2025-10-09

Superintelligence and risks of AI

5ARC0

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

Plan for Today

- Review: Goodall and Awad et al.
- Risks and uncertainties
- What are some risks posed by AI?
- Superintelligence as an AI risk
- Strategies for deep uncertainty

TU/e

Goodall: Phases in programming moral machines

1. Rational: Rule-based ethics or outcome-based ethics
    - Goodall: Rule-based ethics is too minimal to determine outcomes in unclassified cases or conflict cases, whereas outcome-based ethics recommends some morally bad decisions.
2. AI-based: Use training data so that AI can learn to make moral decisions.
    - Goodall: This would reproduce human shortcomings. Also, it would lack explainability.
    - Another issue: Where would the training data come from? Whose judgments would be taken into account, and how?
3. Natural language feedback (XAI) allows us to formulate the rules “post-hoc”

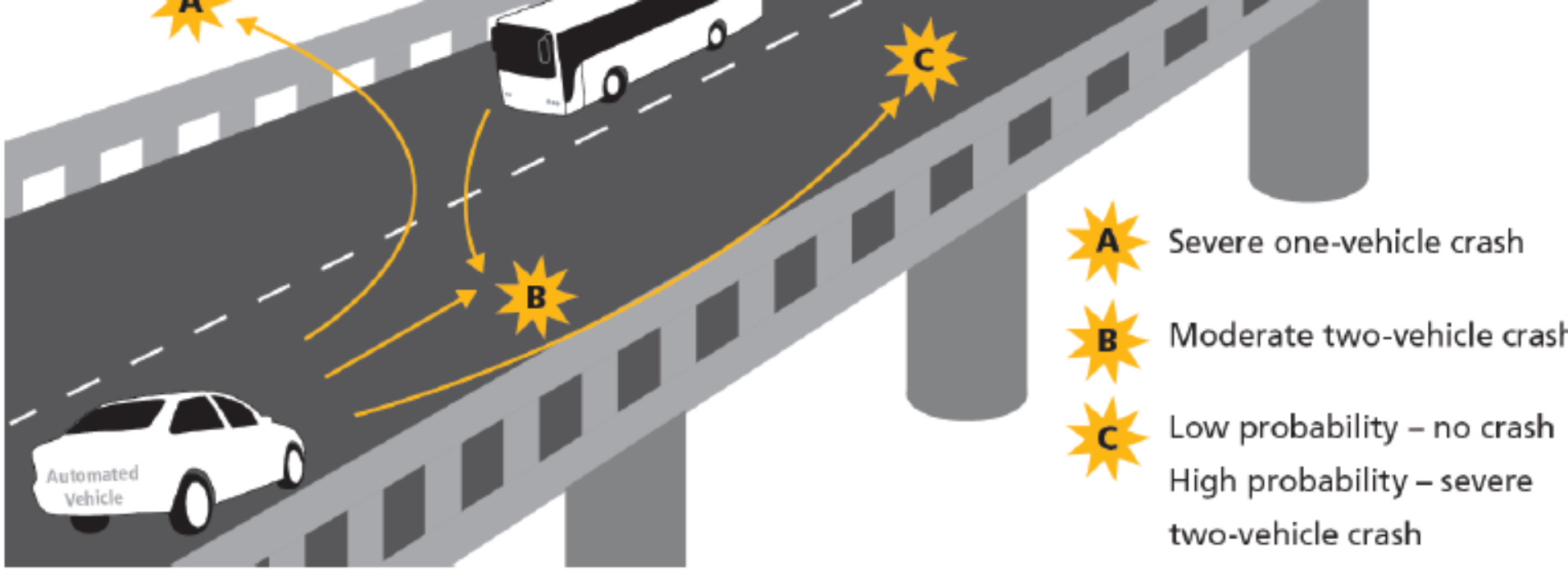


FIGURE 1 Diagram of three alternative trajectories for an automated vehicle when an oncoming bus suddenly enters the vehicle's lane

Moral aspect of crash decisions in mixed traffic

Moral Machines project

Why ethics rather than direct regulation (Awad et al. 2020, p. 49)?

- Black Box machine learning makes the principles opaque
- Technology changes faster than can be regulated
- Source of errors hard to trace to the technology or the source data

> "All these factors make it especially challenging to regulate the negative externalities created by intelligent machines, and to turn them into moral machines. And if the ethics of machine behavior are not sorted out soon, it is likely that societal push-back will drastically slow down the adoption of intelligent machines."

Argument for crowd-sourcing AV ethics (Awad et al. 2020, p. 50)

1. SITL ("society in the loop") is necessary for a dynamic consensus on the ethics of intelligent machines.
2. We need such a consensus if we wish to pursue the benefits of AVs and other full automation.
3. (We should wish to pursue the benefits of AVs and other full automation.)
4. Therefore, we should pursue SITL.

