



CS 182: Ethics, Public Policy, and Technological Change

Rob Reich
Mehran Sahami

Head CA: Roberta Fischli

Today's Agenda

1. **Why we are teaching this course**
 2. Why are you interested in taking it?
 3. Why you *should* take this course
 4. What we are going to do together this quarter
-

TA Introductions

Please stand and briefly introduce yourself

Mehran Sahami



- Professor (Teaching) of Computer Science
- Chair, Computer Science Department
- Spent a decade in tech industry before returning to Stanford

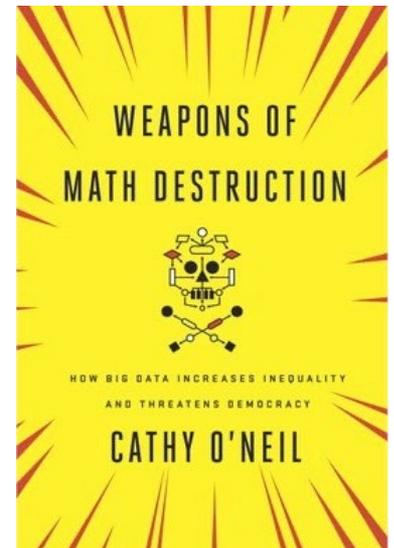
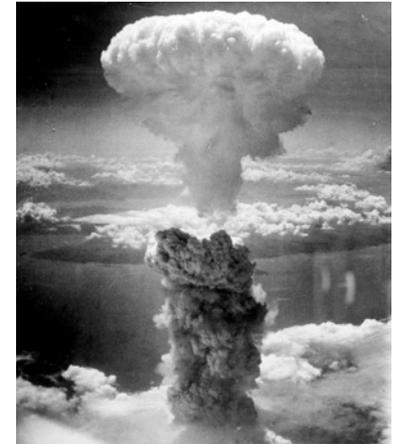
My Motivation

When I worked in industry, I saw two things first hand:

- Many decisions with social consequences result from decisions made in code
 - Rankings of search engine results
 - Recommendations in a social network
 - Objective functions to optimize in machine learning algorithms
 - Often, social consequences of these decisions are not considered (or even identified) when the code is written
 - We don't realize the full implications of our work (e.g., perpetuating biases, creating anti-social behavior, etc.)
 - We only deal with consequences *after* a problem is spotlighted (e.g., Cambridge Analytica scandal, creation of echo chambers, etc.)
-

The “New” Physicists

- After the Manhattan Project, many physicists realized the broader impact of their technical work
 - Some became peace activists
- Some have likened the computer scientists of today to the physicists of the mid-20th century
- In both cases, developing more technology does not provide a complete solution
 - Need to understand the interplay of technology, public policy, and societal impact



Unexpected Consequences



- *“Waymo Collision Illustrates Why Society Might Eventually Ban Human Driving”*
- Forbes, Nov. 7, 2018
- *“Wielding Rocks and Knives, Arizonans Attack Self-Driving Cars”*
- New York Times, Dec. 31, 2018

Not-So-Unexpected Consequences



“Revelations brought to light from whistleblower Frances Haugen, a former data scientist at Facebook, has led to what may be the most threatening scandal in the company's history.

...

Haugen told Congress that Facebook consistently chose to maximize its growth rather than implement safeguards on its platforms...”

Source: Bobby Allyn, NPR.org, Oct. 5, 2021

Being Well-Intentioned is Not Enough



Enter Galactica, an LLM aimed at writing scientific literature. Its authors trained Galactica on "a large and curated corpus of humanity's scientific knowledge," including over 48 million papers, textbooks and lecture notes, scientific websites, and encyclopedias.

<https://arstechnica.com/information-technology/2022/11/after-controversy-meta-pulls-demo-of-ai-model-that-writes-scientific-papers/>



Yann LeCun @ylecun · Nov 25, 2022

Replying to @Grady_Booch

Proper way to use Galactica: start typing your paper; let Galactica predict what you'll type next; accept, fix up, or reject.

But if you worry about naive or ill-intentioned random pple generating lots of nonsense, don't: no one will pay attention.



Grady Booch @Grady_Booch · Nov 25, 2022

Replying to @ylecun

'But if you worry about naive or ill-intentioned random pple generating lots of nonsense, don't'

And yet, I do, Yann.

This is why ethical considerations must be a factor in all software that touches the human experience.

Sadly, something you and @meta seem to always dismiss



This is Not a Spectator Sport

“Colleges are turning students’ phones into surveillance machines, tracking the locations of hundreds of thousands”
- Washington Post, Dec. 24, 2019

“School and company officials call location monitoring a powerful booster for student success: If they know more about where students are going, they argue, they can intervene before problems arise. But some schools go even further, using systems that calculate personalized ‘risk scores’ based on factors such as whether the student is going to the library enough.

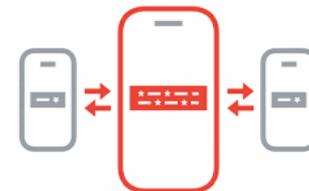
...

The students who deviate from those day-to-day campus rhythms are flagged for anomalies, and the company then alerts school officials in case they want to pursue real-world intervention.”

Multi-Use Technologies



Exposure Notifications: Using technology to help public health authorities fight COVID-19



"Google and Apple jointly created the Exposure Notifications System out of a shared sense of responsibility to help governments and our global community fight this pandemic through contact tracing."

"Your phone and the phones around you will work in the background to exchange these privacy-preserving random IDs via Bluetooth. You do not need to have the app open for this process to take place."

"The Exposure Notifications System does not collect or use the location from your device."

