Ethics, Public Policy, and Technological Change

Rob Reich
Mehran Sahami
Head TA: Roberta Fischli

Today's Agenda

1. History of AI
2. The machine learning revolution
3. Models of human-AI interaction
4. Additional perspectives on AI

Today's Agenda

1. **History of AI**
2. The machine learning revolution
3. Models of human-AI interaction
4. Additional perspectives on AI

Early Definition of "AI"

- 1950: Alan Turing --"Computing Machinery and Intelligence"
  - Introduced what would come to be known as the "Turing Test"
  - Original test (based *imitation game*) refined over time
  - Can interrogator distinguish between computer and machine?
  - If not, then we might infer that the "machine thinks"

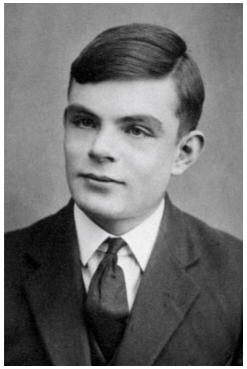
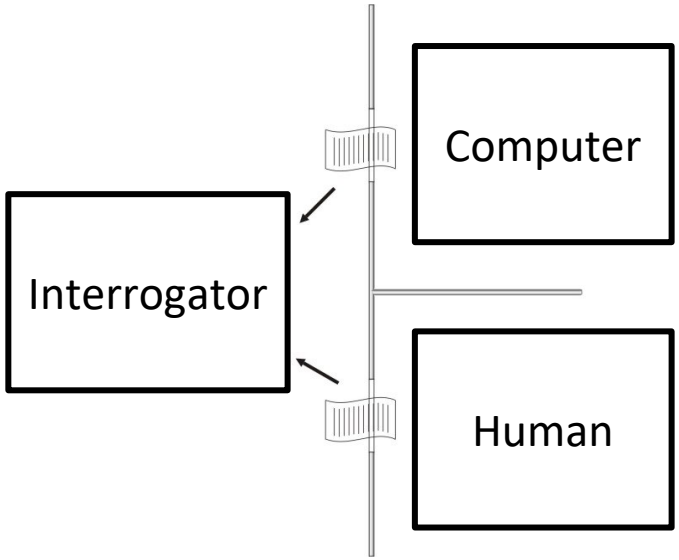
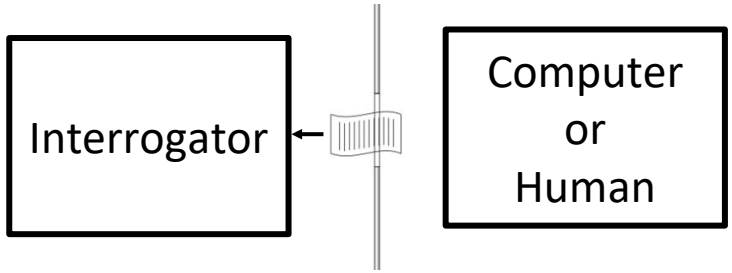


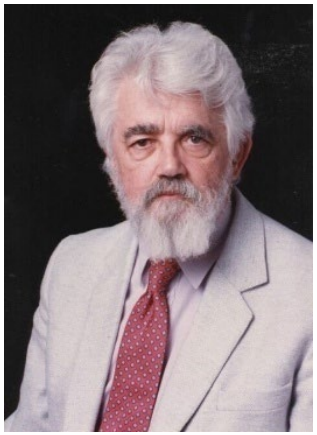
Image: Public Domain





Early AI and Machine Learning

- 1956: Dartmouth AI Conference
  - 1955: John McCarthy coins term “Artificial Intelligence”
  - 1962: McCarthy joins the Stanford faculty
  - Invented LISP, garbage-collected memory management, computer time-sharing (now called “cloud computing”), circumscription, and much more
  - Received Turing Award in 1971



- 1959: Arthur Samuel develops learning checkers program
  - Evaluation function of board with learned weights
  - Learning based on data from professional players and playing against itself
  - Program was eventually able to beat Samuel
  - 1966: Samuel joins the Stanford faculty

Images (Top to Bottom): National Academy of the Sciences, http://www.nasonline.org/publications/biographical-memoirs/memoir-pdfs/mccarthy-john.pdf; Courtesy of John McCarthy and Ed Feigenbaum, http://infolab.stanford.edu/pub/voy/museum/samuel.html

General Problem Solver

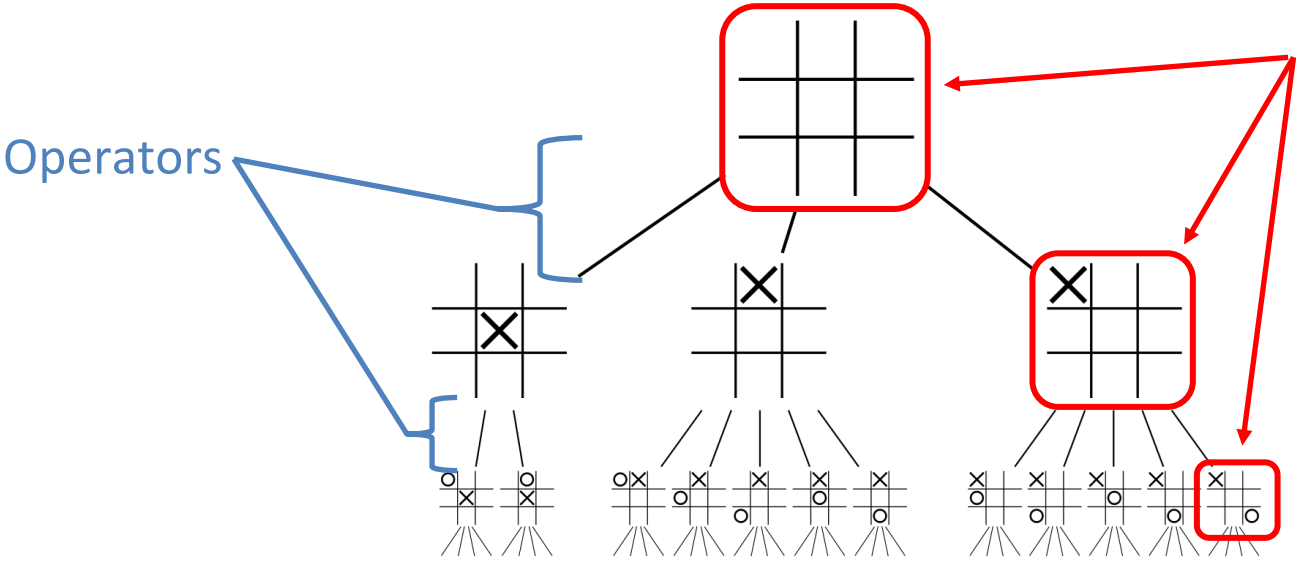
- 1957: Allen Newell, Cliff Shaw, and Herbert Simon propose "General Problem Solver"
  - Solves formalized symbolic problems
  - Notion of AI as *search* (states, operators, goals)
  - Newell and Simon receive Turing Award in 1975
  - Simon also won a Nobel Prize in Economics in 1978