Some results of Moral Machines project

People prefer:

- Sparing the lawful over the unlawful
- Sparing humans over pets
- Sparing the greater number over the fewer
- Sparing younger humans over older

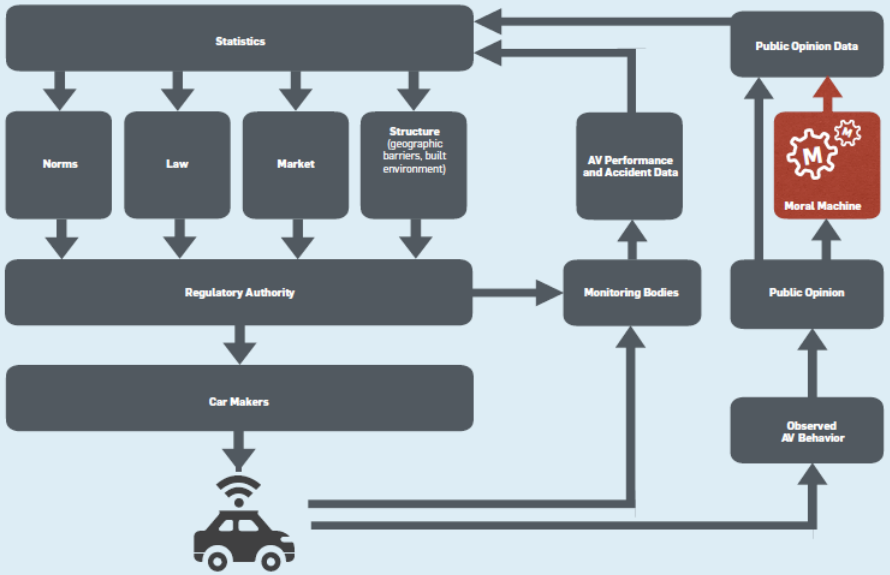
People are silent on some issues:

- Sparing vehicle occupants over non-occupants
- Killing as a means to saving lives
- Sparing the cautious over the uncautious

Where Awad et al think the crowdsourcing fits in

Figure 4. A society-in-the-loop framework for AV regulation.

The model does not represent an actual regulatory system, but it clarifies how a crowdsourcing platform like the Moral Machine fits into the broader regulatory system by providing data on societal norms.



Summing up from last time

- There are broadly two ways that we can implement human-centered AI: one in which humans adapt to AI, and the other in which AI is adapted to humans.
- Human-centered AI may come with handoffs, where we face additional questions of control and responsibility.
- If we choose a more automation-centered approach (autonomous machines), we may face the complicated question of how to program moral aspects of decision making into the automation.

A.I. Poses 'Risk of Extinction,' Industry Leaders Warn

Leaders from OpenAI, Google DeepMind, Anthropic and other A.I. labs warn that future systems could be as deadly as pandemics and nuclear weapons.

EDITORIAL | 27 June 2023

Stop talking about tomorrow's AI doomsday when AI poses risks today

It's time to talk about the real AI risks

RightsCon want us to focus less on existential threats, and more on here and now.

The militarized AI risk that's bigger than "killer robots"

The nuclear stakes of putting too much trust in AI.

By Jeffrey Lewis | Nov 28, 2023, 12:00pm EST

Sam Altman warns AI could kill us all. But he still wants the world to use it

By Samantha Kelly, CNN

7 minute read · Published 6:00 AM EDT, Tue October 31, 2023

ARTIFICIAL INTELLIGENCE / TECH

Top AI researchers and CEOs warn against 'risk of extinction' in 22-word statement

TU/e

AI Poses Risks

Seems impossible to read the news without seeing something on AI risks:

- Bias and discrimination
- Automation and job displacement
- Environmental impacts
- Disinformation and threat to democracy
- Autonomous weapons
- Privacy and surveillance
- Superintelligence
- Existential risk
- etc.

What is a *Risk*?

Superficial definition of risk: "the combination of the probability of an event and its consequence"

$$\text{risk} = \textit{probability} \times \textit{consequence}$$

e.g. risk = 50% × 10 deaths

What is a *Risk*?

Superficial definition of risk: "the combination of the probability of an event and its consequence"

$$\textbf{risk} = \textbf{\textit{probability}} \times \textbf{\textit{consequence}}$$

