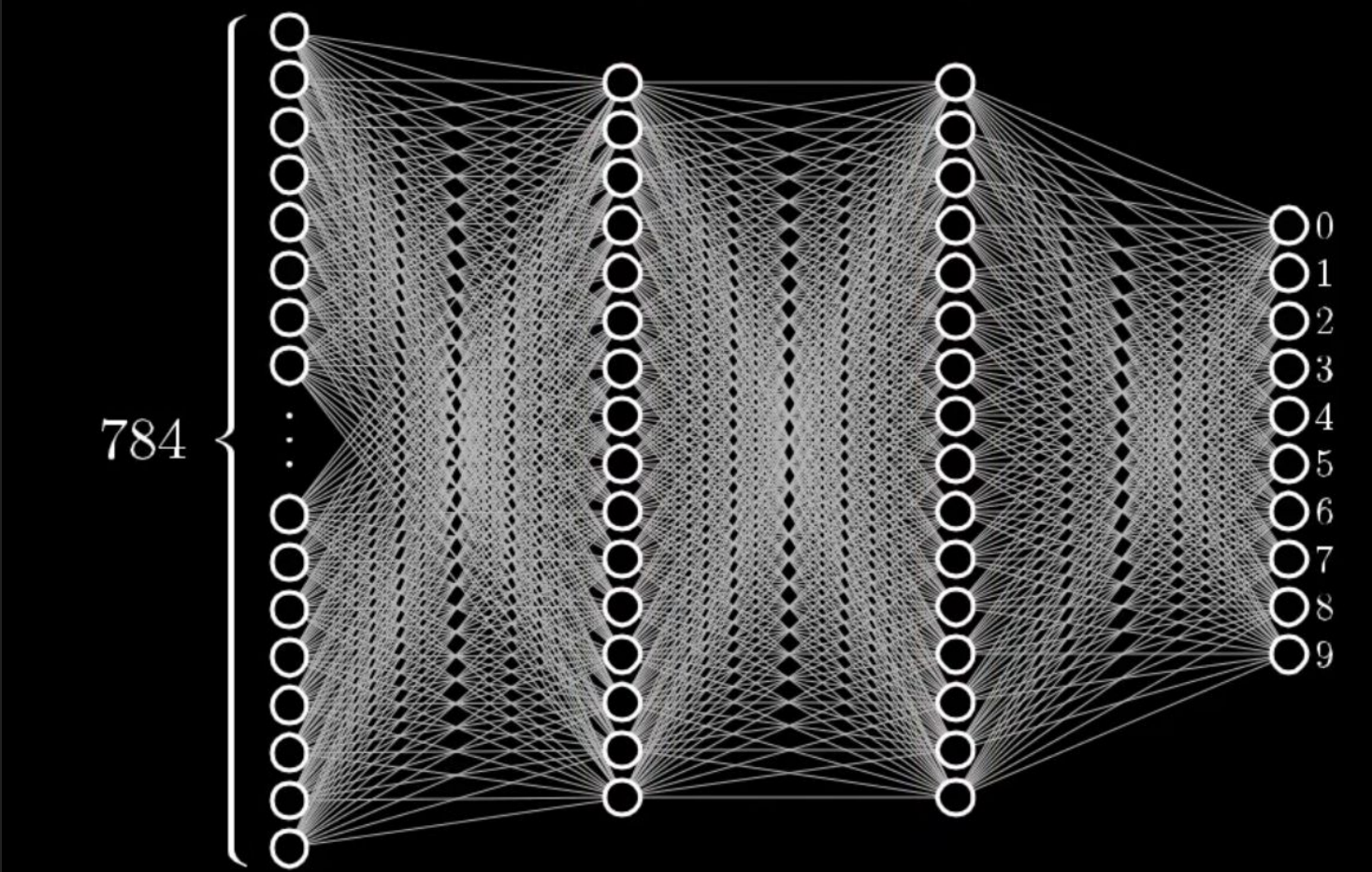
# Neural Networks

Inspired by human brains and the human learning process:

“Neurons that fire together wire together”

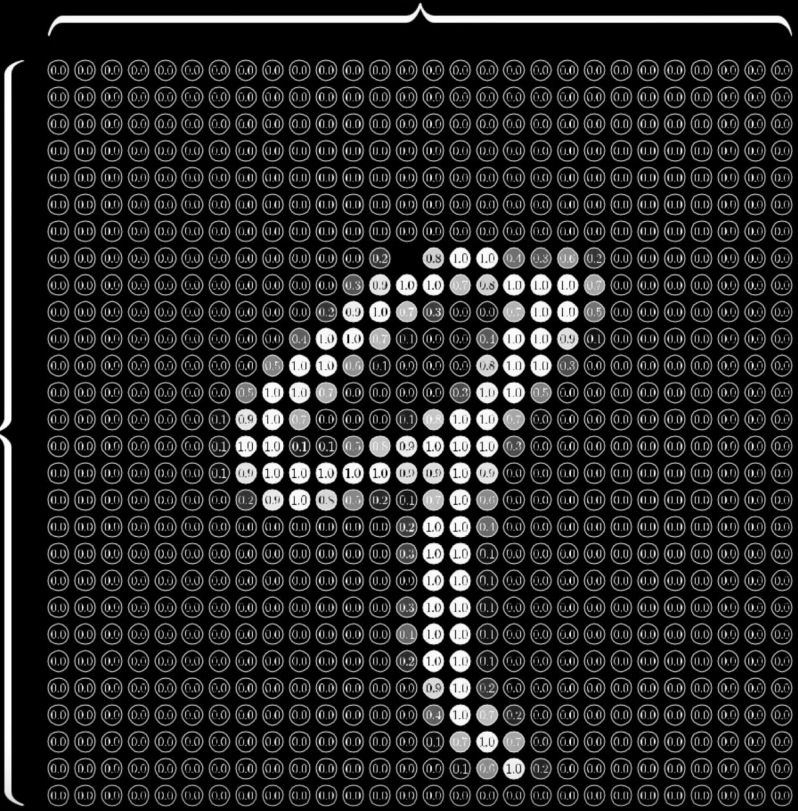
Take use of Hierarchies and hierarchical learning. Each layer supposedly is a different representation of the input. This mirrors what occurs in the human visual system (V1-V5). Each successive layer is encoding and representing more abstract information

Motivated by learning more complex non-linear functions of predictions (stock prediction, game prediction, human vision, cancer detection)



