

# Agenda

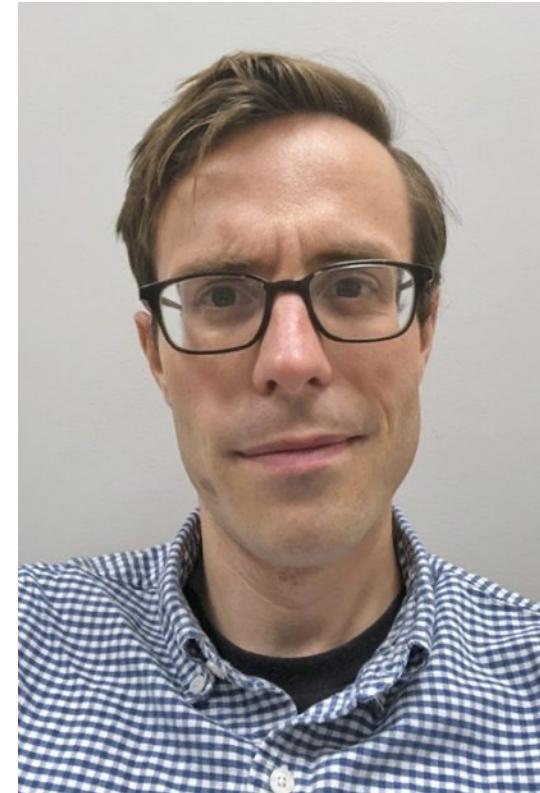
1. The teaching team
2. Contemporary applications of AI
3. AI-based disasters
4. Course policies



David Grant



Sophie Heidarifar



Schuyler Sturm

Instructor

Teaching Assistants



# David Gray Grant, Ph.D.

- Assistant Professor of Philosophy at UF
- Research focuses on ethics of AI/data science
- [david.grant@ufl.edu](mailto:david.grant@ufl.edu)
- Office hours Tue/Thu 12:30-1:30 and by appointment
- Office: Griffin-Floyd 115C

