Introduction to Modern AI
Week 9: Ethical and Policy Considerations for AI

Gavin Hartnett

PRGS, Winter Quarter 2022

Overview

Overview of AI Ethics

Introduction

- How can we ensure that AI and ML systems are designed and employed ethically?
- Broad in scope, includes questions ranging from
  - How to ensure classifiers perform fairly across different sub-populations
  - How to ensure that the benefits of AI are well-distributed across society
  - How to use AI systems ethically in war
  - How to avoid creating a superintelligent AGI that sets about converting all the resources in the known universe into paperclips
- Evolving, interdisciplinary field involving
  - STEM fields (computer science, statistics, etc)
  - Social scientists (psychologists)
  - Humanities (ethicists, philosophers, gender studyist, etc)
  - Corporations, shareholders, general public
  - Public policy researchers (e.g., You!)

Example: Gender Shades paper

- Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." Conference on fairness, accountability and transparency. PMLR, 2018.
- Found that popular facial analysis benchmark datasets were overwhelmingly composed of lighter-skinned subjects
- Introduced a balanced dataset and found that a gender classification system had vaslty different error rates according to the skin-color. Darker-skinned females were the most mis-classified group by far.

Example: Broken Windows Policing

Figure 1
Rate of Enforcement Events per Epoch: Two Subpopulations, Same Crime Rate, Differing Vigilance