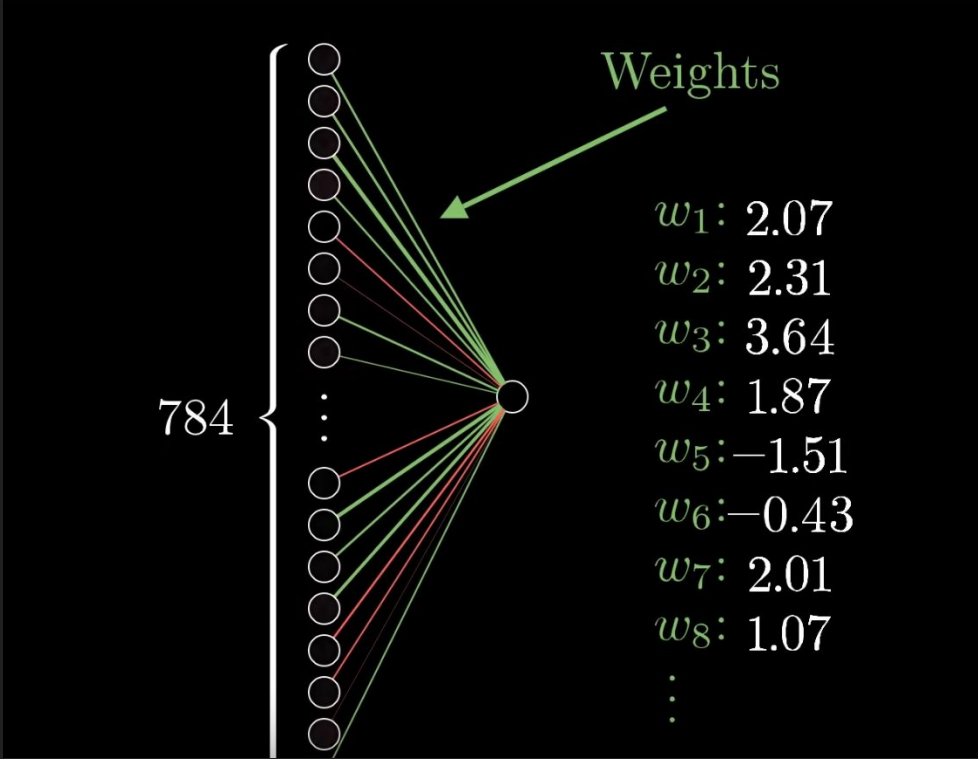
28

28



$$28 \times 28 = 784$$

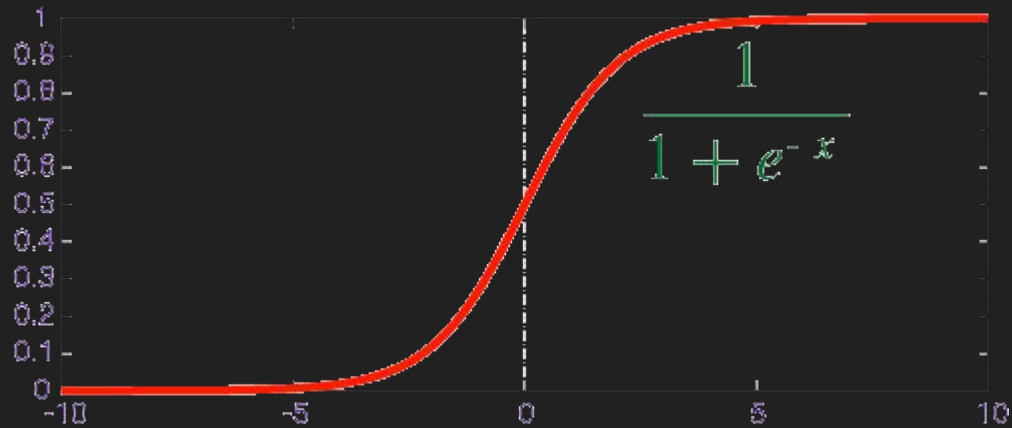
0.91





# Adding Non-linearity

Sigmoid (many others)



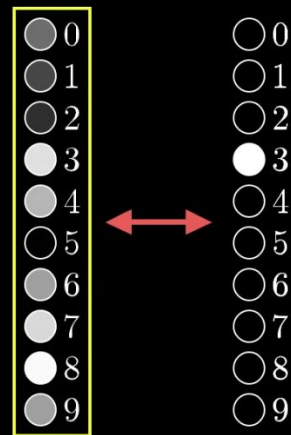
# How does it learn?

