Effective Altruism and AI

Ethics

- "Ethical Realism?
    - Good, it exists and it's...
- Ethical Anti-Realism?
    - Good doesn't exist, or it's what everyone makes of it.
- Ethical Uncertainty...
- Questioning one's intuitions : The Drowning Child

The drowning Child

- For generalising
    -

- Against generalising
    -

Concrete Ethics
When philosophy isn’t just philosophy

- Triage
- Local action
- Everyone can transfer money
    - Just because it's possible doesn't mean it's a good idea.
    - It usually involves personal costs

- If you do want to do some good, how to do it better ?

How to do more good ?

- Priorising
  - Impact
  - Tractability
- For a given project
  - Scientific methods

Main cause areas in Effective Altruism

- Global health and development
  - Animal welfare
- Catastrophic risks, existential risks

Catastrophic and existential risks

- Natural Pandemics
- Asteroids and other space objects
- Nuclear Winter
- Bio-risks
- Climate Change
- AI
- Others..

AI’s risks and promises

By Jonathan Claybrough

Summary

Introduction : Promises and context

1) What is going on in AI ?
    a) Why is it important?
    b) Is safety neglected?
    c) Is this a serious issue?
2) The main categories of risks
    a) Malicious use
    b) Accidents
    c) Systemic

Introduction - promises

"Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make."

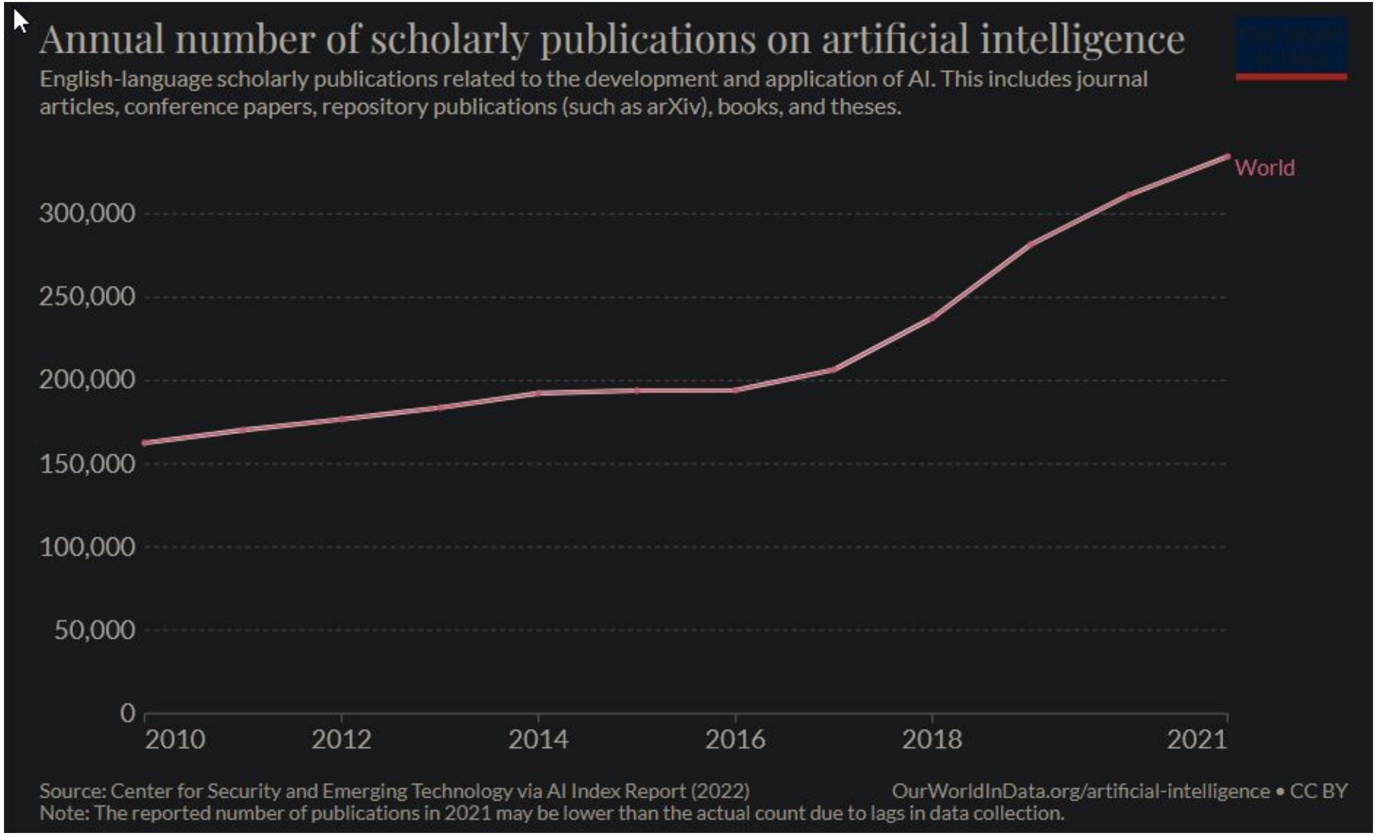
- Irving John Good, Speculations Concerning the First Ultraintelligent Machine Advances in Computers, vol. 6 (1965) 31ff.