Source: <https://www.google.com/covid19/exposurenotifications/>

Rob Reich



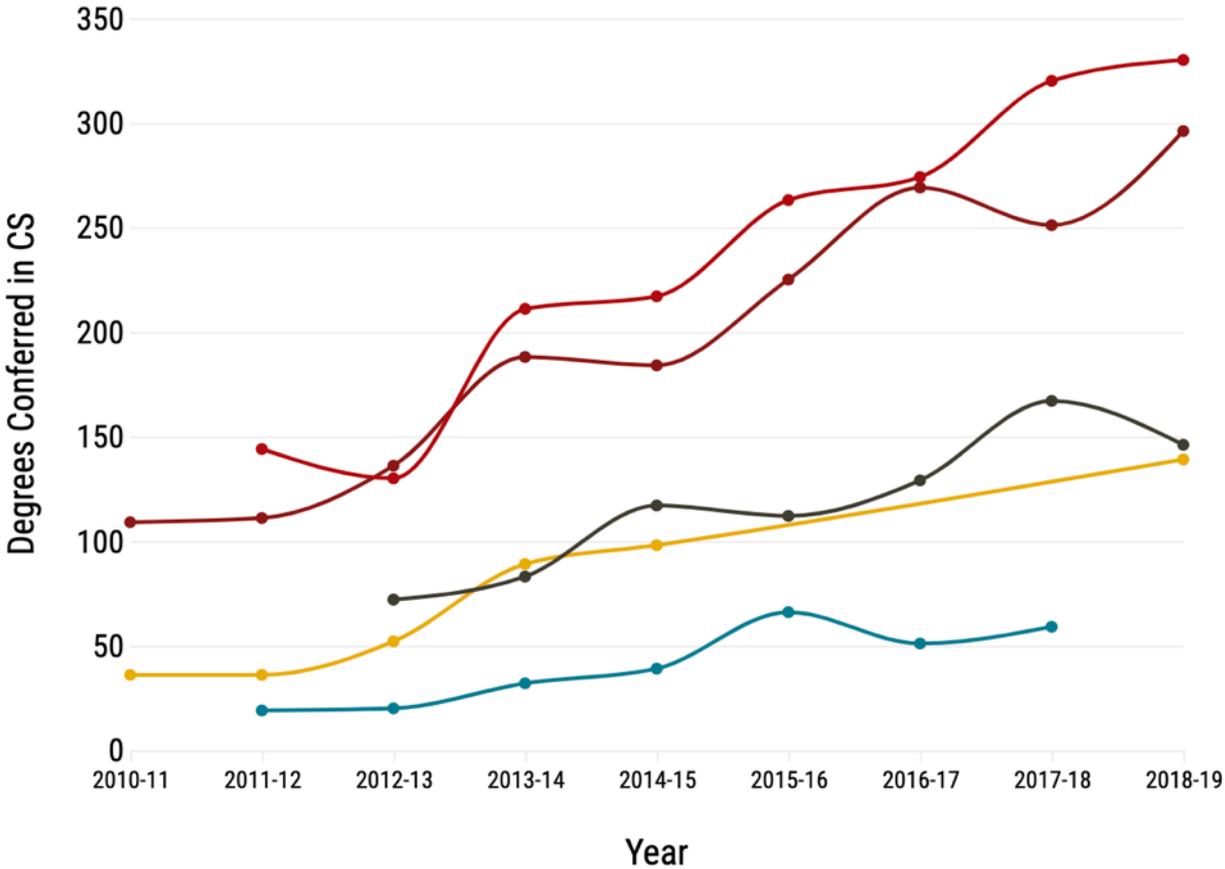
- Professor of Political Science
- Associate Director, Institute for Human-Centered Artificial Intelligence

My Motivation(s)

- Computer Science and Silicon Valley >> the new management consulting and Wall Street. Why?
 - Ethical ambition
 - Stanford says it trains future leaders. Does it? What does 21st Century leadership require?
-

Trends in Computer Science Degrees Over the 2010s

MIT Harvard Princeton Yale Stanford



The Atlantic

TECHNOLOGY

Stanford's Top Major Is Now Computer Science

THE
NEW YORKER

GET RICH U.

There are no walls between Stanford and Silicon Valley. Should there be?

Tech's Ethical 'Dark Side': Harvard, Stanford and Others Want to Address It



Laura Norén, who teaches a data science ethics course at New York University, said, "You can patch the software, but you can't patch a person if you, you know, damage someone's reputation."

Sam Hodgson for The New York Times

By [Natasha Singer](#)

Feb. 12, 2018



[Leer en español](#)

"BRILLIANTLY ARGUED...ESSENTIAL READING FOR ANYONE WORRIED ABOUT MONEY IN POLITICS."—LARISSA MACFARQUHAR

JUST GIVING
WHY
PHILANTHROPY
IS FAILING
DEMOCRACY AND
HOW IT CAN
DO BETTER
ROB REICH

"SPELLS OUT WHAT NEEDS TO BE FIXED."
—Wall Street Journal

"READ THIS."**
—Reed Hastings, CEO of Netflix

"PROFOUNDLY IMPORTANT."***
—Dr. Fei-Fei Li

"BRILLIANTLY CRAFTED."

—Evan Spiegel, CEO of SnapChat

"A TRIUMPH."