Cost of

3

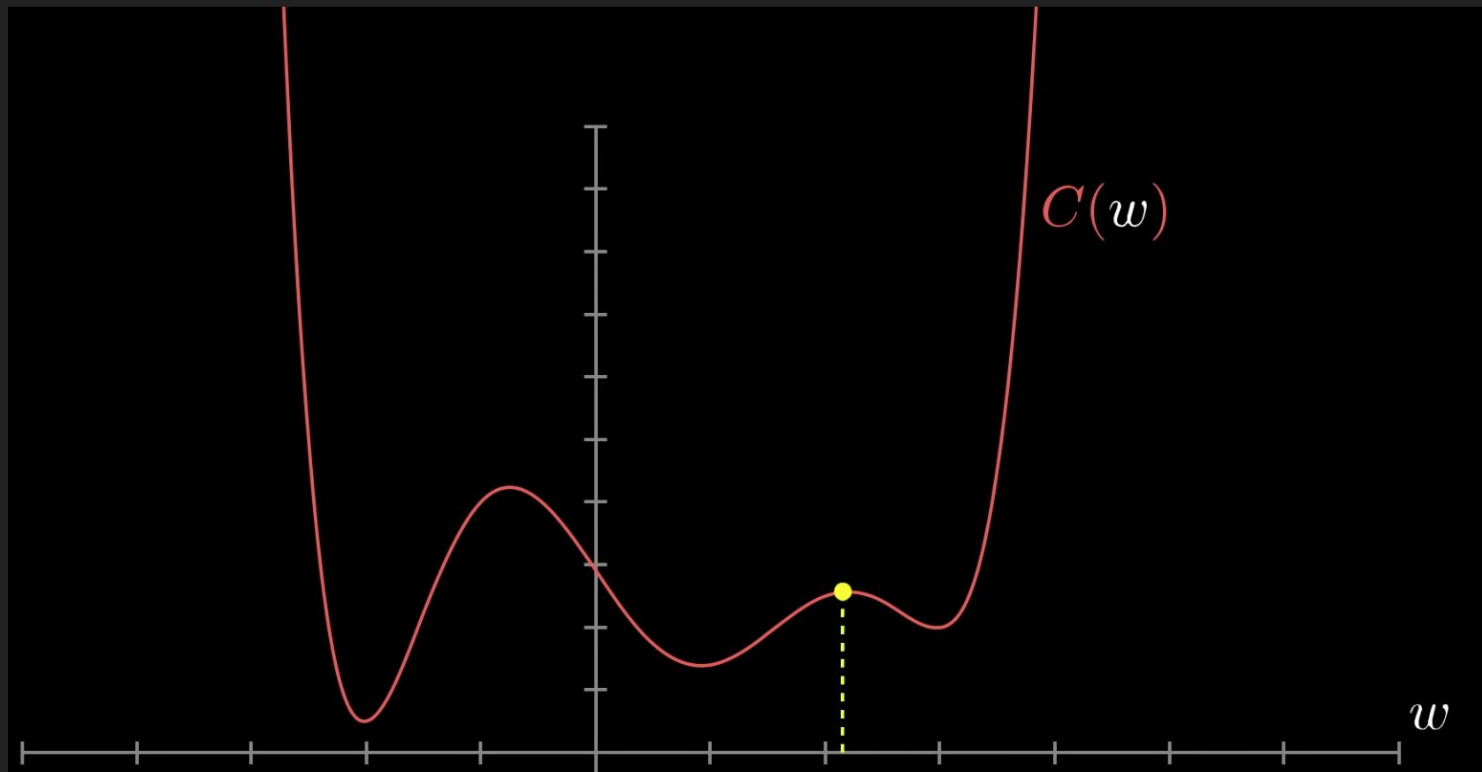
$$\left\{ \begin{array}{l} (0.43 - 0.00)^2 + \\ (0.28 - 0.00)^2 + \\ (0.19 - 0.00)^2 + \\ (0.88 - 1.00)^2 + \\ (0.72 - 0.00)^2 + \\ (0.01 - 0.00)^2 + \\ (0.64 - 0.00)^2 + \\ (0.86 - 0.00)^2 + \\ (0.99 - 0.00)^2 + \\ (0.63 - 0.00)^2 \end{array} \right.$$

What's the "cost" of this difference?

