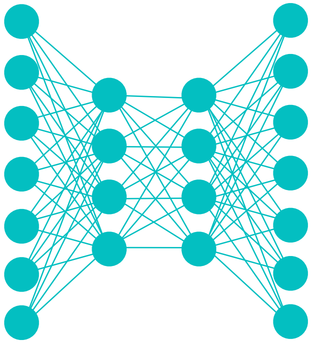
Lecture Notes for

Deep Learning II

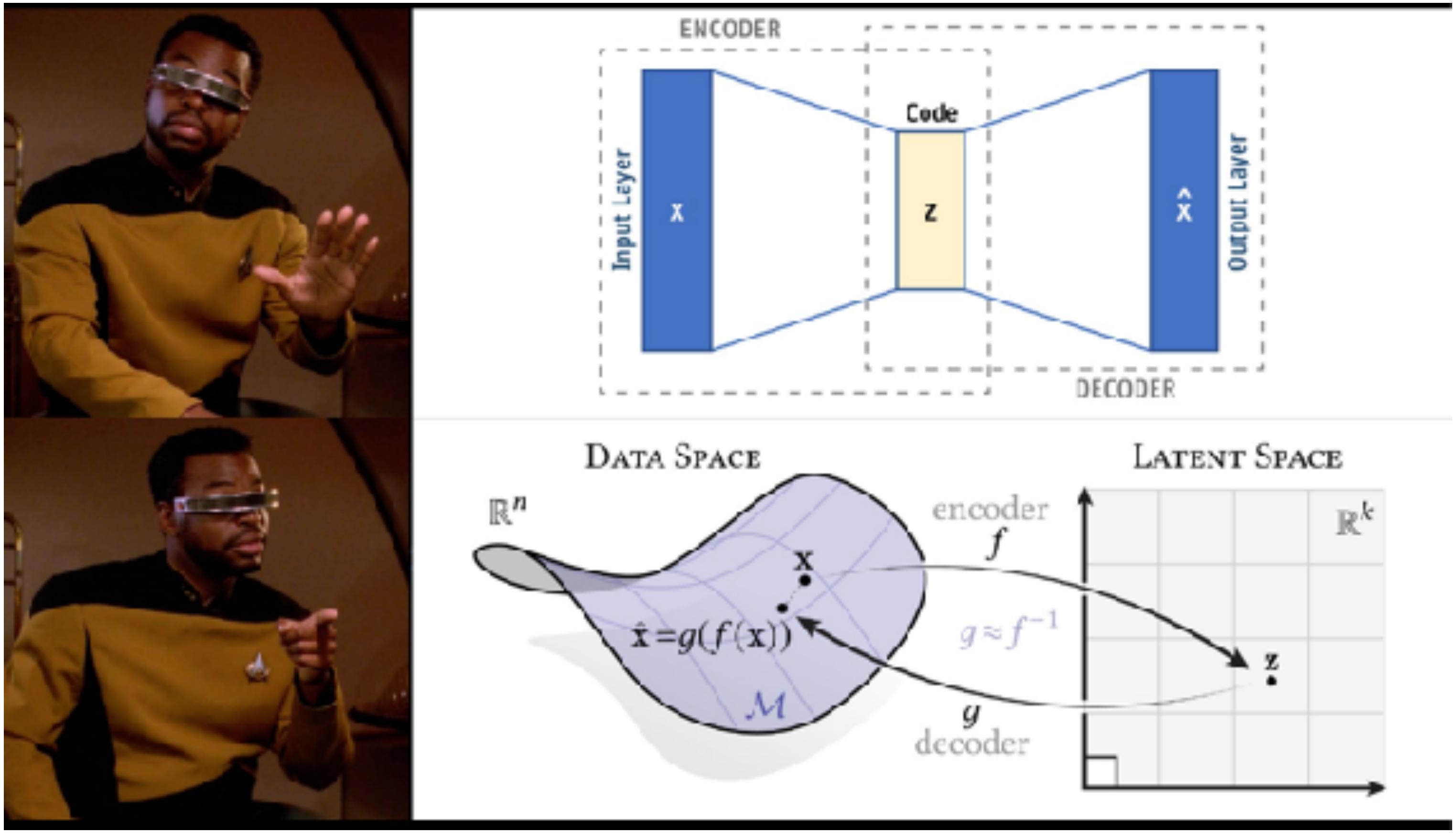
The Ethical AI Principles and Case Studies



