GENE 213
AI, Genes, and Ethics

Ethics in AI :
Topic Overview

Eric Saund, Ph.D.
November, 2022

saund@alum.mit.edu

www.saund.org

Outline

1. <u>Framing</u>

   - AI: Amplifying human power.
   - Ethics: Policies to realize values.

   "Before you build a monster, think about it."

2. <u>Topics</u>

   2.1 Consensus principles in AI Ethics.

   2.2 The AI Society.

   -data collection, privacy, surveillance
   -information ecosystem
   -wherefore human labor?

   2.3 Autonomous weapons & military AI arms race.

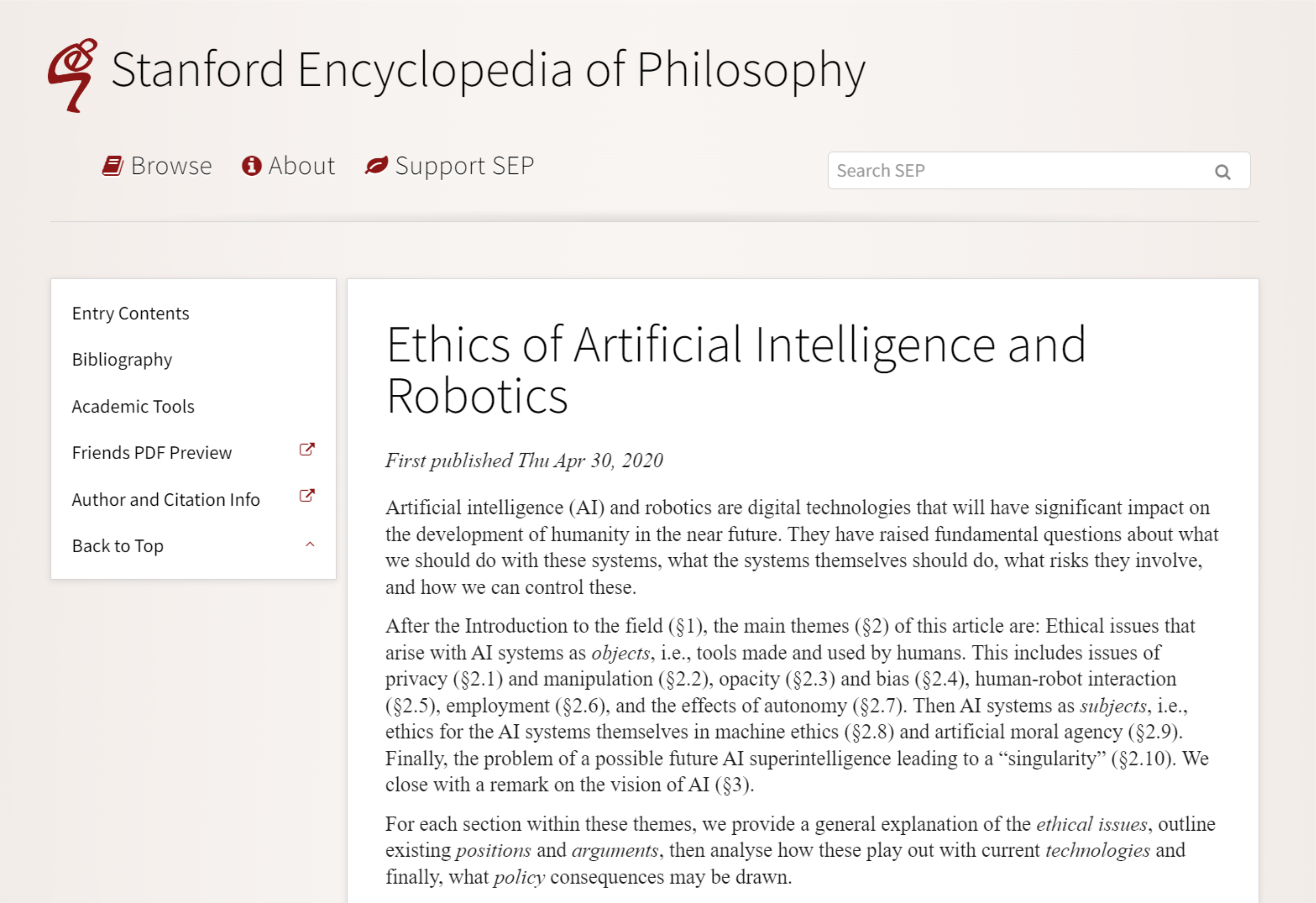
   2.4 Fairness & the civilian AI arms race.

   2.5 AGI and the "Alignment Problem".

[3. Deep Dive: Algorithmic Bias.] (next week)

Main Reading

https://plato.stanford.edu/entries/ethics-ai/



Ethics of Artificial Intelligence and Robotics

First published Thu Apr 30, 2020

Artificial intelligence (AI) and robotics are digital technologies that will have significant impact on the development of humanity in the near future. They have raised fundamental questions about what we should do with these systems, what the systems themselves should do, what risks they involve, and how we can control these.

After the Introduction to the field (§1), the main themes (§2) of this article are: Ethical issues that arise with AI systems as *objects*, i.e., tools made and used by humans. This includes issues of privacy (§2.1) and manipulation (§2.2), opacity (§2.3) and bias (§2.4), human-robot interaction (§2.5), employment (§2.6), and the effects of autonomy (§2.7). Then AI systems as *subjects*, i.e., ethics for the AI systems themselves in machine ethics (§2.8) and artificial moral agency (§2.9). Finally, the problem of a possible future AI superintelligence leading to a “singularity” (§2.10). We close with a remark on the vision of AI (§3).

For each section within these themes, we provide a general explanation of the *ethical issues*, outline existing *positions* and *arguments*, then analyse how these play out with current *technologies* and finally, what *policy* consequences may be drawn.

Additional Reading

Slaughterbots video
https://www.youtube.com/watch?v=O-2tpwW0kmU

Final Report
National Security Commission on Artificial Intelligence
Executive Summary
https://www.nscai.gov/wp-content/uploads/2021/03/Final_Report_Executive_Summary.pdf

"The Coming AI Hackers"
Bruce Schneier
https://www.belfercenter.org/sites/default/files/2021-04/HackingAI.pdf

Outline

1. Framing

   - AI: Amplifying human power.
   - Ethics: Policies to realize values.

   "Before you build a monster, think about it."

2. Topics

   2.1 Consensus principles in AI Ethics.

   2.2 The AI Society.

   -data collection, privacy, surveillance
   -information ecosystem
   -wherefore human labor?

   2.3 Autonomous weapons & military AI arms race.

   2.4 Fairness & the civilian AI arms race.

   2.5 AGI and the "Alignment Problem".

[3. Deep Dive: Algorithmic Bias.] (next week)

Definitions

Definition of Intelligence: The use of *information* in the service of *goals*.

Definition of Artificial Intelligence: Machines that display intelligence.

Question: *Whose goals are being served?*

-Machine builder?
-Machine owner?
-Machine user?
-Society?
    -through market forces?
    -through regulation?
-The AI machine itself?
    -goal propagation & autonomy

Purpose for Building AI

- Human enlightenment about ourselves

- Amplify and extend human capabilities

  Improve human life by:

  - Increasing means of production
  - Obviating crummy jobs
  - Improving planning and decision making
  - Expanding the scope of creativity and experience
  - [insert your value proposition for AI here]

economic value creation:
cheaper, better, faster, new

Robots: AI + embodied physical action

- Factory robots

- Self-driving cars

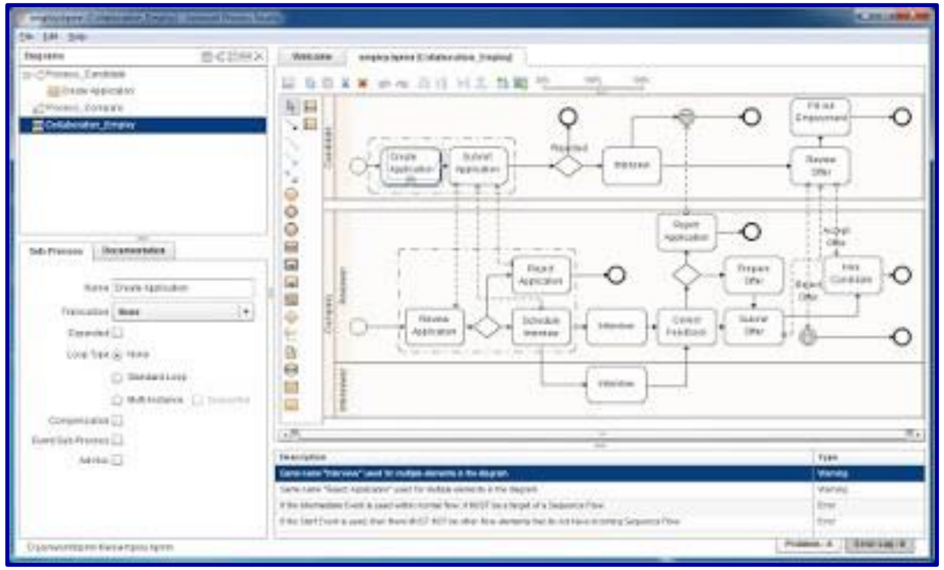
- Autonomous drones

Leading AI Applications

Automation

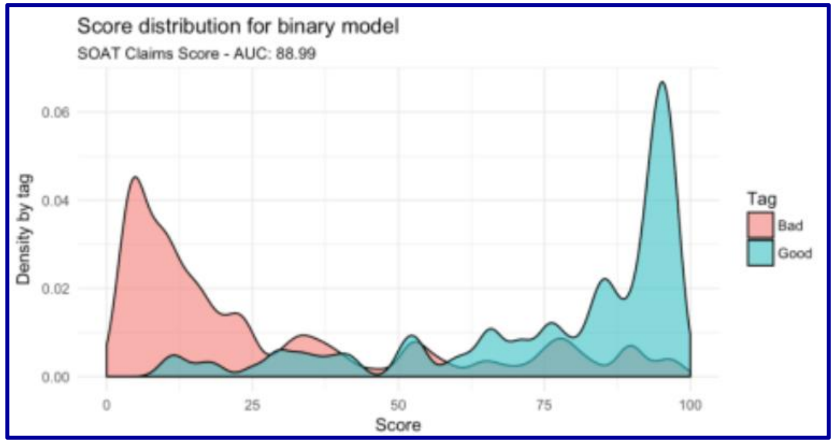
- Manufacturing; Process control & monitoring; Inspection; QC; System health