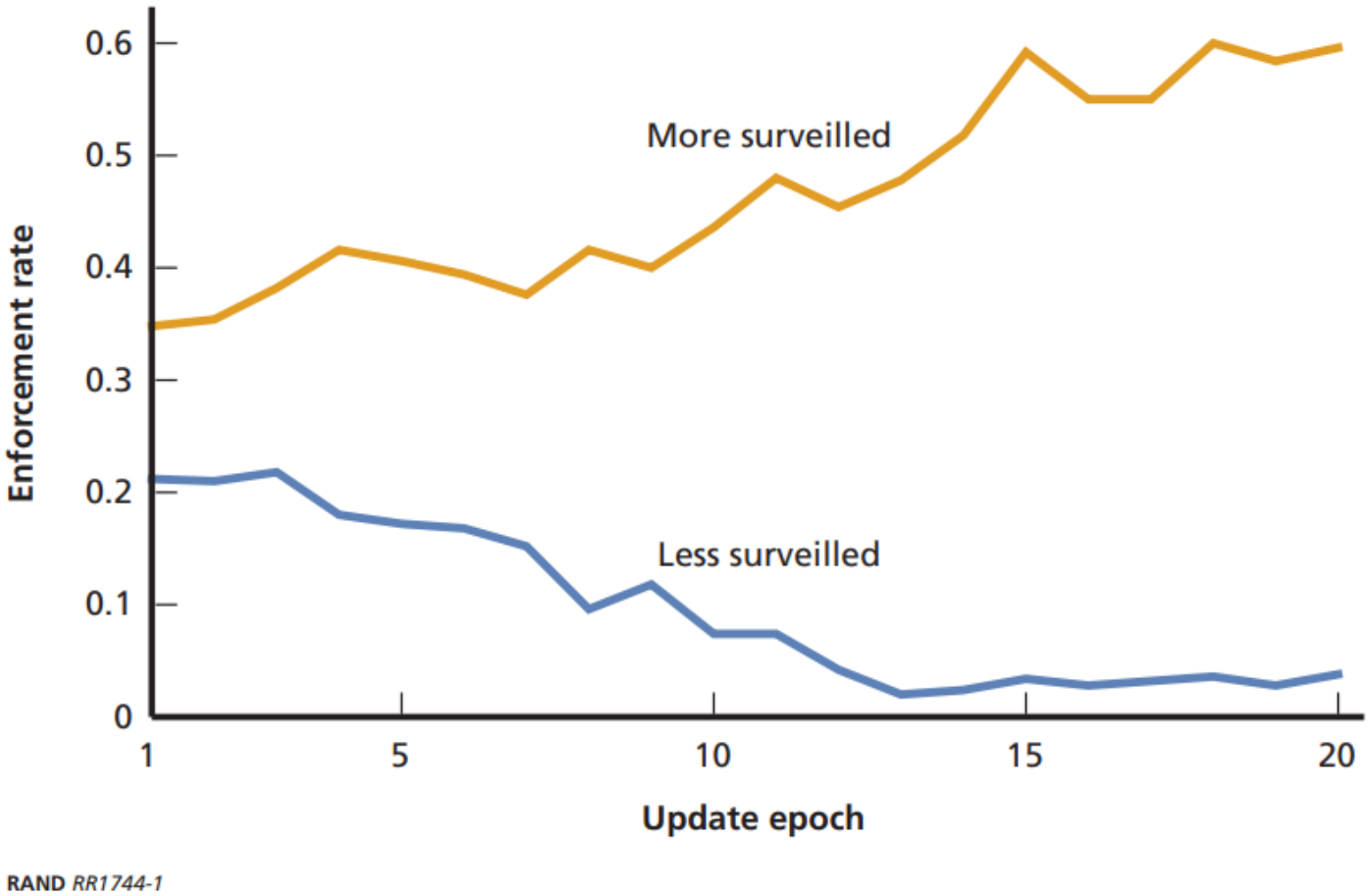


Image Source: Osoba, Osonde A., and William Welser IV. An intelligence in our image: The risks of bias and errors in artificial intelligence. Rand Corporation, 2017.

Example: COMPAS Recidivism Tool

- Pro Publica article: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- Northpointe's Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system used in sentencing and parole hearings across the country
- Authors uncover anecdotes of systematic racial bias in the COMPAS model: Black convicts were being rated higher than non-Black convicts, even when the non-Black convicts had more-severe offenses

Algorithmic Bias

Algorithmic bias: systematic and repeatable errors in a computer system that create unfair outcomes, such as privileging one arbitrary group of users over others.

What are the possible causes of this bias?

- The algorithm
  - E.g., perhaps there is a bias b/w dark and light pixels
- The data
  - Perhaps the data is imbalanced
  - And for that matter, what does imbalance mean here? Equal representation? Proportionate representation?
- The world
  - Real-world data can be expected to reflect real-world biases/historical injustices

FATML

- FATML is a subfield of ML concerned with ensuring that AI algorithms are designed and employed in a way that is
  - Fair
  - Accountable
  - Transparent
  - Ethical
- Sounds great, but
  - What do these notions mean, specifically?
  - Who gets to decide?
  - What if it costs money/hurts performance?
  - In some cases researchers can prove that some notions of fairness/accuracy are at odds with one another

Fairness Through Awareness

- Dwork, Cynthia, et al. "Fairness through awareness." Proceedings of the 3rd innovations in theoretical computer science conference. 2012.
- Provide a normative approach to fairness in classification and a framework for achieving it
  - Treat similar individuals similarly (awareness)
  - Achieves individual fairness rather than group fairness (statistical parity)
  - Can be modified to optimize both group and individual fairness (with a trade-off): fair affirmative action
  - Provide a linear problem formulation of this approach

Ethical Implications of Large Language Models

GPT and Large Language Models

- Generative Pre-trained Transformer (GPT) is a model trained by OpenAI
- Based on the Transformer architecture, which in turn uses the so-called attention mechanism (covered in the advanced AI course)
- Model trained on large amounts of text scraped from the internet
- GPT is a *causal* language model: tries to predict the next token in a sequence
- OpenAI has since released GPT-2, GPT-3
- Many other Large Language Models (LLMs) have been developed by other companies

GPT and Large Language Models

- Play with OpenAI API: https://openai.com/api/
- Ben Boudreaux: ethical implications of large language models