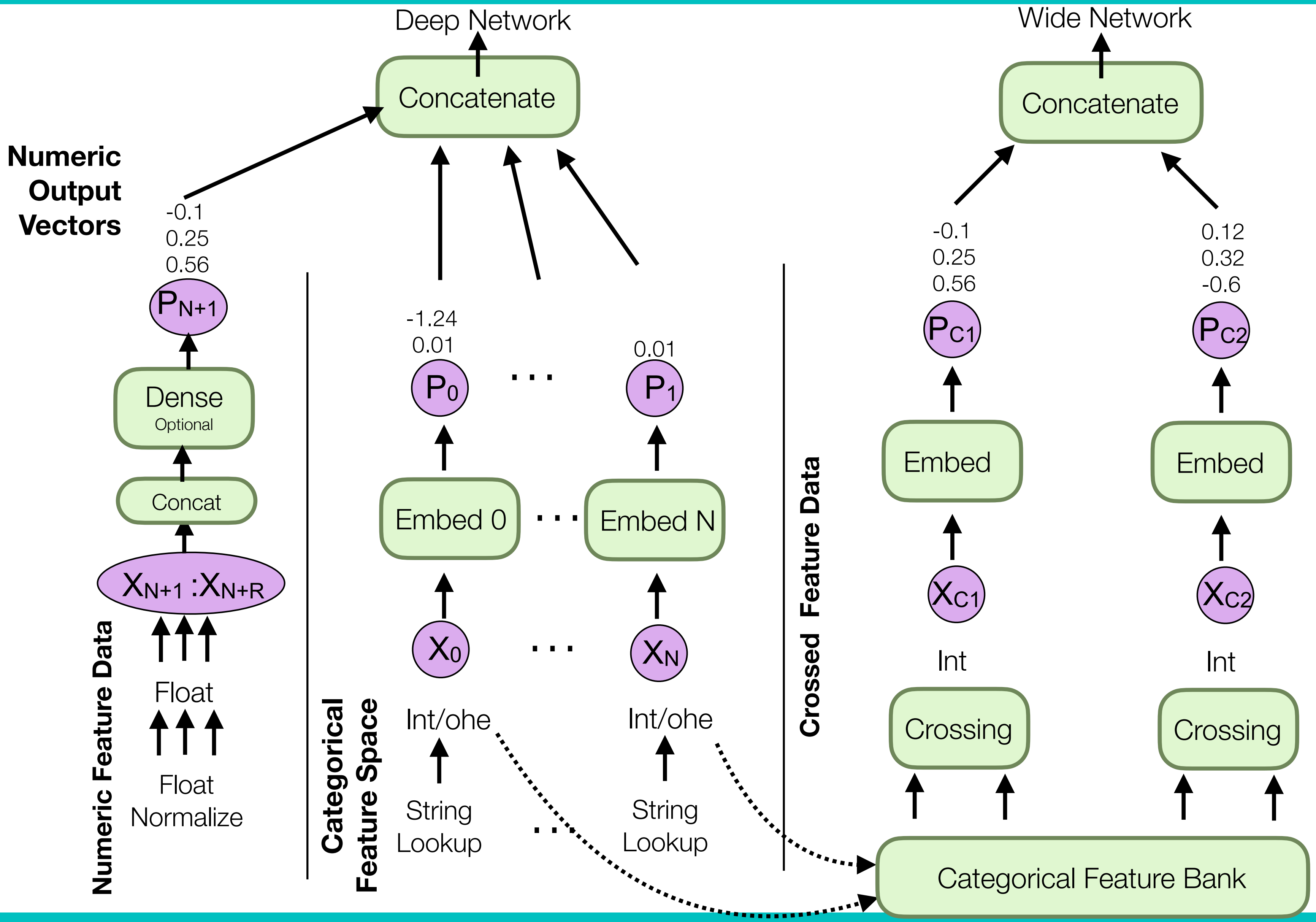
Logistics and Agenda

- Logistics
  - Panopto and course videos
  - No lecture next time
  - First Student Presentation next lecture (1 week) to start lecture
  - Student Presentations
    - Still need responses, for partial credit!
- Last Time:
  - Course Introduction
- Agenda
  - The AI Principles and Fairness measures
  - Case Studies and Discussion
    - Applying the Principles