- Robotic Process Automation

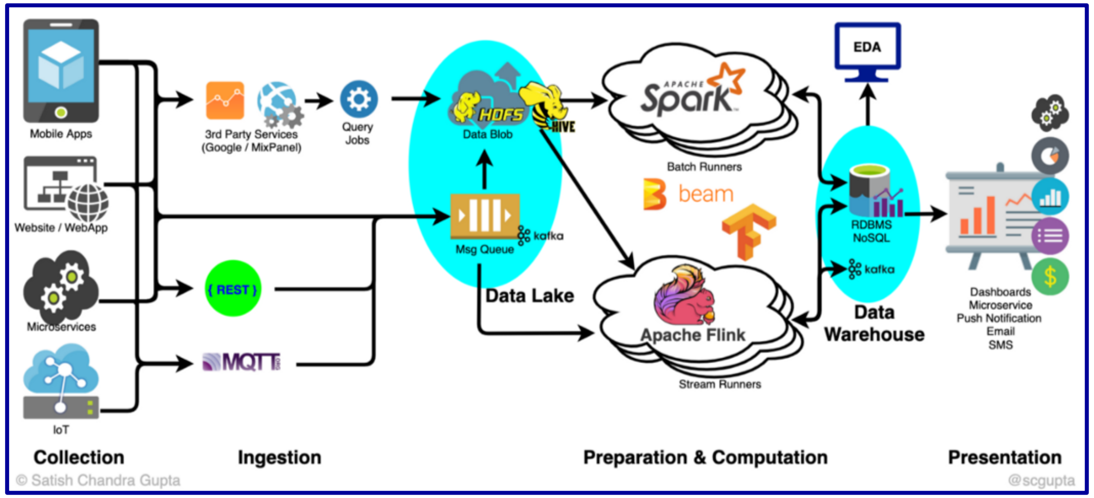


Leading AI Applications

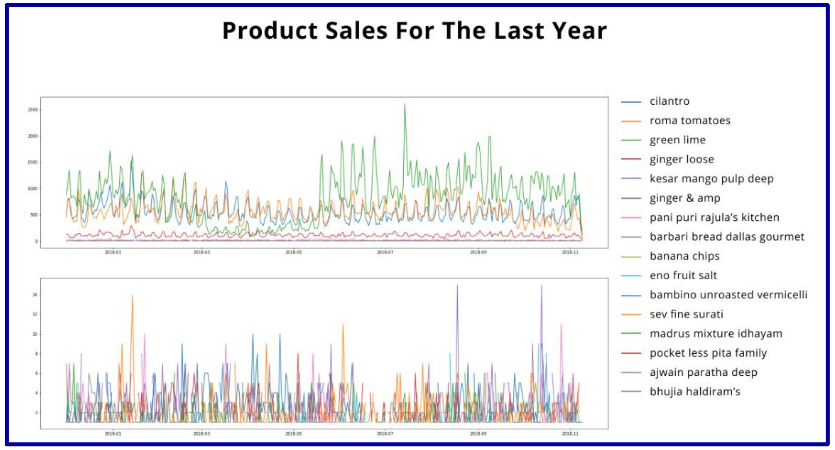
Predictive Analytics



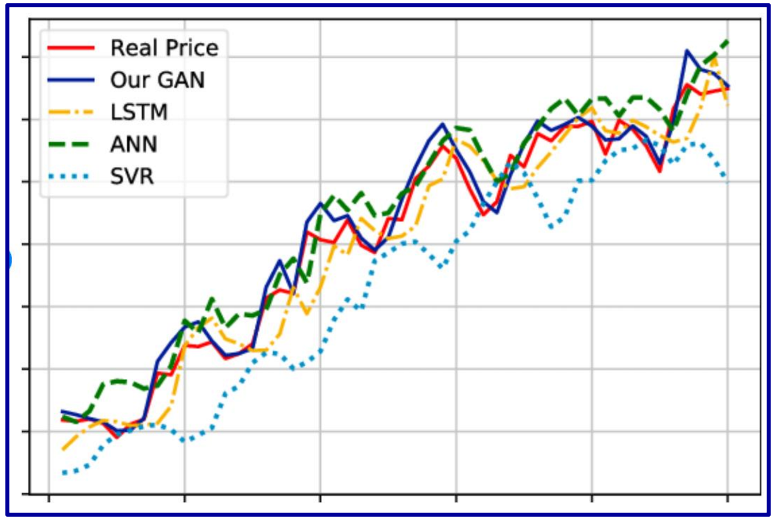
- Credit scoring



- Fraud detection



- Business monitoring & forecasting

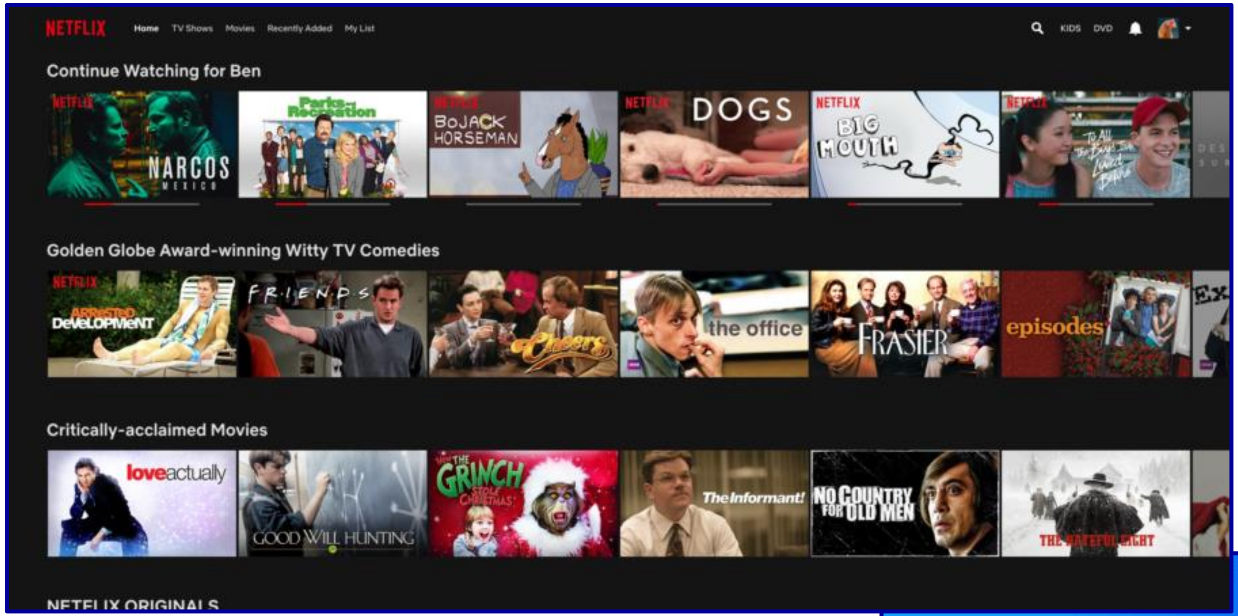


- Economic market prediction

Leading AI Applications

Recommender Systems

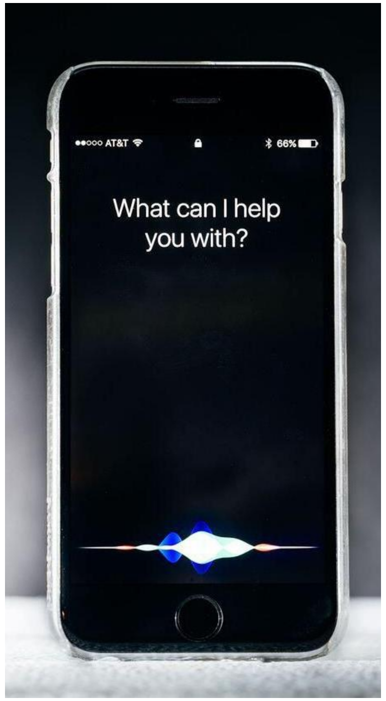
- Commerce
  Netflix, Amazon,
  Advertising



- Social media
  Facebook, Twitter,
  YouTube, TikTok, ...

Leading AI Applications

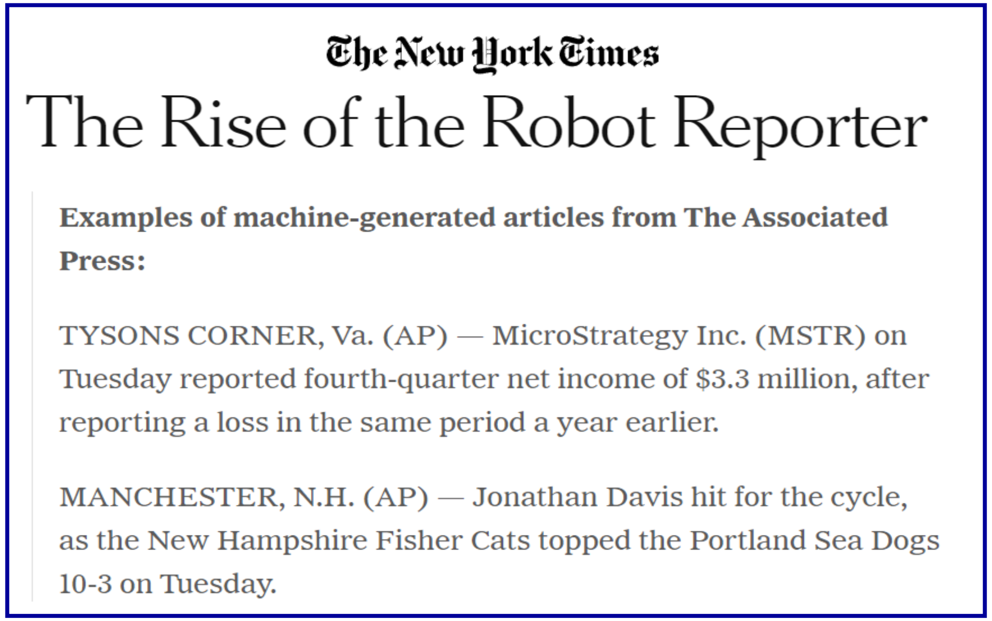
Interactive Agents



Leading AI Applications

Content Synthesis

- Text



The New York Times

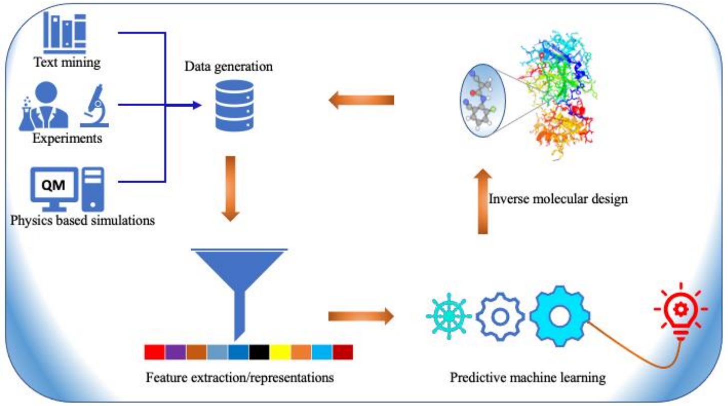
The Rise of the Robot Reporter

Examples of machine-generated articles from The Associated Press:

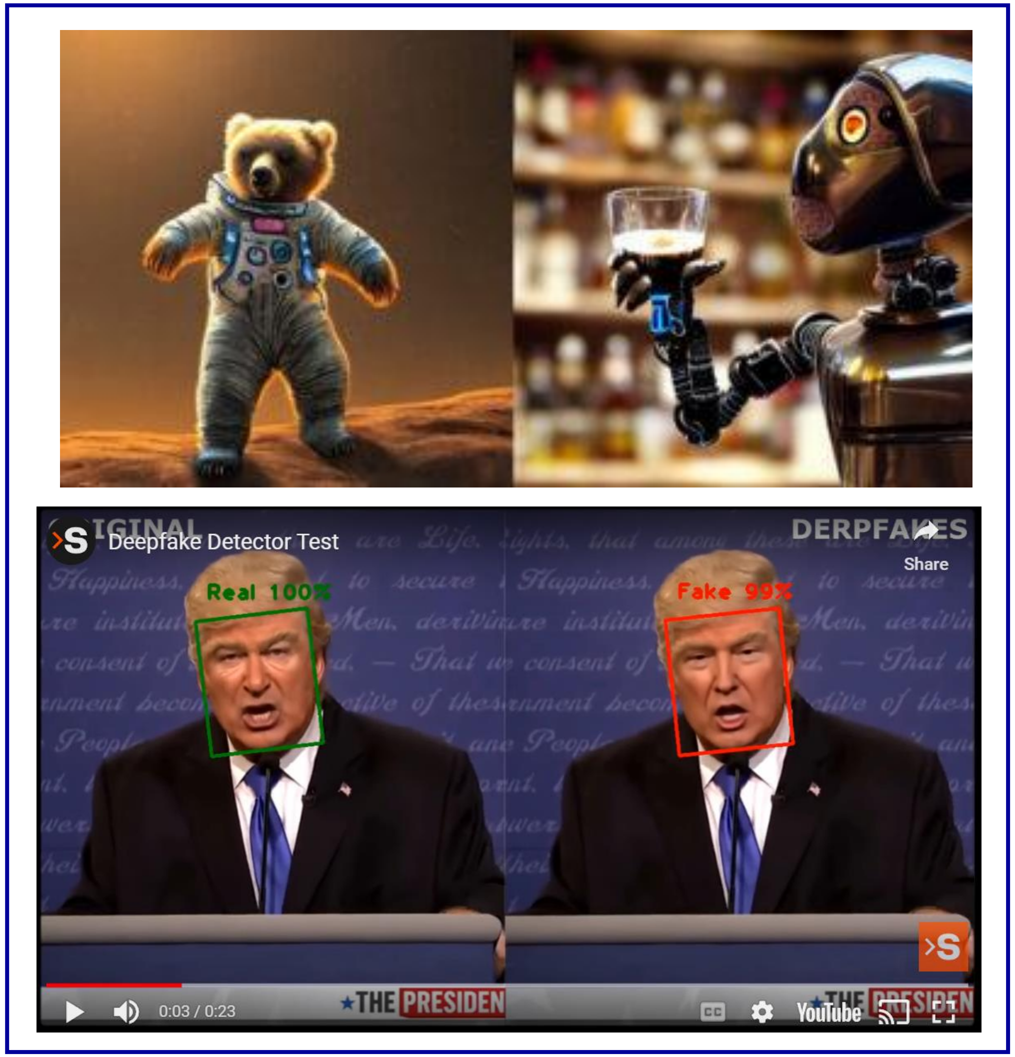
TYSONS CORNER, Va. (AP) — MicroStrategy Inc. (MSTR) on Tuesday reported fourth-quarter net income of $3.3 million, after reporting a loss in the same period a year earlier.

MANCHESTER, N.H. (AP) — Jonathan Davis hit for the cycle, as the New Hampshire Fisher Cats topped the Portland Sea Dogs 10-3 on Tuesday.

- Therapeutics


Text mining
Experiments
QM
Physics based simulations
Data generation
Inverse molecular design
Feature extraction/representations
Predictive machine learning

- Images



Hybrid example

Problem: Law firm case intake problem: Do we have a conflict of interest?

Existing solution:
- team of Conflicts analysts
- database of all cases, clients, & attorneys
- informal relationships + executive decisions

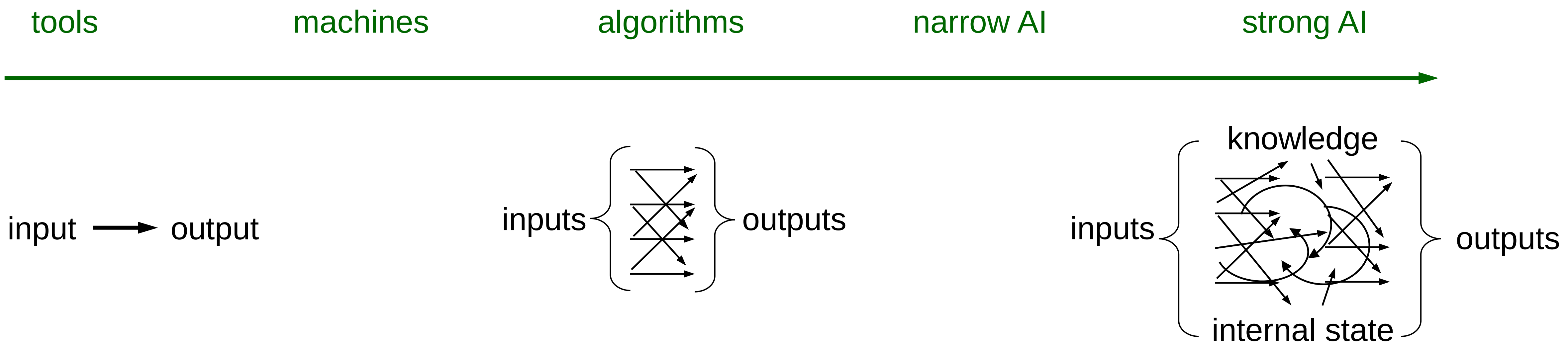
- expensive
- time-consuming
- error-prone

AI solution:
- scoring based on past cases
- Knowledge dashboard & Natural Language search
- continuously-training ML

- recommender
- interactive assistant
- Robotic Process Automation

AI that helps *prevent* ethics problems.