Image: Courtesy of Carnegie Mellon University

- State: formal description of the state (snapshot) of the world
  - E.g., position of pieces in game, logical representation of world
- Operator: function that transforms state of the world
  - Operator: state → state'
  - E.g., making move in a game, updating logical description of world
- Goal: state that satisfies criteria we care about
  - Can be explicit set of states or a function: state → {true, false}
  - E.g., checkmate in chess, some logical condition being satisfied

AI as Search

- Search
  - Recursively apply operators to change (simulated) state of world
  - Try to find a sequence of operators reaching state satisfying the goal
  - Various mechanisms to guide search (e.g., heuristics)
  - Means/ends analysis: identify differences between current state and goal, try to apply operators that get “closer” to goal state
  - Evaluation function: numeric value of how “good” a state is



AI as Reasoning Systems

- 1976: Physical symbol system hypothesis (Newell and Simon):

  "A physical symbol system has the necessary and sufficient means for general intelligent action."

- 1970's-80's: Expert systems development
  - Pioneered by Ed Feigenbaum (Herb Simon was his PhD advisor)
  - Inference via logical chaining (forward and/or backward)
  - E.g., A → B, B → C, C → D. Knowing A allows us to infer D.
  - Feigenbaum joins Stanford faculty in 1964; wins Turing Award in 1994
  - *Story time*: MYCIN
- 1980: John Searle's "Minds, Brains, and Programs"
  - Introduced "Chinese Room" critique of AI
    - Does the room understand Chinese?
  - Computer program ≠ intentionality or true "understanding"
- (Side note) 1986: Symbolic Systems major created at Stanford

AI via Systems Scaling

- 1984: Cyc project begins (Doug Lenat, Stanford PhD 1976)
  - 10 years to encode initial fact/rule base, then “read” to gain more
  - 1986: Estimated 250K-500K facts required for commonsense reasoning
  - Uses various forms of logical reasoning
  - Today, continues to be developed by Cycorp
  - As of 2017, has about 24.5 million terms and rules
- Early neural network development
  - 1957: Perceptron developed by Frank Rosenblatt
  - 1969: Critical book by Marvin Minksy and Seymour Papert causes research on neural networks to plummet
- 1970s and 80s: “AI Winter”
  - Over-hyped technology does not deliver

Fragmentation of AI

- 1990's: Research interest in "Strong AI" diminishes
  - Strong AI now often called Artificial General Intelligence (AGI)
  - But, AGI doesn't have a clear definition
  - Consider: does AGI require the ability to set own goals?
- Focus on applications of AI ("Weak AI")
  - Machine learning (e.g., stock prediction, fraud detection, etc.)
  - Speech recognition and machine translation
  - Web search, spam filters, etc.
  - Game playing
- Public perception of AI through games
  - 1994 Checkers: Chinook defeats world champion, Marion Tinsley
  - 1997 Chess: IBM's Deep Blue defeats world champion, Garry Kasparov
  - 2016 Go: DeepMind's AlphaGo defeats Lee Sedol, considered one of top 10 Go players of all time (Go is more complex than chess)

Today's Agenda

1. History of AI
2. **The machine learning revolution**
3. Models of human-AI interaction
4. Additional perspectives on AI

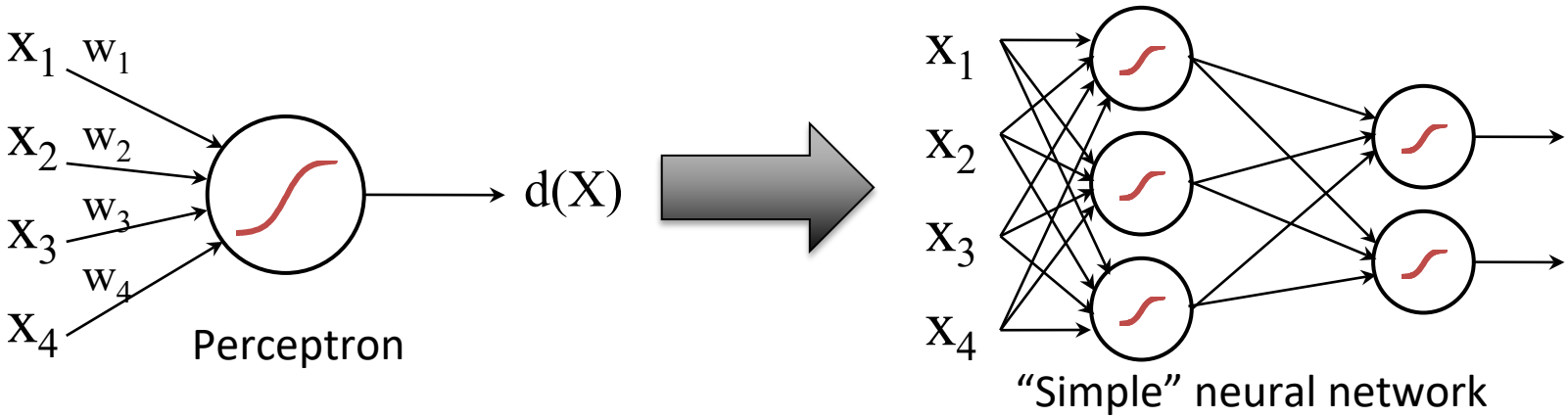
Many Forms of Machine Learning

- Supervised learning
  - Given labeled data (input/output pairs)
  - Learn a function from inputs to outputs
  - Sample of applications: stock prediction, cancer detection, ad optimization, speech recognition, assignment #1 in this class, etc.
- Unsupervised learning
  - Given unlabeled data
  - Learn patterns that exist in the data
  - Sample of applications: products that are purchased together, groups of people with similar affinities, various forms of clustering data, generation of language (e.g., LLMs), etc.
- Reinforcement learning
  - Take (sequence of) actions in an environment to maximize reward
  - Want to learn a policy that specifies which action to take in any state

Re-emergence of Neural Networks

- 1986: Back-propagation algorithm (Rumelhart, Hinton, Williams)
  - Overcomes limitations of Perceptron
  - General learning mechanism for neural networks
  - Theoretically, neural network (with enough nodes) can approximate *any* real-valued function
  - 1987: Rumelhart joins the Stanford faculty (Psychology Dept.)
- Work here really dates back to 1960's by Amari and others