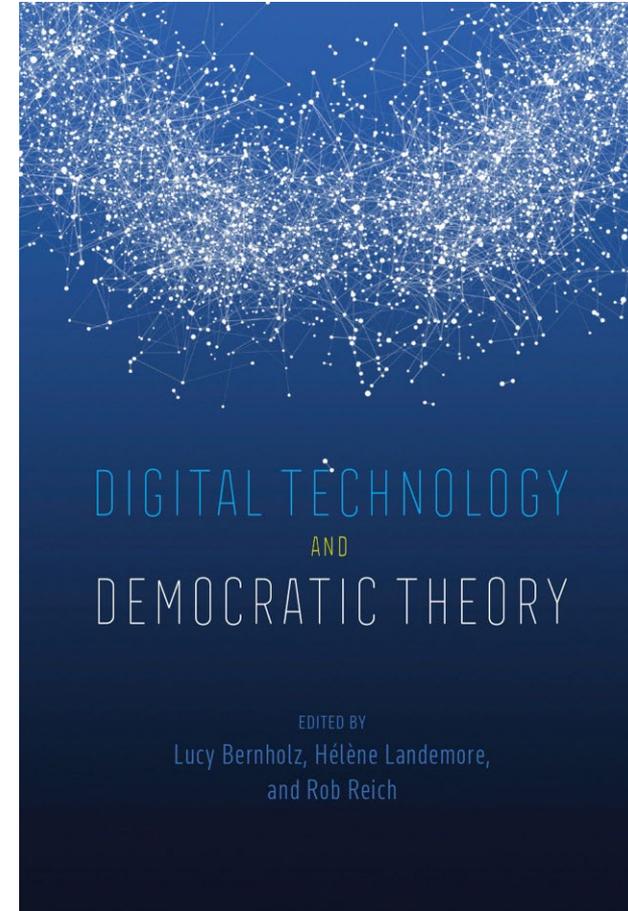
—Anne-Marie Slaughter

SYSTEM

ERROR

**Where Big
Tech Went
Wrong and
How We
Can Reboot**

Rob Reich
Mehran Sahami
**Jeremy
M. Weinstein**





Stanford University
Human-Centered
Artificial Intelligence

Search this site



[About](#) ▾ [Centers](#) ▾ [Research](#) ▾ [Education](#) ▾ [Policy](#) ▾ [News](#) ▾ [Events](#) ▾

Advancing AI research, education, and policy to improve the human condition

 [Sheng Wang | Generative AI for Multimodal Biomedicine](#)

Xiv:2108.07258v2 [cs.LG] 18 Aug 2021

On the Opportunities and Risks of Foundation Models

Rishi Bommasani* Drew A. Hudson Ehsan Adeli Russ Altman Simran Arora
Sydney von Arx Michael S. Bernstein Jeannette Bohg Antoine Bosselut Emma Brunskill
Erik Brynjolfsson Shyamal Buch Dallas Card Rodrigo Castellon Niladri Chatterji
Annie Chen Kathleen Creel Jared Quincy Davis Dorottya Demszky Chris Donahue
Moussa Doumbouya Esin Durmus Stefano Ermon John Etchemendy Kawin Ethayarajh
Li Fei-Fei Chelsea Finn Trevor Gale Lauren Gillespie Karan Goel Noah Goodman
Shelby Grossman Neel Guha Tatsunori Hashimoto Peter Henderson John Hewitt
Daniel E. Ho Jenny Hong Kyle Hsu Jing Huang Thomas Icard Saahil Jain
Dan Jurafsky Pratyusha Kalluri Siddharth Karamcheti Geoff Keeling Fereshte Khani
Omar Khattab Pang Wei Koh Mark Krass Ranjay Krishna Rohith Kuditipudi
Ananya Kumar Faisal Ladhak Mina Lee Tony Lee Jure Leskovec Isabelle Levent
Xiang Lisa Li Xuechen Li Tengyu Ma Ali Malik Christopher D. Manning
Suvir Mirchandani Eric Mitchell Zanele Munyikwa Suraj Nair Avani Narayan
Deepak Narayanan Ben Newman Allen Nie Juan Carlos Niebles Hamed Nilforoshan
Julian Nyarko Giray Ogut Laurel Orr Isabel Papadimitriou Joon Sung Park Chris Piech
Eva Portelance Christopher Potts Aditi Raghunathan Rob Reich Hongyu Ren
Frieda Rong Yusuf Roohani Camilo Ruiz Jack Ryan Christopher Re Dorsa Sadigh
Shiori Sagawa Keshav Santhanam Andy Shih Krishnan Srinivasan Alex Tamkin
Rohan Taori Armin W. Thomas Florian Tramèr Rose E. Wang William Wang Bohan Wu
Jiajun Wu Yuhuai Wu Sang Michael Xie Michihiro Yasunaga Jiaxuan You Matei Zaharia
Michael Zhang Tianyi Zhang Xikun Zhang Yuhui Zhang Lucia Zheng Kaitlyn Zhou
Percy Liang*¹

Center for Research on Foundation Models (CRFM)
Stanford Institute for Human-Centered Artificial Intelligence (HAI)
Stanford University

AI is undergoing a paradigm shift with the rise of models (e.g., BERT, DALL-E, GPT-3) that are trained on broad data at scale and are adaptable to a wide range of downstream tasks. We call these

U.S. ARTIFICIAL INTELLIGENCE SAFETY INSTITUTE

Strategic Vision

Guidelines

Artificial Intelligence Safety Institute Consortium

NIST AI Engagement

AI @ NIST

+

The U.S. AI Safety's Institute's (US AISI) mission is to identify, measure, and mitigate the risks of advanced AI systems so that we can harness the enormous potential of this breakthrough technology. US AISI is tasked with developing the testing, evaluations, and guidelines that will help accelerate trustworthy AI innovation in the United States and around the world – with a keen focus on helping to prevent misuse of this technology by those who seek to undermine our public safety and national security.