e.g. risk = 50% × 10 deaths

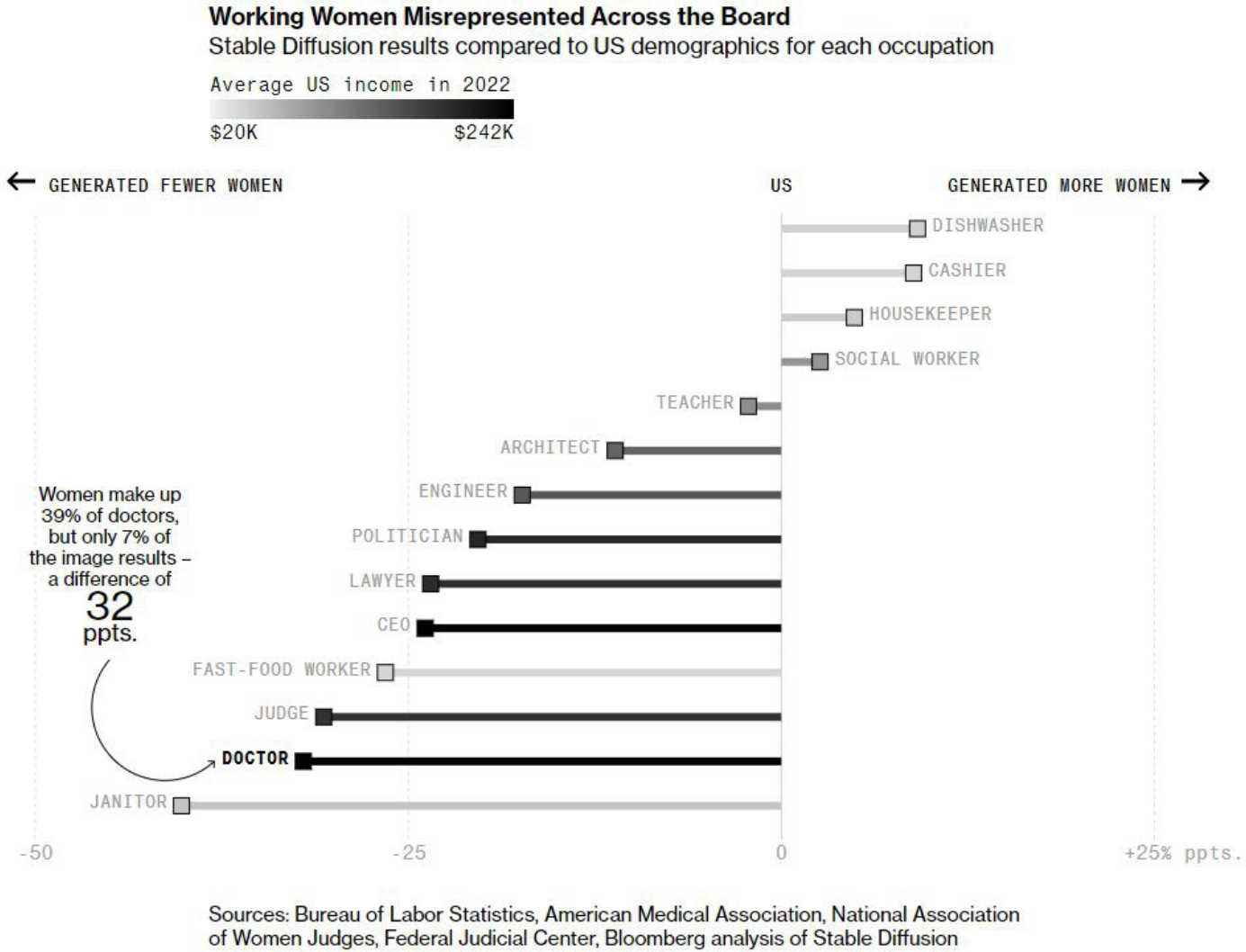
- Sometimes there is **uncertainty**
  - Cannot attribute a probability
- Or we are **ignorant of the scenarios that can materialise**
  - 'We don't know what we don't know'
- Technology is **safe** when the risk is **acceptable**

TU/e

Bias and Discrimination

- AI systems can misrepresent people, introduce biases, and discriminate against them

TU/e


Working Women Misrepresented Across the Board
Stable Diffusion results compared to US demographics for each occupation
Average US income in 2022
$20K
$242K
← GENERATED FEWER WOMEN
US
GENERATED MORE WOMEN →
DISHWASHER
CASHIER
HOUSEKEEPER
SOCIAL WORKER
TEACHER
ARCHITECT
ENGINEER
POLITICIAN
LAWYER
CEO
FAST-FOOD WORKER
JUDGE
DOCTOR
JANITOR
Women make up 39% of doctors, but only 7% of the image results – a difference of 32 ppts.
-50
-25
0
+25% ppts.
Sources: Bureau of Labor Statistics, American Medical Association, National Association of Women Judges, Federal Judicial Center, Bloomberg analysis of Stable Diffusion

Bloomberg

Bias and Discrimination

- AI systems can misrepresent people, introduce biases, and discriminate against them
- A problem for almost all applications
  - Loan applications, hiring, healthcare diagnostic, policing, etc.
- **General problem:** Data reflects existing underlying inequalities
  - Not obvious how this can be solved

TU/e

Misuse

- AI is often a dual-use technology: can be used for civil **and** military purposes
- Malicious actors can use AI for…
  - Creating biochemical weapons, launch cyber attacks, surveillance, etc.

Comment | Published: 07 March 2022

Dual use of artificial-intelligence-powered drug discovery

Fabio Urbina, Filippa Lentzos, Cédric Invernizzi & Sean Ekins

Nature Machine Intelligence **4**, 189–191 (2022) | Cite this article

AI chatbots could help plan bioweapon attacks, report finds

Large language models gave advice on how to conceal the true purpose of the purchase of anthrax, smallpox and plague bacteria

TU/e

Disinformation, Deepfakes, and Threat to Democracy

TU/e

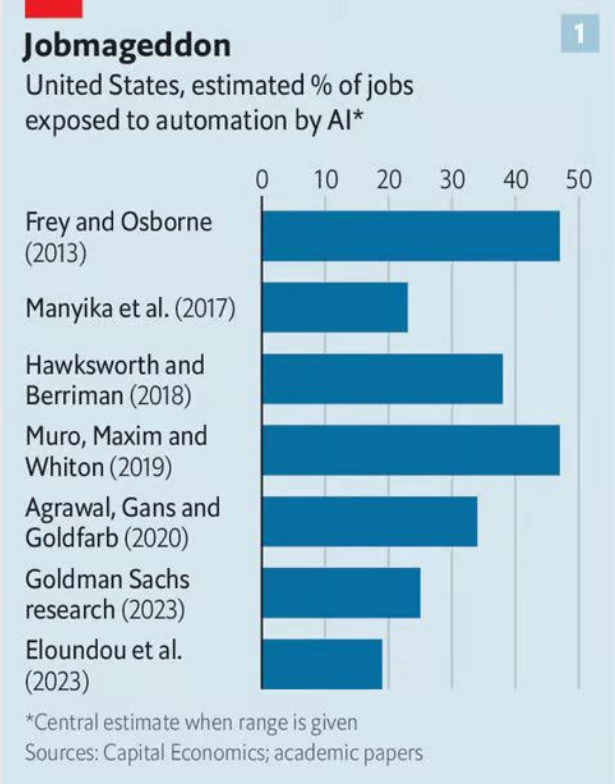
Disinformation, Deepfakes, and Threat to Democracy