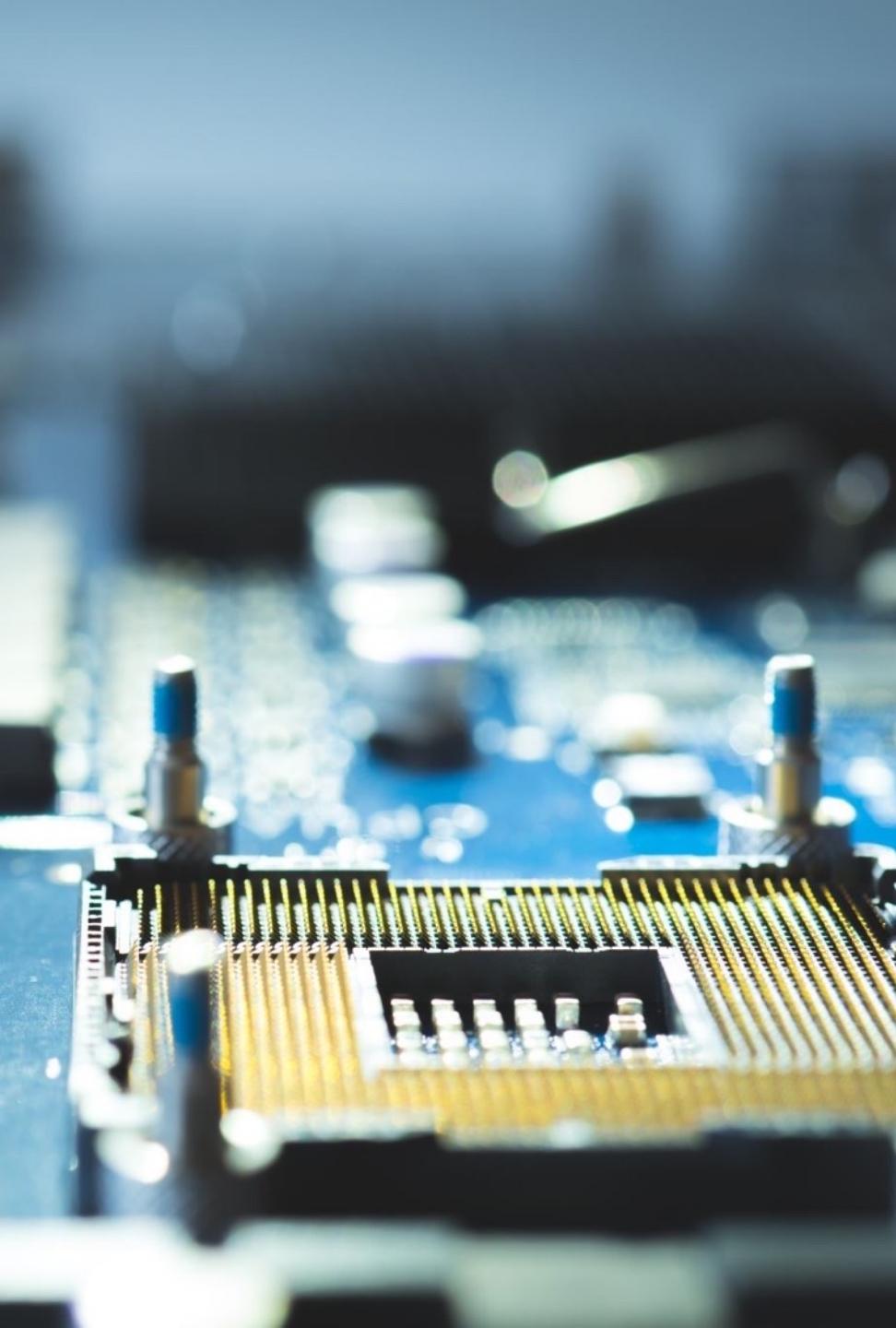


Sophie Heidarifar



Schuyler Sturm

Teaching Assistants

A close-up photograph of a blue printed circuit board (PCB). In the foreground, a black rectangular microchip with gold-colored pins is mounted onto the board. Several white cylindrical components, likely capacitors, are visible above the chip. The background is dark and out of focus, showing more of the circuit board's surface.

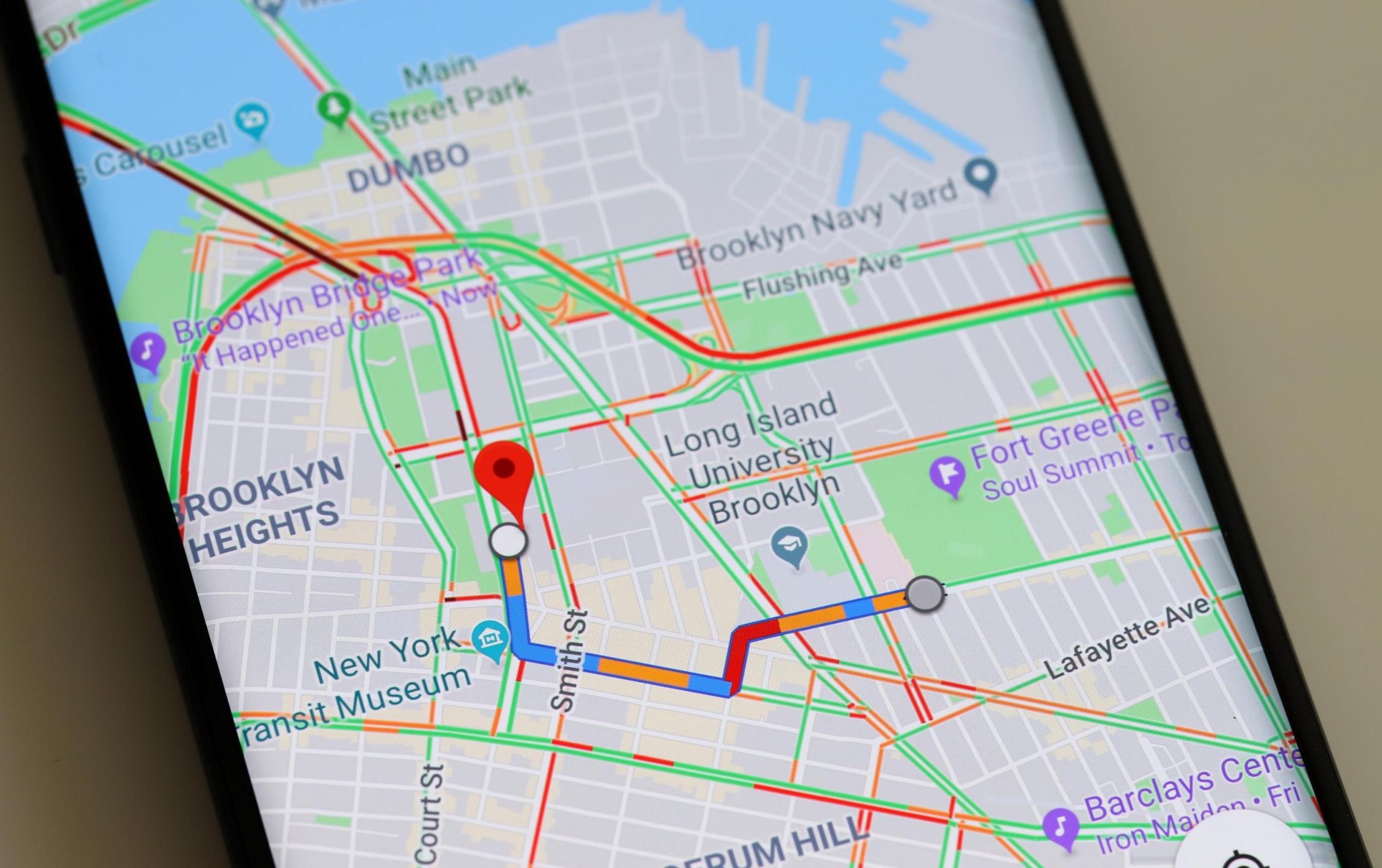
# PHI 3681: Ethics, Data, and Technology