US AISI is housed within the Commerce Department as a part of NIST and draws on NIST's time-tested scientifically grounded processes to facilitate the development of trusted standards around new technologies.

NEWS AND UPDATES



Technical Blog: Strengthening AI Agent Hijacking Evaluations

JANUARY 17, 2025

Large AI models are increasingly used to power agentic systems, or “agents,” which can automate complex tasks

Transluce



Jacob Steinhardt



Sarah Schwettmann



Tiffany Tzeng



Kevin Meng



Kaiying Hou



Dami Choi



Neil Chowdhury



Daniel Johnson



Vincent Huang



Nitarshan Rajkumar



Rob Reich

Transluce is building the public tech stack for understanding and governing AI, in order to ensure reliable and safe rollouts of frontier AI systems.

transluce.org



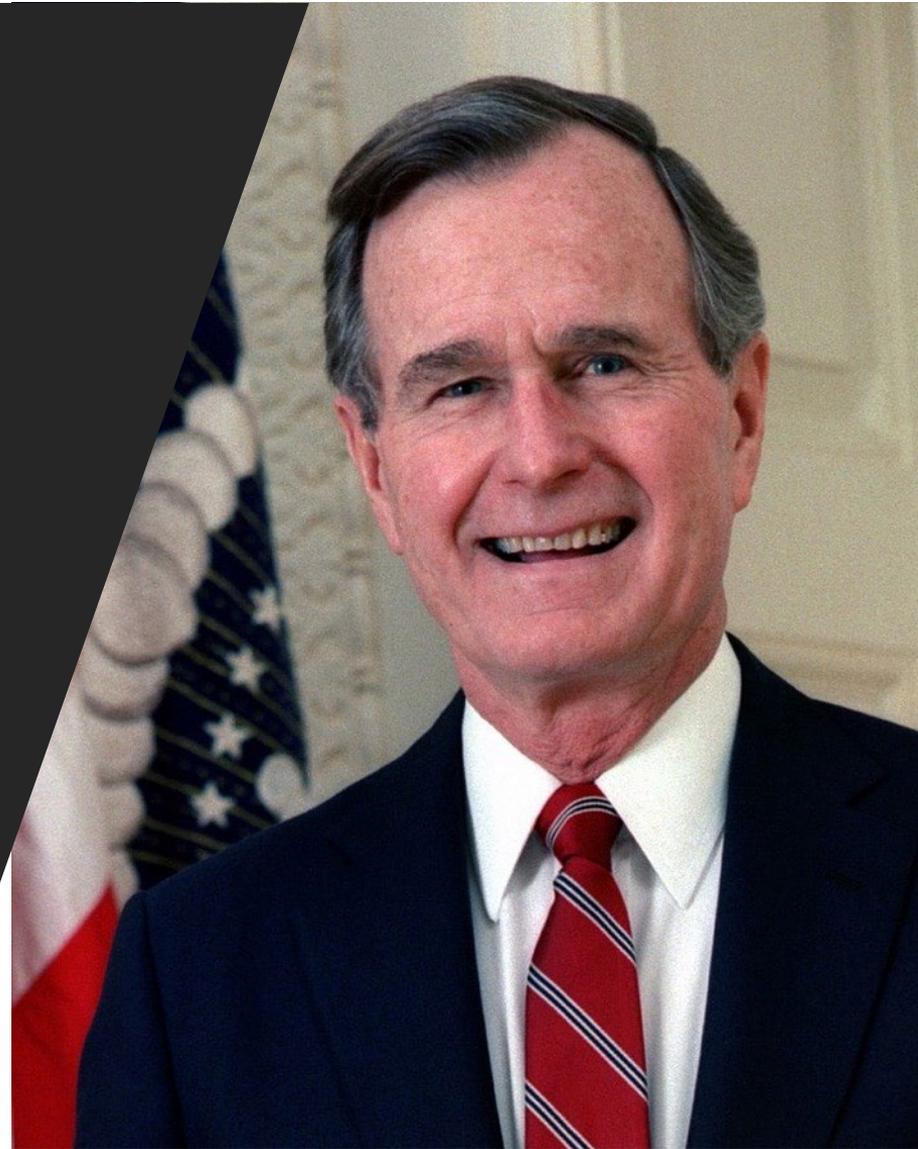
U.S. President Ronald Reagan, 1981-1989

“The Goliath of totalitarianism will be brought down by the David of the microchip.”