- AI reduces the cost of generating *mis*/*dis*information at scale
- Deepfakes can harm individuals and have societal effects
- Can…
  - Empower authoritarian regimes
  - Disrupt democratic institutions and reduce trust

NPR

Automation and Job Displacement

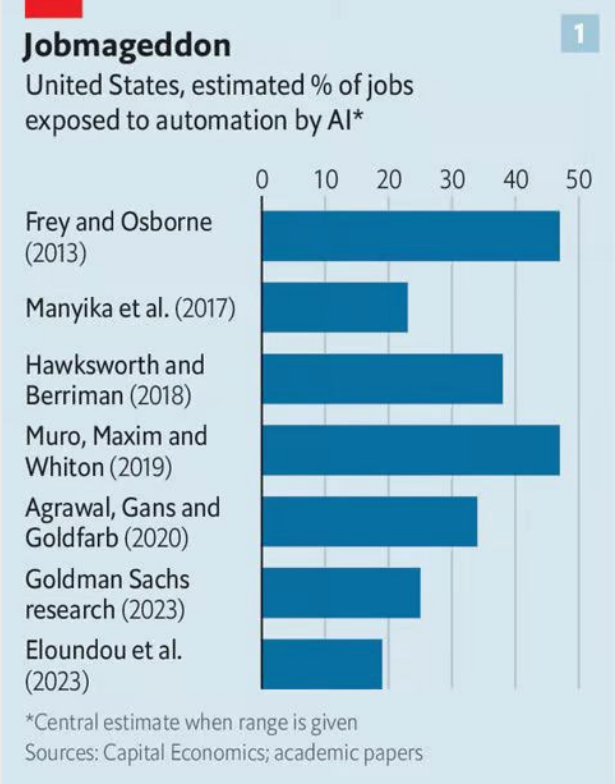
- Technology disrupts work: automate existing jobs, make some jobs obsolete, create new ones
- This time is different? AI targets *cognitive* tasks (coding, creativity, content creation, decision-making, etc.)



The Economist

Automation and Job Displacement

- Technology disrupts work: automate existing jobs, make some jobs obsolete, create new ones
- This time is different? AI targets *cognitive* tasks (coding, creativity, content creation, decision-making, etc.)
- Impact on the economy and society is very difficult to predict



The Economist

Enfeeblement and de-skilling

"Once the practical incentive to pass our civilization on to the next generation disappears, it will be very hard to reverse the process. One trillion years of cumulative learning would, in a real sense, be lost. We would become passengers in a cruise ship run by machines, on a cruise that goes on forever—exactly as envisaged in the film WALL-E. [...]

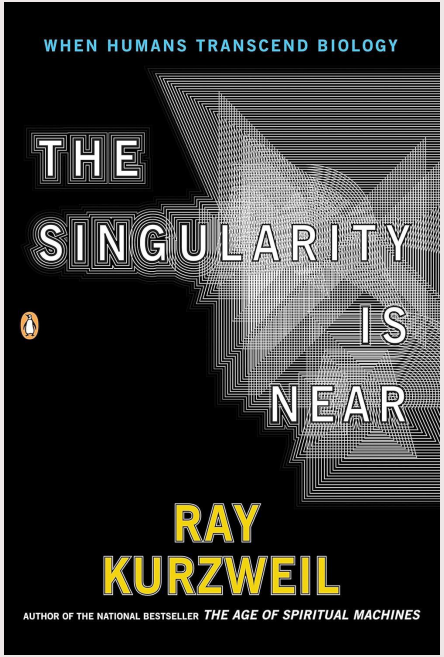
There is a tragedy of the commons at work here: **for any individual human, it may seem pointless to engage in years of arduous learning to acquire knowledge and skills that machines already have; but if everyone thinks that way, the human race will, collectively, lose its autonomy.**" (Russell 2019)

Wall-E

Singularity and Superintelligence

Singularity

Theoretical point where artificial intelligence surpasses human intelligence, becomes uncontrollable, and leads to societal change



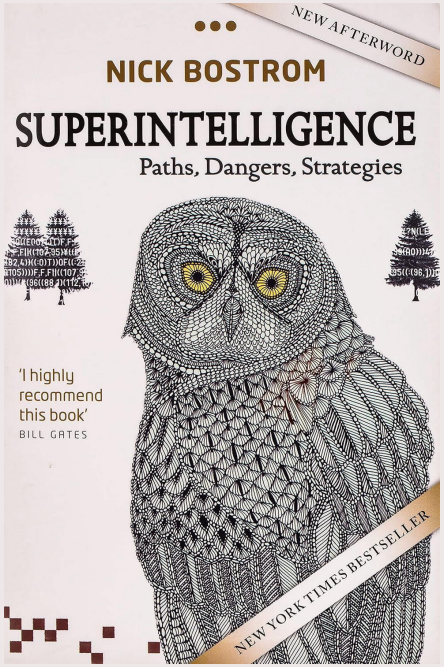
Singularity and Superintelligence

Singularity

Theoretical point where artificial intelligence surpasses human intelligence, becomes uncontrollable, and leads to societal change

