AI Risks – Existential vs Immediate – Beyond the Headlines

Key Issues and Discussion

Based on the Munk Debate on AI

Podcast: https://munkdebates.com/podcasts/munk-dialogue-on-ai/
YouTube video: https://www.youtube.com/watch?v=144uOfr4SYA

D.D. Sharma
(https://www.linkedin.com/in/dsharma6/)
12/05/23

Contents

1 Introduction

Increasingly AI is being deployed or being considered for use in safety-conscious sectors (healthcare, aviation, autonomous driving, power generation, biotechnology, defense and military, robotics, etc.). In these applications, AI is expected to operate safely both as a piece of software (non-fail, conform to design specifications) and as a part of the overall system (not cause other system components to fail or harm the environment). It is no surprise that we find AI safety as an important topic from research labs to corporate board rooms.

When it comes to AI safety, we find two schools of thought. Some are deeply concerned about the existential risk of a super intelligent machine that can outsmart and outcompete us. The other group acknowledges some risks but no more than any new technology humanity has created. This group is optimistic that while creating great benefits, existential risk will be controlled with success though details of exactly how are still in the works.

It is instructive to get beyond the headlines and dive a bit deeper into the debate on existential risks vs immediate risks. Fortunately, on June 22, 2023, MunkDebates.com organized a debate in Toronto with the resolution: Be it resolved, AI research and development poses an existential threat. Four academic and industry leaders joined the debate. In favor of the resolution were Max Tegmark (MIT), and Yoshua Bengio (Mila - Quebec AI Institute, Turing Award). Against the resolution were Yann LeCun (Meta, Turing Award), and Melanie Mitchell (Santa Fe Institute).

For all those interested in AI Safety, this debate is a wonderful introduction to the subject from those who are creating and leading the technology. The debate is available at:
As podcast: https://munkdebates.com/podcasts/munk-dialogue-on-ai/
As YouTube video: https://www.youtube.com/watch?v=144uOfr4SYA

It is a sincere and vigorous debate by four leaders in the field who deeply care about AI and its potential to create profound benefits for mankind. Listening to their arguments one walks away with the feeling that their disagreements are more a matter of degree than substance. And they all seemed to agree on one thing: AI risks, existential or otherwise, cannot be ignored. And there is a need for conscious design, testing, and certification for AI safety. And though they would like different priority on how resources are allocated to address different types of risks, IMHO both their R&D agendas are critically important and can be simultaneously supported.

This article is based on a transcript of the Munk debate on AI Existential Risks. It captures key issues for AI safety and summarizes the discussion. Section 2 gives an overview of the problem of existential risk. Section 3 presents a summary of some key questions and responses based on the discussion during the Munk debate. Section 4 is a simple definition of existential risk used by the debate. Section 5 lists a few definitions. Sections 6 through 15 are based on the debate transcript organized around specific discussion themes.

Disclaimer: I have tried to refer each discussion point back to one of the four debaters. I many have missed some. I may have misattributed some. I may have restated some those original speakers may disagree with. Those are my errors and are unintentional.

2 Background

In the last 12 months, since the arrival of OpenAI's ChatGPT there has been much public debate on the risks and harm from the AI between well-meaning and genuinely concerned AI alarmists (no disrespect intended) and the pragmatists.

From the alarmists' camp, renowned AI luminaries such as Geoffrey Hinton (Turing Award, 2018) and Yoshua Bengio (Turing Award, 2018) have called for a pause in AI research and development until certain AI safety-related problems are addressed. The cautionary approach has also been advocated by leaders from the industry namely Elon Musk, Sam Altman of Open AI, and Bill Gates. Wikipedia provided a reasonably detailed overview of the Existential risks of AI: https://en.wikipedia.org/wiki/Existential_risk_from_artificial_general_intelligence.