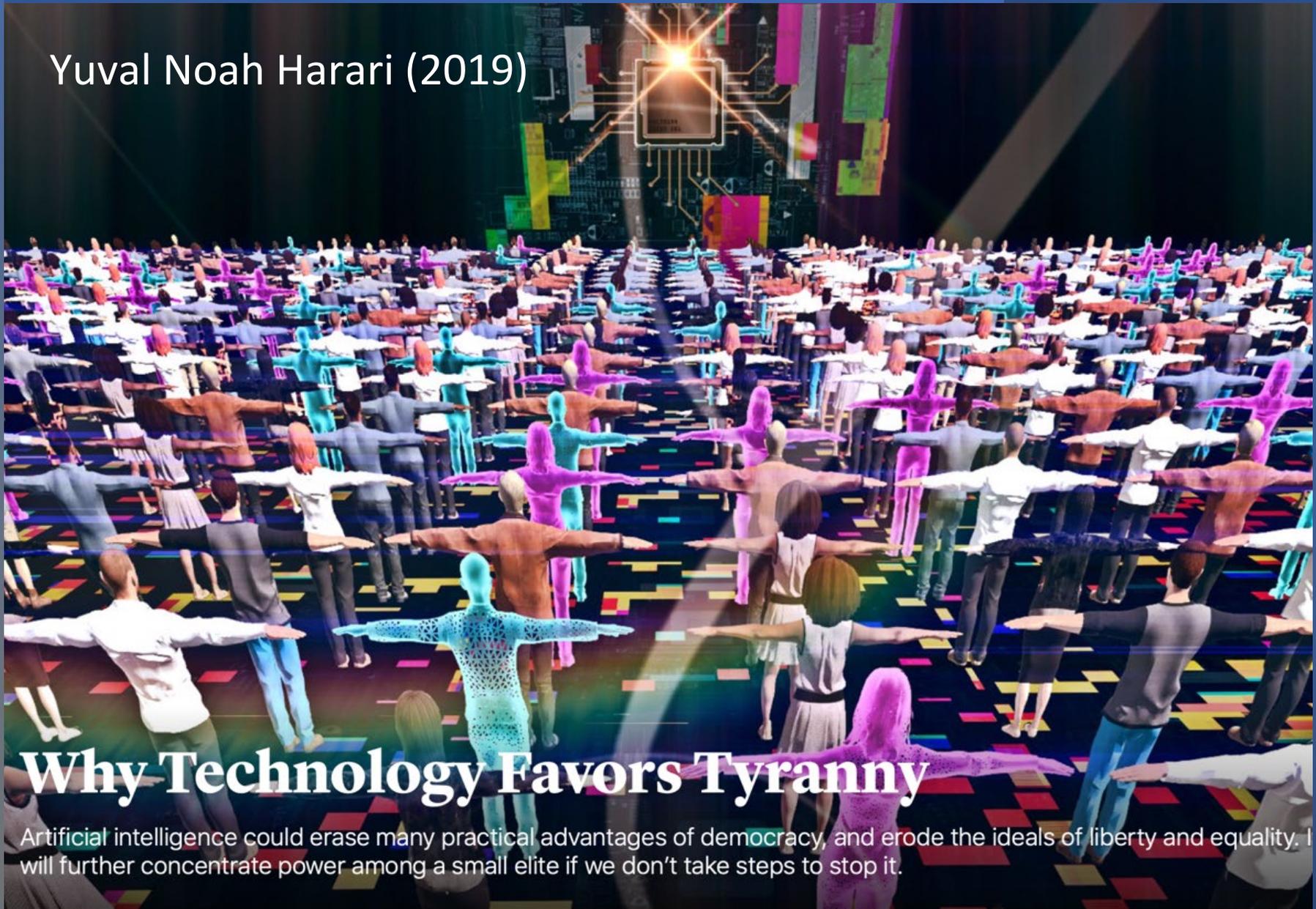


U.S. President George Bush, 1989- 1993

“Imagine if the Internet took hold in China. Imagine how freedom would spread.”



Yuval Noah Harari (2019)



Why Technology Favors Tyranny

Artificial intelligence could erase many practical advantages of democracy, and erode the ideals of liberty and equality. It will further concentrate power among a small elite if we don't take steps to stop it.

Is Big Tech Rotten?

“Ms. Brown said a lot of students criticize Facebook and talk about how they would not work there, but ultimately join. “Everyone cares about ethics in tech before they get a contract,” she said.”

The New York Times

‘I Don’t Really Want to Work for Facebook.’ So Say Some Computer Science Students.



The Cal Hacks 5.0 competition drew students to the University of California, Berkeley, including, from left, Haitao Zhang, Ingrid Wu and Emily Hu, all students at Berkeley. Some students at the hackathon expressed a reluctance to work for big tech firms. Max Whittaker for The New York Times

By Nellie Bowles

Nov. 15, 2018

Source: The New York Times

Finance

What Can We Learn from the Downfall of Theranos?

The health company's plummet carries valuable lessons for Silicon Valley.

December 17, 2018 | by Sachin Waikar



Theranos founder Elizabeth Holmes epitomized Steve Jobs, which attracted Silicon Valley investors who didn't look too closely at the health company's claims, says John Carreyrou, the Wall Street Journal reporter who investigated Theranos. | Reuters/Brendan McDermid

One of the most epic failures in corporate governance in the annals of American capitalism.





SBF 



you said a lot of stuff about how you wanted to make regulations, just good ones - was that pretty much just PR too?

Yesterday, 10:07 PM

there's no one really out there making sure good things happen and bad things don't



Yesterday, 10:07 PM

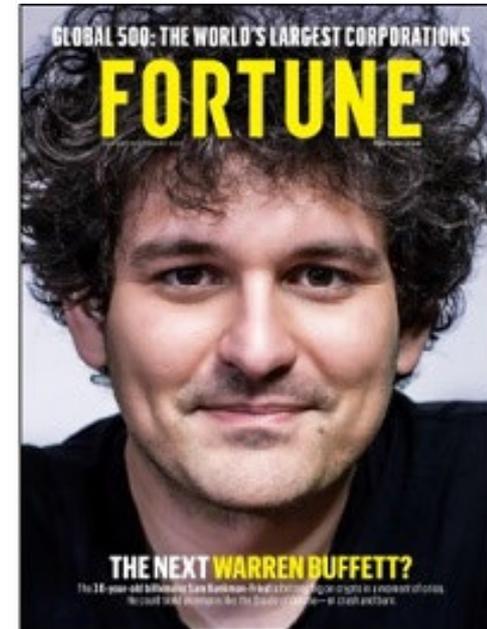
usually there's only one toggle--do more or do less