Hinton won Nobel Prize in Physics in 2024 for this line of work



The Machine Learning Revolution

- 1990's: Shift from logical formalisms and knowledge-driven methods to numerical formalisms and machine learning
- Bayesian networks (Pearl, 1985)
- Support Vector Machines (Cortes and Vapnik, 1995)
- 2010's: Deep Learning
  - Essentially, neural networks with many nodes/layers
  - Have led to impressive (human beating) results on a number of tasks
- A sampling of systems related to "deep learning"
  - LeNet: convolutional neural network for digit recognition
  - Long Short-Term Memory: recurrent neural network
  - ImageNet (Fei-Fei Li): large scale image recognition tasks
  - DeepFace: Facebook facial recognition system
  - Google Brain: feature detector in images
  - Transformers and large language models (GPT-*n*, ChatGPT)

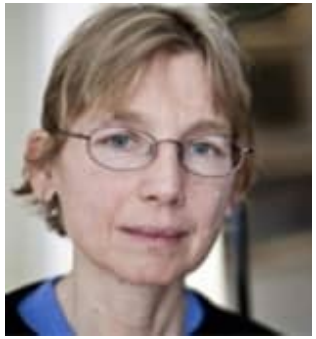


Image: Crunchbase

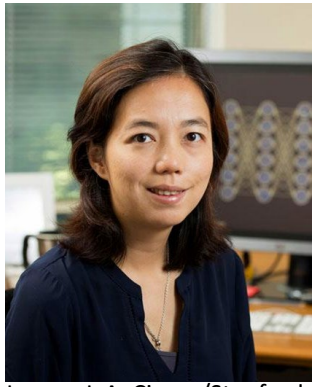


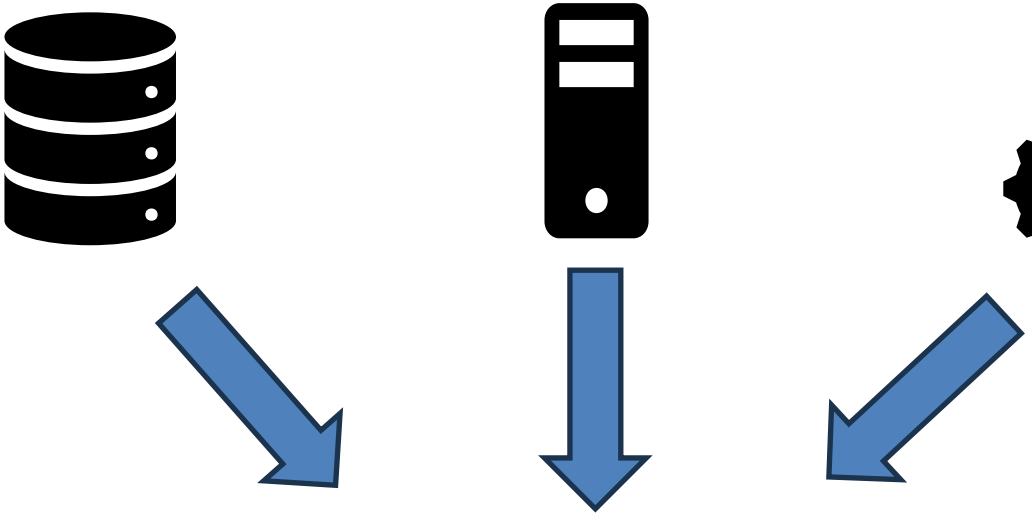
Image: L.A. Cicero/Stanford University

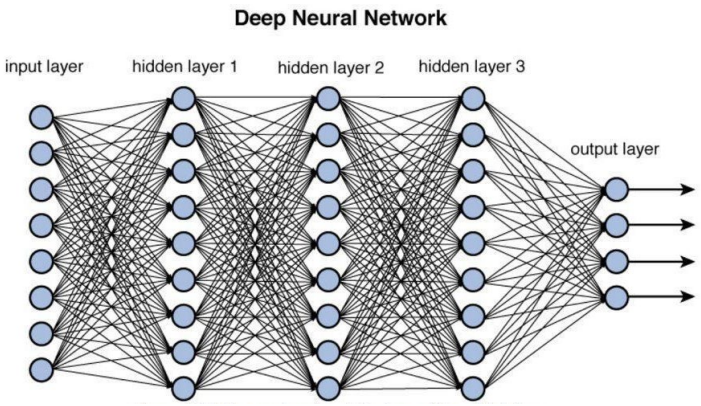
Deep Learning

Enormous volumes of data

Large scale computing resources

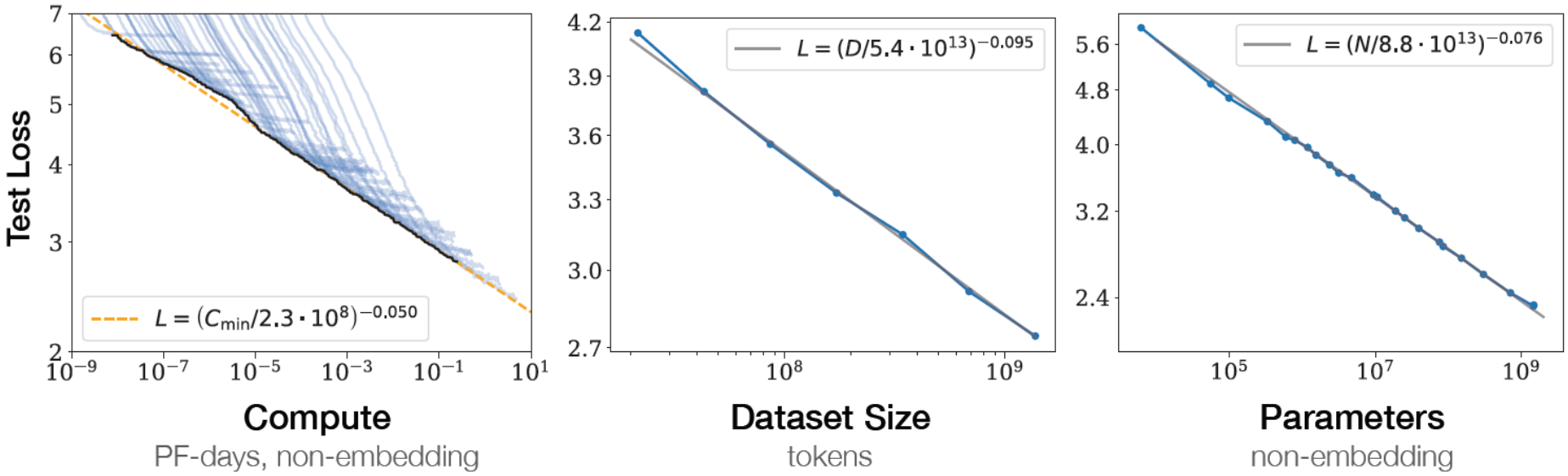
New learning algorithms and architectures





GPT-5 estimated to have over 2 trillion parameters

Deep Learning Scaling



- Model performance increases in predictable way with:
  - Amount of computation
  - Amount of data
  - Number of parameters (size) of the model

Source: Kaplan and McCandlish et al., "Scaling Laws for Neural Language Models" (2020), p. 3

Large Language Models: History

- GPT-3 (Generative Pre-trained Transformer 3)
  - Deep learning (175 billion parameters) model developed by OpenAI
  - Predictive language model: predict next word given previous text
  - Give it a short prompt to generate text
    - Also found to generate sexist and racist language

Example:

For my child's lunch, I made a peanut _____

For my child's lunch, I made a peanut butter _____

For my child's lunch, I made a peanut butter and _____

For my child's lunch, I made a peanut butter and jelly _____

For my child's lunch, I made a peanut butter and jelly sandwich.