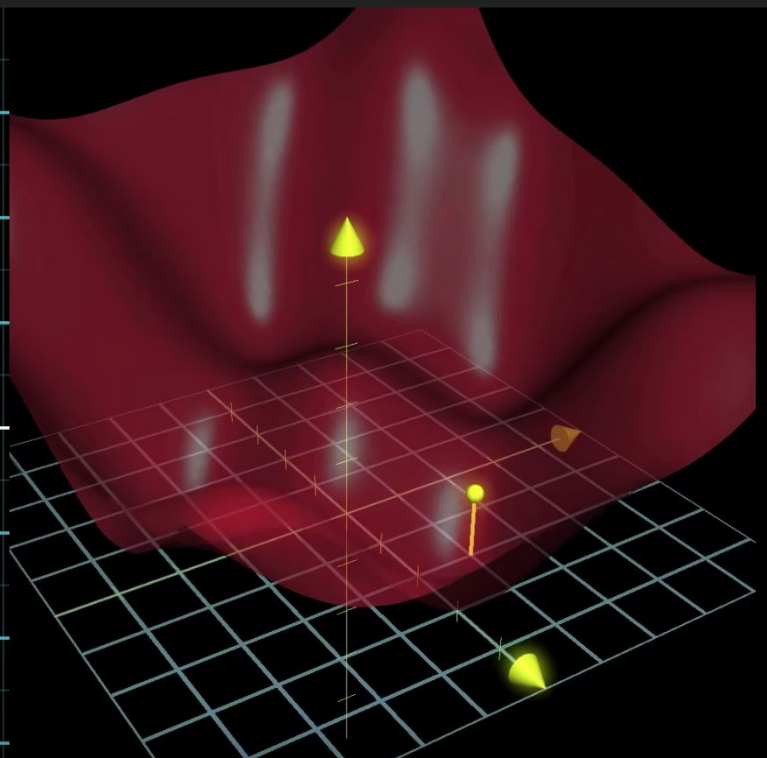
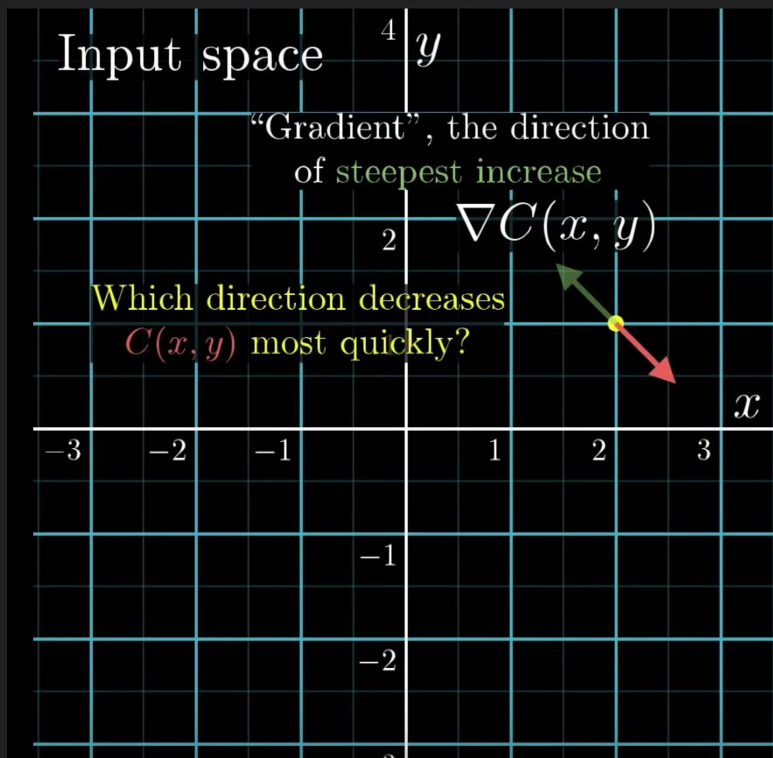