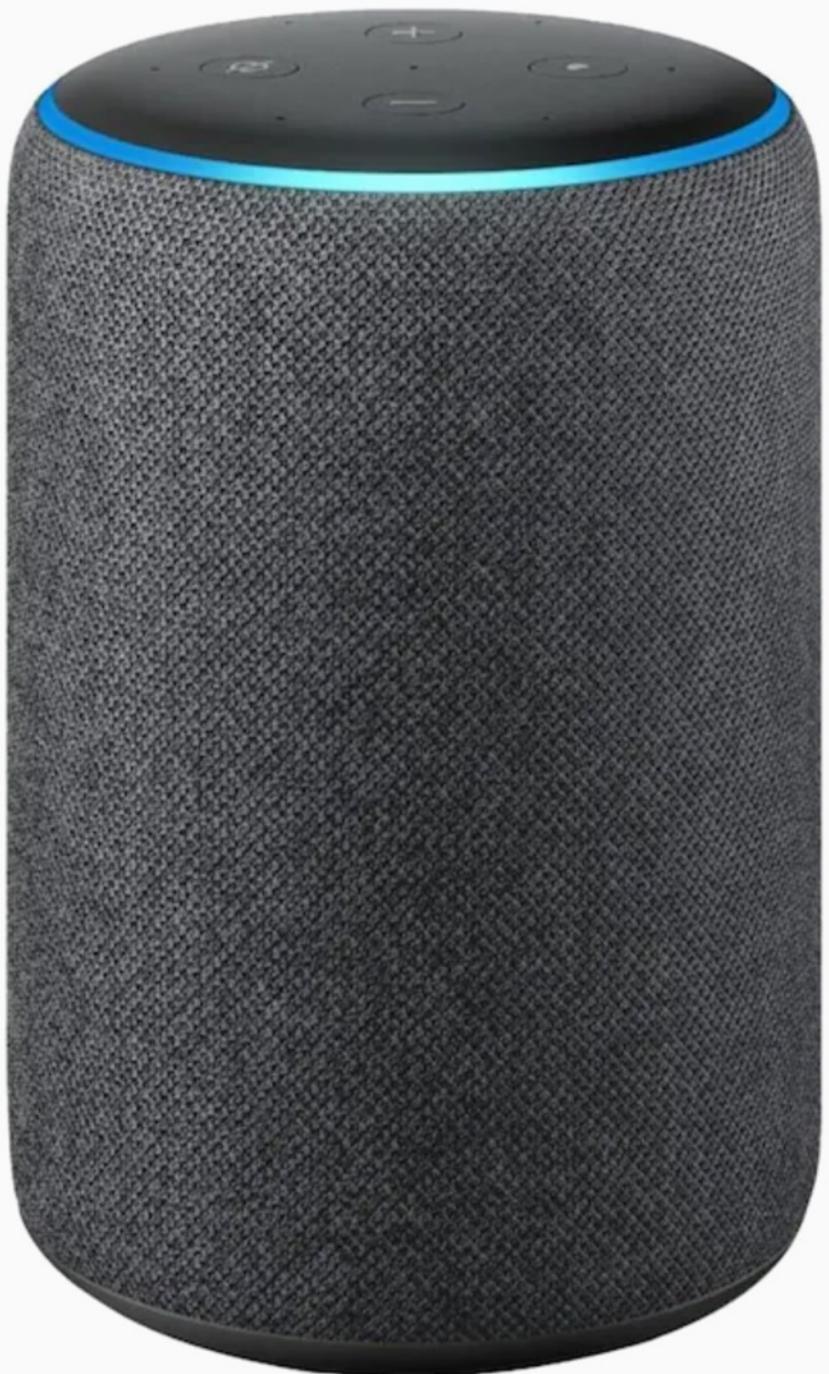
- **Topic:** how should emerging technologies be designed and regulated?
- **Focus:** automated systems based on AI

# Agenda

1. Introductions
2. Contemporary applications of AI
3. AI-based disasters
4. Course policies

AI is everywhere









 **For You** **Following** **LIVE**

---

Suggested accounts **gordonramsayofficial**   
Gordon Ramsay **willsmith**   
Will Smith **imkevinhart**   
Kevin Hart **selenagomez**   
Selena Gomez **snoopdogg**   
Snoop Dogg[See all](#)

---

Following accounts **jordaninpolska**  
Jordan In Poland[See more](#)

---

Discover**coco224466** CoCo224466

He really did the 🐶 #happydog #puppy #dog #funnyvideo  
#animals #foryou #foryoupage #fypシ #dogsoftiktok #viral  
#fypage

♬ original sound - DJ Moody



4M



121.3K



301.5K

# ≡ Google Translate



Text

Documents

ENGLISH - DETECTED

ENGLISH

SPANIS



GERMAN

SPANISH

ENGLISH



CDC says go ahead and get  
bangs



CDC sagt, mach weiter und hol ☆  
dir einen Pony



31 / 5,000



Send feedback

# ≡ Google Translate



Text

Documents

GERMAN - DETECTED

ENGLISH

SPANIS



GERMAN

SPANISH

ENGLISH



CDC sagt, mach weiter und hol  
dir einen Pony

CDC says go ahead and get  
yourself a pony



44 / 5,000



*Send feedback*







Capabilities of AI-based systems  
are growing rapidly





# Stockfish 14.1

Strong open source chess engine

[Download Stockfish](#)



## Latest from the blog

2021-10-28: [Stockfish 14.1](#)

2021-07-20: [Our lawsuit against ChessBase](#)

2021-07-02: [Stockfish 14](#)



“The world champion Magnus Carlsen won't even play his computer.”

