

What is

AI SAFETY

by Gustavo Costa

The way we think about AI is wrong

1

WHY THIS MATTERS

The Nobel Prize in Physics 2024

The Royal Swedish Academy of Sciences has decided to award the Nobel Prize in Physics 2024 to

John J. Hopfield

Princeton University, NJ, USA

Geoffrey Hinton

University of Toronto, Canada

“for foundational discoveries and inventions that enable machine learning with artificial neural networks”

The Nobel Prize in Chemistry 2024

The Royal Swedish Academy of Sciences has decided to award the Nobel Prize in Chemistry 2024 with one half to and the other half jointly to

David Baker

University of Washington, Seattle, WA, USA
Howard Hughes Medical Institute, USA.

Demis Hassabis

Google DeepMind, London, UK

John Jumper

Google DeepMind, London, UK

“for computational protein design”

“for protein structure prediction”

Three Mile Island nuclear reactor to restart to power Microsoft AI operations

Pennsylvania plant was site of most serious nuclear meltdown and radiation leak in US history in 1979



Argentina will use AI to 'predict future crimes' but experts worry for citizens' rights

President Javier Milei creates security unit as some say certain groups may be overly scrutinized by the technology



OCTOBER 24, 2024

Memorandum on Advancing the United States' Leadership in Artificial Intelligence; Harnessing Artificial Intelligence to Fulfill National Security Objectives; and Fostering the Safety, Security, and Trustworthiness of Artificial Intelligence

DISRUPTION

Education

Healthcare

Employment

Art

War

You are already involved

2

WHAT IS GOING ON

Superintelligence will be the most significant development in human history.

...Fixing the climate, establishing a space colony, and the discovery of all of physics...

— Sam Altman, *CEO of OpenAI*

*Deep learning works, and we will solve
the remaining problems.*

— Sam Altman, *CEO of OpenAI*

BLINDSPOT

Natural Science	Social Science
-----------------	----------------

AI Safety	
-----------	--

	?
--	---

- Reasoning
- Understanding
- Human values
- Consciousness
- Algorithmic bias
- Privacy
- Accountability
- Alignment
- Governance
- Social impact
- Monopoly
- Intellectual property

- Surveillance
- National security
- Automation
- Labour disruption
- Deepfakes
- Resource consumption
- Environmental impact

Wicked problems or just inappropriate tools?

Natural Science

AI Safety

Facts

Technical

Resolution

Social Science

AI Ethics

Values

Social

Negotiation

The solution is

INTERDISCIPLINARITY

Philosophy: Debunking Robot Rights Metaphysically,
Ethically, and Legally

Linguistics: Language (Technology) is Power: A Critical
Survey of “Bias” in NLP

Sociology: AI Safety Needs Social Scientists

Ecology: Making AI Less "Thirsty"

Cultural studies: The unseen Black faces of AI
algorithms

Management: 80% of AI projects fail

Political science: Deepfakes did not impact election

3

WHAT TO DO

Natural Science

AI Safety

Social Science

AI Ethics

SYNTHESIS

AI Studies

AI safety & ethics & everything else

Science studies is an interdisciplinary research area that

Gender studies is an interdisciplinary academic field devoted to

Cultural studies is a politically engaged postdisciplinary academic field that

- Science studies - Wikipedia
- Social informatics - Wikipedia
- Technology and society - Wikipedia
- Social construction of technology - Wikipedia
- Algorithmic bias - Wikipedia
- Algorithmic transparency - Wikipedia
- Algorithmic accountability - Wikipedia
- Regulation of algorithms - Wikipedia
- Regulation of artificial intelligence - Wikipedia
- Explainable artificial intelligence - Wikipedia
- Right to explanation - Wikipedia
- Digital rights - Wikipedia
- Automated decision-making - Wikipedia

- Machine ethics - Wikipedia
- Computer ethics - Wikipedia
- Programming ethics - Wikipedia
- Fairness (machine learning) - Wikipedia

DO NOT REINVENT THE WHEEL

ACM FAccT

ACM Conference on Fairness, Accountability, and Transparency

The logo for AINOW features the word "AINOW" in a bold, sans-serif font. The "AI" is rendered in a white, outlined style, while "NOW" is in solid black. The logo is set against a light yellow rectangular background.The logo for MAIEI consists of the letters "MAIEI" in a white, sans-serif font. The "EI" is highlighted in a bright cyan color. The logo is set against a solid black rectangular background.The logo for the Ada Lovelace Institute features a large, stylized black letter "N" on a red square background. To the right of the "N", the words "Ada Lovelace Institute" are written in a black, sans-serif font, stacked vertically.

4.

IN CONCLUSION

The culture around AI needs to change

What is

~~AI SAFETY~~

~~AI ETHICS~~

AI STUDIES

By Gustavo Costa

Slides:



<https://github.com/noah-art3mis/discovery/blob/>

main/Capstone1_assessment1_24000114067.md

RESOURCES

Institutions:

- FAccT - ACM Conference on Fairness, Accountability, and Transparency
- Montreal AI Ethics Institute
- Data and Society
- Home - AI Now Institute
- Ada Lovelace Institute
- AI & SOCIETY

Books

- Artificial Intelligence: A Guide for Thinking Humans (Pelican Books): Amazon.co.uk: Mitchell, Melanie: 9780241404829: Books
- AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference: Amazon.co.uk: Narayanan, Arvind, Kapoor, Sayash: 9780691249131: Books

AI Research

- dl.acm.org/doi/pdf/10.1145/3531146.3533088
- [2406.13843] Generative AI Misuse: A Taxonomy of Tactics and Insights from Real-World Data
- Constitutional AI: Harmlessness from AI Feedback \ Anthropic
- [eleosai.org/papers/20241030_Taking_AI_Welfare_Seriously_web.p](https://eleosai.org/papers/20241030_Taking_AI_Welfare_Seriously_web.pdf)
- [2303.12712] Sparks of Artificial General Intelligence: Early experiments with GPT-4

AI Research

- Constitutional AI: Harmlessness from AI Feedback \ Anthropic
- Embers of autoregression show how large language models are shaped by the problem they are trained to solve | PNAS
- 2404.10072
- Language (Technology) is Power: A Critical Survey of “Bias” in NLP - ACL Anthology
- AI Safety Needs Social Scientists

AI Research

- [2304.03271] Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models
- The unseen Black faces of AI algorithms
- Research shows more than 80% of AI projects fail, wasting billions of dollars in capital and resources: Report | Tom's Hardware
- No evidence that AI disinformation or deepfakes impacted UK, French or European elections results | The Alan Turing Institute

Techno-optimism

- The Intelligence Age
- Dario Amodei — Machines of Loving Grace
- The Techno-Optimist Manifesto | Andreessen
Horowitz

AI News

- Press release: The Nobel Prize in Physics 2024 - NobelPrize.org
- Press release: The Nobel Prize in Chemistry 2024 - NobelPrize.org
- Three Mile Island nuclear reactor to restart to power Microsoft AI operations | Nuclear power | The Guardian
- Google to buy nuclear power for AI datacentres in 'world first' deal | Google | The Guardian

AI News

- Argentina will use AI to ‘predict future crimes’ but experts worry for citizens’ rights | Argentina | The Guardian
- The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023 - GOV.UK
- Memorandum on Advancing the United States’ Leadership in Artificial Intelligence; Harnessing Artificial Intelligence to Fulfill National Security Objectives; and Fostering the Safety, Security, and Trustworthiness of Artificial Intelligence | The

White House