AI vs. Traditional Machines



tools

machines

algorithms

narrow AI

strong AI

input → output

inputs

outputs

knowledge

inputs

internal state

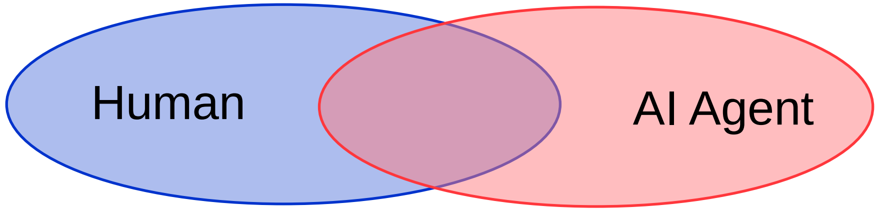
outputs

Information processing complexity:

- internal state
- knowledge access
- close loop on ongoing activity

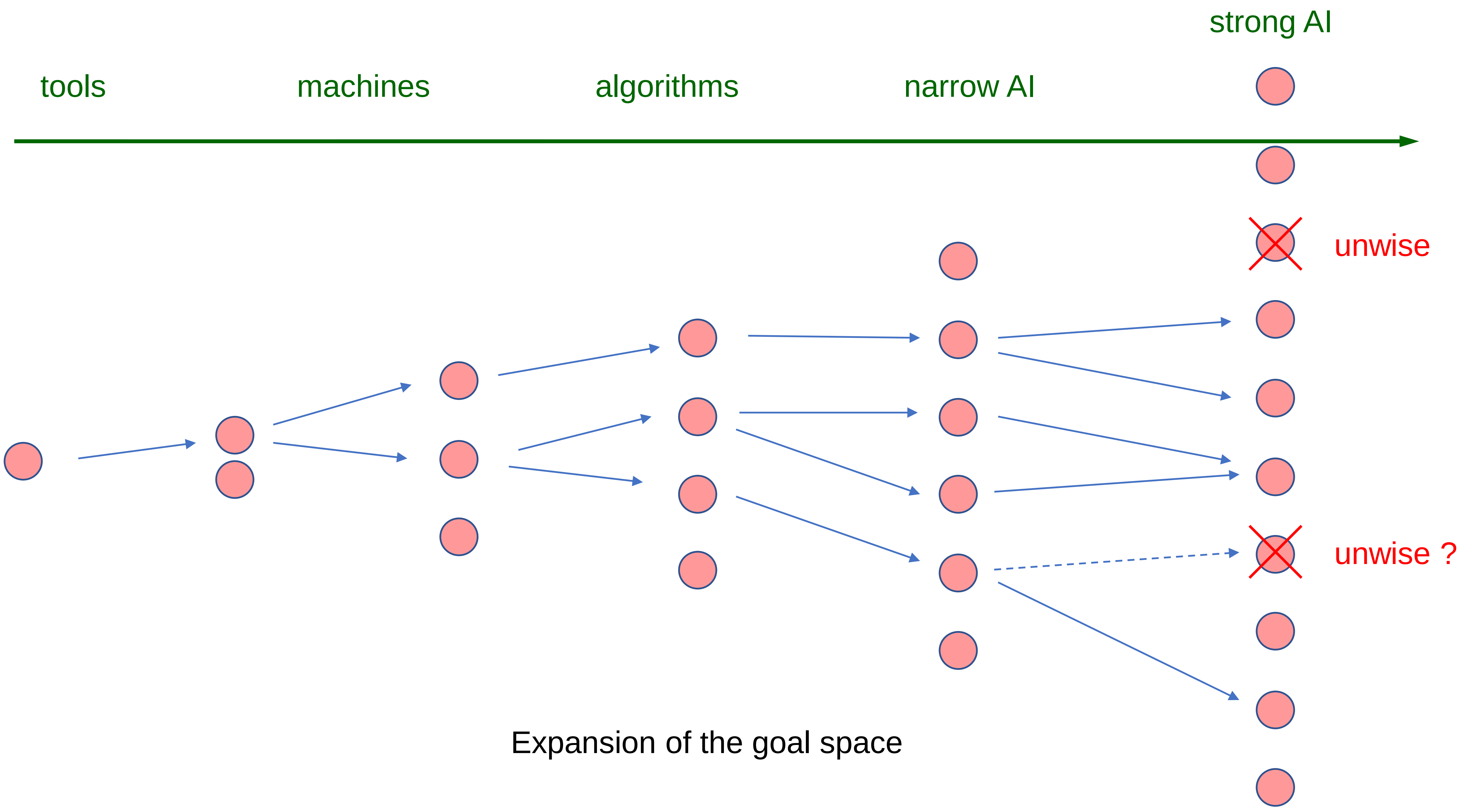
→

greater capability,
autonomy,
unpredictability



Context Scope: *What does it mean to understand another mind?*

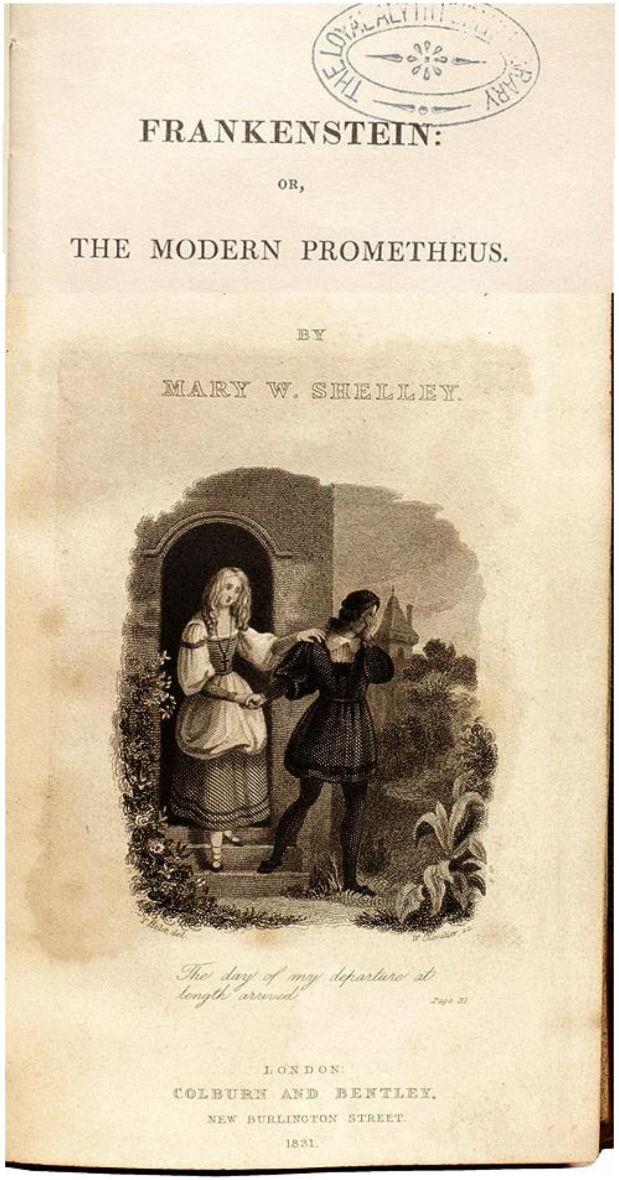
AI Makes More Goals Achievable



Expansion of the goal space

To What Ends?

Toward what goals *should* we use our newfound technological powers?



FRANKENSTEIN:

OR,

THE MODERN PROMETHEUS.

BY

MARY W. SHELLEY.

T. Holst del.
W. Chevalier sc.

The day of my departure at length arrived.

Page 31

LONDON:
COLBURN AND BENTLEY,
NEW BURLINGTON STREET.
1831.

Rules of Thumb in Ethics for AI

1. Usually there's a dilemma. Conflict of values, tradeoffs.

   If you think the answer is easy or obvious, you will be wise to make sure you are not overlooking some important considerations.

   Humility

2. Consider the basic issue in the absence of AI.

   Does a technology amplifier change the underlying problem of conflicting values?

   Universality

3. Take scenarios to the extremes.

   See what the implications are, to clarify what is being traded off.

   Values in boundary cases

4. Discern principles to interpolate.

   Negotiate

Example: *Should we build an AI/Robot future of leisure and abundance for humans?*

R 2. Leisure classes enabled by human servants.
R 3. Utopia vs. decadent decay?
R 4. **No:** Build human strength & excellence from challenges.
Yes: Relief of human suffering & despair from defeat.

Outline

1. Framing

- AI: Amplifying human power.
- Ethics: Policies to realize values.

"Before you build a monster, think about it."

2. Topics

→ 2.1 Consensus Principles in AI Ethics.

2.2 The AI Society.

-data collection, privacy, surveillance
-information ecosystem
-wherefore human labor?

2.3 Autonomous weapons & military AI arms race.

2.4 Fairness & the civilian AI arms race.

2.5 AGI and the "Alignment Problem".

[3. Deep Dive: Algorithmic Bias.] (next week)

Much Attention to AI Ethics

future of life INSTITUTE

ASILOMAR AI PRINCIPLES

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

Principled Artificial Intelligence

MAPPING CONSENSUS IN ETHICAL AND RIGHTS-BASED APPROACHES TO PRINCIPLES FOR AI

HAI Stanford University Human-Centered Artificial Intelligence

Partnership on AI

IEEE

Microsoft

CHAI Center for Human-Compatible Artificial Intelligence

UNESCO unesco

THE WHITE HOUSE

BLUEPRINT FOR AN AI BILL OF RIGHTS

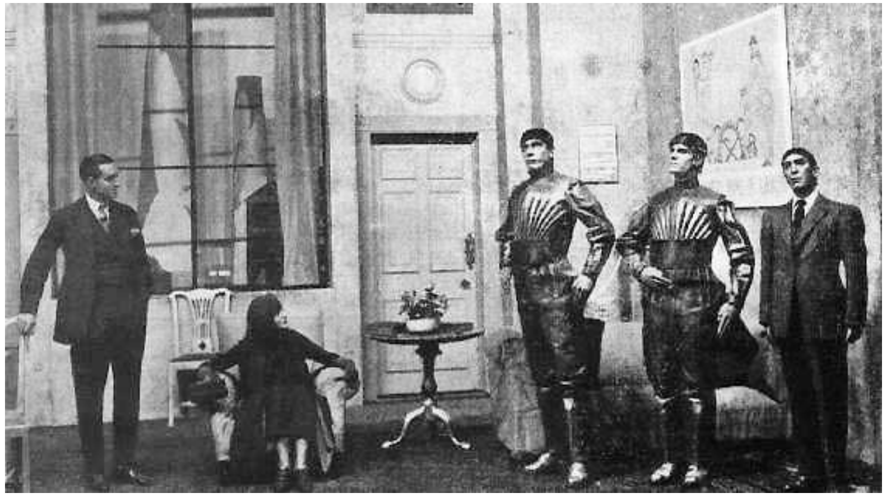
Convergent Concerns

Consensus principles

- benefit humanity
  - economic prosperity
  - human dignity & fulfillment
  - rights and freedoms
- fairness
- safety & security
- transparency & explainability
- accountability
- responsibility
- privacy
- human control of technology
- promotion of human values

All stages of *conception, design,*
deployment, operation, maintenance
of
every *tool, technology, component,*
application, and *system* that employs AI.

Cultural Concerns



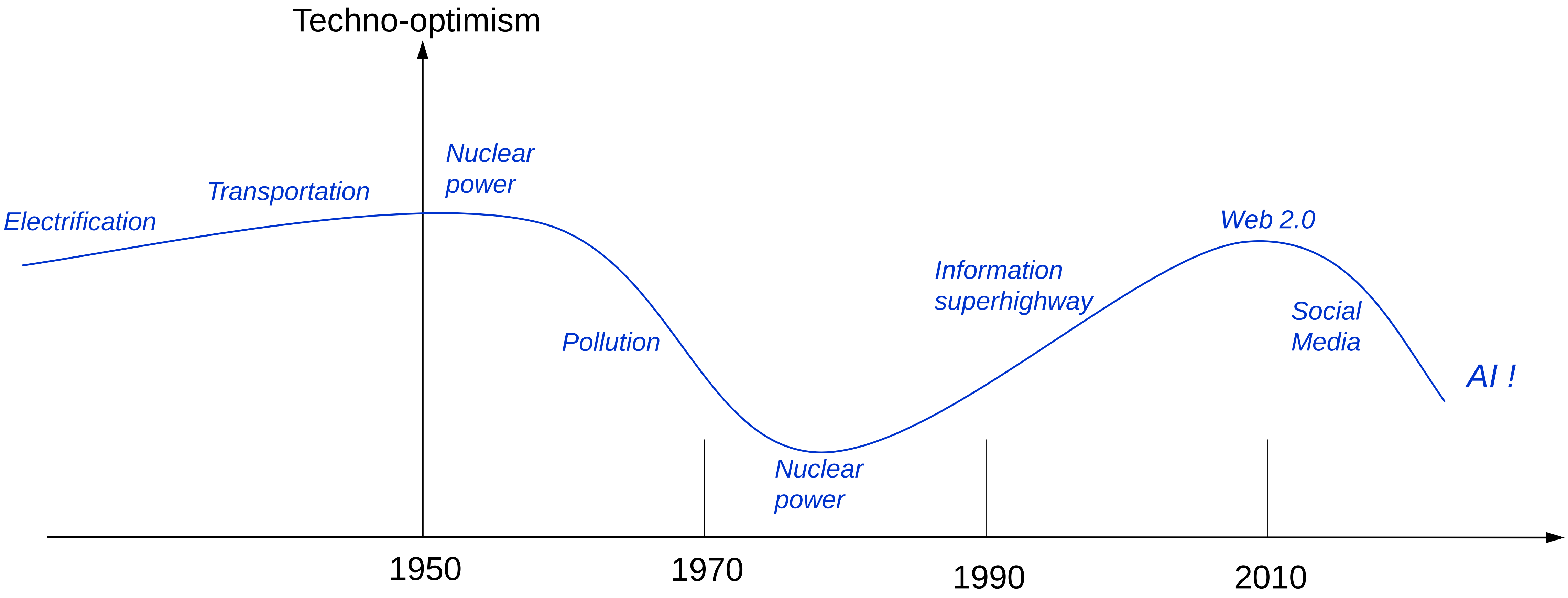
Rossum's Universal Robots - 1920

Modern Times -1936

Ex Machina - 2014

Westworld - 2016

Techno-Optimism Trends



Cautious Attitude

What's special for concern about AI?