AI Safety

Technological Singularities and AI

- The rate of technological improvement seems to be accelerating
- Better technology can assist in designing new technology
- Oft-cited example: Moore's Law: the number of transistors in a dense integrated circuit (IC) doubles about every two years
- Counter-point: at some point the law must fail due to physical limitations
- Counter-counter-point: new paradigms will emerge which will surmount these barriers

Technological Singularities and AI

- Let's try to model different ways that technology might advance
- Technological rate of change is proportional to current technology
  - Exponential growth

$$\frac{dT}{dt} = \alpha T\,, \quad \Rightarrow \quad T(t) = e^{\alpha t}$$

- Technological rate of change experiences a rapid growth past some point
  - Singularity in finite time ($t_* = \ln(1+\alpha)/\alpha$)

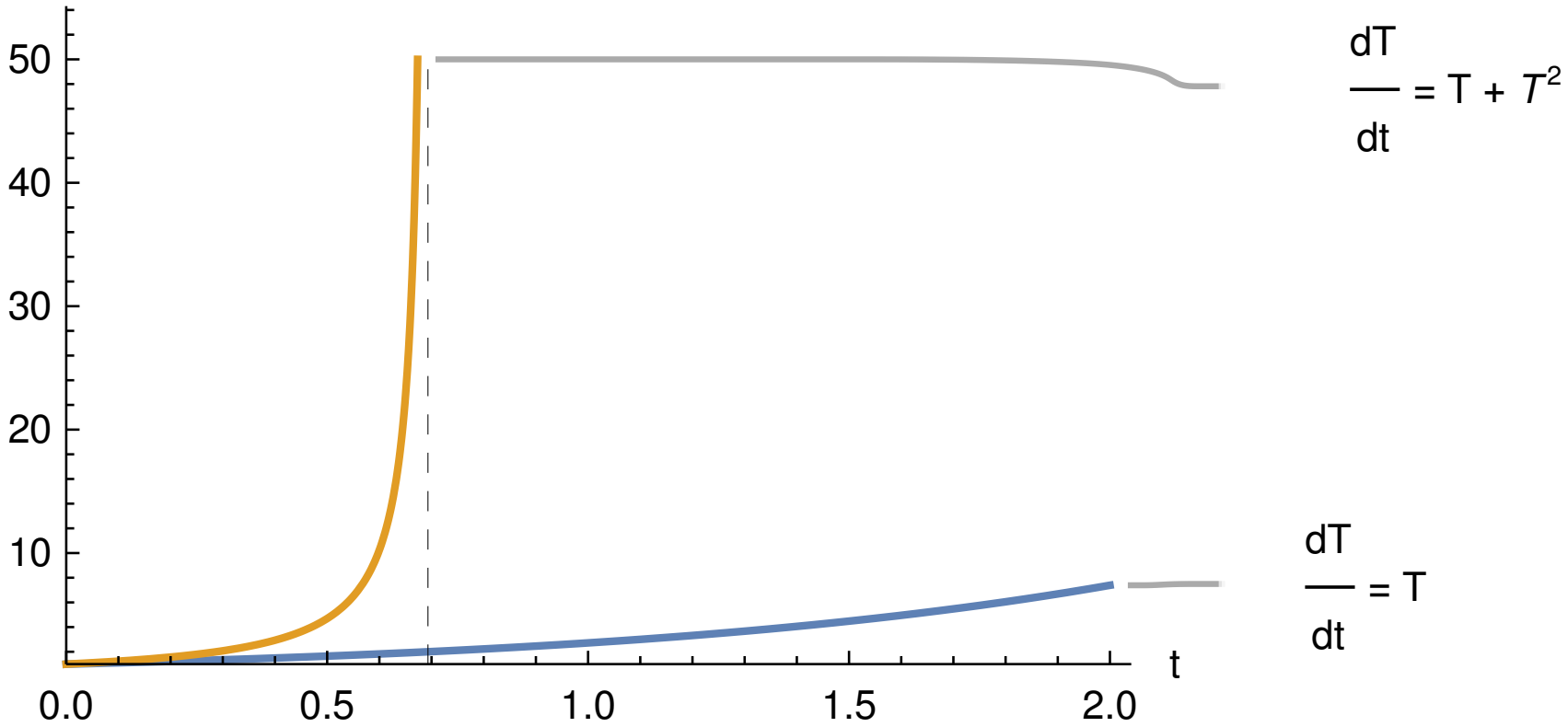
$$\frac{dT}{dt} = \alpha T + T^2\,, \quad \Rightarrow \quad T(t) = \frac{\alpha e^{\alpha t}}{1+\alpha-e^{\alpha t}}\,.$$