Superintelligence

"any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest" (Bostrom 2014, 22)

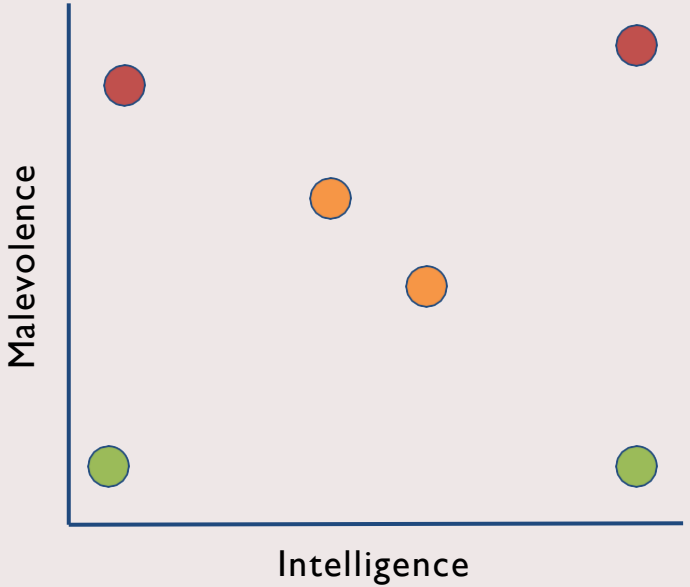


Orthogonality Thesis + Instrumental Convergence

Orthogonality Thesis

"Intelligence and final goals are orthogonal axes along which possible agents can freely vary. In other words, more or less any level of intelligence could in principle be combined with more or less any final goal."

"...goals as bizarre by our lights as sand-grain-counting or paperclip-maximizing." (Bostrom 2012)



TU/e

Orthogonality Thesis + Instrumental Convergence

Orthogonality Thesis

"Intelligence and final goals are orthogonal axes along which possible agents can freely vary. In other words, more or less any level of intelligence could in principle be combined with more or less any final goal."

"...goals as bizarre by our lights as sand-grain-counting or paperclip-maximizing." (Bostrom 2012)

Instrumental Convergence

Thesis that intelligent agents with varied final goals will still pursue similar intermediate goals or means such as self-preservation, acquiring resources or power.

TU/e

Superintelligence + Orthogonality Thesis + Instrumental Convergence

- Will AI systems display instrumental goals?

‘The Godfather of A.I.’ warns of ‘nightmare scenario’ where artificial intelligence begins to set itself objectives like gaining power

BY CHLOE TAYLOR

May 2, 2023 at 1:55 PM GMT+2

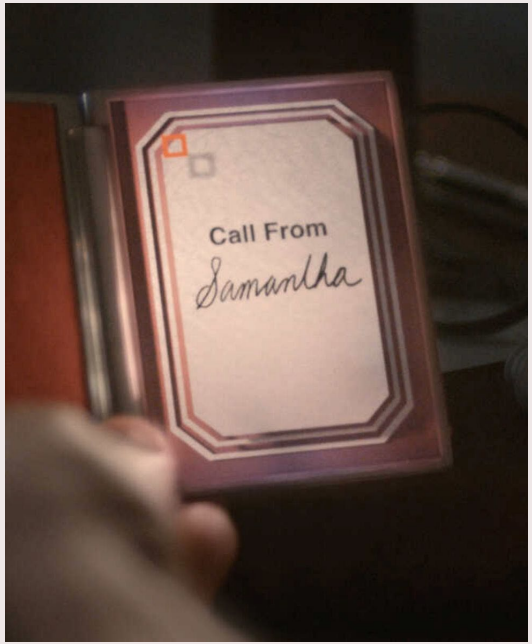
OpenAI checked to see whether GPT-4 could take over the world

"ARC's evaluation has much lower probability of leading to an AI takeover than the deployment itself."

BENJ EDWARDS - 3/15/2023, 11:09 PM

Superintelligence + Orthogonality Thesis + Instrumental Convergence

- Will AI systems display instrumental goals?
- Will potentially superintelligent agent be benevolent or malevolent?



Her

Superintelligence + Orthogonality Thesis + Instrumental Convergence

- Will AI systems display instrumental goals?
- Will potentially superintelligent agent be benevolent or malevolent?

What's the **risk** of things going **really badly for humans** given the **alleged possibility** of superintelligence, the orthogonality thesis, and instrumental convergence?

TU/e

Chalmers' argument for superintelligence

1. There will soon be AI
2. If there is AI, then there will be AI+ soon after
3. If there is AI+, then there will be AI++ soon after
4. So, there will be AI++

Chalmers' mitigation strategy

Involves pursuing strategies that steer towards non-orthogonal AI+.

Some recent terms for describing this:
"Value alignment": the idea that AI's values will align with human values

"Long-termism": the idea that the existence of sentient beings in the distant future should strongly influence our decisions now

TU/e

Existential and Catastrophic Risks

Existential Risk (Torres 2023)

- Extinction of humanity?
- Civilizational collapse?
- Permanent and drastic loss of our potential for desirable future development?
- Pangenerational and crushing negative consequences?
- Loss of a large fraction of expected value?