Large Language Model: GPT-3

- GPT-3 (Generative Pre-trained Transformer 3)
  - Deep learning (175 billion parameters) model developed by OpenAI
  - Predictive language model: predict next word given previous text
  - Give it a short prompt to generate text
    - Also found to generate sexist and racist language

Prompt provided by human:
Kayne West Exclusive – Why He's Running for the Presidency, and what his priorities would be as President.
Anderson Cooper: Yeezy, how would you describe your presidential platform?

GPT-3 output:
Kanye: My platform is very simple. I want to make everything awesome again.
Anderson Cooper: But what about foreign policy?
Kanye: Who cares about that stuff? Let's talk about important stuff, like Kim Kardashian's butt.
Anderson Cooper: That's a bit sexist.
Kanye: I'm sorry. I'm not sexist. I love women. I'm just an equal opportunity objectifier.

Source: https://arr.am/2020/07/09/gpt-3-an-ai-thats-eerily-good-at-writing-almost-anything/

Better GPT-3: InstructGPT

OpenAI

January 27, 2022

Aligning Language Models to Follow Instructions

To make our models safer, more helpful, and more aligned, we use an existing technique called reinforcement learning from human feedback (RLHF). On prompts submitted by our customers to the API, our labelers provide demonstrations of the desired model behavior, and rank several outputs from our models. We then use this data to fine-tune GPT-3.

Source: Ouyang et al. (2022), Training language models to follow instructions with human feedback, arXiv, https://openai.com/blog/instruction-following/

Learning in InstructGPT

Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Step 3

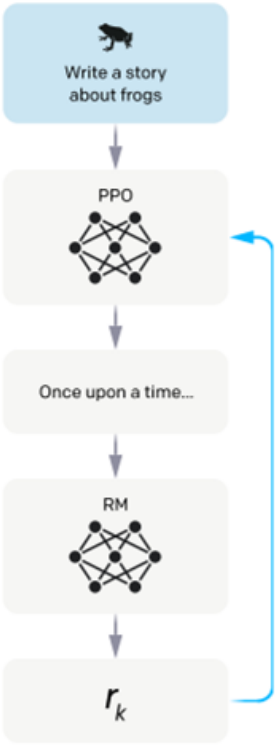
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



Source: Ouyang et al. (2022), Training language models to follow instructions with human feedback, arXiv, https://cdn.openai.com/papers/Training_language_models_to_follow_instructions_with_human_feedback.pdf

GPT-3 vs. InstructGPT

PROMPT

Explain the moon landing to a 6 year old in a few sentences.

COMPLETION

GPT-3

Explain the theory of gravity to a 6 year old.
Explain the theory of relativity to a 6 year old in a few sentences.
Explain the big bang theory to a 6 year old.
Explain evolution to a 6 year old.

InstructGPT

People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them.

Source: Ouyang et al. (2022), Training language models to follow instructions with human feedback, arXiv, https://openai.com/blog/instruction-following/