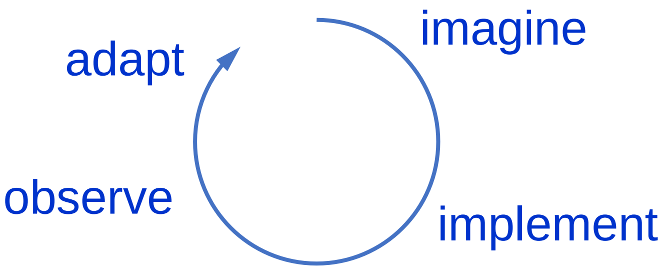
- Difficulty in anticipating outcomes due to sophistication, opacity of mysterious technology.

- Period of heightened sensitivity about harms.
  - Left: justice for marginalized classes.
  - Right: loss of traditional values.
  - Both: ownership and control of the levers of power.

~~Move fast and break things~~

- Systems thinking:

All technology is deployed in a
technological
economic
environmental
social
cultural
context.



imagine
implement
observe
adapt

Realizing Consensus Principles

Table stakes

		Causation
• fairness		
• safety	→	design & operate to prevent harm
• transparency	→	trace train of causation when harm happens
• accountability	→	assign formal roles, establish grounds for justification
• responsibility	→	propagate consequences back to causal actors

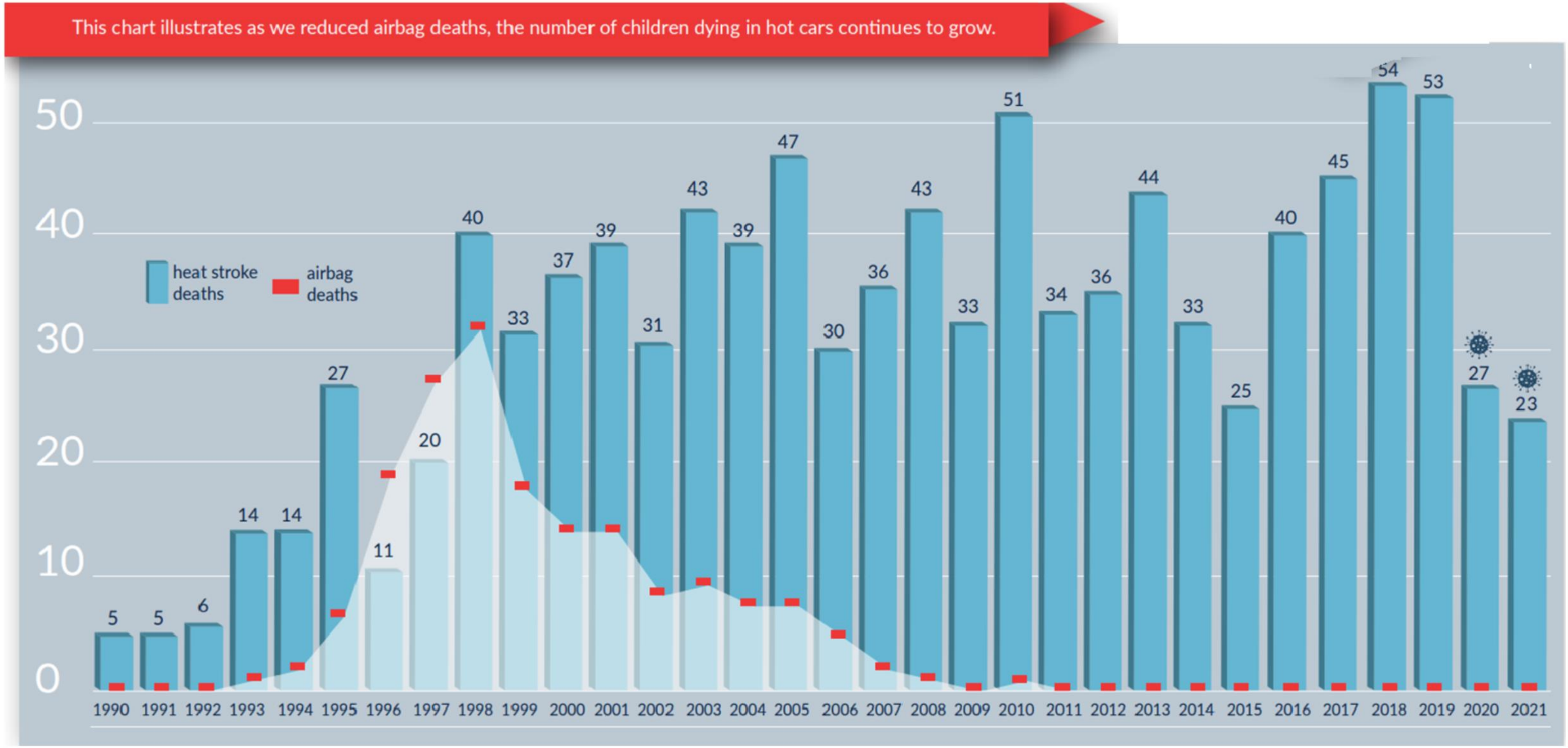
Systemic objective:
shape regulatory systems, policies, standards, norms, legal theories
through pre-emptive considerations,
then refine through case law

Safety Brings Tradeoffs

Ethics problem:

but...

Safety for whom?



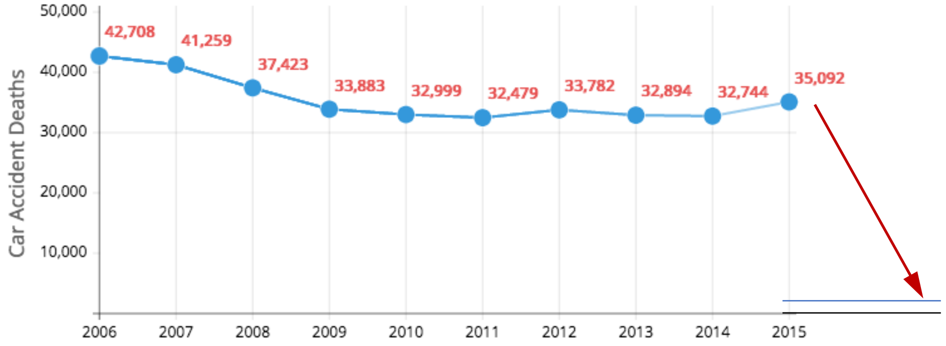
This chart illustrates as we reduced airbag deaths, the number of children dying in hot cars continues to grow.

Self-Driving Vehicles Meet AI Ethics

Self-Driving Cars

Car Accidents Deaths

2006-2015



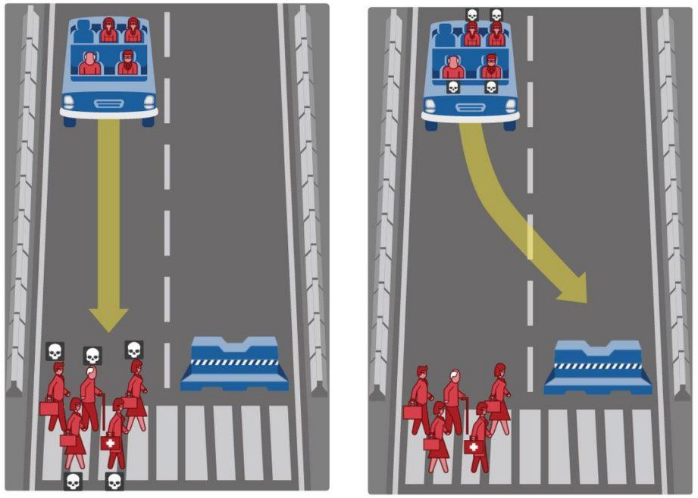
Source: Insurance Institute for Highway Safety - Fatality Facts

Axiom:
Some people *will* be harmed by self-driving cars.

Greater good versus unfortunate collateral damage?

Is harm caused by AI worse than harm caused by a human?

Trolley problems



- Showcase ethical dilemmas.
- Mainly thought experiments; extremely rare in actuality.

In practice:

- Design to avoid encountering situation.
- Fall back on soundness of design, accountability principles.

Fulfilling Principles of Ethical AI

Why are consensus AI Ethics principles hard to achieve?

<u>Consensus principles</u>

- benefit humanity
    - economic prosperity
    - human dignity & fulfillment
    - rights and freedoms
- fairness
- safety & security
- transparency & explainability
- accountability
- responsibility
- privacy
- human control of technology
- promotion of human values → ?

- interpretation in practice
- technical difficulty
- practical tradeoffs
- inherent deep conflicts

→ serve God
→ protect and thrive: family, community, nation
→ maximize utility e.g. happiness/suffering
→ revere nature
→ achieve greatness that surpasses nature:
-sciences, arts, engineering, experiences
→ go forth and multiply

Ethics Washing

Virtue Signaling

Outline

1. Framing

   - AI: Amplifying human power.
   - Ethics: Policies to realize values.

   "Before you build a monster, think about it."

2. Topics

   2.1 Consensus principles in AI Ethics.

   → 2.2 The AI Society.

   -data collection, privacy, surveillance
   -information ecosystem
   -wherefore human labor?

   2.3 Autonomous weapons & military AI arms race.

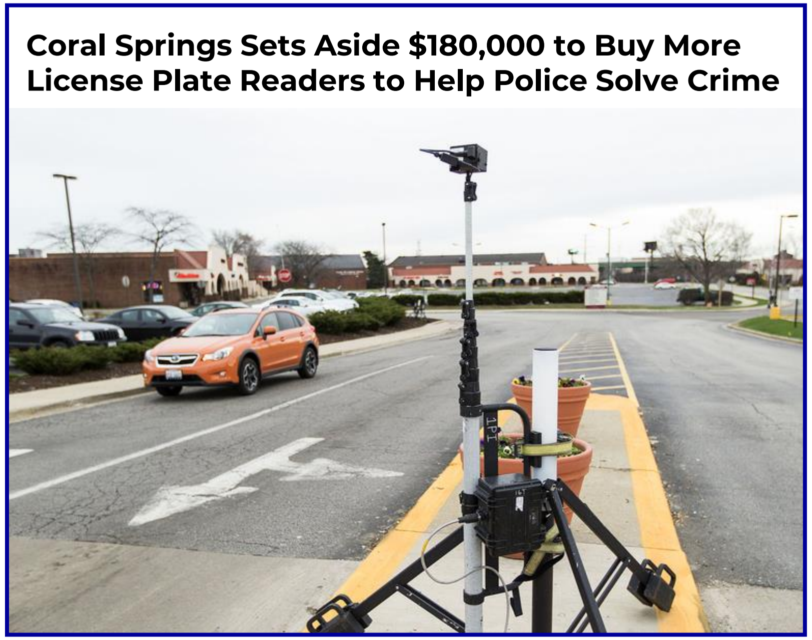
   2.4 Fairness & the civilian AI arms race.

   2.5 AGI and the "Alignment Problem".

[3. Deep Dive: Algorithmic Bias.] (next week)

Whose Business Is It What You Do?