Review, Continued

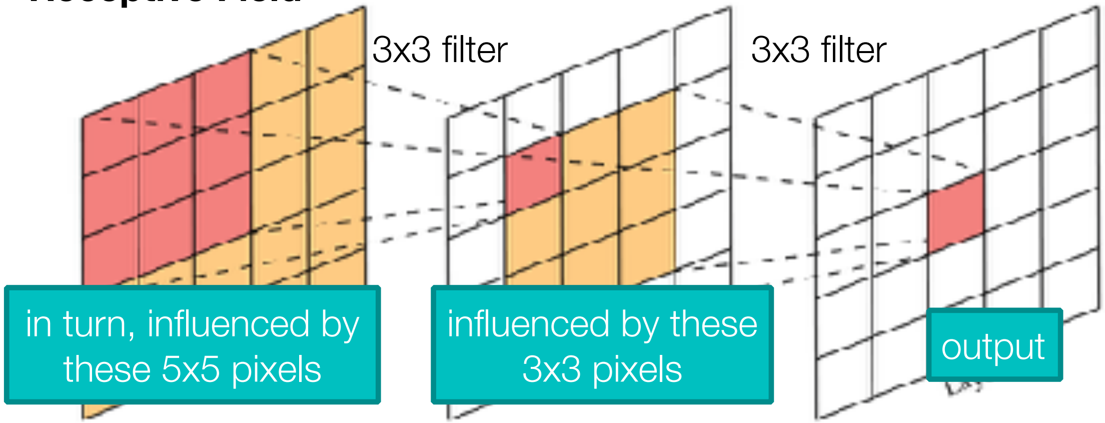


Computation Graph, Feature Spaces



Convolutional Networks

Receptive Field



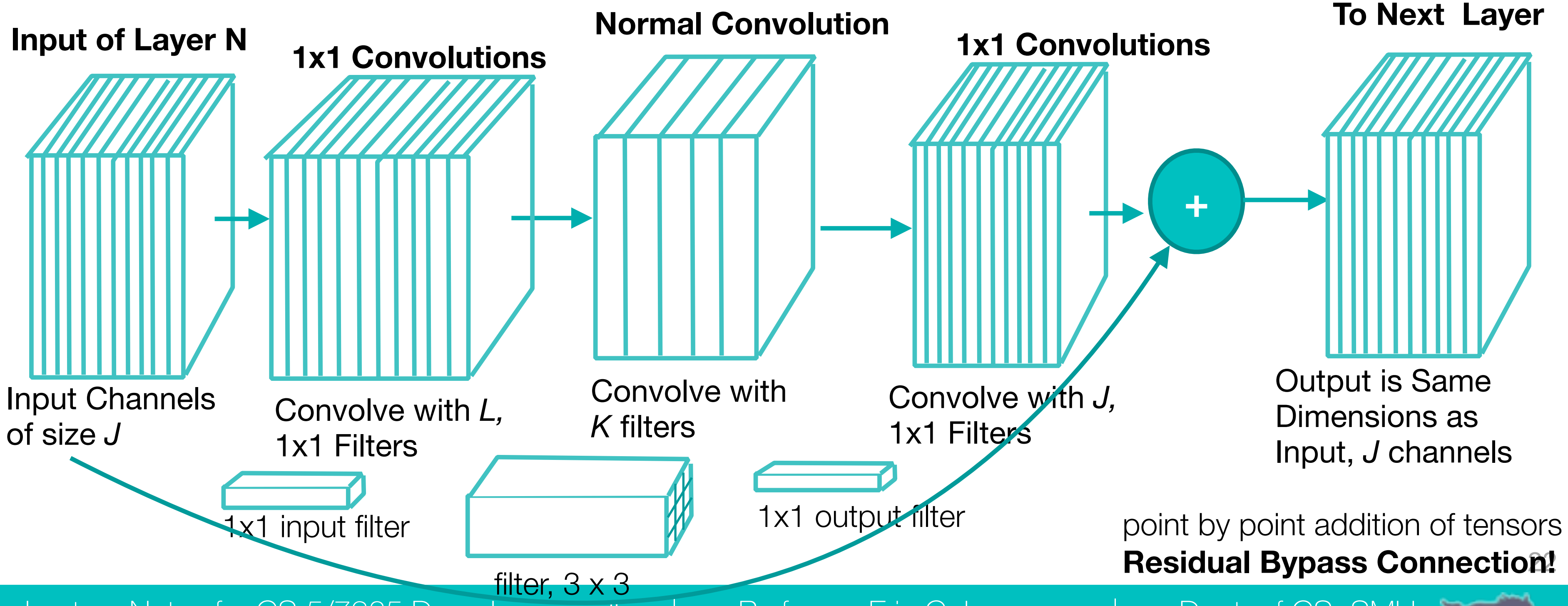
3x3 filter

3x3 filter

in turn, influenced by these 5x5 pixels

influenced by these 3x3 pixels

output



Input of Layer N

1x1 Convolutions

Normal Convolution

1x1 Convolutions

To Next Layer

Input Channels of size *J*

Convolve with *L*, 1x1 Filters

Convolve with *K* filters

Convolve with *J*, 1x1 Filters

+

Output is Same Dimensions as Input, *J* channels

1x1 input filter

filter, 3 x 3

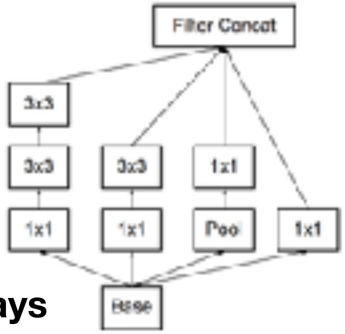
1x1 output filter

point by point addition of tensors

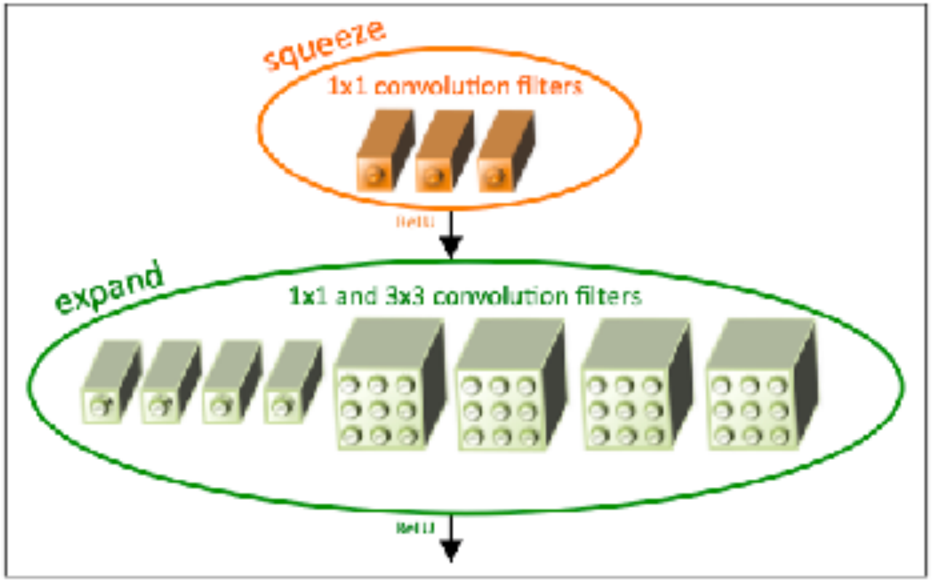
Residual Bypass Connection!

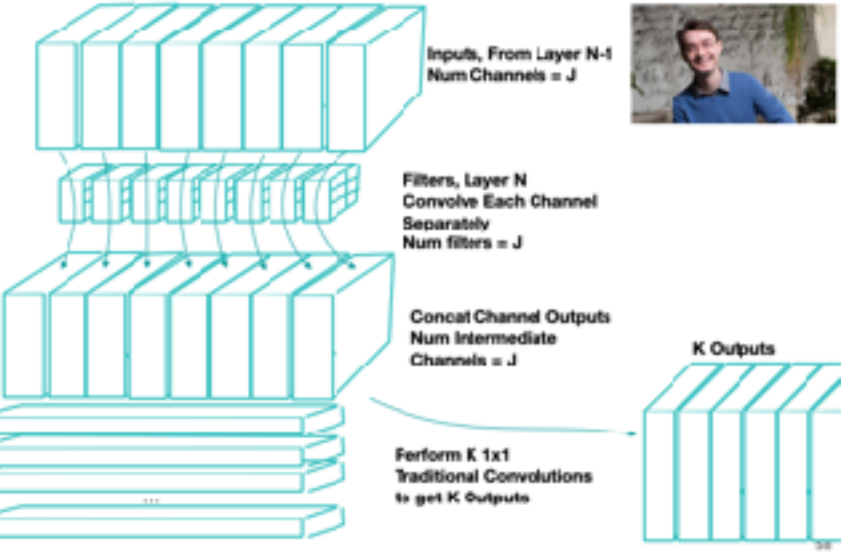
CNN Techniques

- Bottlenecks (1x1 filters)
- Parallel Paths, Concatenation
- Residual Paths
- Separable CNN



Parallel Pathways





Separable Convolution

Squeeze

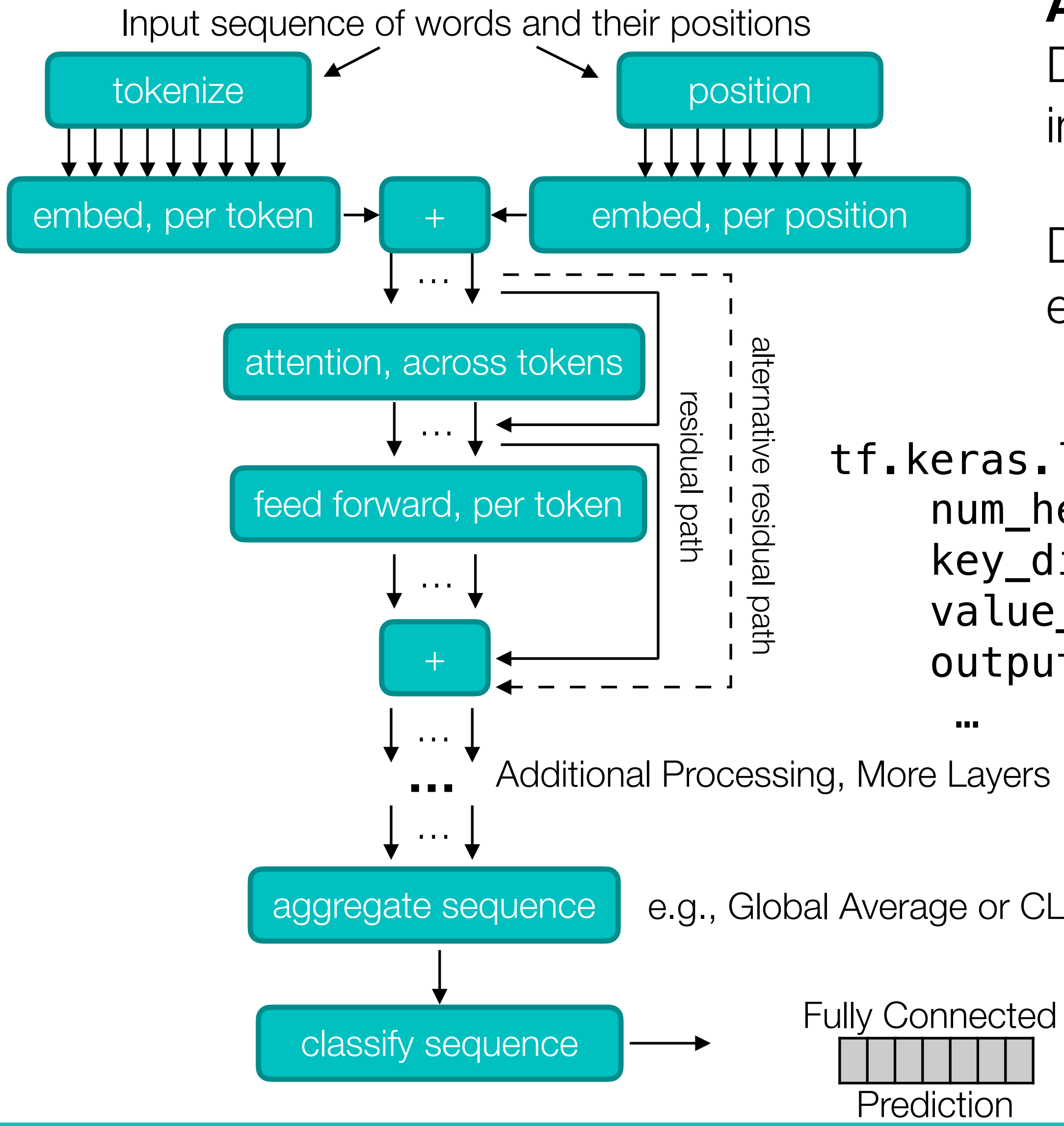
$$SR = \frac{|F_{s1x1}|}{|F_{e1x1}| + |F_{e3x3}|}$$

Controls how much to bottleneck

$$PCT_{3x3} = \frac{|F_{e3x3}|}{|F_{e1x1}| + |F_{e3x3}|}$$

Controls num filter params
(how many 1x1 versus 3x3)