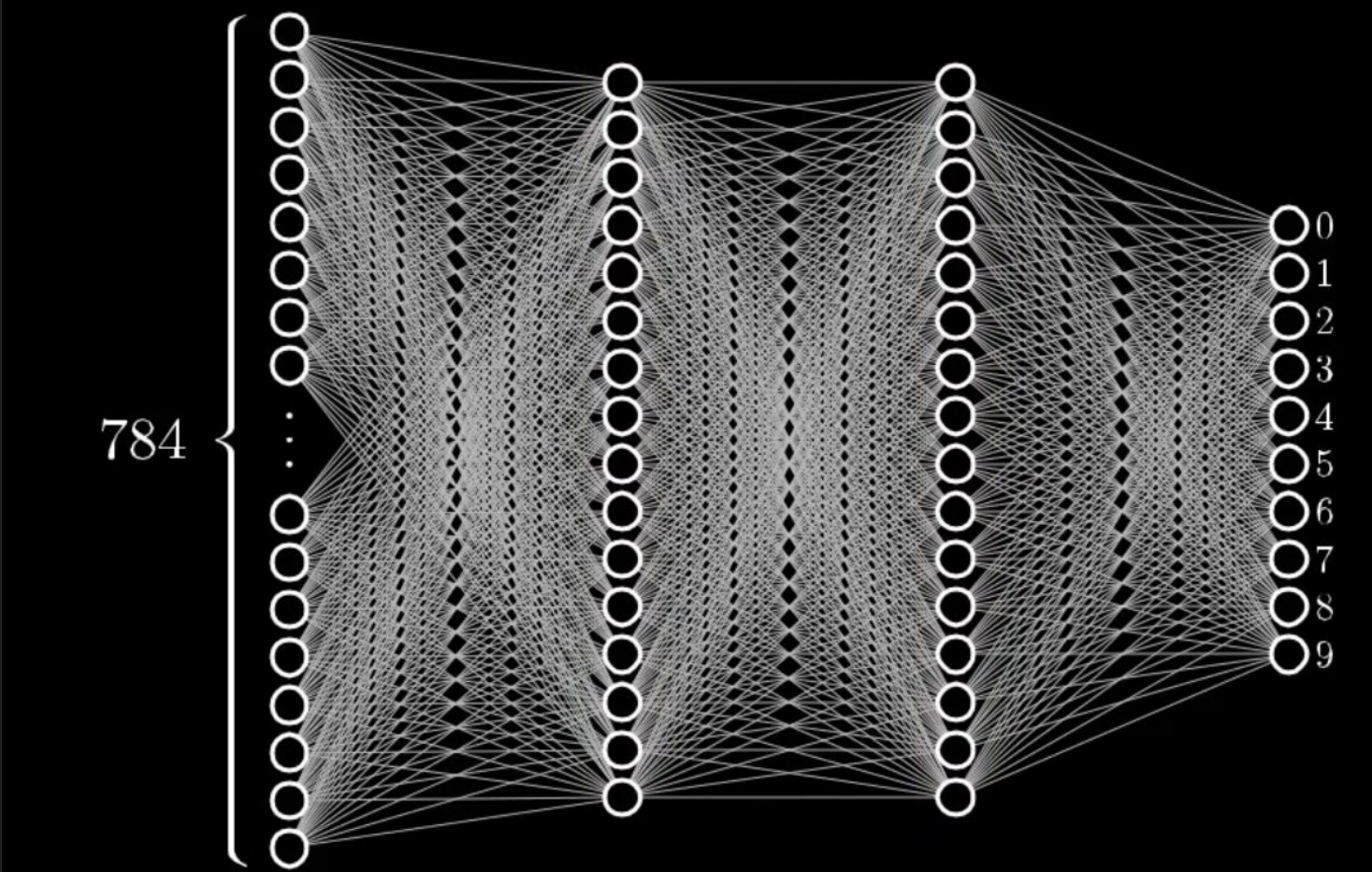
Utter trash

# Brief Calculus



# Gradient Descent





Strong, Weak, AGI & AI

# First Intuitions

- Network continues to act as a function (albeit a more powerful one)
- It does not know when it is wrong, spits each answer with equivalent certainty
- Does it notice its noticing?
- Networks have higher level representations but are they “accessed” by anything? What would be doing the accessing? (Ned Block)
- Would phenomenal consciousness be aiding in accomplishing higher level tasks or is it actually the opposite and just getting in the way? Is it epiphenomenal?
- Seems to lack attention and attentional control. How would it decide how to dedicate its resources