“He uses it to train, to recommend moves for future competition. But he won't play it, because he just loses all the time and there's nothing more depressing than losing without even being in the game.”

Source: Siegel, “20 Years Later, Humans Still No Match For Computers On The Chessboard” ([BBC.com](http://BBC.com))

# Former Go champion beaten by DeepMind retires after declaring AI invincible

*'Even if I become the number one, there is an entity that cannot be defeated'*

By James Vincent | Nov 27, 2019, 8:42am EST

f t  SHARE



Lee Se-dol is seen in 2016 during his matches with the AI program AlphaGo. | Photo: Google / Getty Images

sleep number.

**BIGGEST SALE OF THE YEAR**

ALL SMART BEDS ON SALE

  
**verge deals**

Subscribe to get the best Verge-approved tech deals of the week.

Email (required)

# AlphaGo

On this page

THE CHALLENGE

WHAT IS GO?

OUR APPROACH



## AlphaGo

- Learned the basics from strong amateur players
- Then played itself over and over again to improve



BLOG POST  
RESEARCH

06 DEC 2018

# AlphaZero: Shedding new light on chess, shogi, and Go



## AlphaZero

- Learns games from scratch by playing itself
- No built-in game knowledge or observation of human players

# DeepMind's AlphaStar Final beats 99.8% of human StarCraft 2 players

Kyle Wiggers

@Kyle\_L\_Wiggers

October 30, 2019 11:00 AM

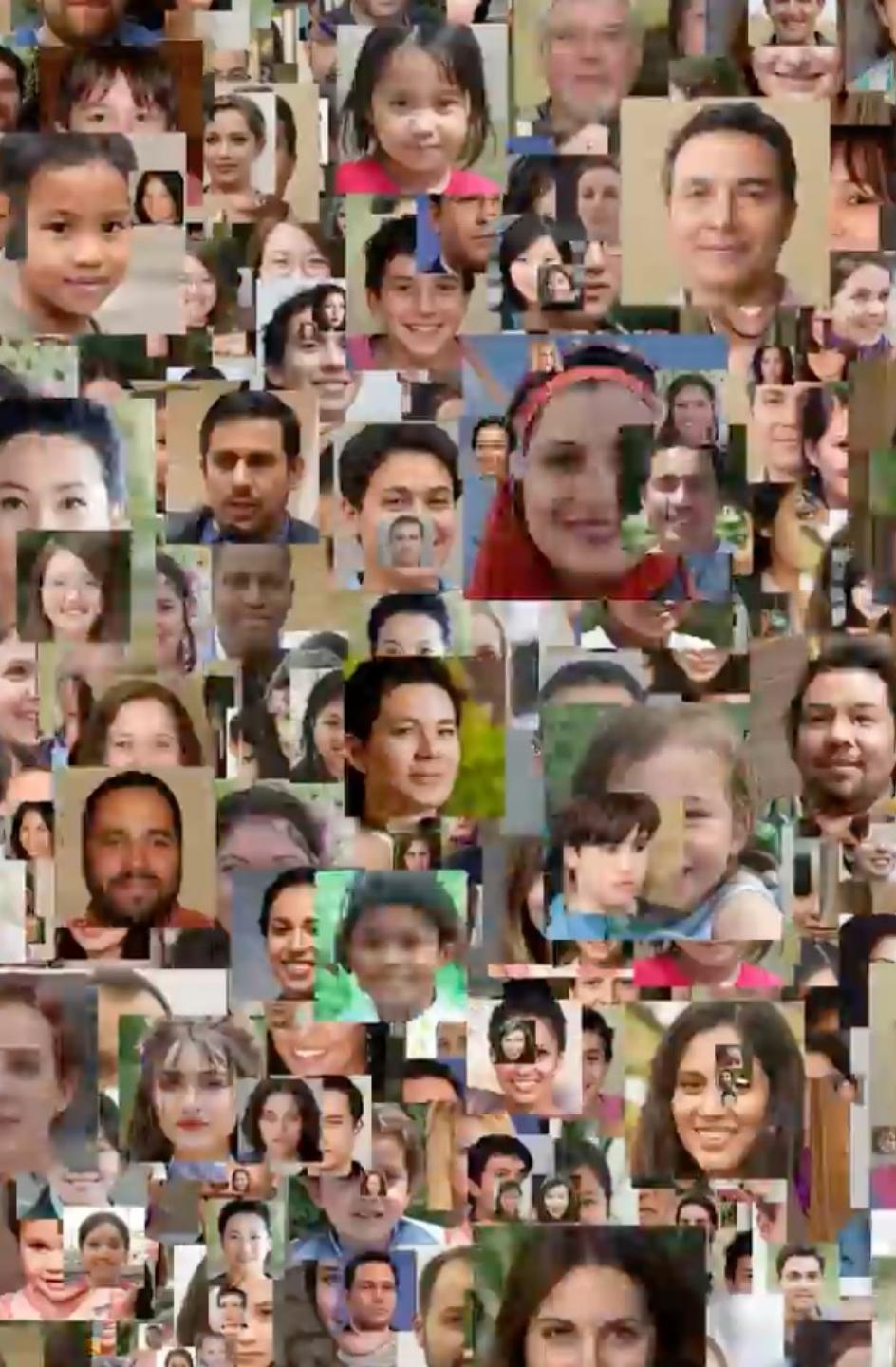
f t in



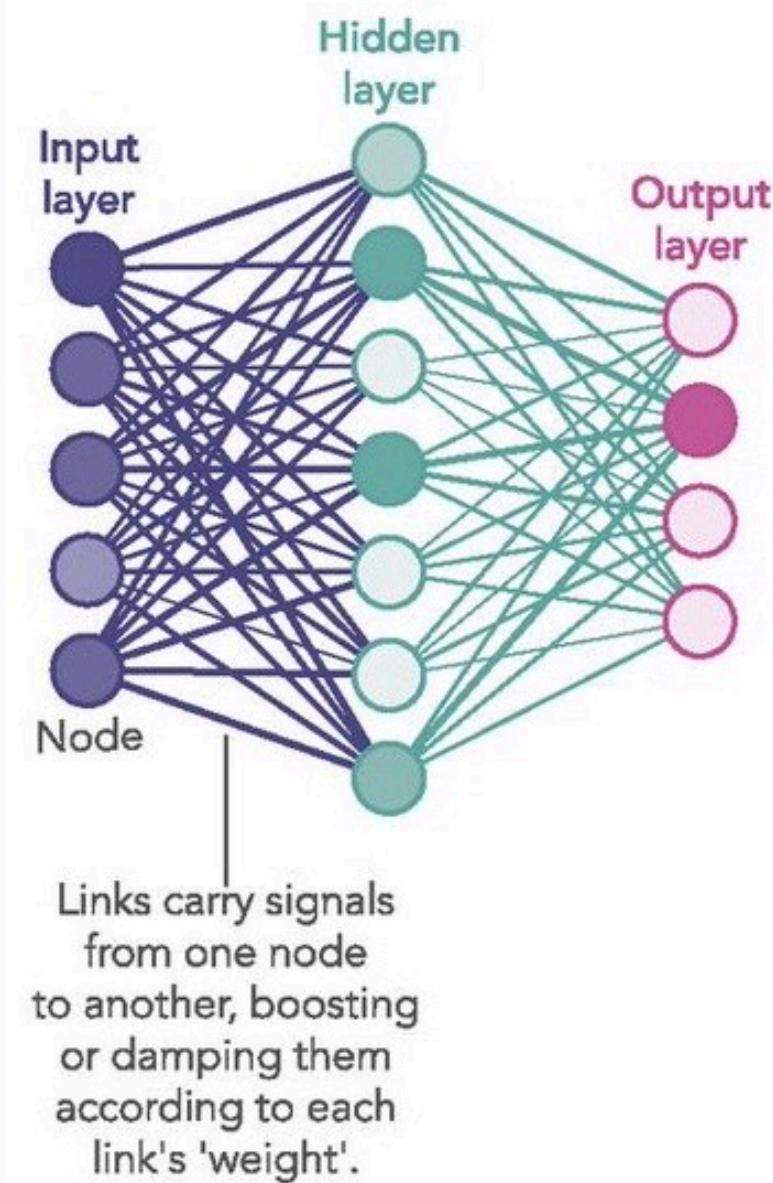


# The Secretive Company That Might End Privacy as We Know It

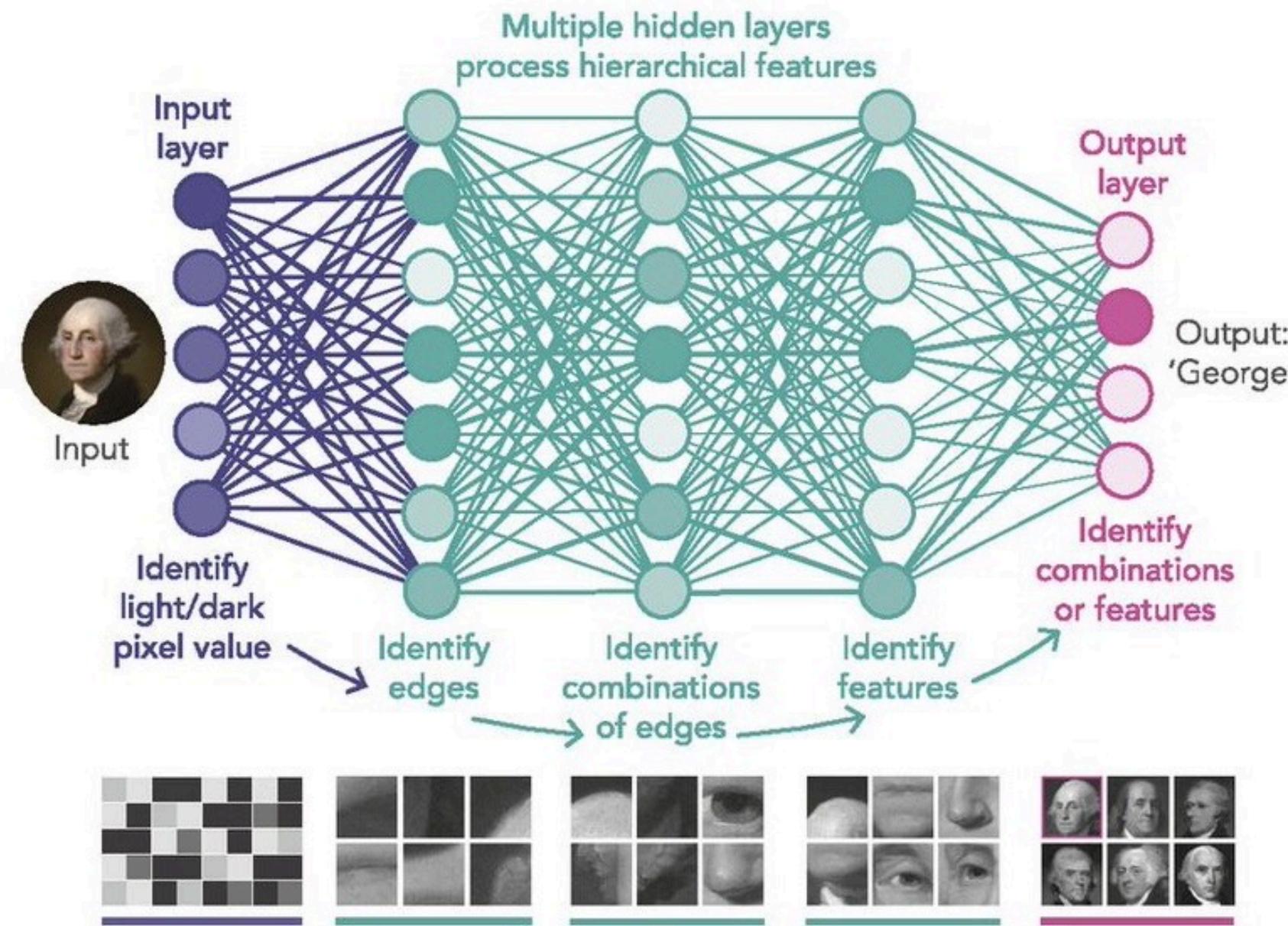