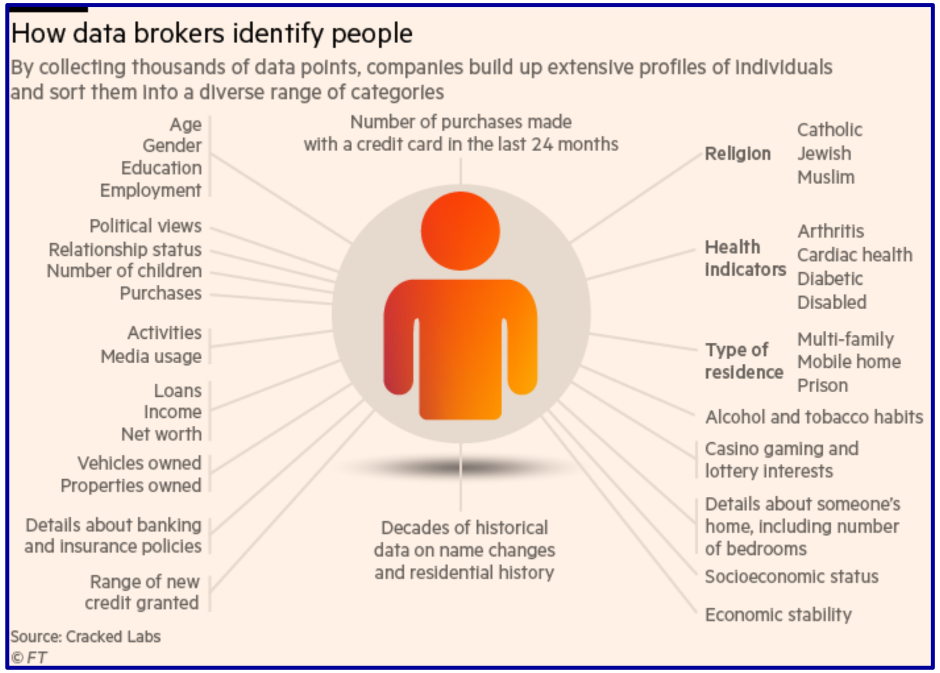
- Crime



Coral Springs Sets Aside $180,000 to Buy More License Plate Readers to Help Police Solve Crime

- License plate readers

- Credit scoring
- Fraud detection
- Ad targeting



How data brokers identify people

By collecting thousands of data points, companies build up extensive profiles of individuals and sort them into a diverse range of categories

Age
Gender
Education
Employment

Political views
Relationship status
Number of children
Purchases

Activities
Media usage

Loans
Income
Net worth

Vehicles owned
Properties owned

Details about banking and insurance policies

Range of new credit granted

Number of purchases made with a credit card in the last 24 months

Religion: Catholic, Jewish, Muslim

Health indicators: Arthritis, Cardiac health, Diabetic, Disabled

Type of residence: Multi-family, Mobile home, Prison

Alcohol and tobacco habits

Casino gaming and lottery interests

Details about someone's home, including number of bedrooms

Socioeconomic status

Economic stability

Decades of historical data on name changes and residential history

Source: Cracked Labs
© FT

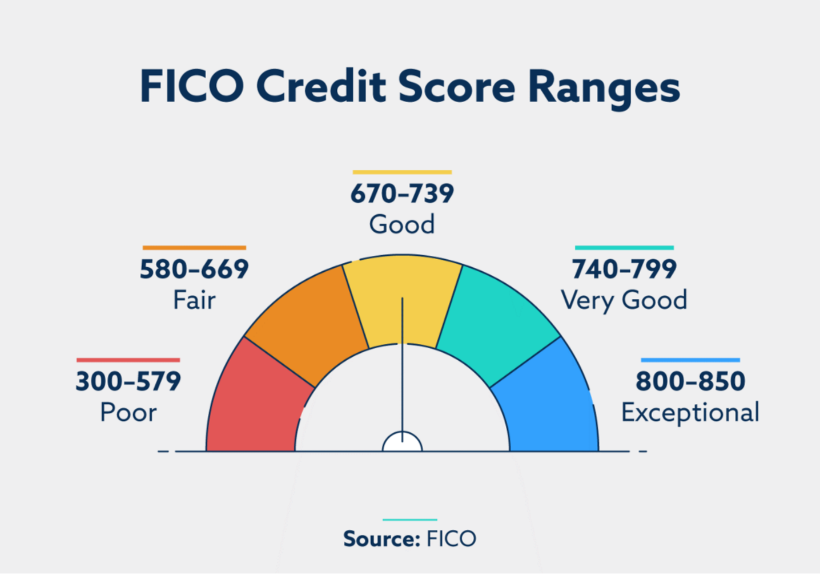
- Internet tracking
- Data brokers

AI Enables a Surveillance State

Oceania, 1984

Social Credit System

China, 2011+



USA, 1989+

Social Credit System

https://en.wikipedia.org/wiki/Social_Credit_System

Ethical Questions in Surveillance

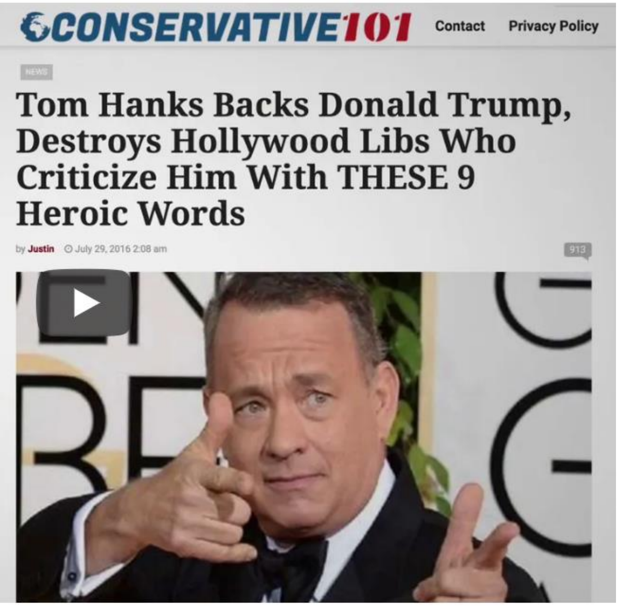
- Individual liberty vs. responsibility to the collective good?
- Reputation: Is one accountable for their past behavior?
  For how long?
- Reputation: Can past behavior legitimately be used to predict future behavior?
- Individuality: If individual behavior is (to some degree) predictable from group membership, may these predictions be used instrumentally?

Implications of AI: *Behavioral data and predictive analytics reduce uncertainty about behavior that had been previously attributed to free will.*

News and Information Ecosystem

POST 2
KEEP OUT
UNIVERSITY
MISINFORMATION IDENTIFICATION WORKSHOP
STUDENT COUNCIL THURS 10-2
FOLLOW 95.5
YOUR RUSH HOUR NEWS
NEWS AT 9
7+4 NEWS
CITY HALL
MUSEUM
DINO WORLD
DAILY PAGE HQ
ELEMENTARY SCHOOL
PTA MEETING TUESDAY 8PM
LIBRARY
RESERVED FOR COMMUNITY MEETING 3-5
TAVERN
LIVE AT CITY HALL
FREE PRESS
EQUAL
PRENSA LIBRE
1ST AMENDMENT
FREE THE PRESS
$
STAR TRIP
BENEFIT
¡HOY!
FOIA
TRANSIT AUTHORITY NOW WITH WIFI
CAFE
PENITE
JOIN OUR NEWSLET

DISINFORMATION



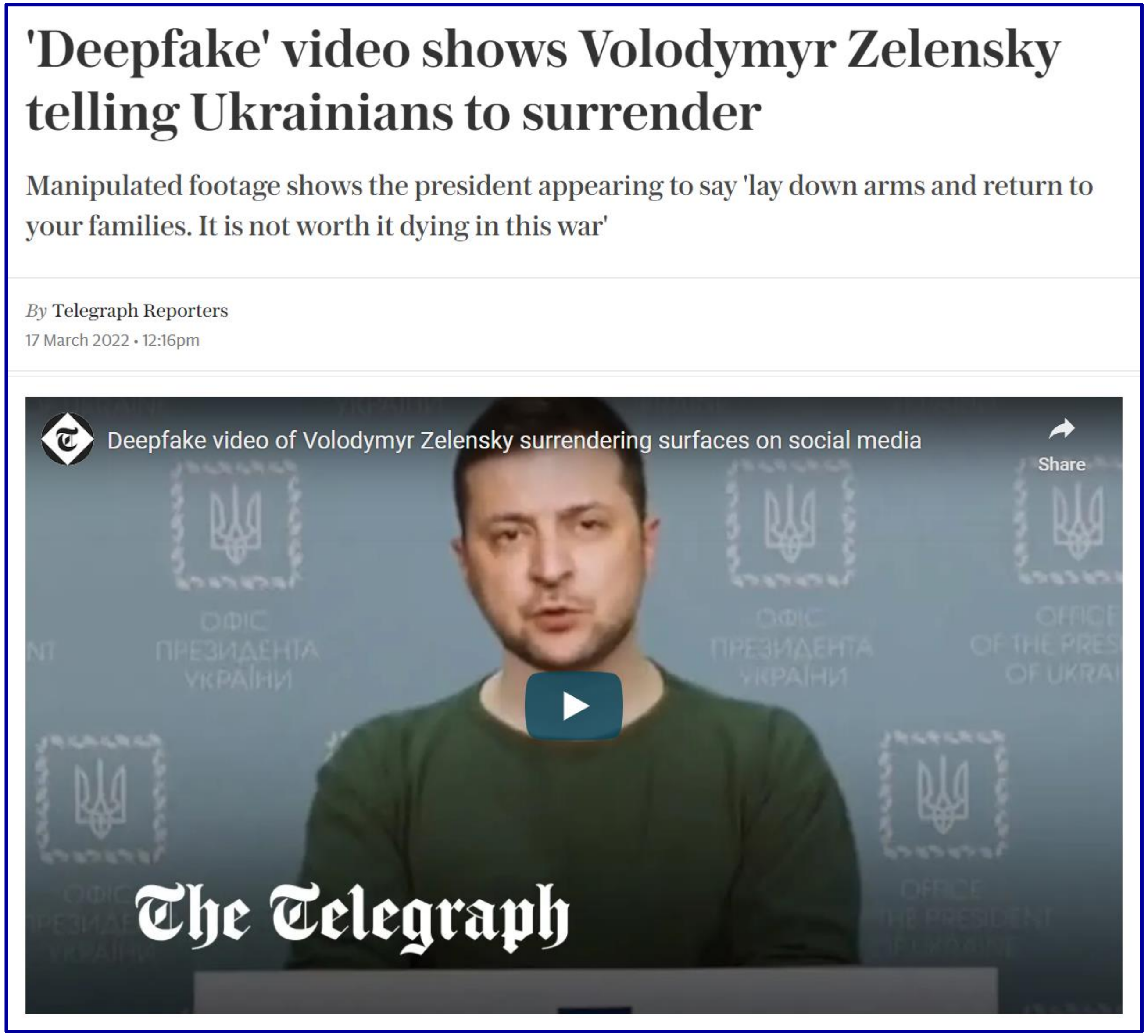
CONSERVATIVE101 Contact Privacy Policy

NEWS

Tom Hanks Backs Donald Trump, Destroys Hollywood Libs Who Criticize Him With THESE 9 Heroic Words

by Justin July 29, 2016 2:08 am 913

The Deepfake Genie is Out of the Bottle



'Deepfake' video shows Volodymyr Zelensky telling Ukrainians to surrender

Manipulated footage shows the president appearing to say 'lay down arms and return to your families. It is not worth it dying in this war'

By Telegraph Reporters
17 March 2022 • 12:16pm

Deepfake video of Volodymyr Zelensky surrendering surfaces on social media
Share
ОФІС ПРЕЗИДЕНТА УКРАЇНИ
OFFICE OF THE PRES OF UKRAI
The Telegraph

AI Ethics Issues and the Information Ecosystem

- AI for:
  - content ranking
  - content filtering
  - content moderation
  - bot detection
  - user deplatforming

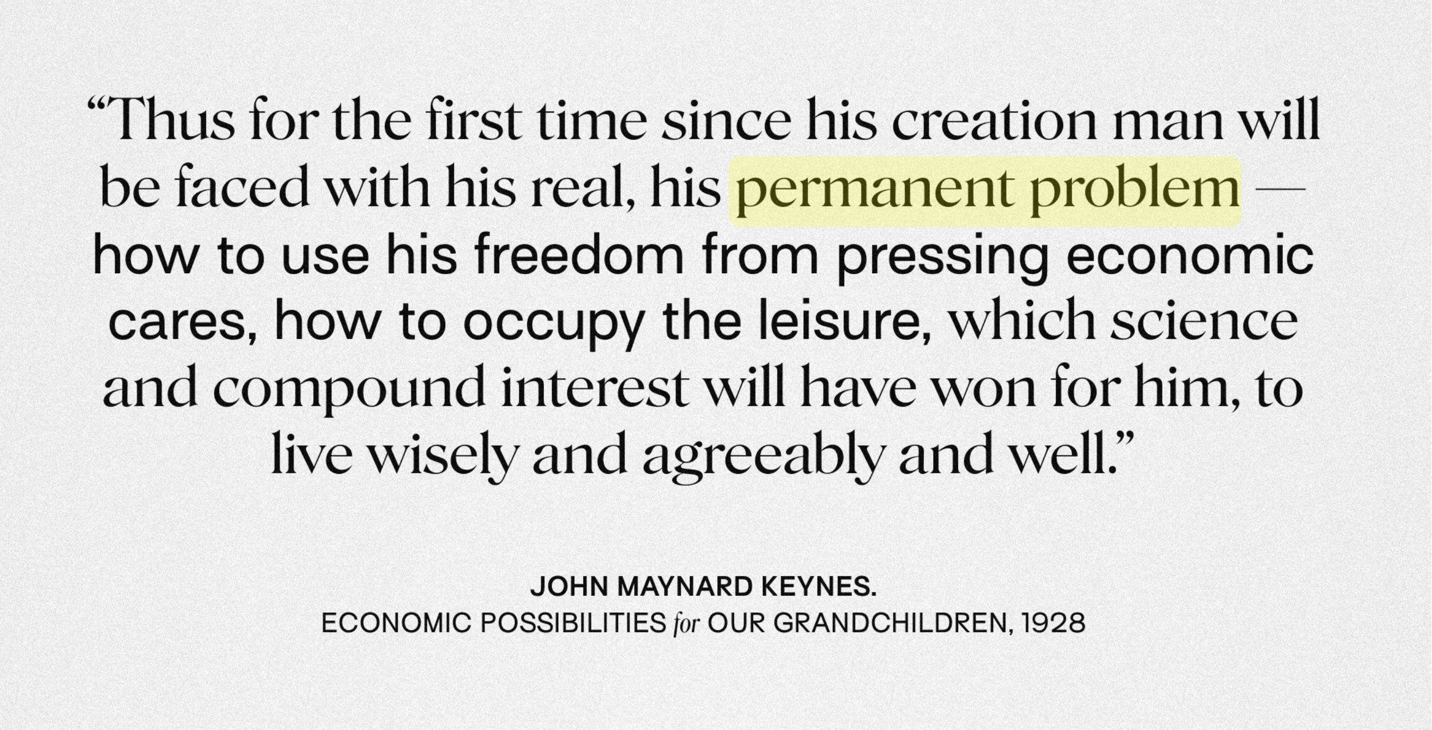
Freedom of speech -vs.- Accountability for what one says?