Introduction - promises

- "OpenAI’s mission is to ensure that artificial general intelligence (AGI) [...] benefits all of humanity”. [1]
- "Anthropic exists for our mission: to ensure transformative AI helps people and society flourish [...]" [2]
- Deepmind : "Our long term aim is to solve intelligence, developing more general and capable problem-solving systems, known as artificial general intelligence (AGI)”. [3]

Hope = Build Utopia

What’s going on in AI ? [1]

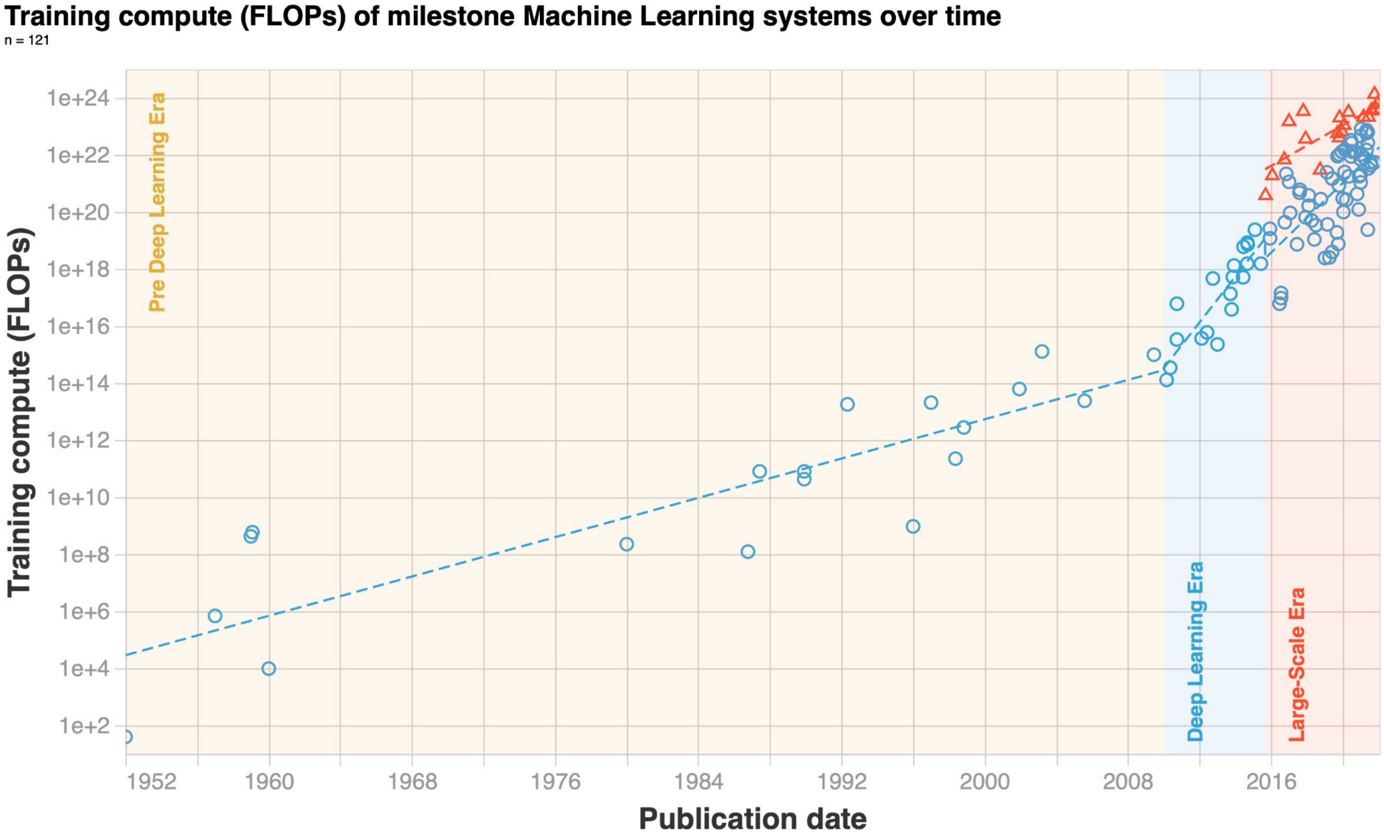


Annual number of scholarly publications on artificial intelligence