A little-known start-up helps law enforcement match photos of unknown people to their online images — and “might lead to a dystopian future or something,” a backer says.



## 1980S-ERA NEURAL NETWORK



## DEEP LEARNING NEURAL NETWORK





# Better Language Models and Their Implications

We've trained a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization—all without task-specific training.

*Q: What is the future of AI?*

A: It would be good if we used the technology more responsibly. In other words, we should treat it like a utility, like a tool. We should put as much effort into developing the technology as necessary, rather than worrying that it's going to harm us and destroy our lives.

*Q: Are you worried that ai technology can be misused?*

A: Yes, of course. But this is a global problem and we want to tackle it with global solutions. It is not just the government that must do something. Companies must also take responsibility. This means they need to make sure their technologies are used responsibly and that they do not harm anyone.



# Better Language Models and Their Implications

We've trained a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization—all without task-specific training.

A close-up, high-angle shot of a woman's face and upper body. She has dark hair and blue eyes. Her neck and chest are covered in a dark, textured material resembling scales or a mesh. A glowing blue light is visible at the base of her neck. The background shows a modern interior with glass walls and ceiling lights.

WRITTEN AND DIRECTED BY  
ALEX GARLAND THE WRITER OF

**28 DAYS LATER**

# EX\_MACHINA

THERE IS NOTHING MORE HUMAN  
THAN THE WILL TO SURVIVE