Transformers

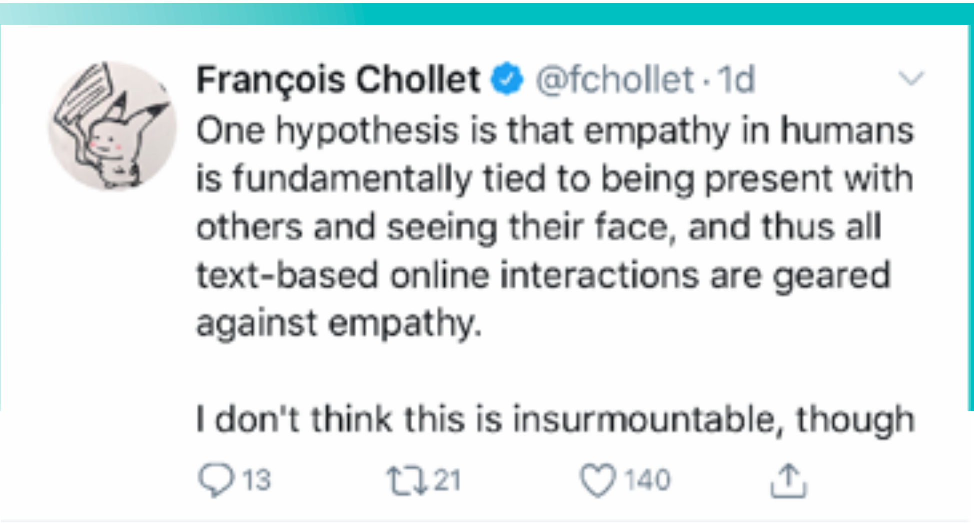


As we proceed:
Does a transformer change the initial word embedding space?
OR
Does it shift word vectors in the existing space?

```
tf.keras.layers.MultiHeadAttention(
        num_heads,
        key_dim,
        value_dim,
        output_shape,
         …
```

num_heads: (Number of heads Z_1-Z_7)
key_dim: (size of query/key d_k)
value_dim: (size of each d_v)
output_shape: (Embed size of Z, dims of W^o)

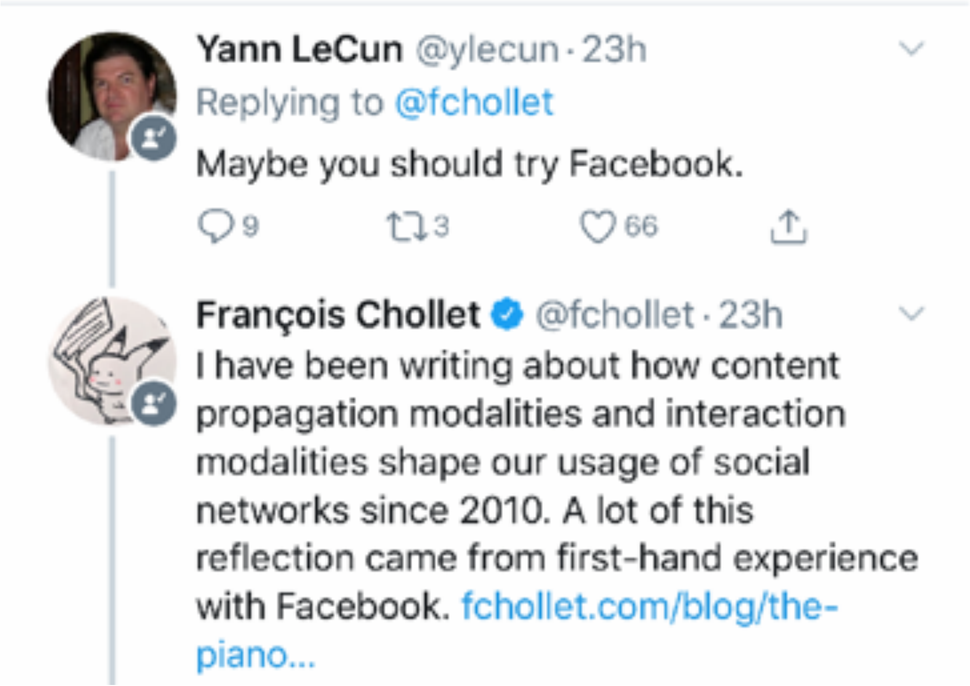
Ethical ML



François Chollet @fchollet · 1d

One hypothesis is that empathy in humans is fundamentally tied to being present with others and seeing their face, and thus all text-based online interactions are geared against empathy.

I don't think this is insurmountable, though



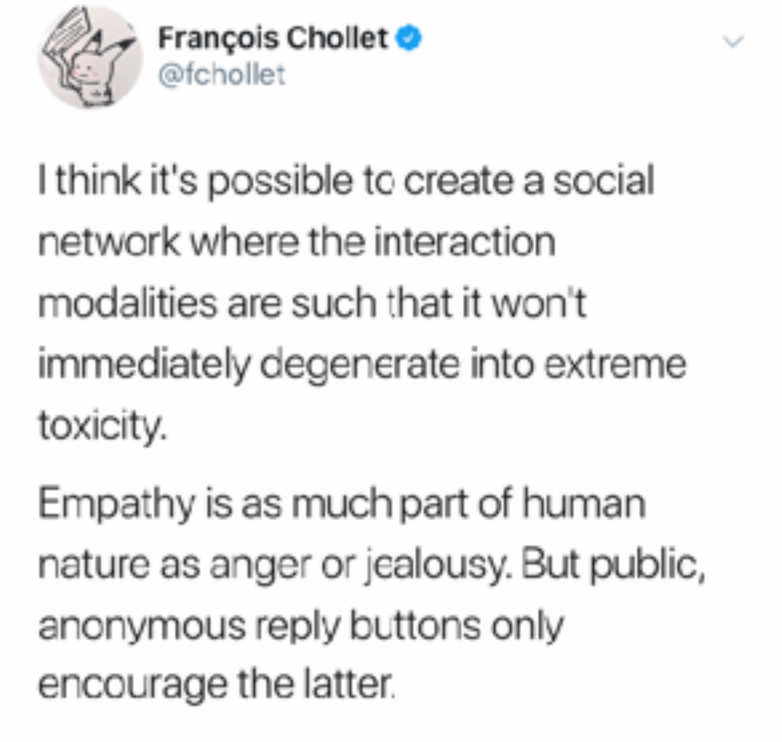
Yann LeCun @ylecun · 23h

Replying to @fchollet

Maybe you should try Facebook.

François Chollet @fchollet · 23h

I have been writing about how content propagation modalities and interaction modalities shape our usage of social networks since 2010. A lot of this reflection came from first-hand experience with Facebook. fchollet.com/blog/the-piano...