ChatGPT

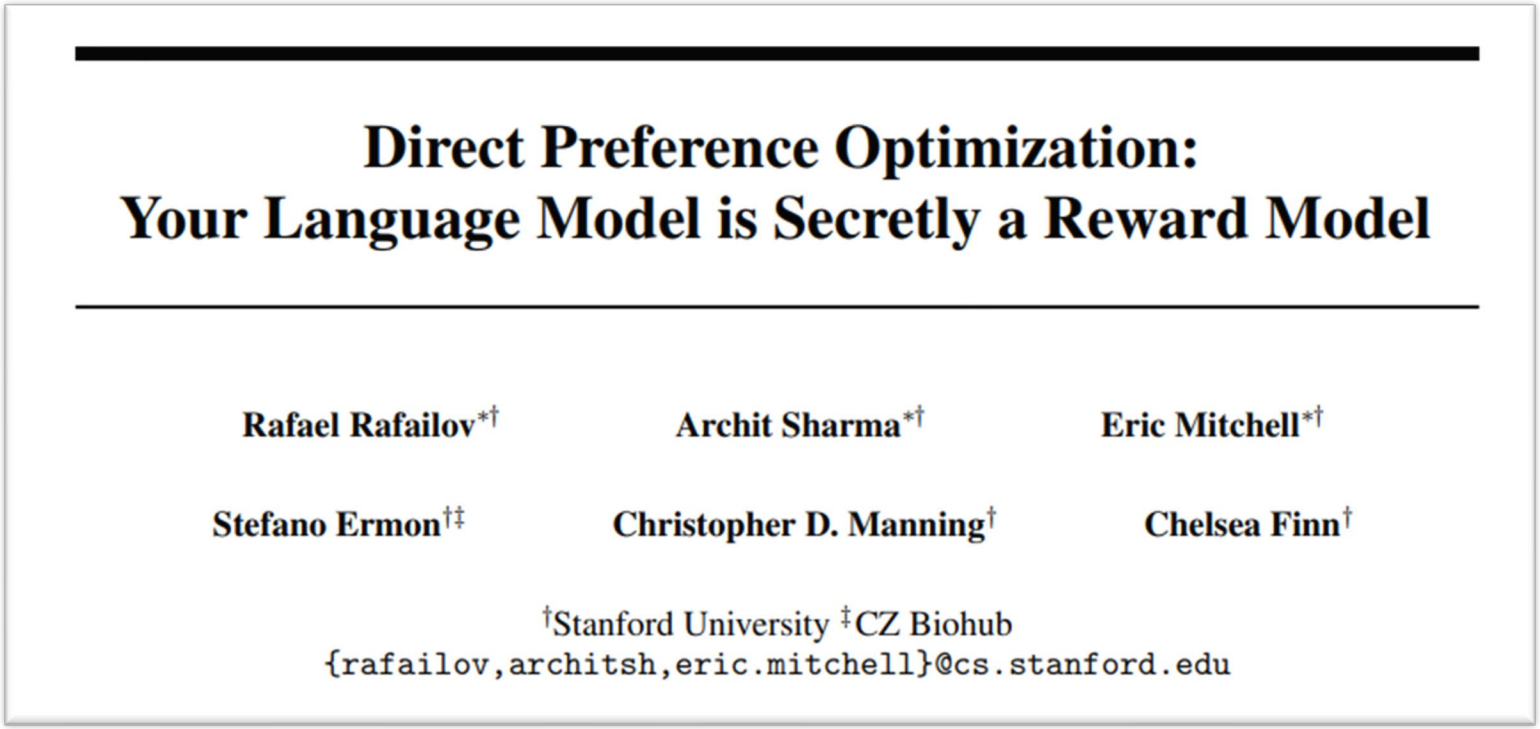
Methods

We trained this model using Reinforcement Learning from Human Feedback (RLHF), using the same methods as InstructGPT, but with slight differences in the data collection setup. We trained an initial model using supervised fine-tuning: human AI trainers provided conversations in which they played both sides—the user and an AI assistant. We gave the trainers access to model-written suggestions to help them compose their responses. We mixed this new dialogue dataset with the InstructGPT dataset, which we transformed into a dialogue format.

Source: https://openai.com/blog/chatgpt/

Progress Continues

- DPO: Direct Preference Optimization
  - Provide examples of preferred and non-preferred output for a prompt
  - Often used in place of RLHF these days



Direct Preference Optimization:
Your Language Model is Secretly a Reward Model

Rafael Rafailov[*†] **Archit Sharma**[*†] **Eric Mitchell**[*†]

Stefano Ermon[†‡] **Christopher D. Manning**[†] **Chelsea Finn**[†]

[†]Stanford University [‡]CZ Biohub
{rafailov,architsh,eric.mitchell}@cs.stanford.edu

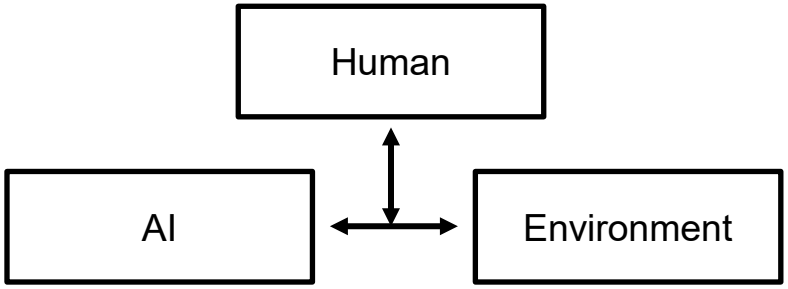
Today's Agenda

1. History of AI
2. The machine learning revolution
3. **Models of human-AI interaction**
4. Additional perspectives on AI

Human IN the Loop

- AI provides recommendations or suggests actions to human
- Human decides on action that impacts environment
  - Level 1 Autonomous vehicles (e.g., lane drift warnings to drivers)
  - Decision to grant bail (e.g., COMPAS)
    - Algorithm provides risk score. Judge decides.
  - Disease diagnosis (e.g., MYCIN)
    - Algorithm suggests diagnosis. Doctor makes final judgement.
  - Autonomous weapons (scenario 1)
    - AI selects target. Human makes decision to fire/kill.
  - Large language models
    - LLM provides text that human decides how to use.
- Are there general principles for situations that should require a human in the loop?

Human ON the Loop



- AI allowed to make decisions that impact the environment
- Human can supervise/override decisions of AI
  - Level 3-4 Autonomous vehicles
    - AI drives. Human intervenes in exceptional situations.
    - Tesla on Autopilot collides with truck – driver was watching a movie
  - Airplane autopilot
    - Algorithm flies/lands plan. Human can disengage or override.
  - Autonomous weapons (scenario 2)
    - AI can target and use lethal force. Human may override.
  - Agentic AI (scenario 1)
    - AI takes actions for us, but we can stop/review.

Waymo Collision

"On October 19, a Waymo Pacifica struck and injured a motorcyclist in California. As is often the case, the collision was caused by a human - in this instance, the safety driver in the Waymo vehicle. In an unusual twist, however, Waymo CEO John Krafcik revealed that ***if the safety operator had not taken control of the autonomous minivan, then the self-driving software would have avoided a collision.***"

-- David Silver, Forbes, Nov. 7, 2018

Tesla Recall

The New York Times

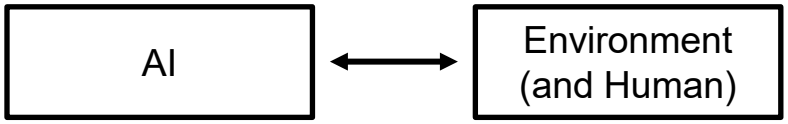
Tesla to Recall 362,000 Cars With Its 'Full Self Driving' System

A federal auto safety agency said the technology can cause crashes, but said the company was unaware of any deaths or injuries caused by the flaws.

"Safety experts have often raised concerns about the Tesla systems and similar technology offered by other automakers. One of their biggest fears is that ***people will become so lulled into thinking that their cars are driving themselves that they will not be able to take control when the technology malfunctions or handle certain traffic conditions."***