  - This example used by Bostrom in *Superintelligence* to illustrate how an "intelligence explosion" could occur

Technological Singularities and AI

$$\frac{dT}{dt} = \alpha T\,, \quad \Rightarrow \quad T(t) = e^{\alpha t}$$

$$\frac{dT}{dt} = \alpha T + T^2\,, \quad \Rightarrow \quad T(t) = \frac{\alpha e^{\alpha t}}{1 + \alpha - e^{\alpha t}}\,.$$



Technological Singularities and AI

- AI is a special type of technology
- Once the AI can improve itself, it can become better at improving itself (and so on)
- Lots of very smart people have convinced themselves this is a serious possibility:
  - *The development of full artificial intelligence could spell the end of the human race. . . .It would take off on its own, and re-design itself at an ever-increasing rate. Humans, who are limited by slow biological evolution, couldn't compete, and would be superseded.*
    - Stephen Hawking
  - *It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. . . They would be able to converse with each other to sharpen their wits. At some stage, therefore, we should have to expect the machines to take control.*
    - Alan Turing

Institutions Focused on the Possible Existential Risk Posed by AI

- Future of Life Institute (FLI)
- Future of Humanity Institute (FHI)
- OpenPhilanthropy
- OpenAI (very debatable)
- Machine Intelligence Research Institute (MIRI)

Are We There Yet?

Ilya Sutskever
@ilyasut

it may be that today's large neural networks are slightly conscious

3:27 PM · Feb 9, 2022 · Twitter Web App

404 Retweets **348** Quote Tweets **2,831** Likes

Tweet your reply

Reply

Yann LeCun @ylecun · Feb 12
Replying to @ilyasut
Nope.
Not even for true for small values of "slightly conscious" and large values of "large neural nets".
I think you would need a particular kind of macro-architecture that none of the current networks possess.

45 38 772

Judea Pearl @yudapearl · 5h
Rushing to gleefully agree with @ylecun on this point. Before a system can lay claims to consciousness it must exhibit "deep understanding" of some domain, which large NN's have yet to exhibit by answering questions at all three levels of the reasoning hierarchy.