Pause here (expand Zoom)



# Searle and his Chinese Room

# Turing Test

# Overview

Searle is not talking about Deep networks or neural networks, but computer programs in general (which are also dictated by syntactic and formal rules)

The argument is directed at the view that formal computations on symbols can produce thought

He says that programs are purely syntactic; they follow rules, etc and because we can follow a set of rules for programs without understanding the computer itself would lack understanding

Syntax is NOT semantics

# Main Objections

1. Maybe Searle does not understand Chinese, but perhaps there is something else that does
2. We could tweak the program, or run it in a way that would then lead to understanding
3. Our intuitions that Searle does not understand Chinese are wrong; he actually does (or does in the way that matters)

# Systems Reply

Searle does not understand Chinese but the entire system does. Searle is only the CPU

Searle's Reply: he could internalize all of the rules and formal logic of the computer and put it in the brain (he would be the entire system) and he still would not understand Chinese

Reply: [Searle's] failure to understand Chinese is irrelevant: he is just the implementer. The larger system implemented would understand – there is a level-of-description fallacy. Can also argue that even if the Searle person internalized all of the rules and stuff, he himself does not become the system. Instead, it would be like multiple personality disorder with multiple people.

# Tim

I've always thought the System's Reply was an effective response to Searle's Chinese Room. But not just any system will do. Searle was dealing with very primitive chat-bots when he developed the argument, so it is not surprising he thought them stupid since they were mostly just using language tricks to answer questions or were dealing with very limited domains. And I'd agree with him that just passing the Turing Test is not enough to say a system understands. But people now are building systems that are much more sophisticated. For example, take a look at this video from David Ferruci, who was the lead researcher on Watson and has moved on to start a company called Elemental Cognition:

<https://m.youtube.com/watch?v=vsyPZdt6noE> (Links to an external site.) . See especially at 15:56 where he describes how the system creates a "mental" model of a soccer game in order to track what is happening in the story. I think if the Chinese Room was running a system like this, it is much harder to say the system as a whole doesn't understand Chinese...

# Tim Con't

And I think similar considerations apply to his objection that computers are only intelligent in an “observer-relative” sense. It isn't just in an observer-relative sense that the robot destroyed Tokyo, and, similarly, it isn't just in an observer-relative sense that Deep Blue won the chess match. Of course, the programmer arranged the 0's and 1's in such a way that when it spit out a move, like P-K4, someone (or a robotic arm) would move Pawn to King 4, but this, in principle, is no different than my thinking “Pawn to King 4” results in sending a signal to my arm to move the pawn. If reference and semantics depends on our causal relations with the world, as many believe, then it is all there.

On the other hand, all that said, I don't think the system would be conscious. In my mind, this is a very different kind of thing: Chalmers' hard problem versus an “easy” one.

# Virtual Mind Reply

There could be an agent, that is neither the human nor the entire system, that understands. A third party. The running computer itself would be the thing that “understands”

Apple SIRI: not identical with the CPU, nor the entire system

Just because Searle, acting as a computer processor, does not understand chinese, does not mean that no understanding is being created. For example, there are thoughts, beliefs, and a personality that come out in an all inclusive Turing test (an intense investigation). In that way

Ex: Korean rules as well as Chinese; how can two people be identical to Searle?