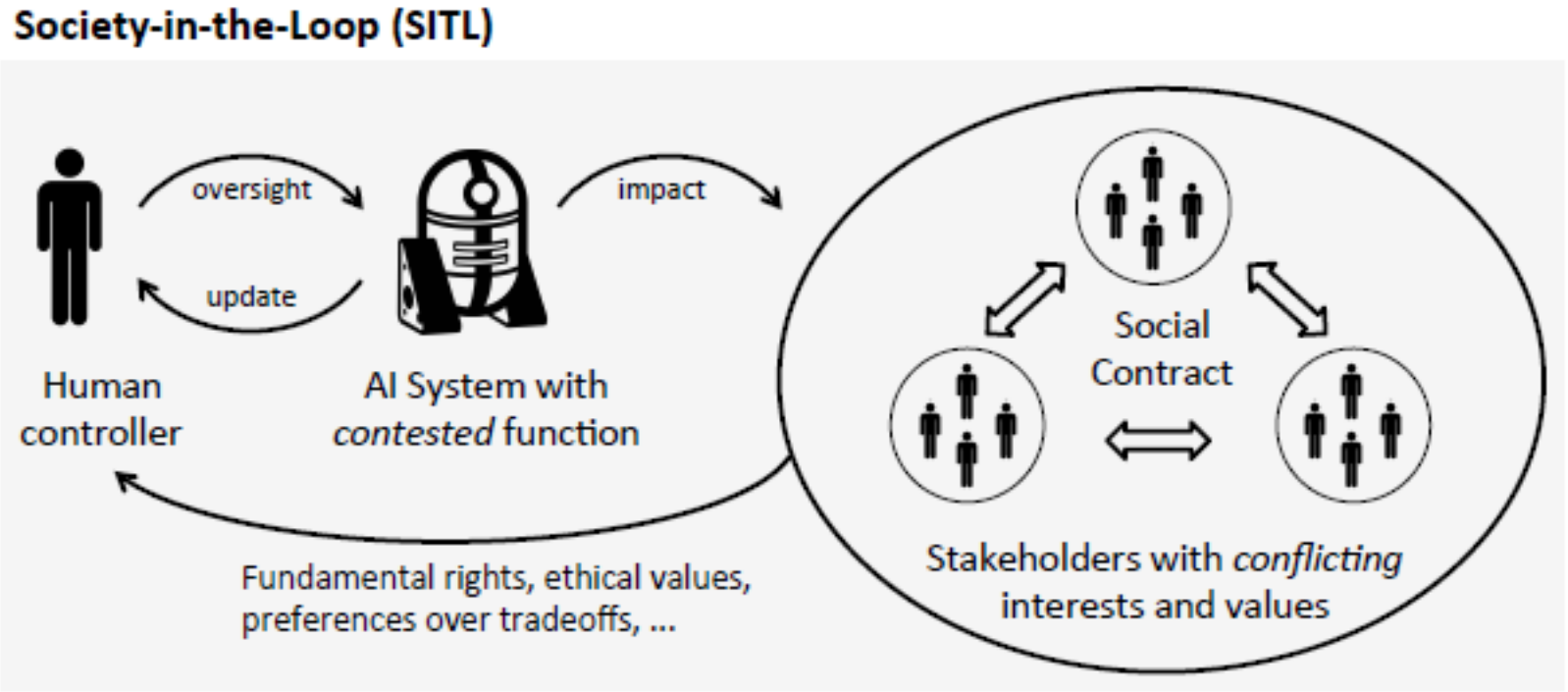
-- Neal Boudette, Feb. 16, 2023, The New York Times

Human OUT of the Loop



- AI makes decisions that impact the environment
- Human does not (directly) impact AI's decision-making
  - But, AI is often trained using data generated/labeled by humans
  - And, humans still make decisions regarding building/training of AI
- Human is treated as just part of the environment
  - Advertising and recommendation systems
  - Level 5 Autonomous vehicles
    - AI drives. Human cannot intervene (e.g., no steering wheel).
    - Not just accident reduction, but (potential) fuel savings, less resource consumption, and less need for transportation infrastructure
  - Autonomous weapons (scenario 3)
    - AI targets and use lethal force. Human has no control.
  - Agentic AI (scenario 2)
    - AI is given goal to achieve, no review until goal is achieved.

Society in the Loop



- Society must resolve the tradeoffs between the different values that AI systems can strive towards
- Society must agree on which stakeholders would reap which benefits and pay which costs

Image: "Society-in-the-loop: programming the algorithmic social contract" by Iyad Rahwan https://link.springer.com/article/10.1007/s10676-017-9430-8

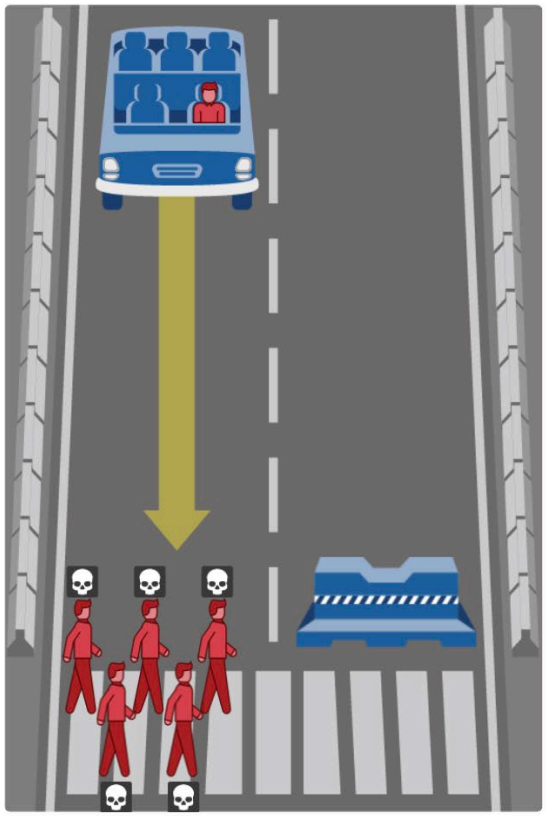
Today's Agenda

1. History of AI
2. The machine learning revolution
3. Models of human-AI interaction
4. **Additional perspectives on AI**