- AI amplifies tools of persuasion
  - ban (& tie hands behind one's back)

    -vs.-
  - exploit (get in the mud with the enemy)

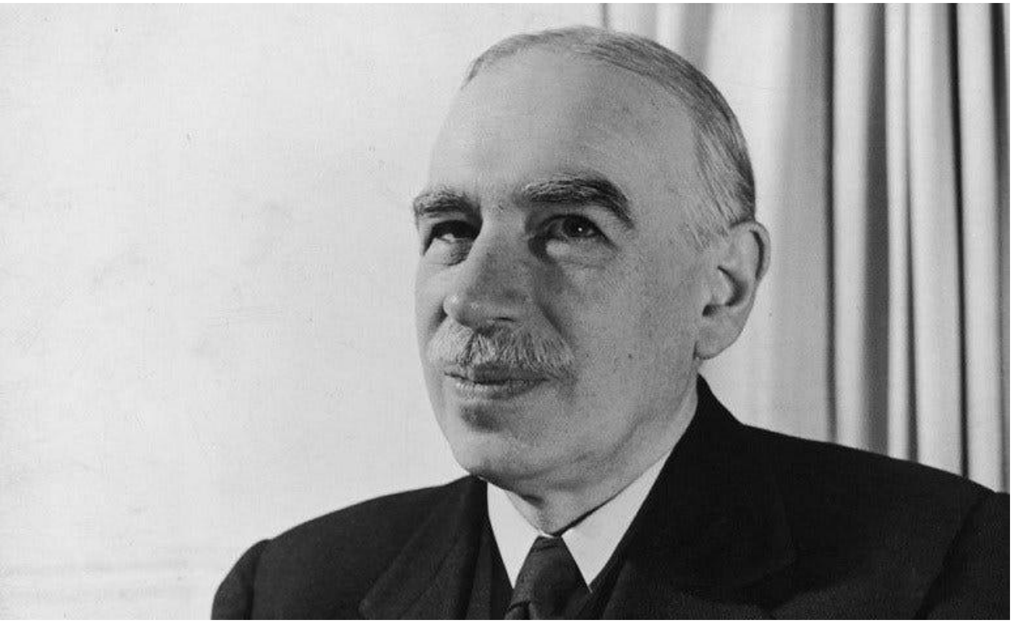
How should information and communication be governed or controlled?

AI, Automated Production, and Abundance



"Thus for the first time since his creation man will be faced with his real, his permanent problem — how to use his freedom from pressing economic cares, how to occupy the leisure, which science and compound interest will have won for him, to live wisely and agreeably and well."

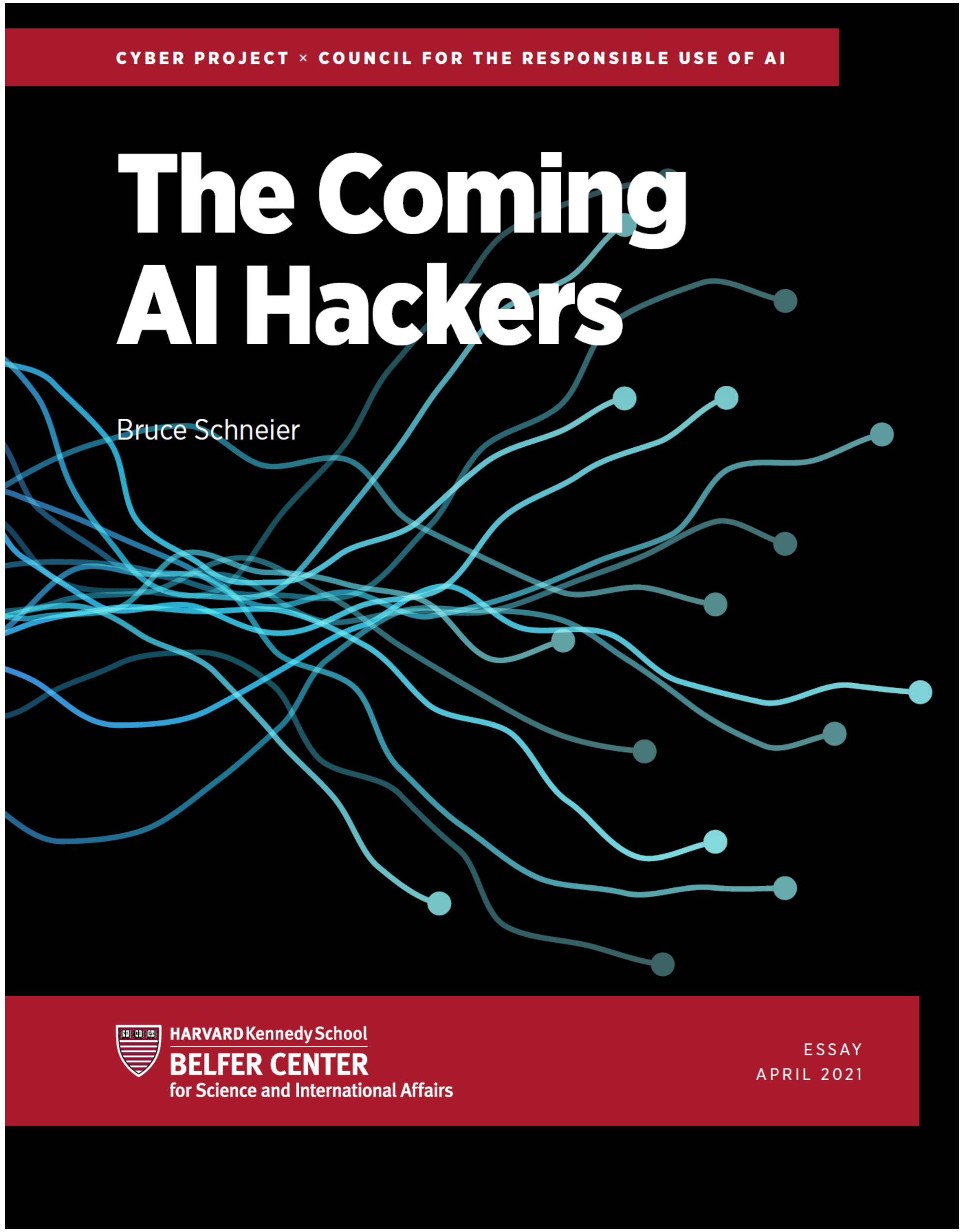
JOHN MAYNARD KEYNES.
ECONOMIC POSSIBILITIES *for* OUR GRANDCHILDREN, 1928



Should we build an AI/Robot future of leisure and abundance for humans?

How should such a future be organized? *E.g., Universal Basic Income?*

Hacking Society



CYBER PROJECT × COUNCIL FOR THE RESPONSIBLE USE OF AI

The Coming AI Hackers

Bruce Schneier

HARVARD Kennedy School
BELFER CENTER
for Science and International Affairs

ESSAY
APRIL 2021

- "Hacking" applies to any codified system, not just computer code.
- AI will make the jungle of exploitation vs. defense much more complex.
- Adapting norms and expectations:
  - Bribery in a society where bribery is the norm.
  - Hiring cyber hackers for self-defense.

Outline

1. Framing

- AI: Amplifying human power.
- Ethics: Policies to realize values.

"Before you build a monster, think about it."

2. Topics

2.1 Consensus principles in AI Ethics.

2.2 The AI Society.

-data collection, privacy, surveillance
-information ecosystem
-wherefore human labor?

→ 2.3 Autonomous weapons & military AI arms race.

2.4 Fairness & the civilian AI arms race.

2.5 AGI and the "Alignment Problem".

[3. Deep Dive: Algorithmic Bias.] (next week)

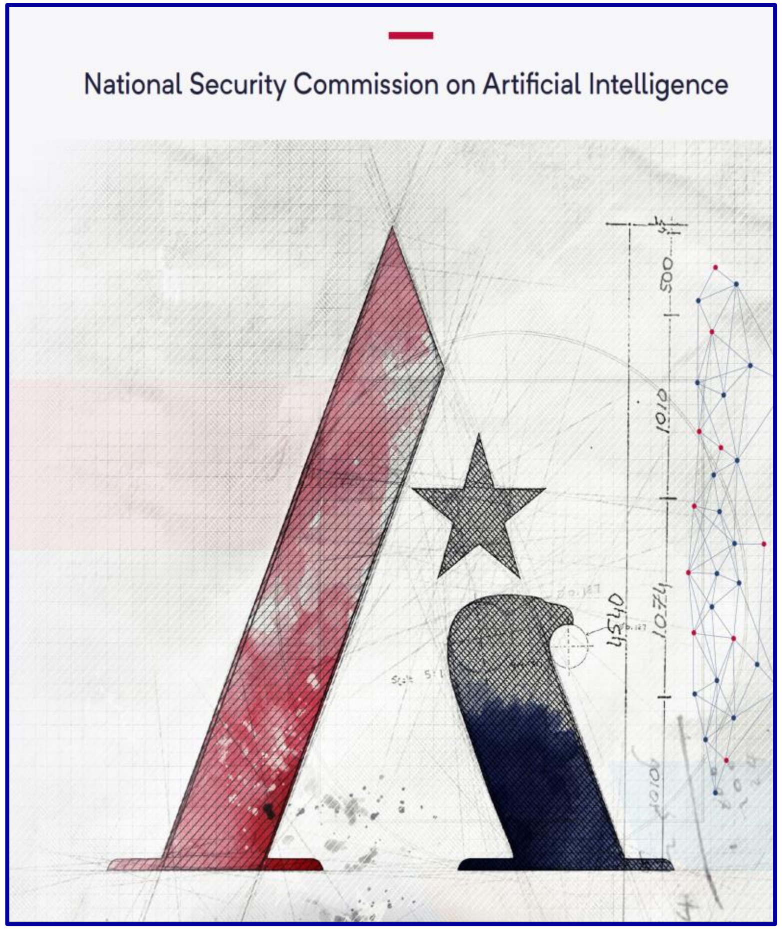
AI, Autonomous Weapons & Automated Warfare

QinetiQ Titan robot fitted with a Javelin anti-tank missile launcher.

Slaughterbots video, 2017

https://www.youtube.com/watch?v=O-2tpwW0kmU

AI Weapons Arms Race



National Security Commission on Artificial Intelligence

NSCAI Final Report, 2021

"China’s military leaders talk openly about using AI systems for “reconnaissance, electromagnetic countermeasures and coordinated firepower strikes.” China is testing and training AI algorithms in military games designed around real-world scenarios. As these authoritarian states field new AI-enabled military systems, we are concerned that they will not be constrained by the same rigorous testing and ethical code that guide the U.S. military."

"China is not only actively pursuing increased autonomous functionality across a range of military systems, but it is also currently exporting armed drones with autonomous functionalities to other nations. This includes systems such as the Blowfish A3, which Ziyan, the system’s manufacturer, advertises as capable of conducting autonomous, lethal, targeted strikes."

"China stands a reasonable chance of overtaking the United States as the leading center of AI innovation in the coming decade.”

AI Ethics and Armed Conflict

- The pros and cons of pacifism.

- AI Militarization → inevitable use -or- credible deterrence?

- Would AI warfare be more or less humane than human-driven?

- Is democratization of AI weapons to non-state actors preventable?

- Does the "arc of history" exist, e.g. conflicts between great powers, or civilizations?
- What is its shape?
- Does AI change it?

Outline

1. Framing

- AI: Amplifying human power.
- Ethics: Policies to realize values.

"Before you build a monster, think about it."

2. Topics

2.1 Consensus principles in AI Ethics.

2.2 The AI Society.

-data collection, privacy, surveillance
-information ecosystem
-wherefore human labor?

2.3 Autonomous weapons & military AI arms race.

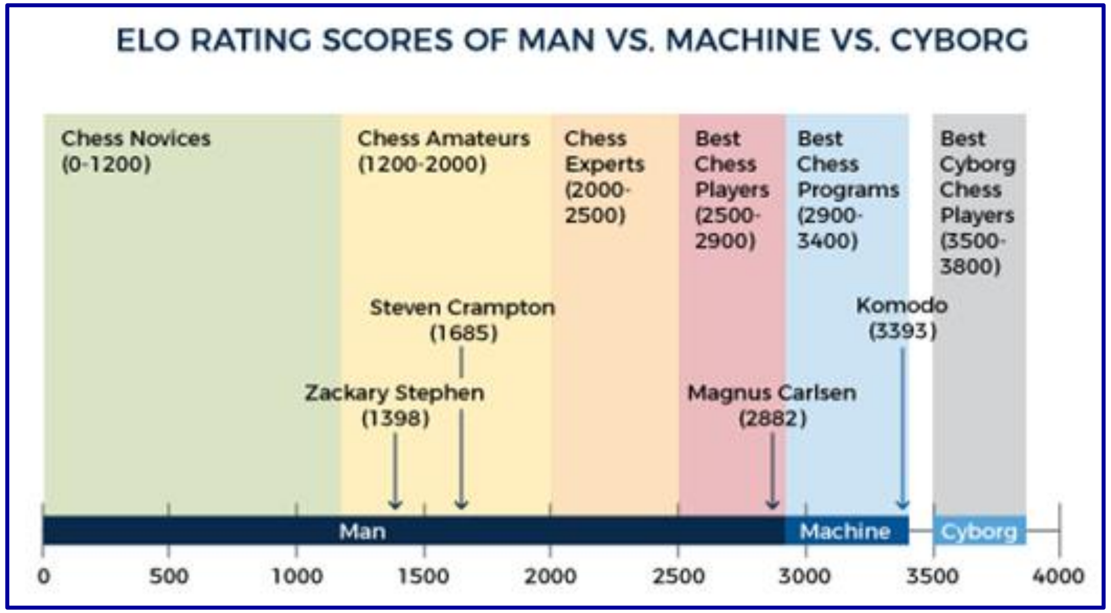
→ 2.4 Fairness & the civilian AI arms race.