yeah just PR

fuck regulators

they make everything worse

they don't protect customers at all



Sam Bankman-Fried
Source: Fortune Magazine

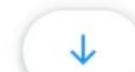
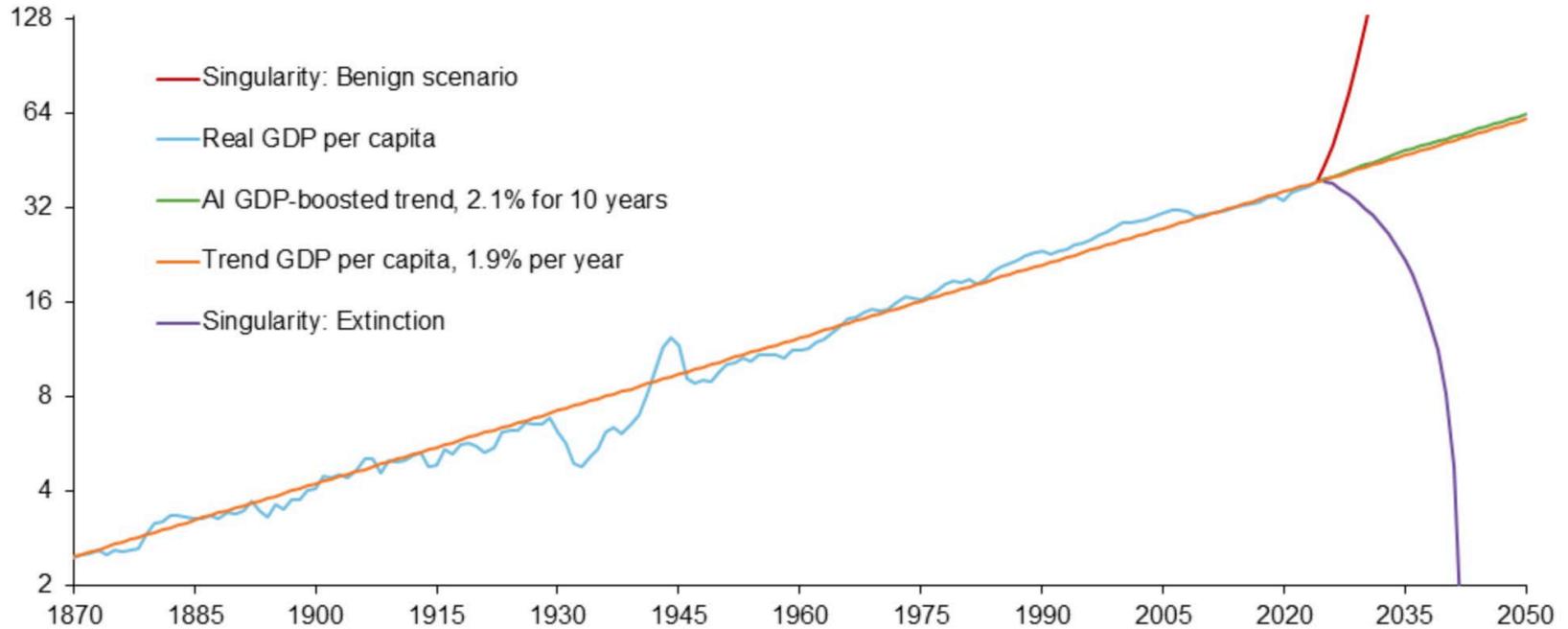


Chart 1 AI scenarios

1990 dollars (thousands), log scale



NOTES: The blue line is real gross domestic product (GDP) per capita in 1990 dollars. The orange line is a trend line fitted to the data for 1870–2024 with a trend growth rate of 1.9 percent per year. The red, green and purple lines are hypothetical paths for per capita GDP based on different scenarios.

SOURCES: Bureau of Economic Analysis; Haver Analytics; MacroeconomicHistory.net; United Nations; authors' calculations.

Federal Reserve Bank of Dallas

Today's Agenda

1. Why we are teaching this course
 2. **Why are you interested in taking it?**
 3. Why you *should* take this course
 4. What we are going to do together this quarter
-

Your Motivation

Why are you interested in taking this class?

- Think about a moment when you saw/experienced/learned about/anticipated the impacts of a new technology that gave you pause about the costs to society of this technological change.
 - What was the technology?
 - What were your concerns?
 - Do these concerns outweigh the benefits? Why?
 - Type up one paragraph sharing your thoughts and submit it now. There is a form to submit your write-up on the class website: <https://web.stanford.edu/class/cs182/>
-

Today's Agenda

1. Why we are teaching this course
 2. Why are you interested in taking it?
 3. **Why you *should* take this course**
 4. What we are going to do together this quarter
-

This Moment

- We are experiencing the societal consequences of many new technologies
 - These consequences are raising critical questions about how these technologies are designed, and whether and how new technologies should be governed and by whom
 - You have a role to play in answering these questions – as an engineer, a corporate executive, a policymaker, a citizen, or simply a user
-

Central Themes

- The impacts of technology are not fixed. They reflect a set of “design” choices. Those design choices encode a set of values.
 - The impacts also reflect choices about what policies and regulations society chooses to put in place.
 - When competing values are at stake, they must be weighed against one another. Who weighs these values and how? These are critical questions of governance, politics, and power.
 - Going forward, you will be a central participant in this drama. Understanding your role(s) and exploring/debating the values you want to see encoded are a modern form of civic duty.
-

Navigating the Moment

We want you to prepare you for this moment by:

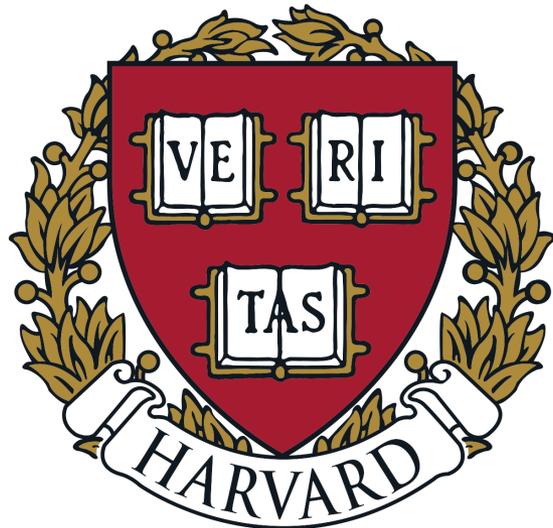
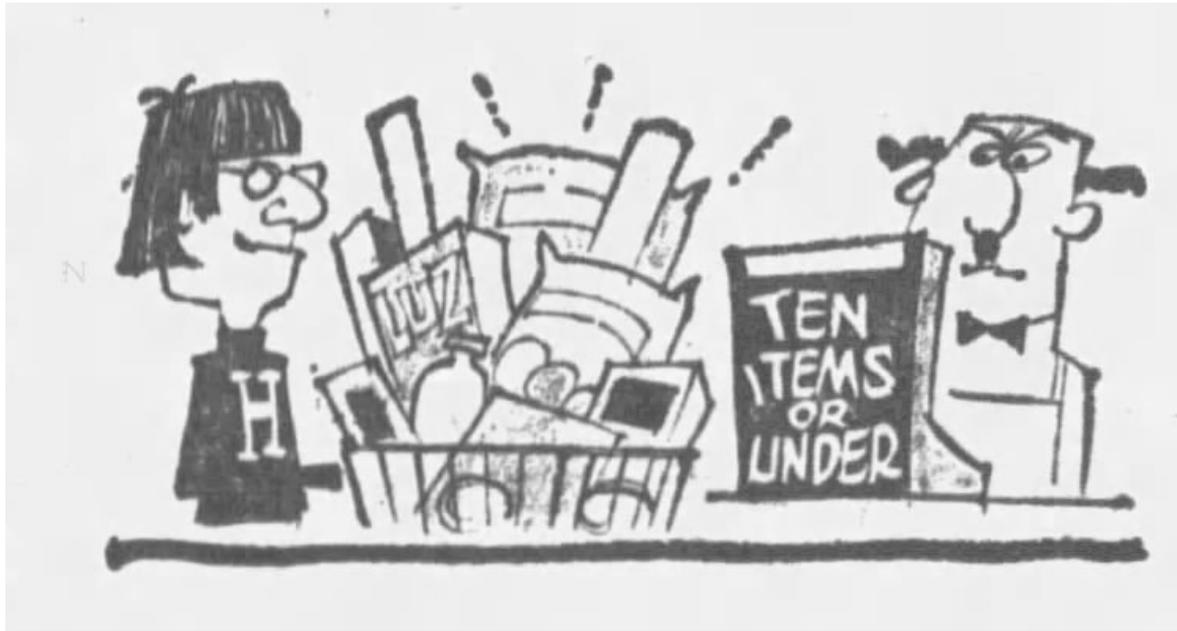
- Exploring technological frontiers that surface difficult trade-offs and require us to grapple with competing values
 - Making those competing values explicit and thinking about why we might prioritize some over others
 - Investigating the underlying technologies to understand how design choices can produce different outcomes
 - Thinking hard about how we should choose the values we want new technologies to encode
 - Grappling with the role of regulation and policy in mitigating the potential harms of new technologies
-

Today's Agenda

1. Why we are teaching this course
 2. Why are you interested in taking it?
 3. Why you *should* take this course
 4. **What we are going to do together this quarter**
-

Expectation Setting

- We are going to ask you to read and write! These will be key ways in which you gain familiarity with the ethical and policy dimensions of new technologies.
 - We will encourage you to share your views in lecture and discussion, and you can expect that we will challenge you in an effort to sharpen your views.
 - You will leave the class with more questions than answers, because the issues we are tackling do not have a right answer.
-



**Massachusetts
Institute of
Technology**

Modules

The course will focus on four frontiers that (a) you are likely to play a role in shaping over the next decade and (b) where engagement with material from philosophy, social science, and public policy is likely to be helpful.

1. Algorithmic Decision-Making
 2. Data Collection, Privacy, and Civil Liberties
 3. Power of Private Platforms
 4. Generative AI
-

Three Lenses and Course Design



Technologist

Philosopher



Policymaker

Course Design

Four Modules

Criminal risk
assessment

Facial
recognition

Social media

Generative
AI

Course Design

Four Modules



Criminal risk
assessment



Clearview.ai

Facial
recognition



Social media



Generative
AI

Course Design

Four Modules

Ethics, values, and rights

Technical approaches

Policy interventions



Criminal risk assessment



Clearview.ai

Facial recognition



Social media



Generative AI

Course Design

Four Modules

Ethics, values, and rights

Technical approaches

Policy interventions



Criminal risk assessment

Fairness



Clearview.ai

Facial recognition



Social media



Generative AI

Course Design

Four Modules

Ethics, values, and rights

Technical approaches

Policy interventions

Criminal risk assessment

Fairness

Algorithmic fairness

Due process, equal protection



Clearview.ai

Facial recognition



Social media



Generative AI

Course Design

Four Modules

Ethics, values, and rights

Technical approaches

Policy interventions



Criminal risk assessment

Fairness

Algorithmic fairness

Due process, equal protection



Facial recognition

Privacy

Differential privacy

California Consumer Privacy Act



Social media

Speech

Content moderation

Communications Decency Act Reform



Generative AI

Openness

Benchmarking

California SB 1047

Themes

- Promise and Perils
 - Technical Solutions
 - Rights and Responsibilities
 - Tensions and Trade-offs
 - Distributive Effects
 - Making Choices
-

Tensions and Trade-Offs

Each discussion of tensions and trade-offs will be organized around a case study we have developed specially for this course.