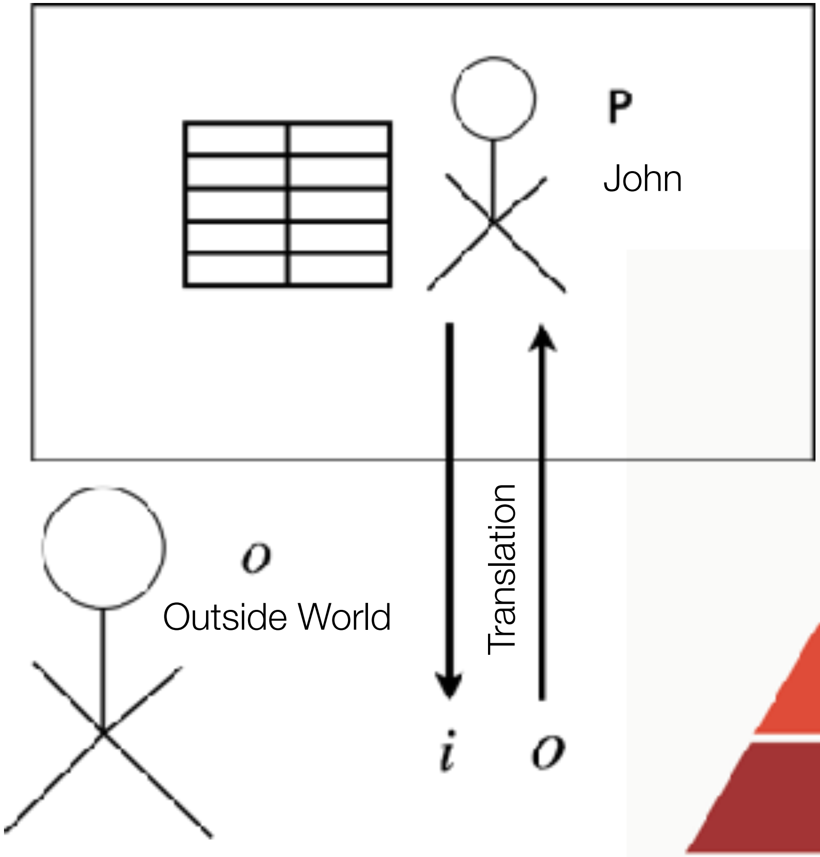
François Chollet
@fchollet

I think it's possible to create a social network where the interaction modalities are such that it won't immediately degenerate into extreme toxicity.

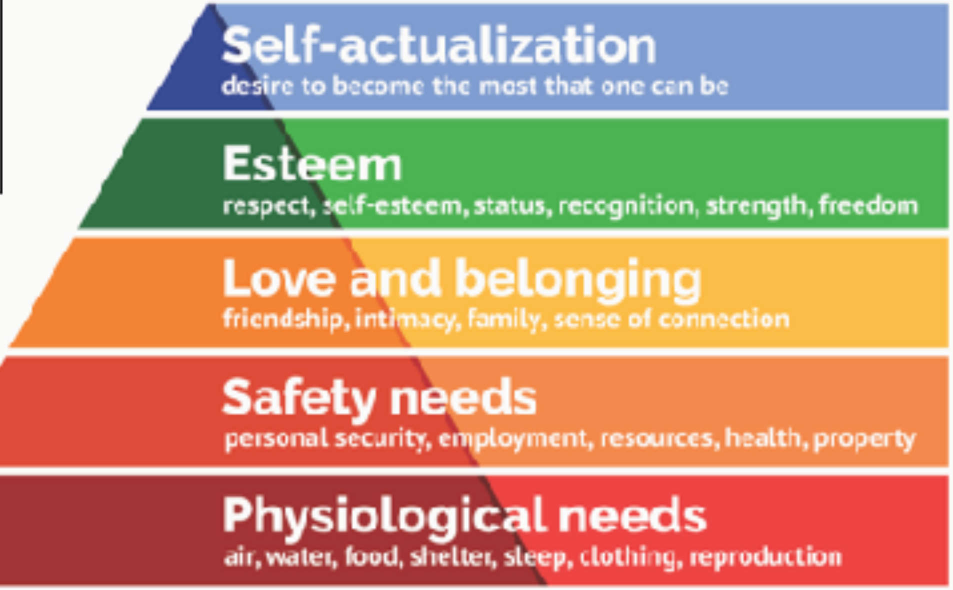
Empathy is as much part of human nature as anger or jealousy. But public, anonymous reply buttons only encourage the latter.

Strong AI, i.e., machines and thinking

- John Searle’s Foreign Room Argument:
  - Can John ever understand what he is saying?



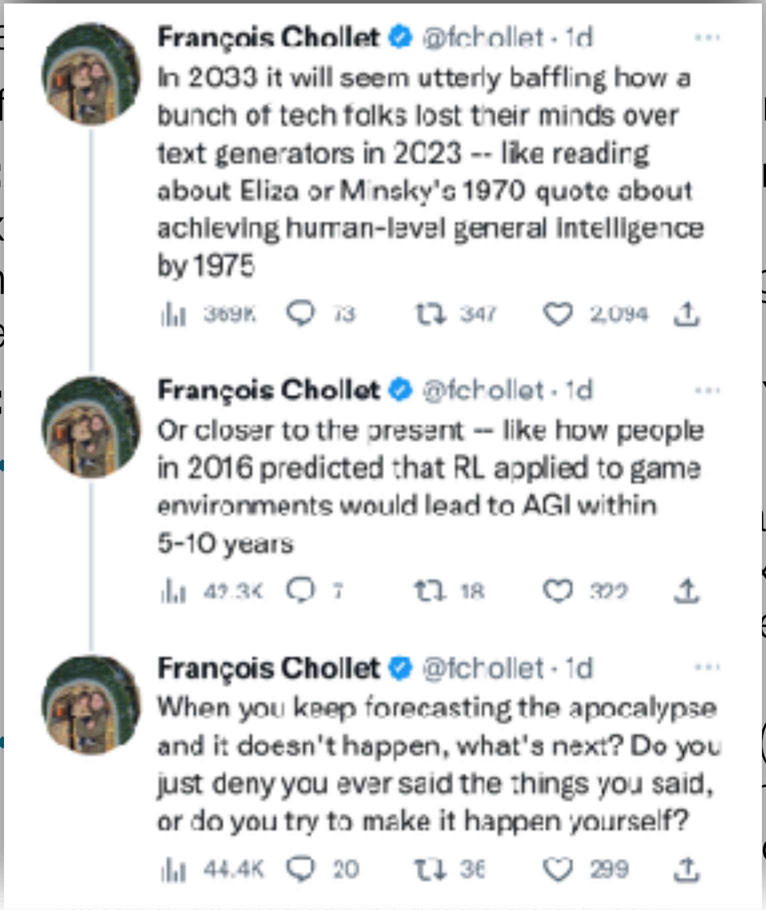
- If always translating without mistakes, even then we cannot be sure if what is inside truly understands what the output is
  - Humans share a need that drives our communications and interactions:


Self-actualization
desire to become the most that one can be
Esteem
respect, self-esteem, status, recognition, strength, freedom
Love and belonging
friendship, intimacy, family, sense of connection
Safety needs
personal security, employment, resources, health, property
Physiological needs
air, water, food, shelter, sleep, clothing, reproduction

Maslow’s Pyramid of Human Need

Can machines think?

- 🦜LLMs generate similar patterns from patterns they have se
- Is that f ns do?
  - **A:** nd ex kn gent, ne
  - **B:** **Yes.**
    - ve no k ern
    - (not ng of om similar patterns in language, not understanding
- We impose sentience on machines. Human brains are **nothing like neural networks**.