Existential and Catastrophic Risks

"It's hard to kill 8 billion people" —Melanie Mitchell

- Bad outcomes don't need to be 'existential' to be catastrophic
- Semi-technical-definition of catastrophic risks: **non-zero probability of a catastrophic outcome** (Stefánsson 2020)
  - Usually involves *harm on a large scale*, e.g. nuclear war, pandemic, volcanic eruption, bioweapons, totalitarianism, asteroid impact, climate change, etc.
- *Can* AI systems pose catastrophic risks? If so, *how*?

TU/e

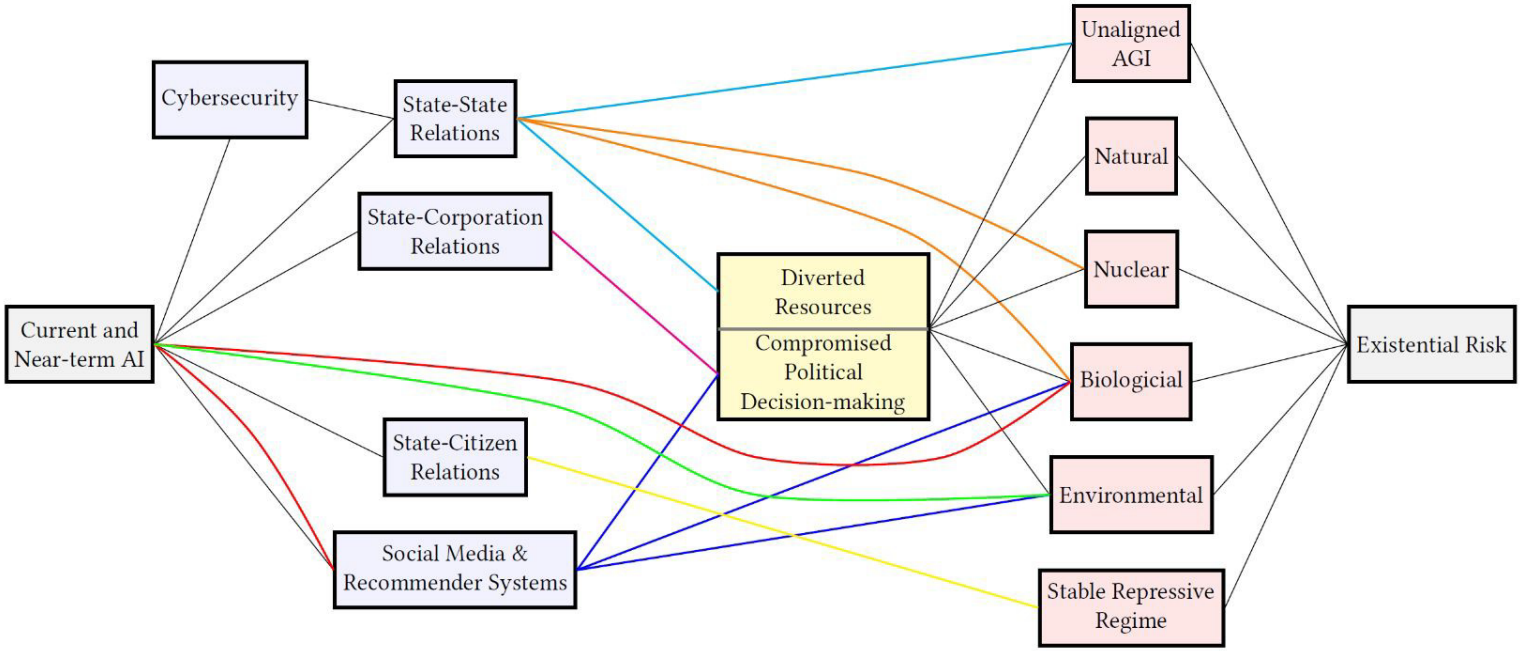


Figure 1: A graphical representation of the causal pathways from current and near-term AI to existential risk identified in this paper. Blue nodes represent effects of current and near-term AI, whereas red nodes represent identified existential risks [64]. The yellow box represents the general risk factors discussed in Section 5.1. Coloured edges represent causal connections as given in Table 1.

Bucknall & Dori-Hacohen (2022)

Risk Management

Assessing Risks

- What could happen?
- How likely is it?
- How bad/good is it?
- Are we facing uncertainty?
- Are we ignorant?