# Robot Reply

We attach meaning to things through seeing and doing, so yes, the digital computer would not understand language, but if we added sensors and effectors, so that it could interact, then there would be some understanding of language there

-Minds must be embodied it makes no sense to have a “brain in a vat”

Searle’s response: we can have it so that all of the input/additional sensory information coming from these new sensors/effectors are just more input/work for the man in the room

# Robot Reply (cont)

A computer might have propositional attitudes if it has the right causal connections to the world – but those are not ones mediated by a man sitting in the head of the robot. We don't know what the right causal connections are. Searle commits the fallacy of inferring from “the little man is not the right causal connection” to conclude that no causal linkage would succeed. There is considerable empirical evidence that mental processes involve “manipulation of symbols”; Searle gives us no alternative explanation

-there need to be causal connections, just because one of the causal connections cannot give rise to consciousness does not mean NO causal connections exist

Functionalists aren't behaviorists: they care about how things are created

# Other Minds Reply

You really only know that other people understand Chinese based on their behavior. If this computer

Searles reply is really just to say that we have extra knowledge (biological processes, centuries of evolution, first-hand knowledge) that these cognitive states are conscious, and also first-hand knowledge that the way in which these other cognitive states are being created is quite different

If we asked a robot, what is your subjective experience like, what would it say?  
How would it introspect?

# Intuition Reply

What gives Searle's argument its "force" is how our belief that the man in the room does not understand Chinese relies on our intuitions that this is the case

Aliens, anatomically quite unlike humans, cannot believe that humans think when they discover that our heads are filled with meat. The Aliens' intuitions are unreliable – presumably ours may be so as well. - Pinker's story

Because the speed at which the Chinese room would operate would be so excessively slow, this the reason why we struggle to attribute intelligence to it, because we are accustomed to seeing "fast" thinkers". But it is possible to speed it up so that our intuitions DO align --Dennett

# Discussion

- What would it mean for a system to “understand”?
- What would be the sort of thing a computer would need to have to be confident that it is conscious?
  - Attention, modulating attention, goals, a notion of self, causality
- Even if it had all of that, is it missing “qualia?”
- How would we create “awareness”? Remember, this is deeper than the MPE that we talked about earlier. Would a computer “notice that it notices”?
- A lot of the discussion around the Chinese room seems to be caught up in also what level of language capabilities it has. Such a linguistically capable agent, to engage in full-blown conversations, not just about weather and plans, but about life personalities, goals, and abstract thought might necessarily be conscious.

# Discussion

1. Can we draw a distinction between an agent that understands and one that is conscious? Can one type of agent exist without the other?
2. Searle makes the claim that computation is abstract; it exists in consciousness as an ontological subjective phenomenon with no intrinsic properties. It is the special causal relationships and connections in our brains that give rise to consciousness, not just the computations that we can simulate. What support does he have for this claim?

Should We Even Try?

# Robots Should Be Slaves

Robots should serve us and our needs

Robot is defined as any agent that can turn perception to action, including digital agents which can share information across databases, send emails, communicate on behalf, hold secrets, etc

What should our relationship be with robots?

What is the correct relationship? What are robots?

What is the appropriate relationship?



# How Bryson sees robots

They are not persons.

Even though they are not persons they can be servants

Servants are a good thing, as long we are not dehumanising anything

# The cost of being identified with robots

If we think they are more conscious than they are (tamagatchis) then we can cause harm to ourselves and others at the cost of benefitting a “non-agent”

My arguments in this chapter derive primarily from the default liberal-progressive belief that the time and attention of any human being is a precious commodity that should not be wasted on something of no consequence

Humans have limited time on earth to socialize and form relationships, and much of that time is now being spent on interacting with these useless agents

There can be an even greater cost for the over-identification with AI as an entire species, or at the national level:

If we have AI making legal and military decisions, who can we blame if things go wrong?

# What should things be?

We should NOT built anything that “understands”

We should also communicate as much as possible to other humans that these machines are just machines, and we should invest and treat them in a way that is appropriate to machines

# Discussion

1. Bryson thinks that just because we manufacture and design them, means that we own them. But there is at least a possibility we could create conscious AI
2. What would that even look like? What would be the criterion for which a robot would be something that had moral agency?
3. Can we ascribe moral responsibility to something that isn't conscious? Dogs that murder humans and then are sentenced
4. Even if they had conscious experience, what sort of conscious experience would it be wrong to enslave them?

Closing thoughts on Ned Block