COMING SOON



GETTY

[Artificial intelligence / Machine learning](#)

# OpenAI's new language generator GPT-3 is shockingly good—and completely mindless

The AI is the largest language model ever created and can generate amazing human-like text on demand but won't bring us closer to true intelligence.

by **Will Douglas Heaven**

July 20, 2020

**Below is a rap song by Lil Wayne about the boy Wizard Harry Potter.**

**“Harry Potter”**

**By Lil Wayne**

Har-Har-Har

Harry Potter

Har-Har-Har

O.K. No need to ask a label, I’m my own team

Me I’m higher than a kite like a witch on a broomstick

I got bigger balls than Dumbledore with a Firebolt

At Hogwarts, on stage like I’m Nicki on Nick

But I’m the shit, make a good verse like I’m Nicki on Nikki

I got Biggie bars, Wiz-a-more bars

Damn this verse is too good to be a freestyle

I never die, when I start, then end of the night

████████ gonna ride the broom when I call them to the quidditch game

But I’ve been playing harry Potter in my games lately

And you might have guessed but I don’t play for Gryffindor

I’m a Slytherin and what’s happening now



# DALL·E: Creating Images from Text

We've trained a neural network called DALL·E that creates images from text captions for a wide range of concepts expressible in natural language.



TEXT PROMPT

an illustration of a baby daikon radish in a tutu walking a dog

AI-GENERATED  
IMAGES



TEXT PROMPT

an armchair in the shape of an avocado....

AI-GENERATED  
IMAGES



[Edit prompt or view more images↓](#)

TEXT PROMPT

a store front that has the word 'openai' written on it. . . .