Narrative case studies include primary source materials for you to review.

Case study discussions will be highly participatory and will take place during your weekly discussion section.

Lecture

- Attendance at lectures and sections is mandatory.
 - Lectures will be highly participatory. We will ask you questions, invite your questions, encourage you to talk to your neighbors, and think through the critical issues together.
 - If you are unable to attend a lecture because of COVID or another health emergency, please email your teaching assistant to stay on track.
-

Section

- We have a terrific, interdisciplinary team of teaching assistants from computer science, philosophy, political science, law, and sociology.
 - They will meet with you in small groups once a week to discuss critical issues raised in lecture and the readings.
 - Section attendance is mandatory and active participation is essential to success in the course.
 - You will receive section assignments via email. You do not need to sign up for sections on Axess.
-

WIM

- There is an enrollment limit of 100 students for CS182W
 - This limit is set by the Technical Communications Program based on how many WIM students they can support
 - If CS182W reaches enrollment limit, waitlist will be activated on Axxess
 - If any enrolled students drop CS182W, others will be enrolled in the class from the waitlist (in the order they signed up)
 - You can add yourself to the CS182W waitlist on Axxess if you are not already enrolled and are trying to get into CS182W
-

Your Role

- Come to lecture and section having done and digested the readings
 - Engage actively in discussion
 - Complete the five required assignments
 - Algorithmic Decision-Making (Technical)
 - Privacy (Essay)
 - Generative AI (Policy Memo)
 - Platforms and Social Networks (Technical)
 - Final Reflection Assignment
-

Agency

Building a National AI Research Resource:
A Blueprint for the National Research Cloud

WHITE PAPER

Daniel E. Ho
Jennifer King
Russell C. Wald
Christopher Wan

OCTOBER 2021

HAI
Stanford University
Human-Centered
Artificial Intelligence

Stanford
LawSchool

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF MANAGEMENT AND BUDGET
WASHINGTON, D.C. 20503

THE DIRECTOR

September 24, 2024

M-24-18

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: Shalanda D. Young *Shalanda D. Young*

SUBJECT: Advancing the Responsible Acquisition of Artificial Intelligence in Government

I. OVERVIEW

The use of artificial intelligence (AI) in the Federal Government presents tremendous opportunity for modernizing agency operations and improving the delivery of government services to the public, provided that the risks presented by the use of AI technology are mitigated. Realizing this goal involves recognizing that AI poses novel types of risk, and proactively integrating considerations for AI risk management into agency acquisition planning. This memorandum builds on previous efforts to harness the power and utility of AI in service of agency missions while protecting the public from potential risks or harms.

Consistent with the Advancing American AI Act ("the Act"),¹ Executive Order 14110, *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, and Office of Management and Budget (OMB) Memorandum M-24-10, *Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence*,² this memorandum directs agencies to improve their capacity for the responsible acquisition of AI. It contains new requirements and guidance for agencies on establishing meaningful cross-functional and interagency collaboration to reflect new AI responsibilities, managing AI risk and performance, and promoting a competitive AI market with innovative acquisition. As required by the Act, interagency consultation was conducted throughout the development of this memorandum.

Ensuring Cross-functional and Interagency Collaboration. Cross-functional collaboration throughout the acquisition lifecycle has long been a foundational principle of the Government's acquisition practices, and is no less critical for the acquisition of AI. This memorandum requires agencies to create or update acquisition policies, procedures, and practices to reflect new responsibilities and governance for AI, as established by OMB

¹ Pub. L. No. 117-263, div. G, title LXXII, subtitle B, § 7224(d)(1) (codified at 40 U.S.C. 11301 note), <https://www.congress.gov/117/plaws/publ263/plaw-117/publ263.pdf>.

² OMB Memorandum M-24-10, *Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence* (March 28, 2024), <https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf>.

1

Stanford LawSchool
Law and Policy Lab
Policy Practicum: Creating a Social Media Oversight Board

Recommendations for the Facebook Content Review Board

2018-19 PRACTICUM RESEARCH TEAM:

Shaima BAKR, Ph.D. Electrical Engineering '20	Madeline MAGNUSON, J.D. '20
Fernando BERDION-DEL VALLE, J.D. '20	Shawn MUSGRAVE, J.D. '21
Isabella GARCIA-CAMARGO, B.S. '20	Ashwin RAMASWAMI, B.S. '21
Julia GREENBERG, J.D. '19	Nora TAN, B.S. '19
Tara IYER, B.S. '19	Marlena WISNIAK, LL.M. '19
Alejandra LYNNBERG, J.D. '19	Monica ZWOLINSKI, J.D. '19

INSTRUCTORS:

Paul BREST, Faculty Director, Law and Policy Lab
Daniel HO, William Benjamin Scott and Luna M. Scott Professor of Law
Nathaniel PERSILY, James B. McClatchy Professor of Law
Rob REICH, Faculty Director, Center for Ethics in Society

TEACHING ASSISTANT:

Liza STARR, J.D./M.B.A. '21

POLICY CLIENT:

Project on Democracy and the Internet, Stanford Center on Philanthropy and Civil Society

SPRING 2019

559 Nathan Abbot Way Stanford, CA <https://law.stanford.edu/education/only-at-stsls-law-policy-lab/>