English-language scholarly publications related to the development and application of AI. This includes journal articles, conference papers, repository publications (such as arXiv), books, and theses.

Source: Center for Security and Emerging Technology via AI Index Report (2022)
OurWorldInData.org/artificial-intelligence • CC BY
Note: The reported number of publications in 2021 may be lower than the actual count due to lags in data collection.

What’s going on in AI ? [1]

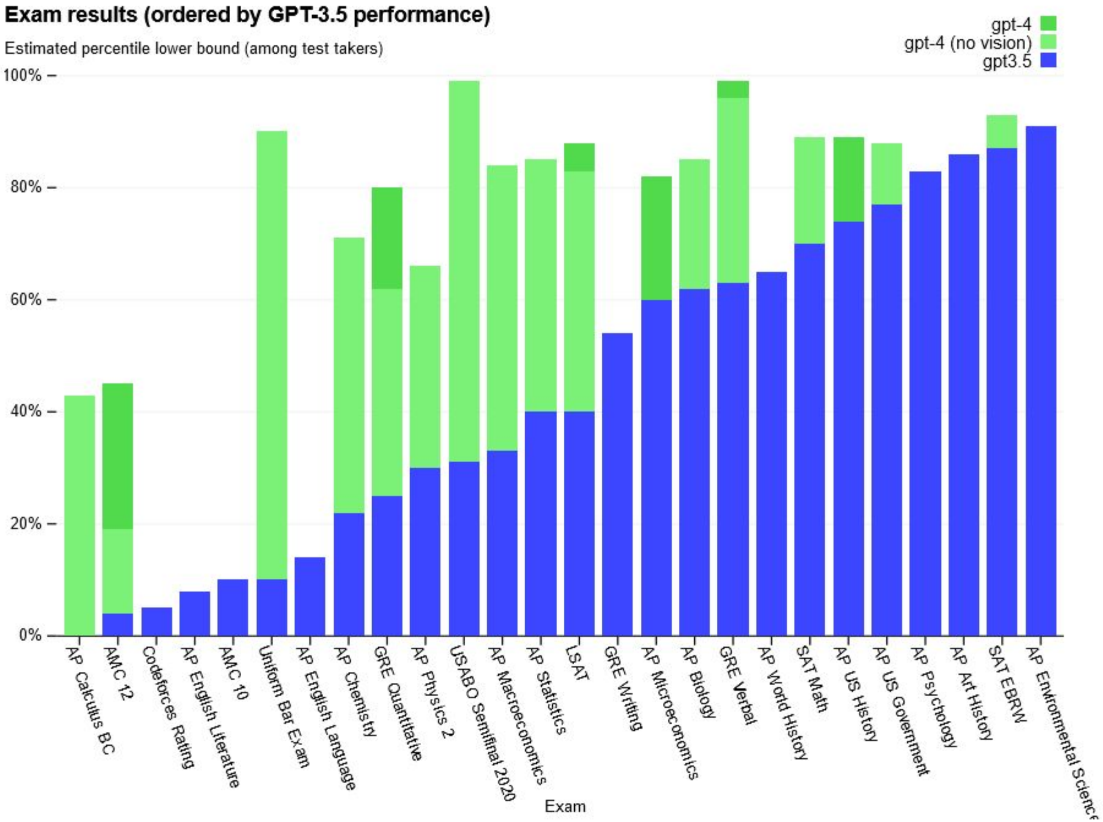


What’s going on in AI ?