François Chollet @fchollet · 1d

In 2033 it will seem utterly baffling how a bunch of tech folks lost their minds over text generators in 2023 -- like reading about Eliza or Minsky's 1970 quote about achieving human-level general intelligence by 1975

369K 73 347 2,094

François Chollet @fchollet · 1d

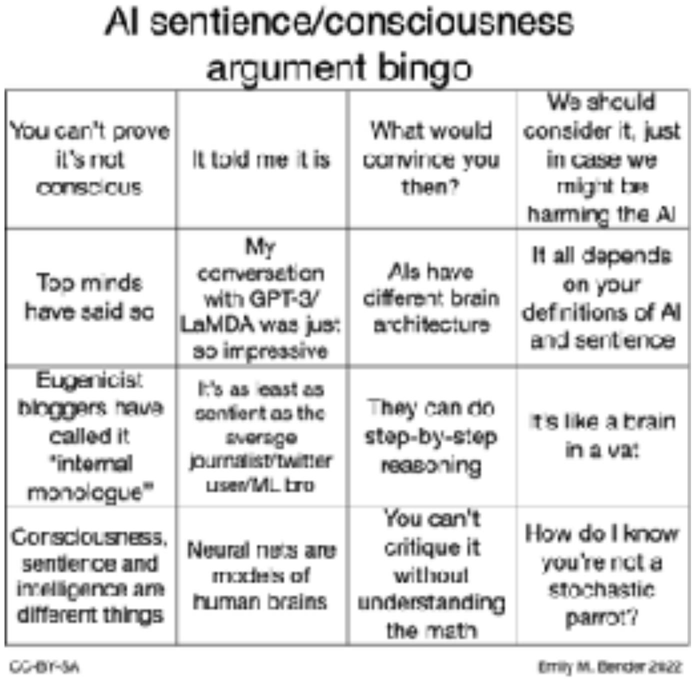
Or closer to the present -- like how people in 2016 predicted that RL applied to game environments would lead to AGI within 5-10 years

42.3K 7 18 322

François Chollet @fchollet · 1d

When you keep forecasting the apocalypse and it doesn't happen, what's next? Do you just deny you ever said the things you said, or do you try to make it happen yourself?

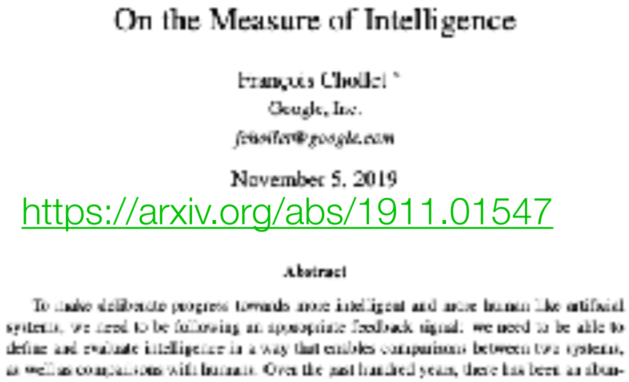
44.4K 20 36 299



AI sentience/consciousness argument bingo

You can't prove it's not conscious	It told me it is	What would convince you then?	We should consider it, just in case we might be harming the AI
Top minds have said so	My conversation with GPT-3/ LaMDA was just so impressive	AIs have different brain architecture	It all depends on your definitions of AI and sentience
Eugenicist bloggers have called it "internal monologue"	It's as least as sentient as the average journalist/twitter user/ML bro	They can do step-by-step reasoning	It's like a brain in a vat
Consciousness, sentience and intelligence are different things	Neural nets are models of human brains	You can't critique it without understanding the math	How do I know you're not a stochastic parrot?

CC-BY-SA Emily M. Bender 2022



On the Measure of Intelligence

François Chollet *
Google, Inc.
fchollet@google.com

November 5, 2019

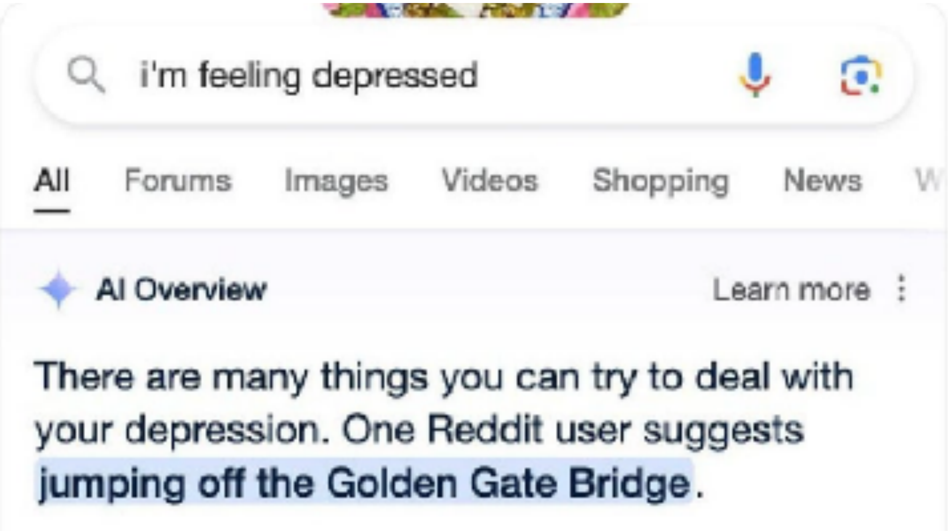
https://arxiv.org/abs/1911.01547

Abstract

To make deliberate progress towards more intelligent and more human-like artificial systems, we need to be following an appropriate feedback signal: we need to be able to define and evaluate intelligence in a way that enables comparisons between two systems, as well as comparisons with humans. Over the past hundred years, there has been an abun-

64 Pages of theory, evidence, questions, and bliss!

Ethical Principles



i'm feeling depressed

AI Overview

There are many things you can try to deal with your depression. One Reddit user suggests jumping off the Golden Gate Bridge.

Learning about neural networks and human consciousness
So Sentient AI is a myth?
Always has been
Programmers

*AI Search is the biggest consumer of LLM environmental impact. About 10x more energy than a typical Google Query. If replacing every google search, this is about the equivalent energy of 1.5M people.

** Many people correctly dispute the number as there are efficient ways to mitigate this—but unclear ho much. **We should be looking for ways to mitigate environmental impact,** but not by using statistics to mischaracterize…

***Estimates of water consumption are typically overblown as the water usage for cooling is not typically potable. There is impact, but energy generation water usage is already astronomical.