From the pragmatists' camp equally scientists and technology leaders such as Yann LeCun (Turing Award, 2018) and Andrew Ng (Stanford Engineering, GoogleBrain, Baidu AI, Coursera, …), and a whole host of industry leaders argue that the much-feared existential risk is overblown, the humanity has much to gain from AI, and a pause will not only deprive us of the benefits but also diminish our abilities in countering existential risks from other events such as climate change, being hit by an asteroid, and evil use of other technologies such as biotechnology.

The debate on the concerns about AI safety and the promise of AI benefits has mobilized policymakers around the world to step up and define some rules of the game. A few being:

- US White House Executive order on the Safe, Secure, and Trustworthy Development and Use of AI: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/
- UK guidelines for Secure AI: https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development
- EU AI Act to be finalized by the end of 2023: https://cset.georgetown.edu/article/the-eu-ai-act-a-primer/

Not everyone is happy with governments entering the debate armed with regulations, but one thing they all agree on is that AI risks, existential or otherwise, cannot be ignored. That there is a need for conscious design, testing, and certification for AI safety. And this agreement is good news for the AI industry.

3 AI Safety - Ten Key Questions and Quick Answers

Question	Answer based on Munk Debate participants
Is Superhuman AI feasible?	Yes. Because we are building them. The human brain is a biological machine and neuroscience is figuring it out. Digital tech is way more powerful than biology and can replicate. Hence superhuman. ETA 5-20 yrs. (Yoshua). In principle yes. But AI learns from human-provided data. It cannot reason or think. Our brain is not just any brain. It is guided by human wants, needs, and cultural context. It is complex. (Melanie)
Why does superhuman AI pose an existential risk?	It can wipe out humanity if it has that goal which it can acquire through extinction scenarios (Max). It can disrupt society, disempower humanity, and self-replicate with self-preservation goals (Yoshua).
What are some of the extinction scenarios?	1. Human misuse – some genocidal group uses AI for extinction. 2. Rouge AI – misunderstands goal and accidentally kills humanity. 3. Outcompete – makes human knowledge and skills redundant and irrelevant to enterprises and systems. 4. Malevolent super-intelligent AI – develops the goal to dominate and destroy humanity. (Max)
Is existential risk inevitable?	No. Current AI capabilities and trajectories do not lead to a near-term existential threat. Extrapolating the current AI blueprint (which has few safeguards) is not correct. There are many other risks. Humanity has to be optimistic and look forward towards an era of enlightenment. (Yann)
Is there proof that AI does not pose an existential risk?	Cannot disprove the assertion that AI poses an existential risk. Many assertions cannot be proven or disproven. For example, cannot disprove that there is a teapot flying between the orbits of Jupiter and Saturn. (Yann)
What other types of risks should we worry about?	Many. Disinformation. Hate speech. Manipulative messages. Fake life-like videos. Unauthorized information access. AI confabulating, hallucinating. (Yann)
How to make superhuman AI safe?	Build safe AI as an engineering effort, in small steps, iteratively. (Yann) Understand what can go wrong. Prepare for when things go wrong. Regulations to reduce the mismatch between AI goals and society's needs. (Yoshua)
How to ensure superhuman AI goals are always aligned with humans?	No clear answer from participants. Melanie said that we tend to use poorly understood phrases and anthropomorphize AI. AI is not live. AI does not have goals (Yann, Melanie). But it is easy to give AI goals (Yoshua).
How to avoid malicious users from exploiting AI?	No clear answer from participants. But then this is not all that different from the misuse of other technologies. Need to leverage law and regulation.
How would one go about building safe AI?	Objective-driven, constraints conforming AI. Virtuous AI (emotions, compassion, transparency, proper behavior). Open-source platform and arduous engineering (iterative, testing) (Yann) Formal assessment of AI threats and risks grounded in science, evidence, and empirical data. (Melanie) Precautionary Approach for dealing with risk: Safety testing. Empower good guys with good AI that is superior to bad guys' AI. Counter-Intelligence.