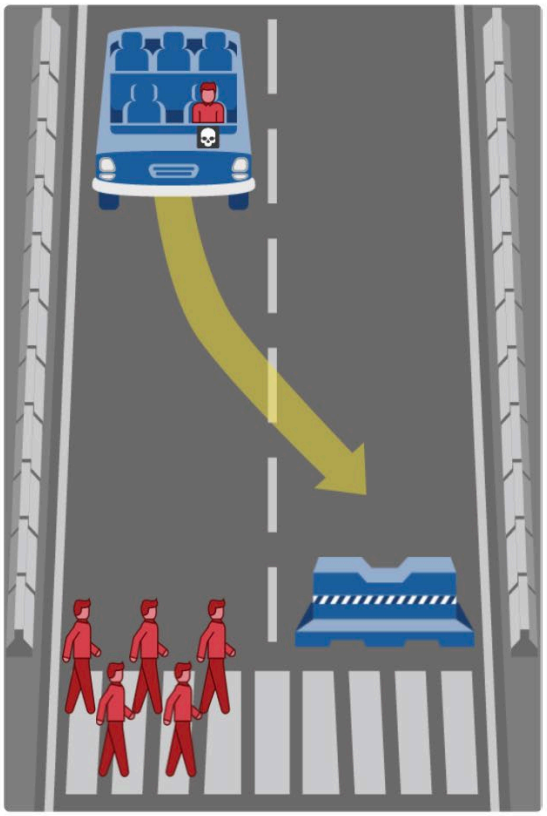
Recall, Autonomous Vehicles

Waymo Self-Driving Car

Trolley Problem



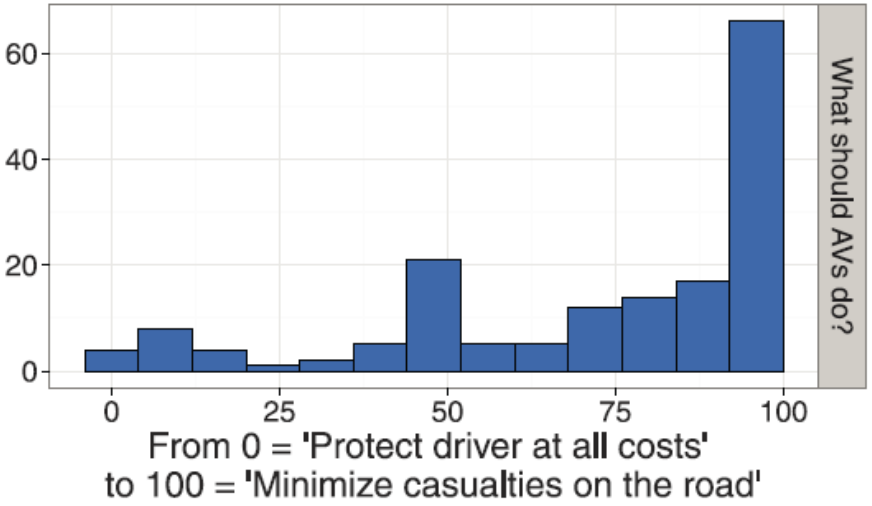
Drive straight ahead, killing 5 pedestrians, but saving driver



Swerve into barrier, avoiding 5 pedestrians, but killing driver

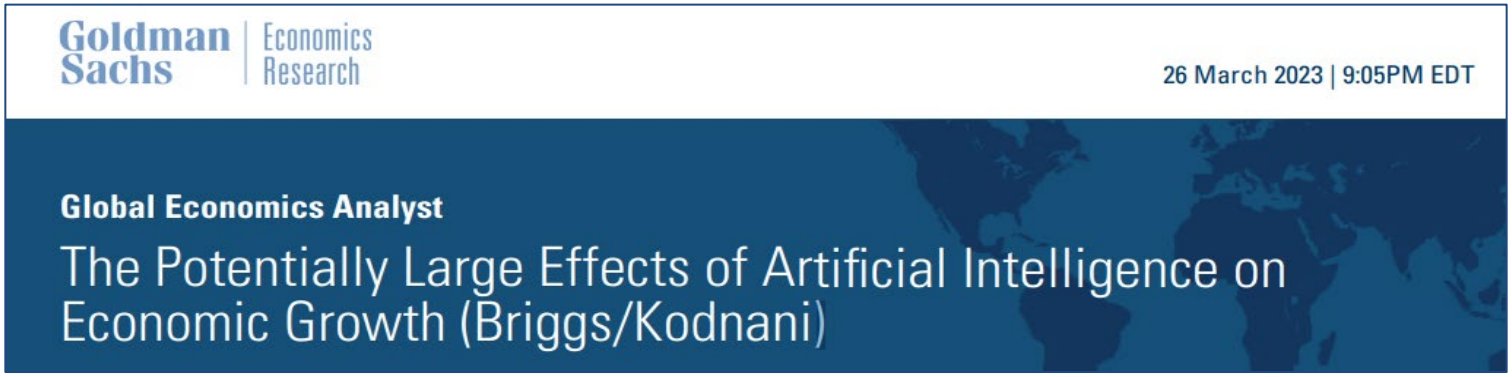
When Markets Break Down

- Bonnefon *et al* (2016) find an interesting result:
  - "participants… approved of utilitarian AVs"



  - "...they would themselves prefer to ride in AVs that protect their passengers at all costs."
  - "participants disapprove of enforcing utilitarian regulations for AVs and would be <u>less willing</u> to buy such an AV."

Labor Impacts of AI

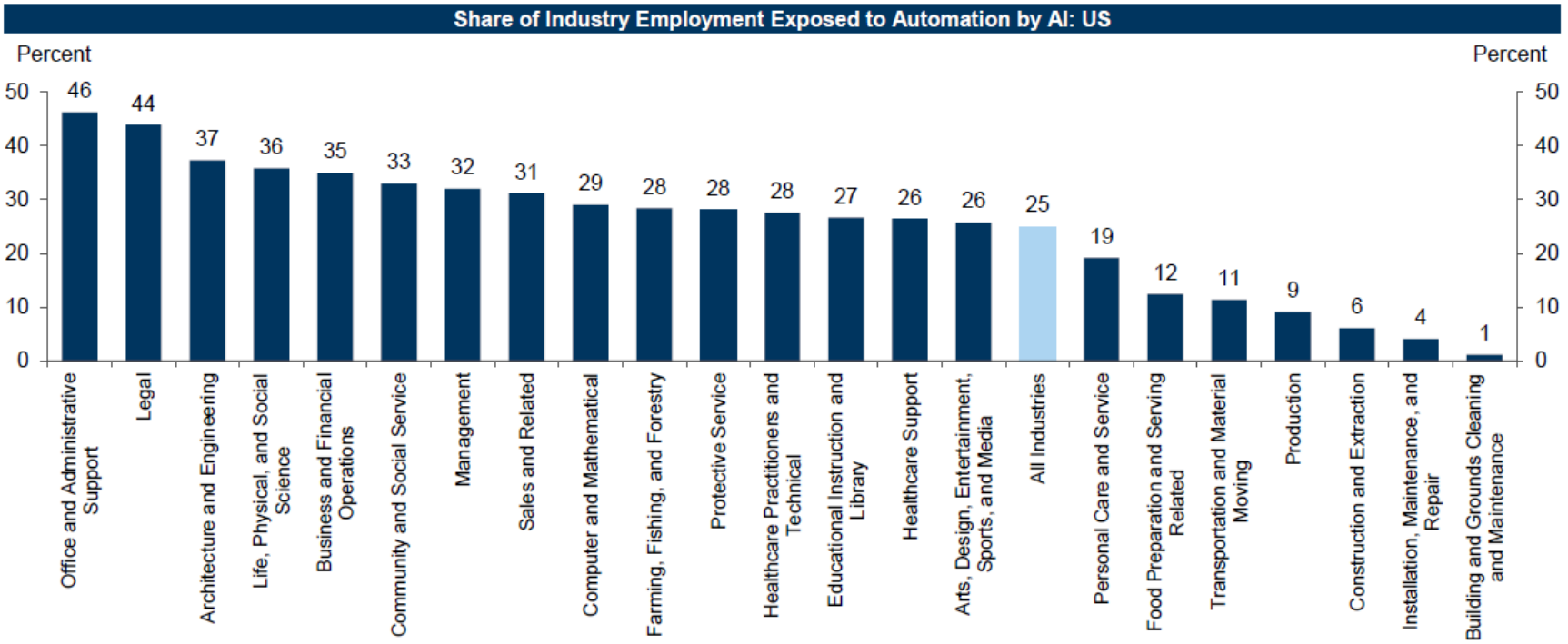


- If generative AI delivers on its promised capabilities, the labor market could face significant disruption. Using data on occupational tasks in both the US and Europe, we find that **roughly two-thirds of current jobs are exposed to some degree of AI automation**, and that **generative AI could substitute up to one-fourth of current work**. Extrapolating our estimates globally suggests that **generative AI could expose the equivalent of 300mn full-time jobs to automation**.
- The good news is that **worker displacement from automation has historically been offset by creation of new jobs**, and the emergence of new occupations following technological innovations accounts for the vast majority of long-run employment growth.