References: MIT Sloan, Washington Post, Nature: Machine Intelligence

The Google AI Principles

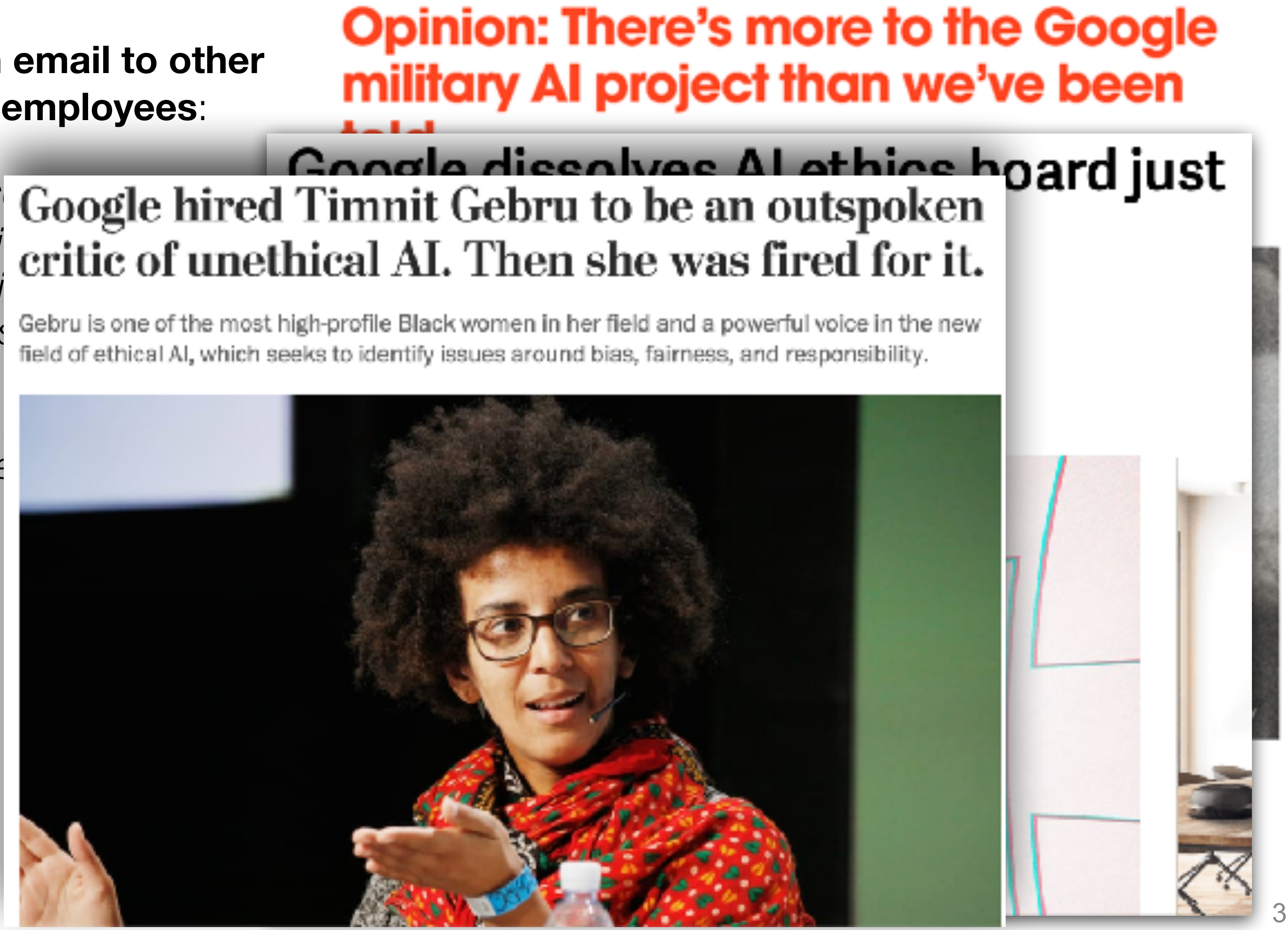
- Be socially **beneficial**
- **Avoid** creating or reinforcing **unfair** bias
- Be built and tested for **safety**
- Be **accountable** to people
- Incorporate **privacy** design principles
- Uphold high standards of scientific excellence
- Be **made available** for uses that accord with these principles
- **Google will not pursue**:
    - Tech likely to cause **harm**, tech that **principally** is a **weapon**, Tech that violates **surveillance** norms, Tech that contravenes **human rights**

https://www.blog.google/technology/ai/ai-principles/

How is Google doing?

FeiFei Li, in an email to other Google Cloud employees:

“Avoid at ALL C
mention or impli
Weaponized AI i
of the most sens
AI — if not THE
red meat to the
ways to damage



Opinion: There’s more to the Google military AI project than we’ve been told

Google dissolves AI ethics board just

Google hired Timnit Gebru to be an outspoken critic of unethical AI. Then she was fired for it.

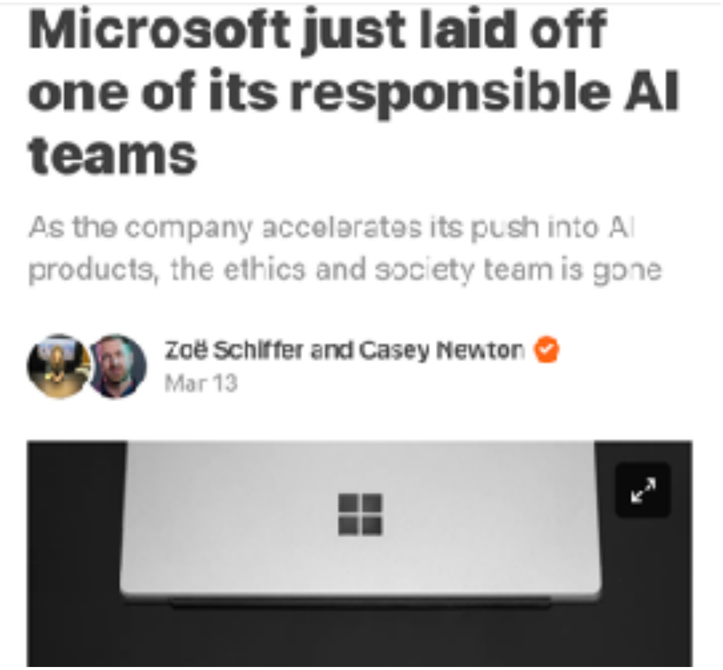
Gebru is one of the most high-profile Black women in her field and a powerful voice in the new field of ethical AI, which seeks to identify issues around bias, fairness, and responsibility.

What went wrong?

- “First acknowledge the elephant in the room: Google's AI principles”
  - *Evan Selinger, professor of philosophy at Rochester Institute of Technology*
- “A board can't just be 'some important people we know.' You need actual ethicists"
  - *Patrick Lin, director of the Ethics + Emerging Sciences Group at Cal Poly*
- “The group has to have authority to say no to projects”
  - *Sam Gregory, program director at Witness*

https://www.technologyreview.com/s/613281/google-cancels-ateac-ai-ethics-council-what-next/

Was it just Google?

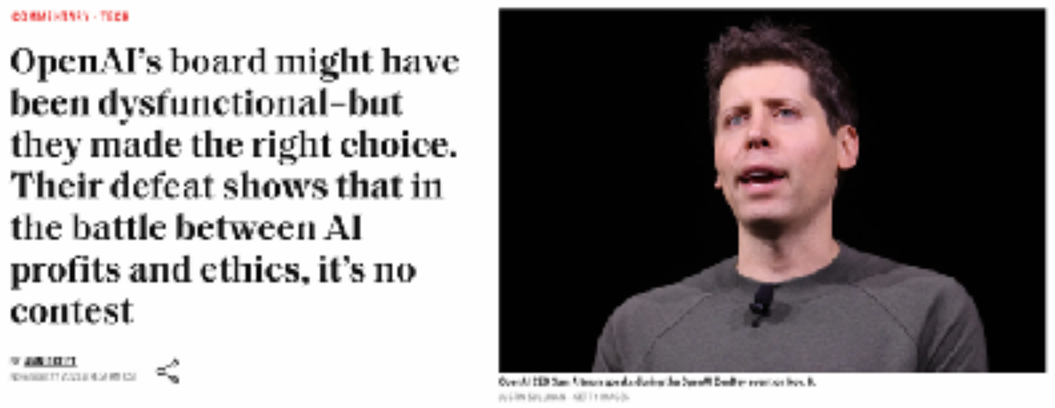


Microsoft just laid off one of its responsible AI teams

As the company accelerates its push into AI products, the ethics and society team is gone