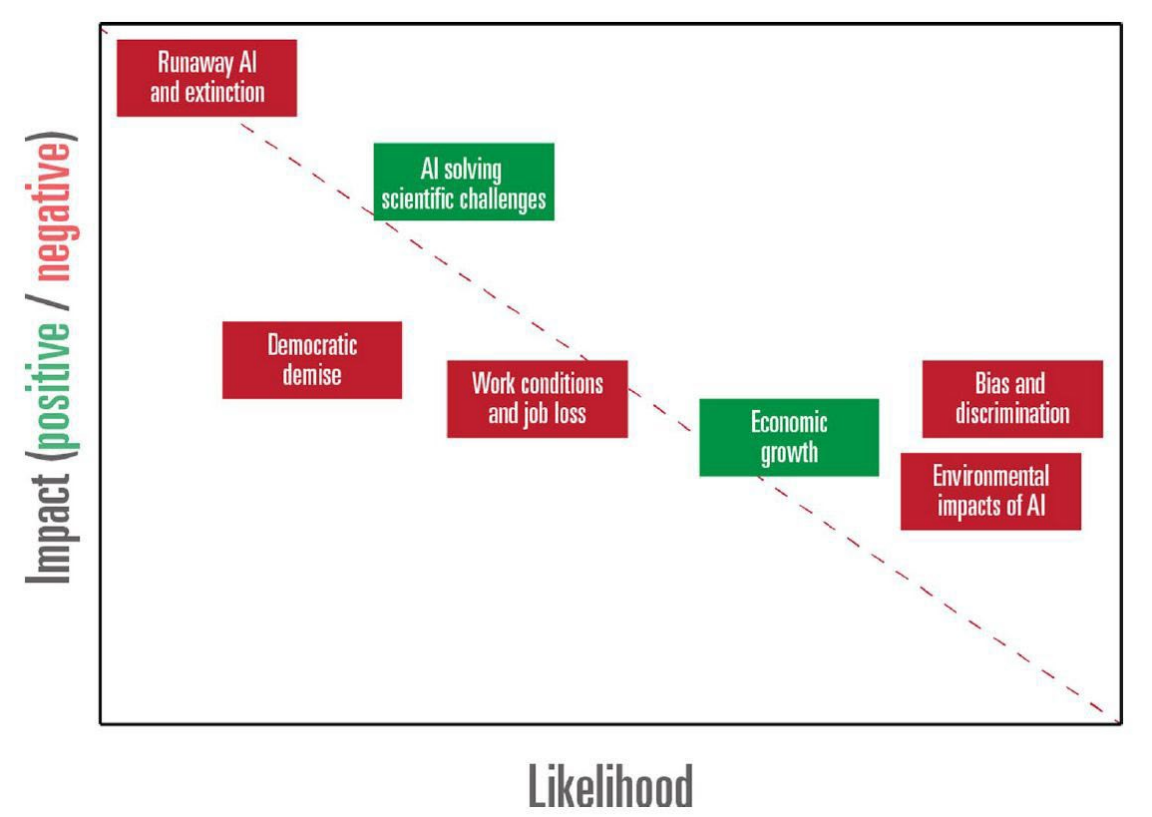
Two types of Errors

Type-I: Mistake of assuming that a statement is true while it is false

Type-II: Mistake of assuming that a statement is false while it is true

TU/e

Risk Management



Skaug Sætra & Danaher (2023)

Ethics of Risk

When is it *morally permissible* to impose risks?

Some relevant considerations (see Hayenhjelm & Wolff 2012):

- Do the benefits **outweigh** the costs?
- Are the risks **fairly** distributed?
- Would people **agree** to be subject to risks?
- What are the alternatives?

TU/e

Fair Distribution of Risks

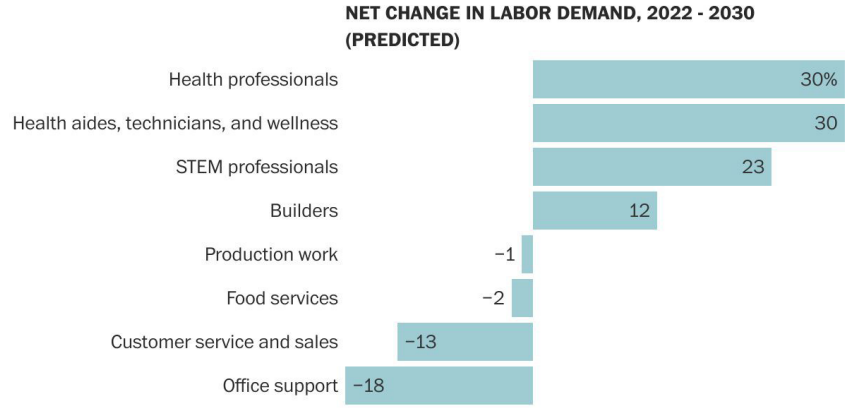
BUSINESS

AI will take more jobs from women than men by 2030, report says

By Annabelle Timsit

July 26, 2023 at 11:09 a.m. EDT

Health care, STEM fields and construction are expected to boom. Workers in customer and food services, office support and production roles could suffer.