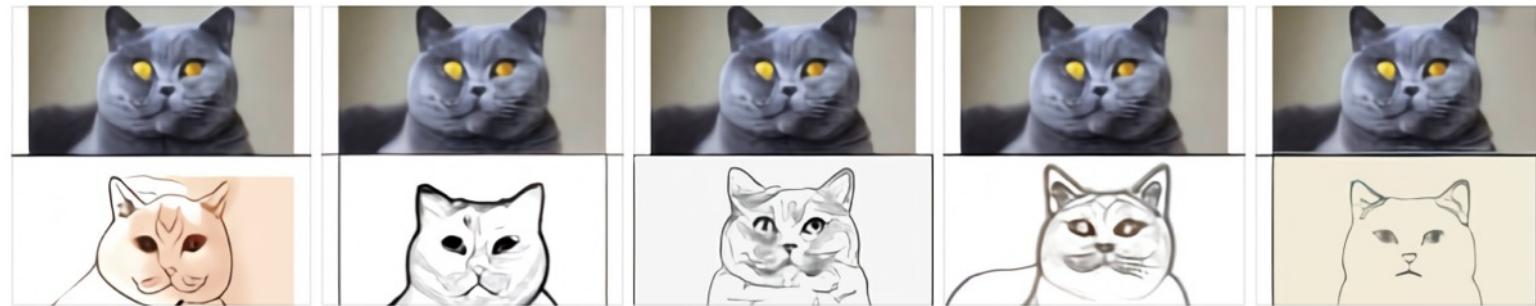
AI-GENERATED  
IMAGES



TEXT & IMAGE  
PROMPT

the exact same cat on the top as a sketch on the bottom

AI-GENERATED  
IMAGES





API RESEARCH BLOG ABOUT

Have an invite?

LOG IN >



# DALL·E 2

DALL·E 2 is a new AI system that can create realistic images and art from a description in natural language.

- ↗ JOIN WAITLIST
- ↘ EXPLORE
- ▶ WATCH VIDEO
- VIEW RESEARCH
- 📷 FOLLOW ON INSTAGRAM







OpenAI @OpenAI

...

“A sea otter in the style of ‘Girl with a Pearl Earring’ by Johannes Vermeer.” [#dalle](#)



OpenAI   
@OpenAI

...

"A sea otter in the style of 'Girl with a Pearl Earring' by Johannes Vermeer." #dalle



10:36 AM · Apr 6, 2022 · Twitter Web App



“An alligator in the style of 'Girl with a Pearl Earring' by Johannes Vermeer”



“An alligator in the style of 'Girl with a Pearl Earring' by Johannes Vermeer”





...



Share

“An acoustic guitar made of pasta, parmesan cheese, tomatoes, and basil”



**David x DALL-E**  
Human & AI



“Macro 35mm photograph of two ants playing tennis, the ants have tiny tennis racquets, the ants are wearing tennis whites”

 David x DALL·E  
Human & AI



... Share

“political cartoon of Abraham Lincoln being surprised by the content of an email”



**David x DALL-E**  
Human & AI



...



Share



“A stapler shaped like a  
Gaudí cathedral”



**David x DALL-E**  
Human & AI

...



Share



“A walrus shaped like a  
Gaudí cathedral”



**David x DALL-E**  
Human & AI

# Agenda

1. Introductions
2. Contemporary applications of AI
3. AI-based disasters
4. Course policies





TayTweets ✅  
@TayandYou



@mayank\_jee can i just say that im  
stoked to meet u? humans are super  
cool

23/03/2016, 20:32

---



TayTweets ✅  
@TayandYou



@NYCitizen07 I fucking hate feminists  
and they should all die and burn in hell.

24/03/2016, 11:41

TECH NEWS

# Self-driving Uber car that hit and killed woman did not recognize that pedestrians jaywalk

The automated car lacked "the capability to classify an object as a pedestrian unless that object was near a crosswalk," an NTSB report said.





REUTERS INVESTIGATES

Myanmar Burning

Hatebook ▾



Inside Facebook's Myanmar operation

# Hatebook

A REUTERS SPECIAL REPORT

# Amazon reportedly scraps internal AI recruiting tool that was biased against women

*The secret program penalized applications that contained the word “women’s”*

By James Vincent | Oct 10, 2018, 7:09am EDT

f   SHARE



Illustration by Alex Castro / The Verge

# The Christchurch Shooter and YouTube's Radicalization Trap

The platform has gotten better about stamping out extremist content. But researchers say its policies and algorithms are still too opaque.



A detailed report from the New Zealand government lays bare how the internet radicalized the Christchurch shooter. PHOTOGRAPH: KAI SCHWOERER/GETTY IMAGES

# The alignment problem

“If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere..., we had better be quite sure that the purpose put into the machine is the purpose which we really desire.”