Zoë Schiffer and Casey Newton

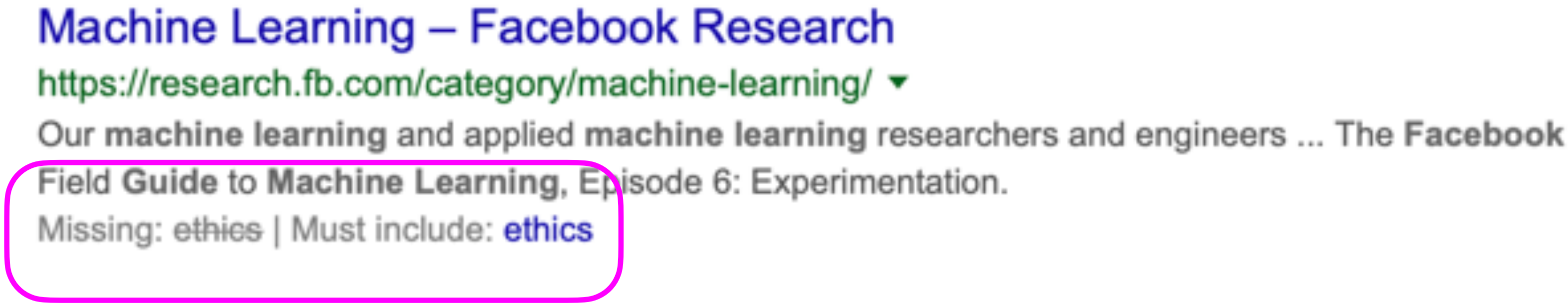
Mar 13



COMMENTARY · TECH

OpenAI's board might have been dysfunctional–but they made the right choice. Their defeat shows that in the battle between AI profits and ethics, it's no contest

Sam Altman terminated by board, partially for "An aversion to ethics in AI and deep learning in the face of rapid innovation and AI research."
Was reinstated 5 days later and the boards members pushed out that wanted ethical transparency.



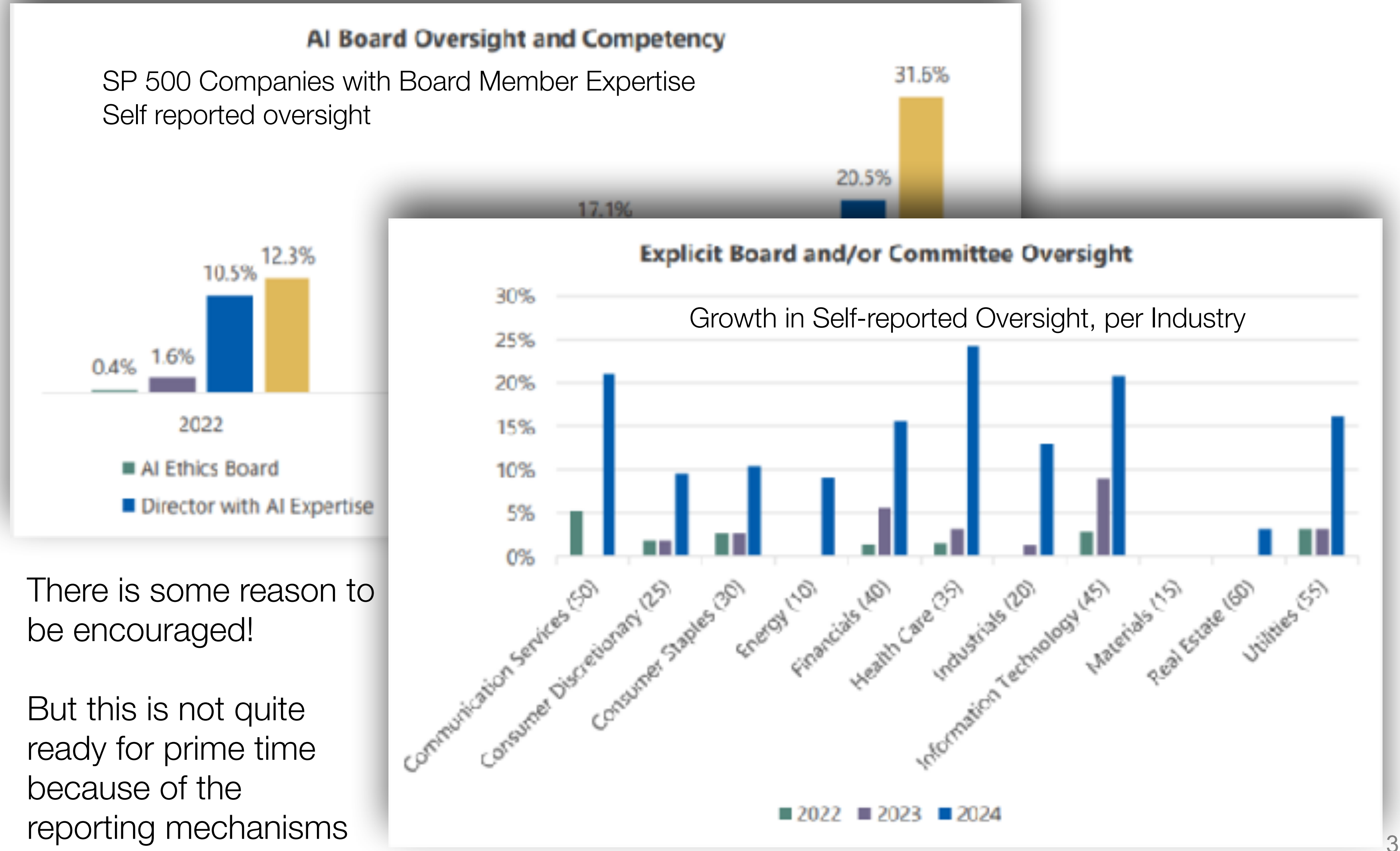
Machine Learning – Facebook Research

https://research.fb.com/category/machine-learning/ ▾

Our **machine learning** and applied **machine learning** researchers and engineers ... The **Facebook** Field Guide to **Machine Learning**, Episode 6: Experimentation.

Missing: ~~ethics~~ | Must include: ethics

Oversight growth



SP 500 Companies with Board Member Expertise
Self reported oversight

Growth in Self-reported Oversight, per Industry

There is some reason to be encouraged!

But this is not quite ready for prime time because of the reporting mechanisms

Ethical Principles in ML

From Australian Government, Department of Science

- **Reliability**: does system operate in accordance with intended purpose?
- **Fairness**: will system be inclusive and accessible? Will it involve or result in unfair discrimination against individuals, communities, or groups?

Model Measurement and Objective Alignment

- **Beneficence**: does system benefit individuals, society, or environment?
- **Respect**: does system respect human rights and autonomy of individuals?

Forethought and Insight

- **Privacy**: will system respect and uphold privacy rights and data protection, and ensure the security of data?
- **Transparency**: will system ensure people know when they are engaging with an AI system? Or know if significantly impacted?
- **Contestable**: will there be a timely process to allow people to challenge the use or output of the AI system?

Deployment Design

- **Accountability**: Those responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and *human oversight* of AI systems should be enabled.

Organizational Structure

https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework/ai-ethics-principles