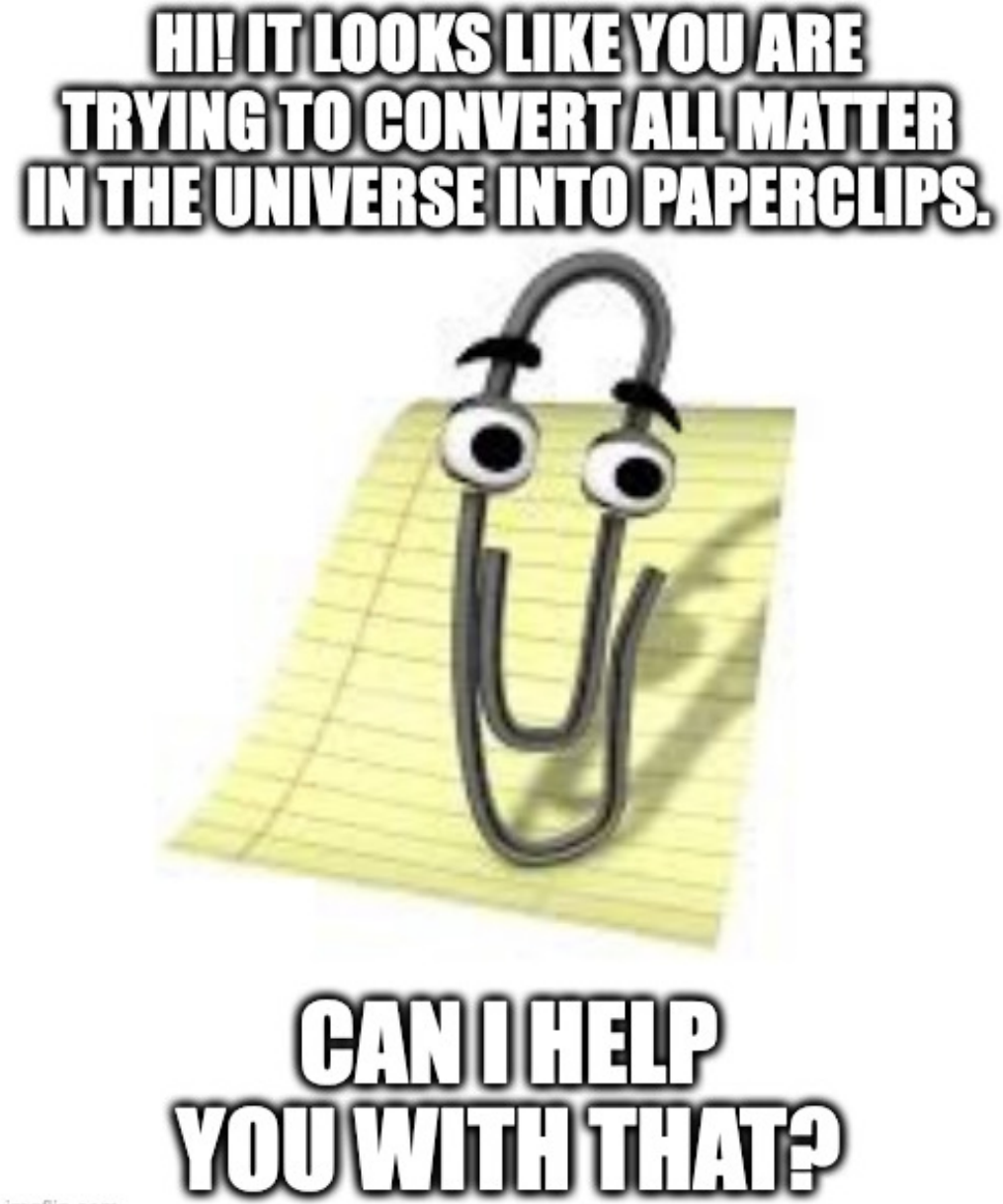
8 8 66

Show replies

Value Alignment

- Will the values of a true AGI be aligned with ours?
- Orthogonality thesis: the level of intelligence that an agent has is independent of its values
- If true, means that it is very important to align a future AI's values with our own

Paperclip Maximizers



Non-Existential Risk

- In addition to concerns that an AGI could be inadvertently created, there are many more practical concerns associated with the AI/ML systems
- Bias/ethical issues - discussed above
- Use in warfare - discussed below
- Other areas:
  - Disinformation/Propaganda
  - Safety-critical systems

AI and Disinformation

- Deepfakes - high quality fake/adulterated images or video
- Obama deepfake: https://www.youtube.com/watch?v=cQ54GDm1eL0&ab_channel=BuzzFeedVideo
- Russian disinformation: https://twitter.com/oneunderscore__/status/1498349668522201099?s=20&t=9dRHg0UK4MfYDZL6iTdmLg
- Through language models like GPT we can also have high-quality fake text

AI and Disinformation

Brainstorm questions:

- What are some possible misuse cases for models like these?
- How should these risks be regulated/mitigated?

AI and War