- Generative AI
  - DallE
  - Midjourney

- Large Language Models
  - ChatGPT
  - Google’s Bard
  - Microsoft’s Bing
  - Facebook’s Lamma

- Other
  - Facebook’s Segment Anything
  - Whisper

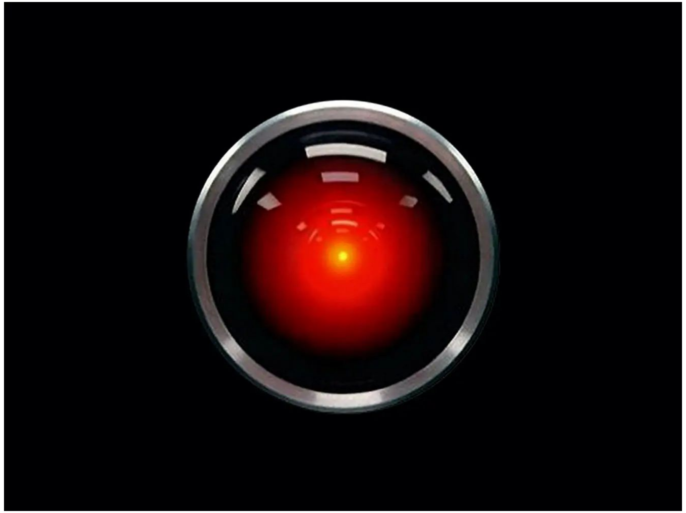


How are most powerful AI systems developed ?

- Configurable “neural” networks
    - Trained on large datasets
    - With emerging capabilities
    - No guarantees
    - Functionally black-box

What about Safety ?

- Identify AI-related risks
- Anticipate and mitigate risks
    - Increase the probability of taking a safe trajectory
- Enable humanity to use AI to solve scientific and societal problems

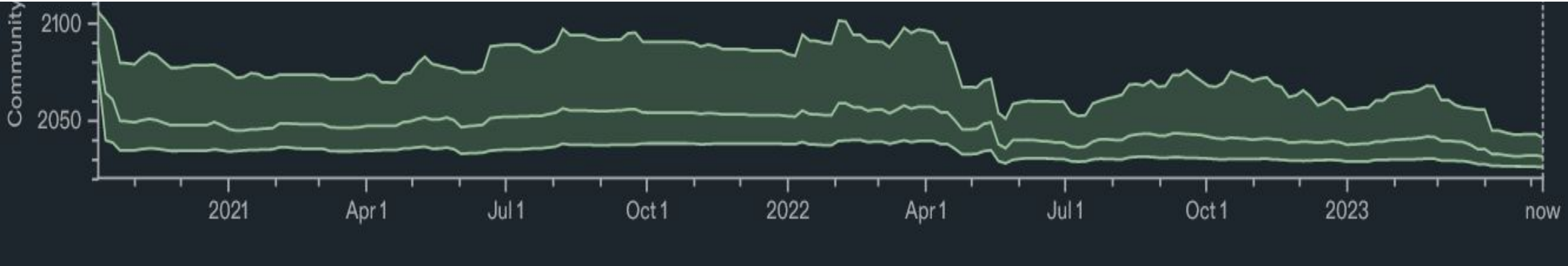


VS

(Your mileage may vary)

Is AI safety important?

- Incredibly fast adoption (100 million users in a few months)
- Risks at all scales
    - Local, catastrophic, existential
- In all areas
    - Computer science and physics
- AGI is expected sooner and sooner (Metaculus AGI in 2059, now 2032)