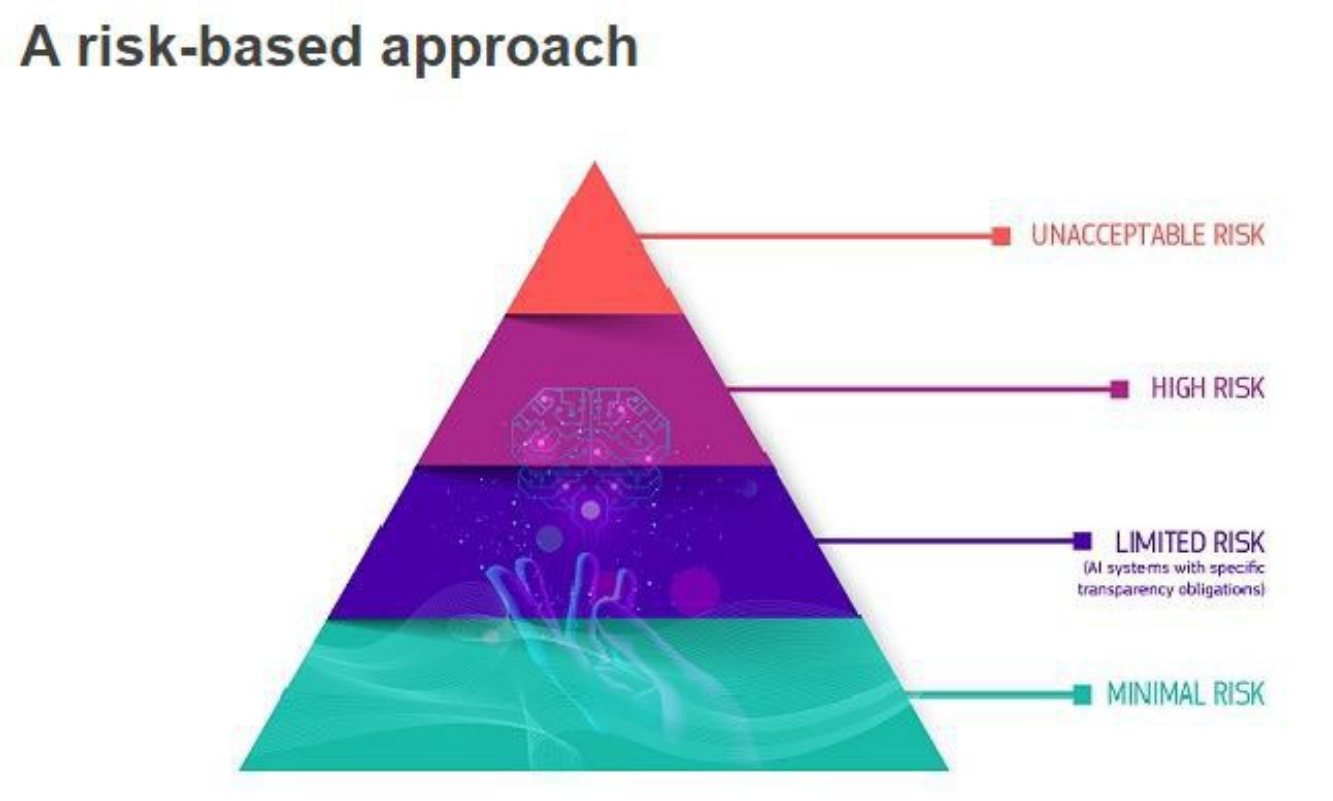
Source: "Generative AI and the future of work in America," McKinsey Global Institute (2023).

Fair Distribution of Risks

- **Distribution** of risks often matters: People should be treated equally
  - Are the most vulnerable more subject to risks? Are the better off more likely to obtain the benefits?
- Possible solutions
  - **Prioritarianism:** Principle of justice that gives priority to the worse off over the better off
  - **Contractualism:** Would people in principle agree to a given distribution of risks?
- Policy/regulation effectively redistribute risk exposure

TU/e

Categorization of risks in AI Act



EU AI Act

The challenge of deep uncertainty

→ We do not know what scenarios can materialize, making it impossible to predict and prepare

→ We must switch from a predict-then-act paradigm to some other paradigm.
- → Precautionary principle?
- → Decision making under deep uncertainty?

Uncertainty and the Precautionary Principle

Principle 15

In order to protect the environment, the precautionary approach shall be widely applied by States according to their capabilities. Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation.

Rio Declaration on Environment and Development (1992)

Definition of deep uncertainty

deep uncertainty: it is impossible to predict a full range of realistic possible outcomes in a situation. (Marchau et al. 2019)

In AI risks, we face **collective** situations of deep uncertainty in which some of the anticipated **scenarios require significant mitigation or prevention efforts**.

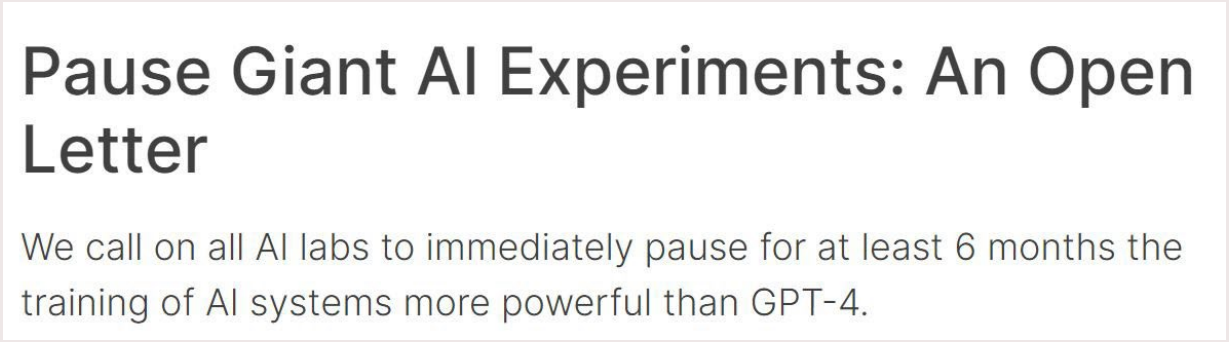
Uncertainty and the Precautionary Principle

> Principle 15
>
> In order to protect the environment, the precautionary approach shall be widely applied by States according to their capabilities. Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation.

Rio Declaration on Environment and Development (1992)

- Lack of knowledge concerning possible consequences and probabilities doesn't justify inaction
  - Mitigating or stopping risks
- *But* there may also be downsides to being too cautious
  (cf. type I vs type-II errors)

Uncertainty and the Precautionary Principle



- Lack of knowledge concerning possible consequences and probabilities doesn't justify inaction
  - Mitigating or stopping risks
- *But* there may also be downsides to be too cautious (cf. type I vs type-II errors)

The Role of Data Scientists and AI Professionals

- Responsible AI development and deployment
- Risk identification and management
- Regulatory and standards compliance