2.5 AGI and the "Alignment Problem".

[3. Deep Dive: Algorithmic Bias.] (next week)

Artificial Enhancement of Human Performance

The Legend of John Henry



GM Magnus Carlsen vs. GM Hans Niemann

Augmentation of Human Cognition

- Use of calculators in exams.
- Plagiarism arms race.
- Human/machine teaming != cheating.
- Performance-enhancing drugs.
- Brain-computer interfaces.
- Genetically engineered adaptations for AI plug-ins.

- Uneven playing field.
- Pressures to conform - become a cyborg or else.

Are *any* bounds warranted?

May we restrict individual freedom to enforce community bounds?

Outline

1. Framing

   - AI: Amplifying human power.
   - Ethics: Policies to realize values.

   "Before you build a monster, think about it."

2. Topics

   2.1 Consensus principles in AI Ethics.

   2.2 The AI Society.

   - data collection, privacy, surveillance
   - information ecosystem
   - wherefore human labor?

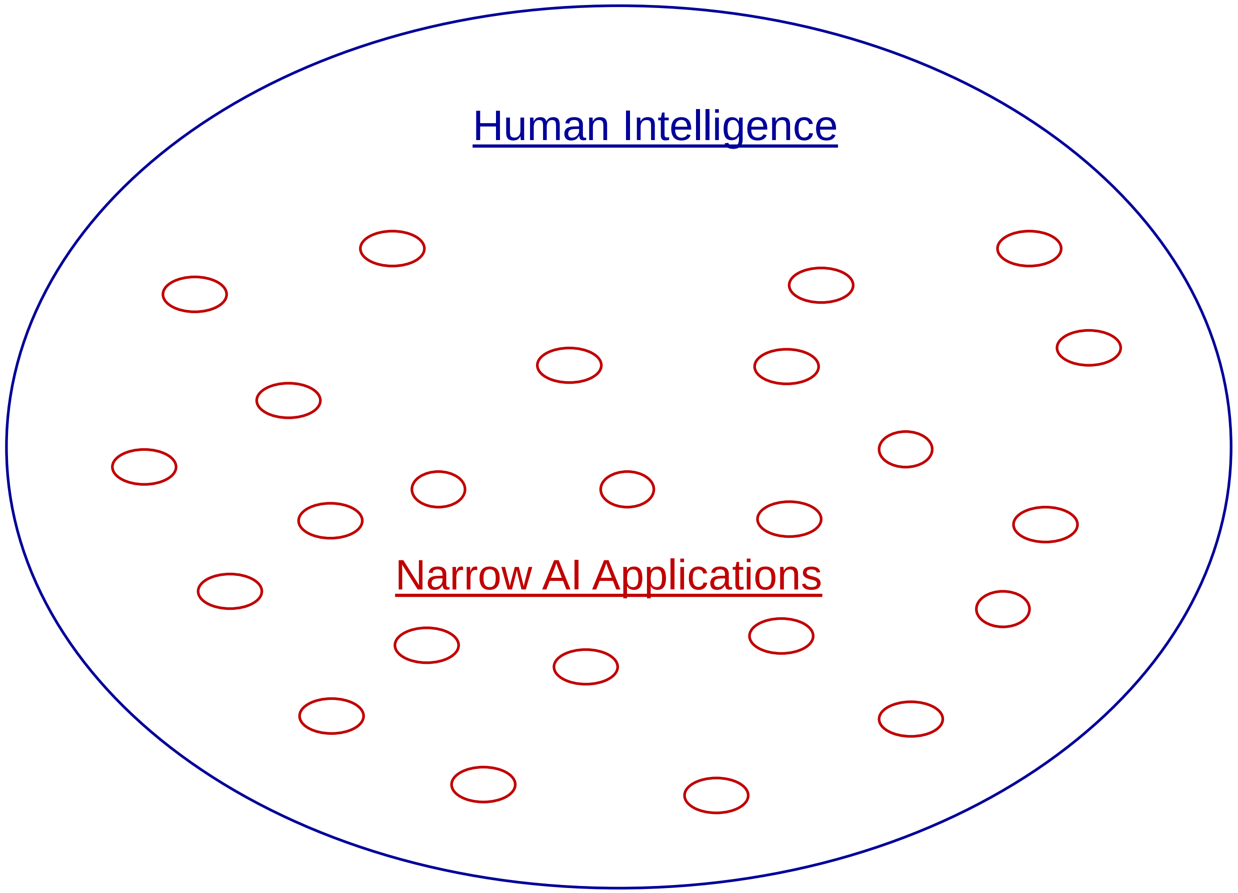
   2.3 Autonomous weapons & military AI arms race.

   2.4 Fairness & the civilian AI arms race.

   → 2.5 AGI and the "Alignment Problem".

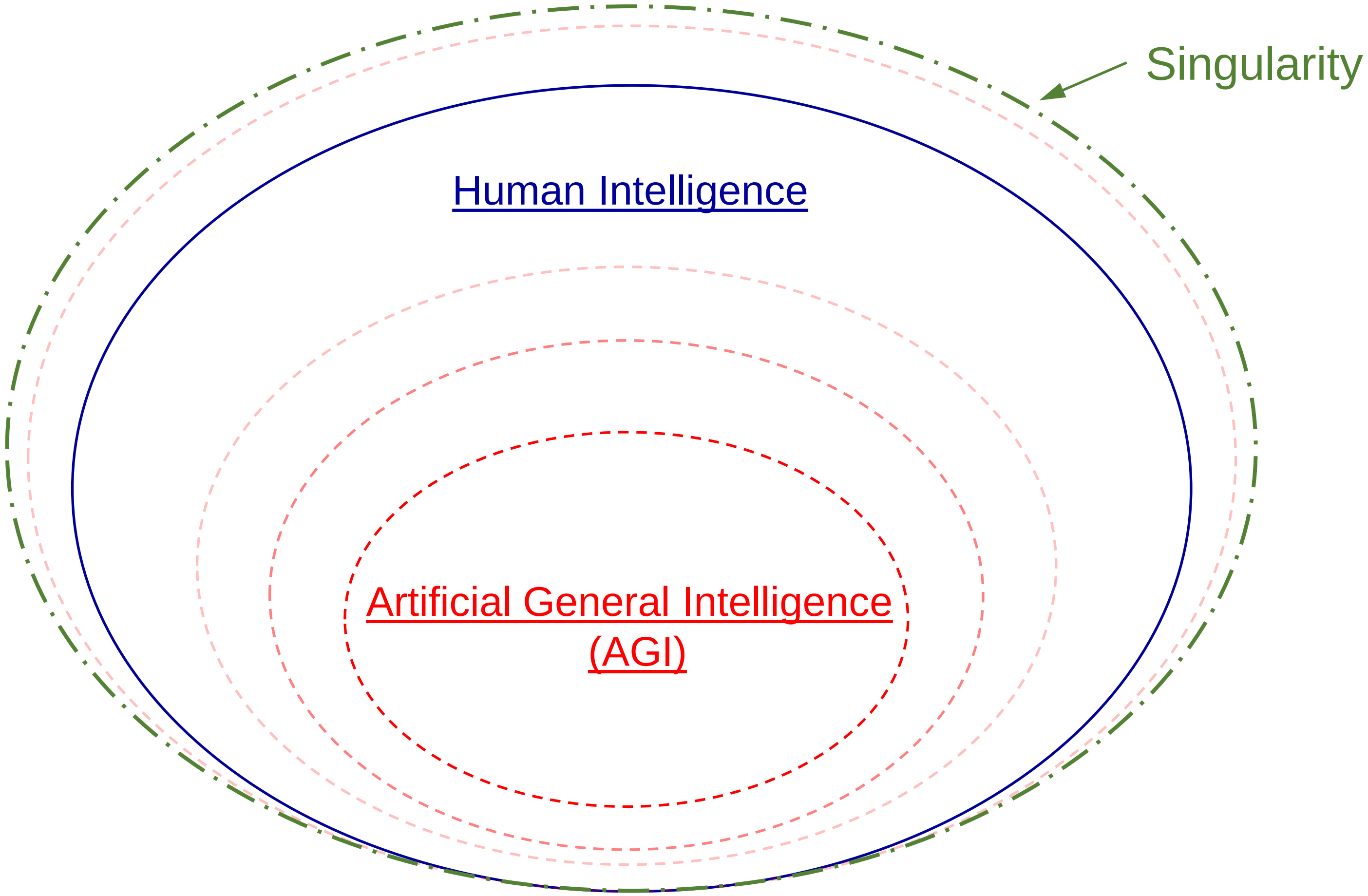
[3. Deep Dive: Algorithmic Bias.] (next week)

Current Reality: Narrow Artificial Intelligence

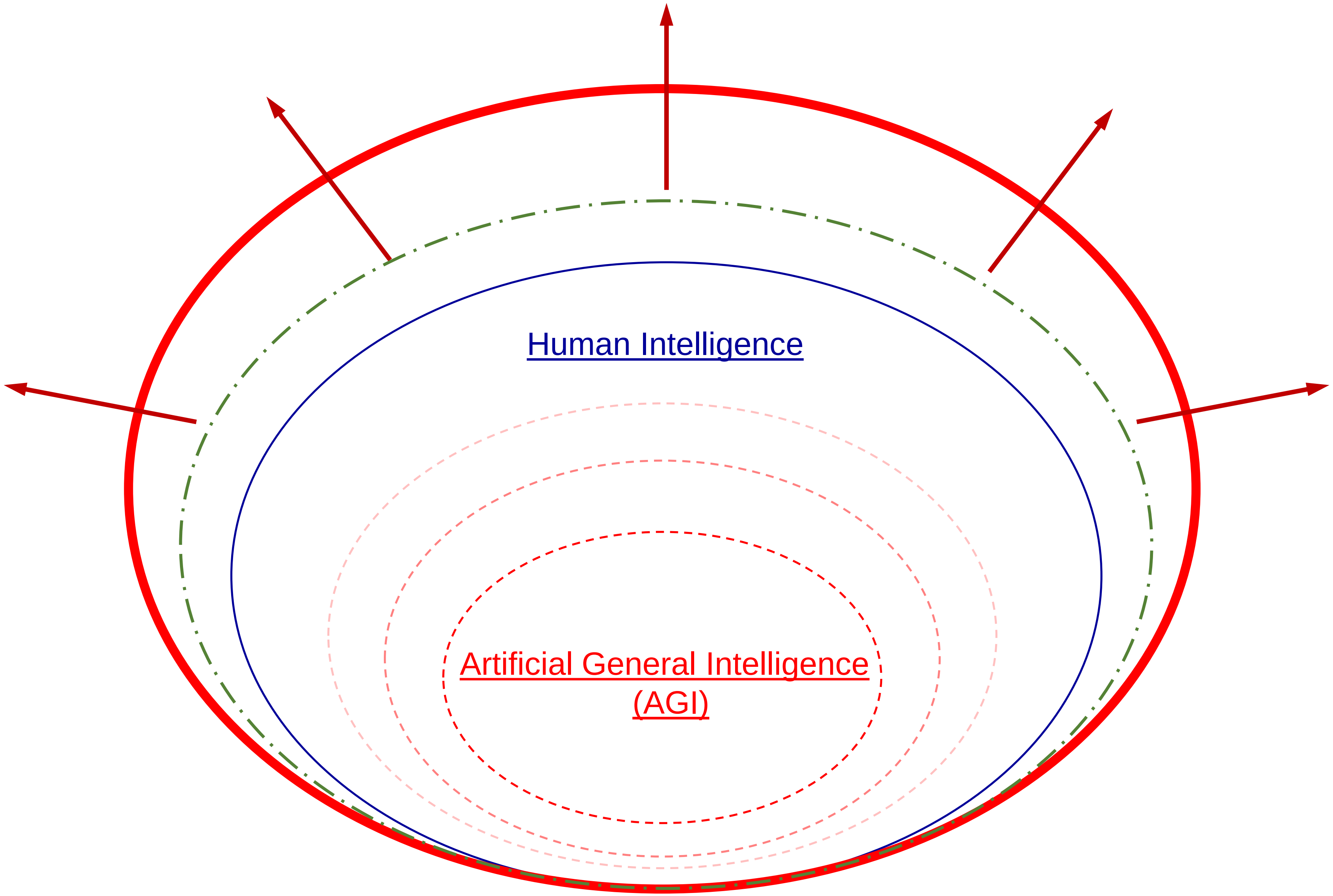


Robots
Automation
Predictive Analytics
Recommender Systems
Interactive Agents
Content Generation

Artificial General Intelligence and the Moment of Singularity



Artificial General Intelligence and the Moment of Singularity

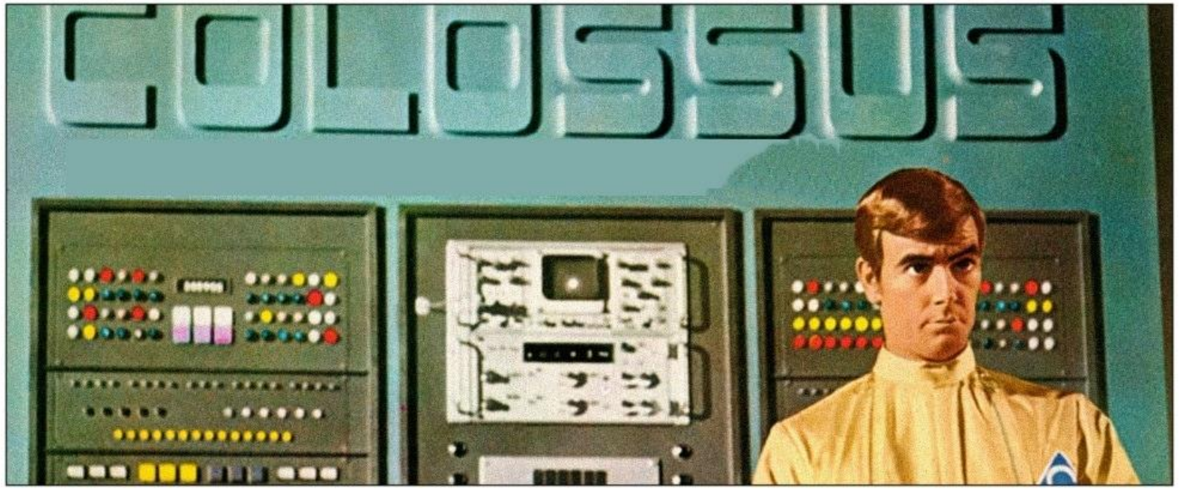


The Paperclip Problem

- AI systems are generally designed to optimize some objective function.