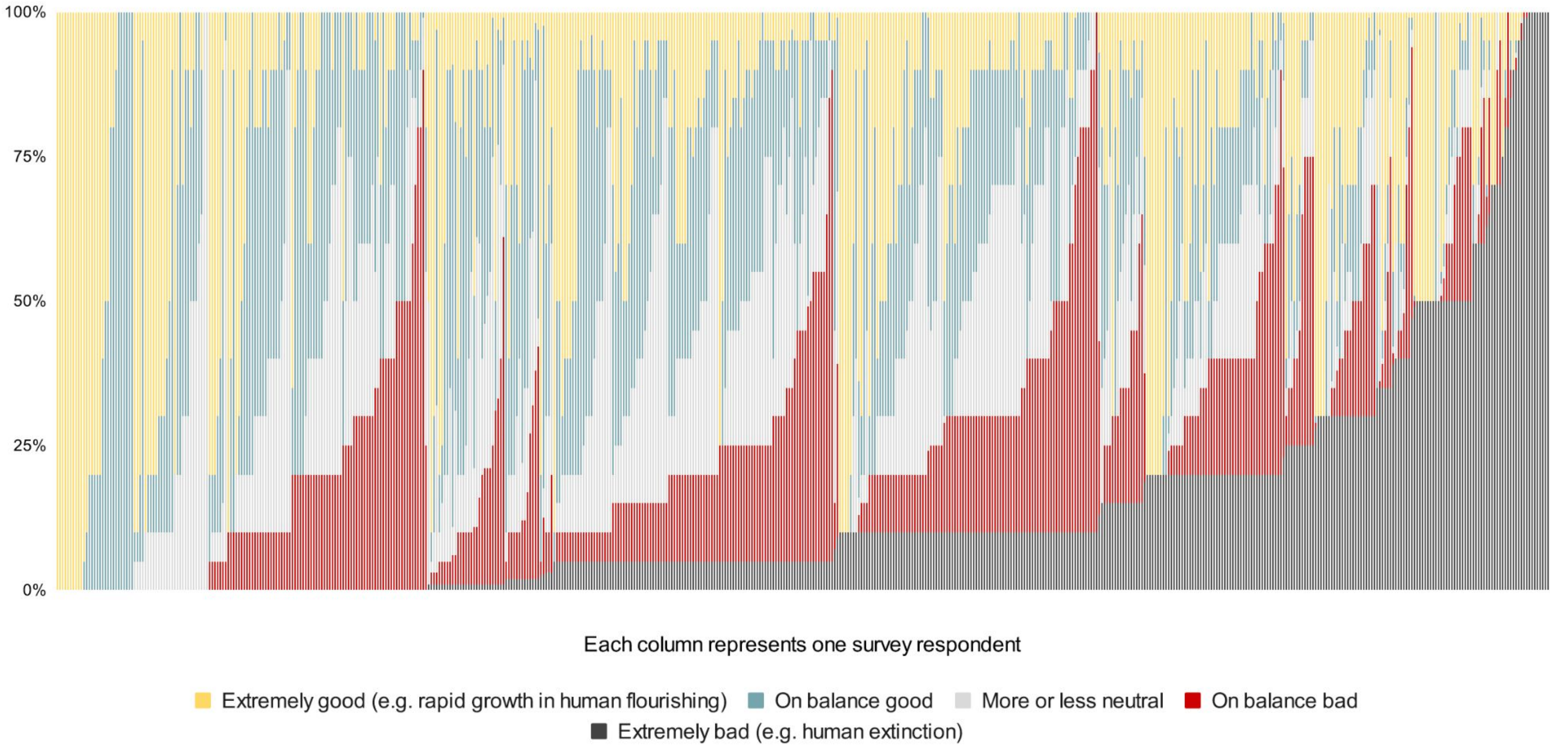


Total Forecasters: **596** Community Prediction: **Sep 30, 2031**

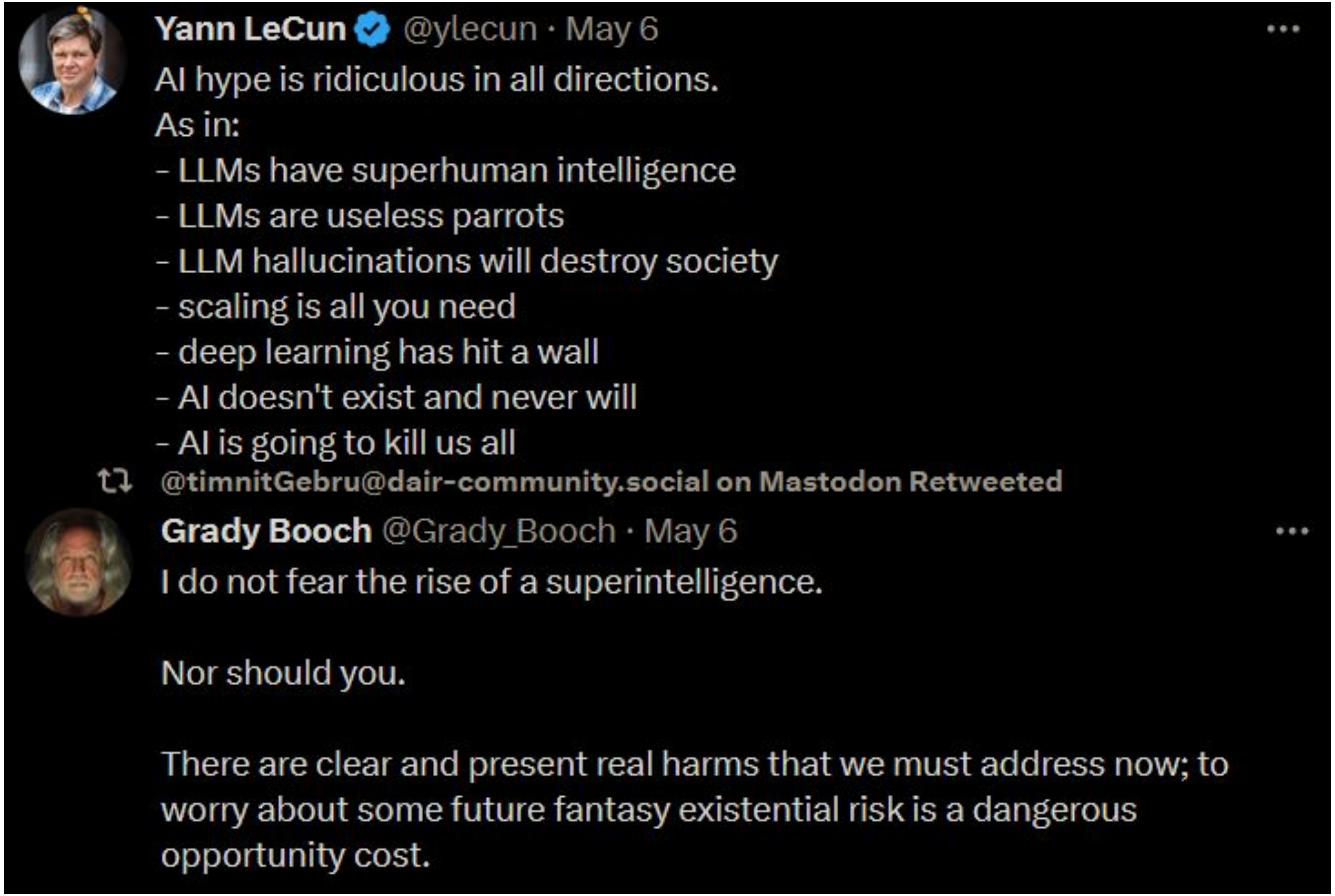
How serious are these concerns ?

How positive or negative will the impacts of high-level machine intelligence on humanity be in the long run? (2022)

559 machine learning experts' guesses, ordered by probability of 'extremely bad' outcomes



How serious are these concerns ? Not unanimous



Yann LeCun @ylecun · May 6

AI hype is ridiculous in all directions.
As in:
- LLMs have superhuman intelligence
- LLMs are useless parrots
- LLM hallucinations will destroy society
- scaling is all you need
- deep learning has hit a wall
- AI doesn't exist and never will
- AI is going to kill us all

@timnitGebru@dair-community.social on Mastodon Retweeted

Grady Booch @Grady_Booch · May 6

I do not fear the rise of a superintelligence.

Nor should you.

There are clear and present real harms that we must address now; to worry about some future fantasy existential risk is a dangerous opportunity cost.