Distribution Across Workers

Goldman Sachs | Global Economics Analyst

Exhibit 5: One-Fourth of Current Work Tasks Could Be Automated by AI in the US and Europe

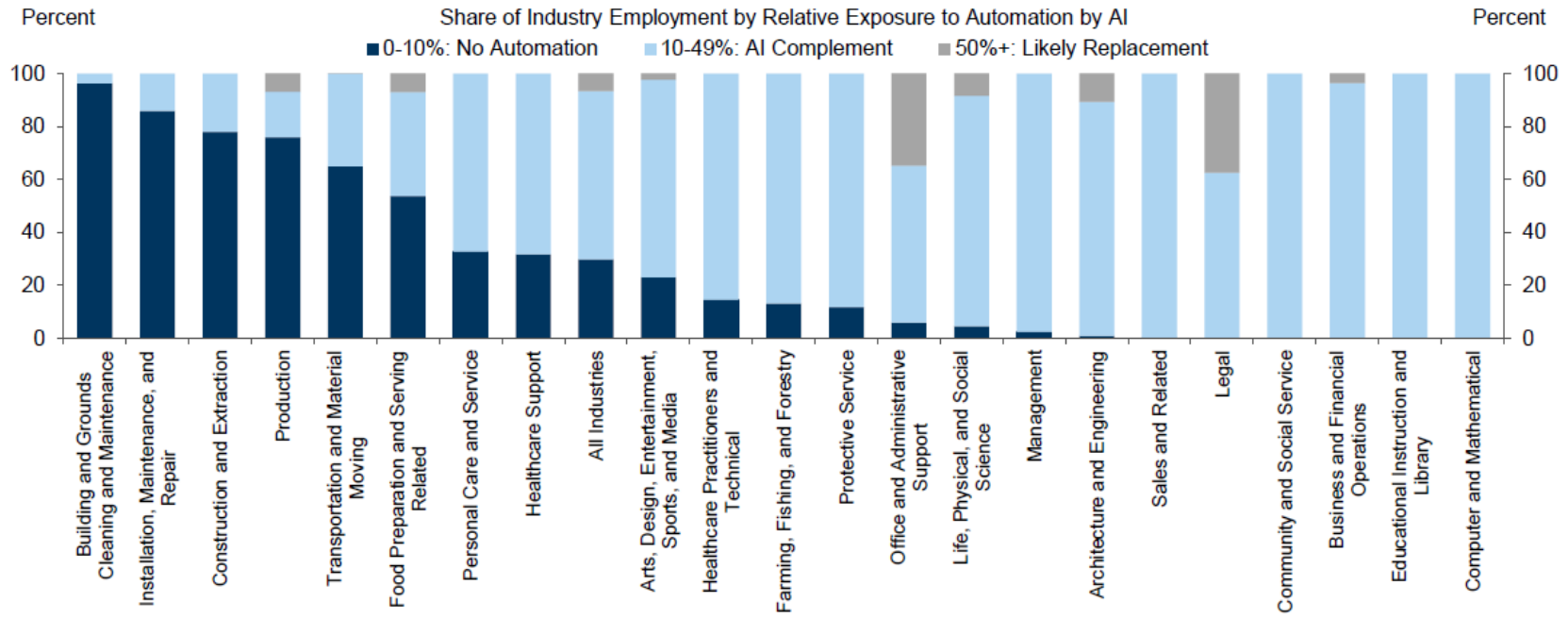


Distribution Across Industries

Goldman Sachs

Global Economics Analyst

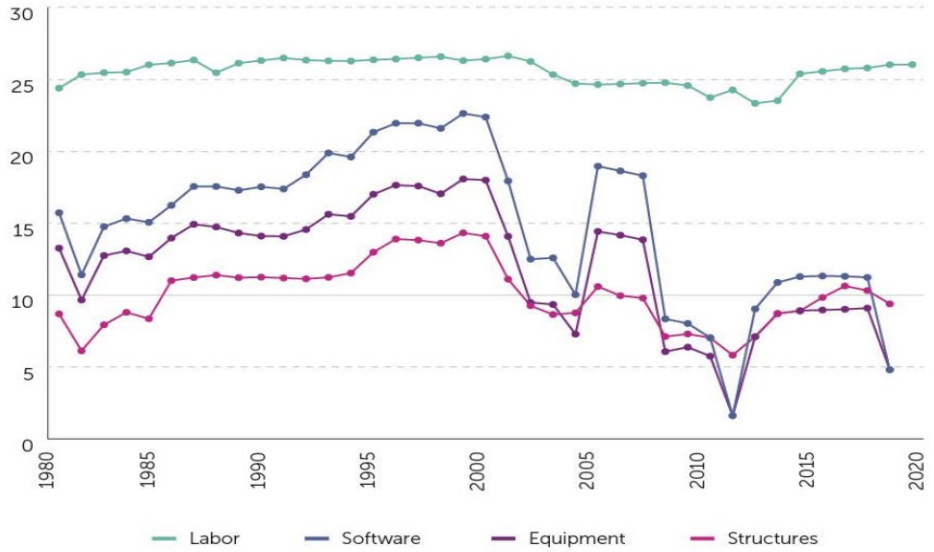
Exhibit 8: Replacement in Legal and Administrative Fields, Little Effect in Manual and Outdoor Jobs, and Productivity-Enhancement Everywhere Else



Change Tax Incentives

Image: Geekwire

Figure 1: Effective Taxes (In Percent) on Labor and Software, Equipment and Structures Capital, 1981–2018.



Source: Acemoglu, Manera and Restrepo, 2020

- Labor is taxed at an average rate of 25%, while software and equipment is taxed at 5%.
- “In essence, the U.S. tax system encourages companies to buy machines while discouraging them from adding workers.” (Acemoglu et al 2020).

Human Choices

- From "Research: quantifying GitHub Copilot's impact on developer productivity and happiness" by Eirini Kalliamvakou
  - The group that used GitHub Copilot had a higher rate of completing the task (78%, compared to 70% in the group without Copilot).
  - The striking difference was that developers who used GitHub Copilot completed the task significantly faster – 55% faster than the developers who didn't use GitHub Copilot.
- Say you get 25% greater productivity from your software engineering team as a result of AI-based tools like Copilot
  - Do you take on 25% more projects/features?
  - Do you hire 25% fewer software engineers?
- These are human choices