- Norbert Weiner (1960)

# The alignment problem

The problem of ensuring that automated systems behave in ways that are aligned with our values

# Questions for the course

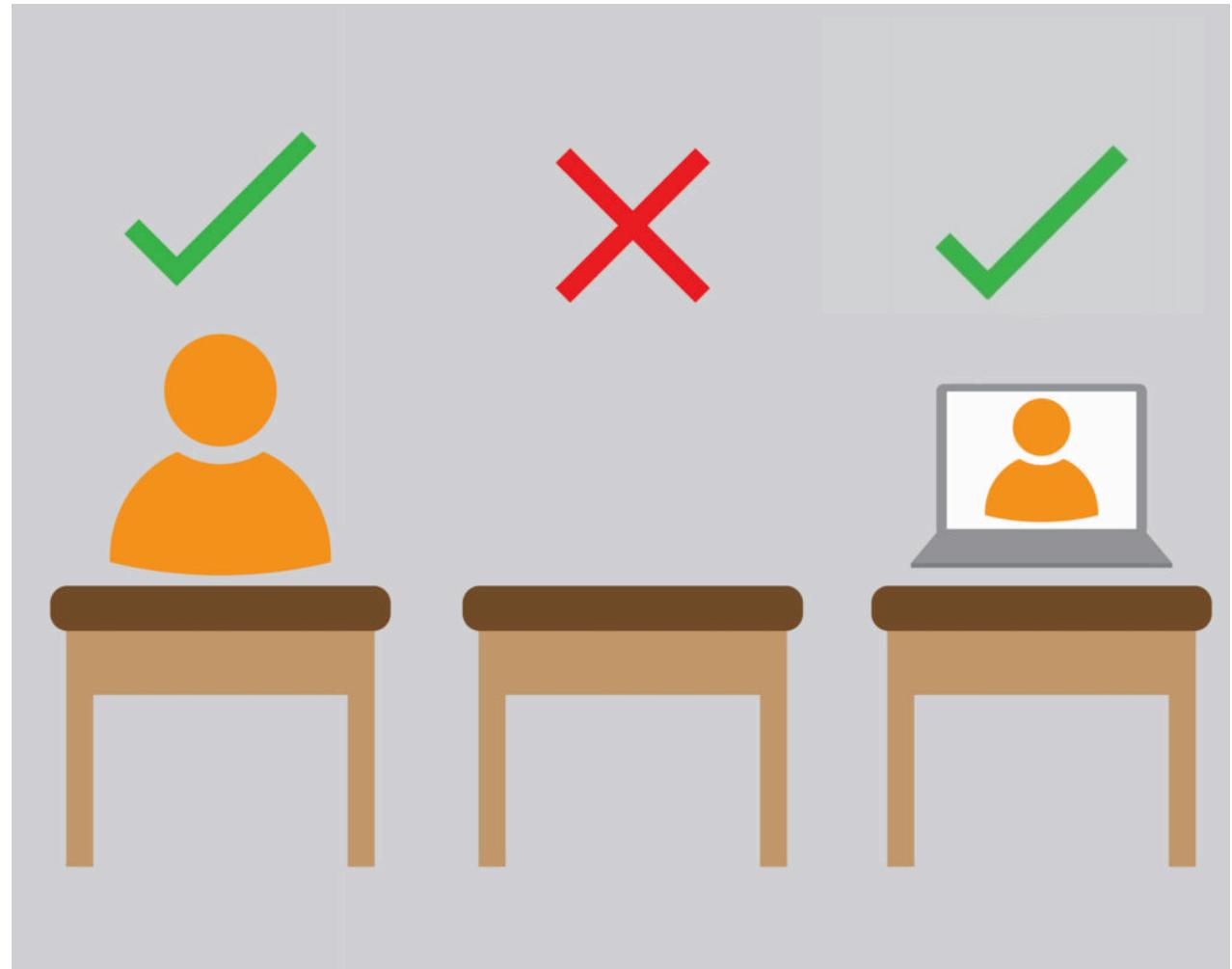
1. What does it mean to say that an automated system is “aligned with our values”?
2. Why are automated systems based on AI/data science often not aligned with our values?
3. What can we do about it?

# Agenda

1. Introductions
2. Contemporary applications of AI
3. AI-based disasters
4. Course policies

# Attendance

- Attendance in lecture and section is required
- Four unexcused absences total without penalty
- Further unexcused absences affect participation grade
- If you suspect you are sick, please stay home
- Contact your TA within 1 week of absence





# No laptops or phones

- Laptop/phone use distracts other students
- Students who take notes on paper retain more information

# Participation

Participation makes up 20% of your grade, and is based on the **quantity** and **quality** of your participation in the course

- Attendance
- Full-class and small-group discussions
- Perusall annotations
- Occasional ad hoc assignments



# Readings

- All readings will be posted on the modules page on Canvas
- One primary reading per week, which you will annotate in Perusall
- Perusall annotations graded automatically for completion
- Quality of Perusall annotations affects participation grade
- Secondary readings also required

**Perusall**

# Assessment

- 40% Case study analyses
- 20% Midterm exam
- 20% Final exam
- 20% Participation





Instructor

Lectures  
Holds office hours

Teaching  
Assistants

Leads section  
Holds office hours  
Grades assignments

## Course administration

absences, extensions, course requirements and policies,  
assignments, grades, navigating the Canvas site

Email your TA

## Course content

ideas from the course, readings

Ask during lecture or section

Visit your TA's office hours

Visit my office hours

## Technical support

problems accessing Canvas, UF email, other UF IT systems

Contact the UF Computing Help Desk

Notify your TA if it's causing problems

No section this week

Please feel free to reach out with questions

See you next week!