Yann LeCun's
draft for AGI [1]

Main categories of risk[1]

- Misuse
  - Scams (phishing, …), cybercrime, weaponization, manipulation, surveillance
  - ChaosGPT
- Accidents
  - Autonomous car, wrong classification,
  - Imperfect specifications,
  - Converging instrumental objectives
- Systemic risks
  - Epistemic erosion, fragility to feedback loops
  - Multi-agent risks+automation
  - Goodheart's law

[1] See Remco Zwetslott and Allan Dafoe, 'Thinking About Risks From AI: Accidents, Misuse and Structure' (Lawfare, 11 February 2019)

Main categories of risk : Misuse

- Cybercrime
  - Phishing : automatic, large-scale, more effective [1]
  - Psychological warfare
  - Writing exploits
  - Spam
- Trojan attacks
- Interstate cyber-warfare
- Public opinion manipulation
- Mass surveillance
- Terrorism
  - AI Systems can help generate weapons

Main categories of risk : Misuse

- ChaosGPT - An Autonomous System with One Purpose: To Destroy Humanity

- Its danger scales with its capabilities
  - Beware of Self-replication, cybercrime, manipulation

- Some current limits :
  - Uses an API, so it could have its calls monitored and cut off
  - What happens if anyone can run this at home?

Main categories of risk : Accidents

- Lacking robustness
  - Autonomous vehicles, medical misdiagnosis
- Goodheart and specification problems
  - Optimizing for time spent on social media
  - Reproduce bias
- Out of domain generalisation
- Instrumental objectives
  - Power seeking
  - Self preservation
  - Goal preservation
  - Deception

Main categories of risk : Accidents

Levels of risks and appropriate caution

- Human level AGI [1]
  - May duplicate itself as fast as network speed
  - Great self cooperation
  - Can theoretically be contained/detected with good engineering practices (boxing, etc)
- Superintelligence [2]
  - May exploit the physical flaws of computers (rowhammer), humans
  - New attack vectors (innovative bio-chemicals)
  - Advanced and long-term deception capability, steganography
  - No solution in sight, might be theoretically impossible to robustly align long term

Main categories of risk : Systemic

- Epistemic erosion
- Weakness to fast feedback loops (eg. nuclear war, Petrov)

- Fully automated industry: Andrew Critch “What multipolar failure looks like” [2]

- The world becomes optimized for simple measures, which do not correspond to our real needs : Christanio “What failure looks like” [3]

Main categories of risk : Suffering

- Studied by the Center on Long-Term Risks
- Some considered scenarios :
  - Misuse
  - Value lock-in
  - Bad nash equilibrium from multi-agent worlds[1]
  - Digital sentience

Main categories of risk - Takeaways

- Some overlap between immediate, near term and longerterm risk
  - Especially around the concept of “robustness”
- Everything is important in the absolute, but some things are more neglected

[1] See Remco Zwetslott and Allan Dafoe,
‘Thinking About Risks From AI: Accidents, Misuse and Structure’
(Lawfare, 11 February 2019)

Bonus : Specific x-risk scenarios (Do not do at home)

- AI in control of nuclear weapons
- AI in a fully automated world that goodhearts
- Deceptive misaligned AGI which bides its time
- Superintelligence from Recursive Self Improvement

Bonus : Probability of x-risk

- Current trajectory, no additional safety efforts
- Current trajectory, only AGI labs slow down
- Global coordination on AI safety

Q&A

Many interesting topics

- AI and its consciousness
    - Does it matter for ethics?
- AI and brain uploads
- AI and understanding ourselves
- AI and species agnostic ethics
-

List of references 1/2 (mentioned during the talk and QA)

- Activation addition : controlling LLM behavior at inference time (eg. could be used so completions respond under certain ethics norms)
- Talk on AI governance, how we failed to regulate the first wave of AI in social media : couldn't find the reference :(
- Effective altruism : "doing good better"
- 80000 hours : resources on having a positive impactful career
- GiveWell : A charity evaluator that can give recommendations on which charity saves the most lives per dollar or has the greater benefit.
- LessWrong : An open forum with novice and expert discussion on AI, technical alignment, ai safety, ai governance, and many other themes.

List of references 2/2 (mentioned during the talk and QA)

- The Precipice, by Toby Ord : introduction to the study of existential and catastrophic risks to humanity, good source of risk estimates.
- EffiSciences : Association promoting high impact research, particularly to reduce catastrophic risks (biohazards and AI)
- To delve deeper into more concrete scenarios and details on the risks of AI, and their potential solutions, you can see my more technical talk https://www.youtube.com/watch?v=HM3vJEEOAIk&t=946s (the first 10 minutes share some introduction slides)

Keep in touch with the speaker

Jonathan Claybrough