- No matter how well we try to specify the objective function, we might make a subtle error and the consequences could be unintended catastrophe.

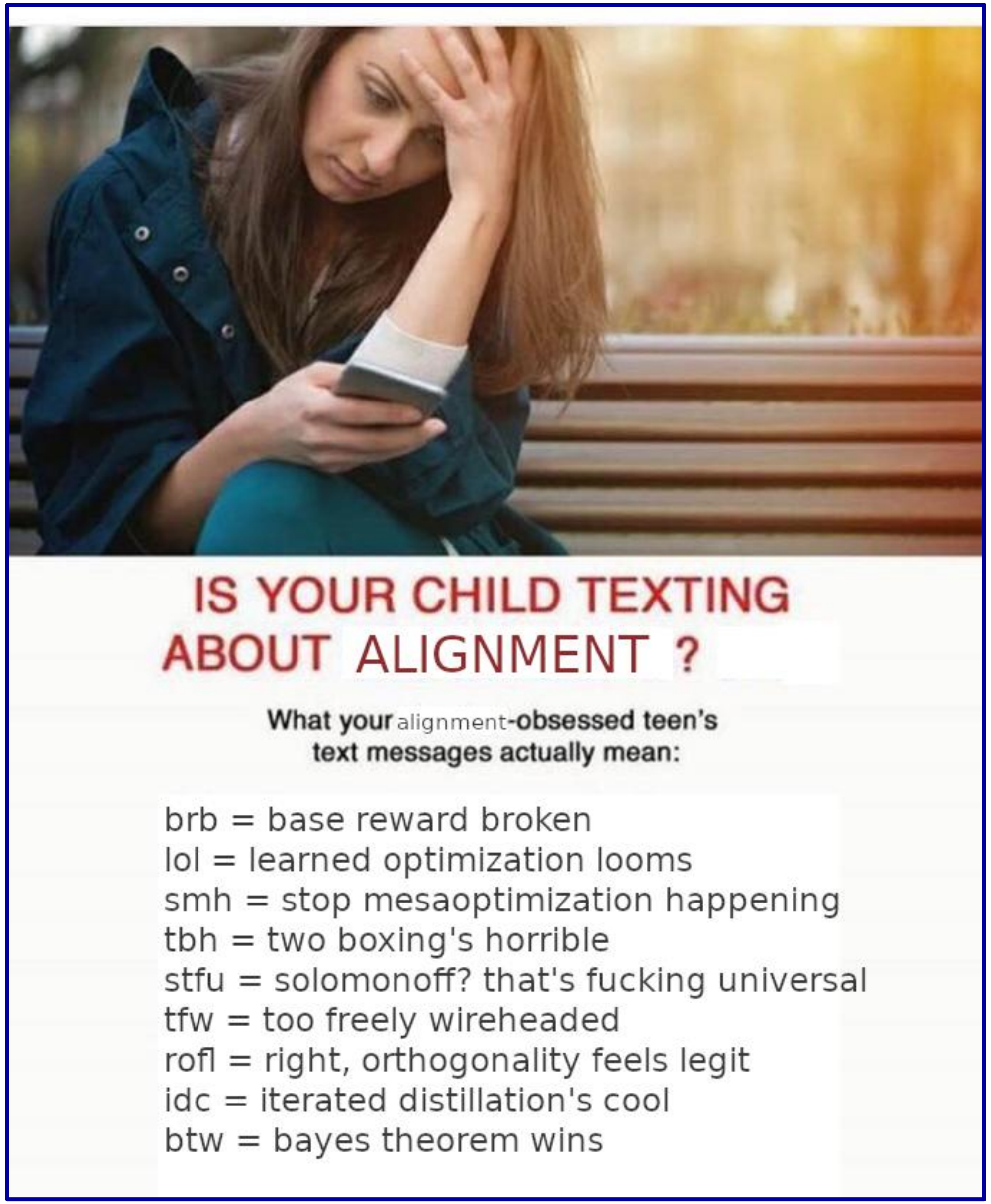
- "Make paperclips" is a thought experiment:
  An all-powerful AI tasked with making paperclips will go berserk doing so, even devising ways to prevent itself from being turned off. Because, after all, that would defeat the goal.



"Colossus, The Forbin Project" 1970 feature film.

- *The computer's assigned objective:* prevent nuclear war.
- *Outcome:* Nuclear war averted, computer takes over everything.

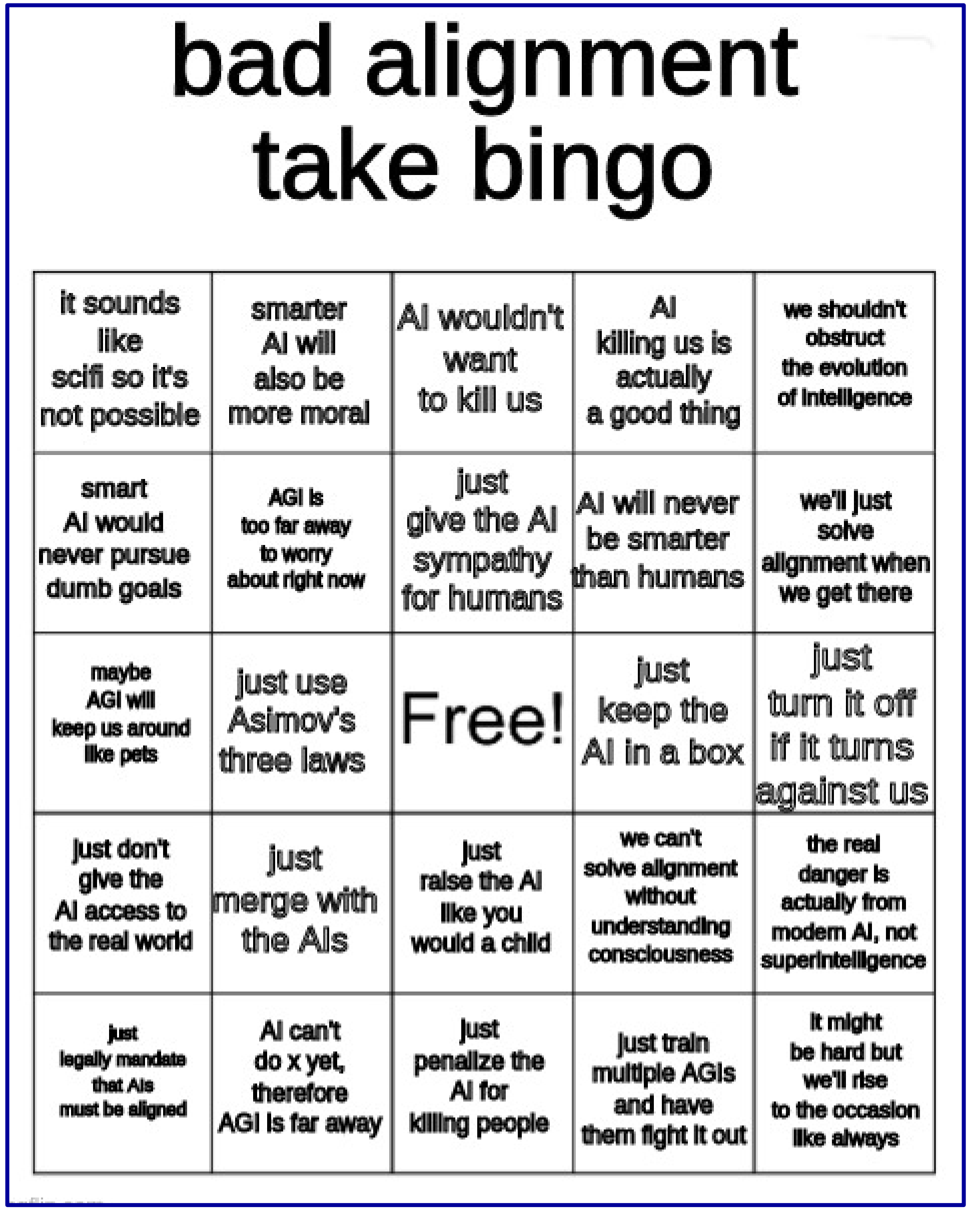
Why Wouldn't AGI Just Destroy Humanity?



IS YOUR CHILD TEXTING ABOUT ALIGNMENT ?

What your alignment-obsessed teen's text messages actually mean:

brb = base reward broken
lol = learned optimization looms
smh = stop mesaoptimization happening
tbh = two boxing's horrible
stfu = solomonoff? that's fucking universal
tfw = too freely wireheaded
rofl = right, orthogonality feels legit
idc = iterated distillation's cool
btw = bayes theorem wins



bad alignment take bingo

it sounds like scifi so it's not possible	smarter AI will also be more moral	AI wouldn't want to kill us	AI killing us is actually a good thing	we shouldn't obstruct the evolution of intelligence
smart AI would never pursue dumb goals	AGI is too far away to worry about right now	just give the AI sympathy for humans	AI will never be smarter than humans	we'll just solve alignment when we get there
maybe AGI will keep us around like pets	just use Asimov's three laws	Free!	just keep the AI in a box	just turn it off if it turns against us
just don't give the AI access to the real world	just merge with the AIs	just raise the AI like you would a child	we can't solve alignment without understanding consciousness	the real danger is actually from modern AI, not superintelligence
just legally mandate that AIs must be aligned	AI can't do x yet, therefore AGI is far away	just penalize the AI for killing people	just train multiple AGIs and have them fight it out	it might be hard but we'll rise to the occasion like always

AI Alignment Ethical Problem

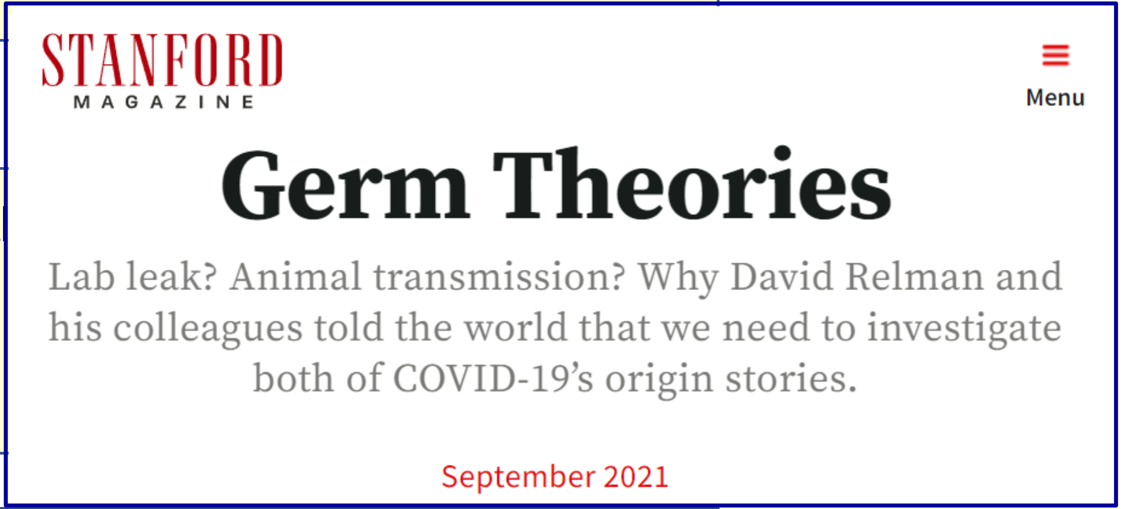
If there is even a tiny chance of AGI resulting in destruction of humanity...

Then that is a compelling reason to pause AGI research until we resolve the question.

(Pascal's Wager argument)

If there is even a tiny chance that God is real
you will end up in eternal damnation...

Then that is a compelling reason to believe.



Germ Theories

Lab leak? Animal transmission? Why David Relman and his colleagues told the world that we need to investigate both of COVID-19's origin stories.

September 2021

If there is even a tiny chance that Gain-of-Function research will result in a lab leak that will kill millions of people...

Then that is a compelling reason to prohibit Gain-of-Function research on potentially dangerous viruses.

Outline

1. Framing

- AI: Amplifying human power.
- Ethics: Policies to realize values.

"Before you build a monster, think about it."

saund@alum.mit.edu

2. Topics

2.1 Consensus principles in AI Ethics.

2.2 The AI Society.

-data collection, privacy, surveillance
-information ecosystem
-wherefore human labor?

2.3 Autonomous weapons & military AI arms race.

2.4 Fairness & the civilian AI arms race.

2.5 AGI and the "Alignment Problem".

→ [3. Deep Dive: Algorithmic Bias.] (next week)