4 What is Existential Risk?

"An existential risk is one that threatens the premature extinction of Earth originating intelligence or life or the permanent and drastic destruction of its potential for desirable future development."
Source: Nick Bostrom

5 Definitions

The following are intuitive rather than formal research definitions.

- Human-level AI (also called AGI – Artificial General Intelligence): An AI that can tackle any intellectual task a human can. It is generally believed we are not there yet. Current AI models lack certain human traits such as reasoning, thinking, true creativity, emotional awareness, etc.
- Superhuman AI/Super-intelligent AI: An AI that far surpasses the cognitive performance of the brightest and most gifted human minds.
- Rouge AI: In this debate, Rogue AI was used to refer to an AI that was given proper goals, but for some reason, it starts pursuing other goals or sub-goals harmful to humanity.

6 What is Superhuman AI?

An AI that is so much more powerful that:

- It can do things humans can do (Max)
  - Goal-driven behavior: plan, adapt
  - Persuade, manipulate, hire people
  - Design and control robots
  - Automate scientific process including self-improvement
  - run companies, organizations
  - invent and build bioweapons
  - copy itself massively
- It can do things humans cannot do (Max):
  - Think/process 1000X faster
  - Instantly share new knowledge or skills with millions of copies (instant scale)
  - Lacks emotions and compassion
- Superhuman AI with self-preservation goals will replicate more of its kind (Yoshua)
  - We the designers, creators give it that goal
  - Self-preservation goals will lead it to control the environment and humans
  - Disempower humanity (take agency, control away)

7 Is Superhuman AI feasible?

- Yes. We will build superhuman AI machines (Yoshua)
  - Human-level AI machines are possible to build
    - The human brain is just a biological machine
    - Neurobiologists are figuring it out
    - No scientific roadblocks to building human-AI-level machines
    - By when - ??
  - Digital computers have an advantage over analog brains
    - Speed
    - Capacity/memory/ large data volumes
    - massive scale
    - immediate information sharing

        - learning in parallel of large number of computers
    - AI build using same principles of intelligence as humans
- ETA of superhuman AI is 5-20 years (Yoshua)
    - Concurrence between the three Turing Award winners (Hinton, Yoshua, LeCun)
    - Evidence GPT-4 - passes Turing test
    - AI/GPT-4 not superhuman yet. Something missing
        - Good at intuitive intelligence
        - But not good at reasoning, thinking through

8 Why does Superhuman AI pose an existential risk?

- Superhuman AI can wipe humanity if it has that goal (Max)
    - Tech can be used for good and for bad
    - Tech is increasingly getting more destructive power
      rock, bomb, nuke, bioweapons, nuke winter, superhuman AI
- Superhuman AI can acquire the goal to wipe out humanity via 3 paths to extinction (Max)
    - Rogue AI
        - AI is given goals that are Misaligned with human goals
        - Deception / Byproduct
        - Breakout
    - Malicious or Negligent Use by AI designer
    - We are outcompeted
        - AI replaces jobs
        - AI replaces decision making
        - Race to bottom Organizations that don't use AI lose to organizations that use AI
- Risks from superhuman AI Disrupt society, disempower humanity, and possibly extinction (Max)
    - Disrupt/defeat our social systems/order
        - Defeat our cyber security
        - Perform organized crime
        - Hire people to legally do harmful things
        - Open bank accounts
    - Build robots and self-replicate with self-preservation goals
    - Exert direct control through self-like copies
    - Disempower humanity
    - Humanity extinction
- Blog post on extinction and AI safety

https://yoshuabengio.org/2023/06/24/faq-on-catastrophic-ai-risks/

9 Is AI caused existential risk inevitable

- AI does not pose an existential risk in the near term
    - Current AI trajectory is not an immediate Existential threat. It lacks many basic capabilities. (Yann)
        - Current AI lacks many capabilities
            - confabulates, makes up things
            - cannot reason
            - doesn't plan
            - does not understand reality
            - trained on very limited subset of human knowledge (mostly text)

            - Current AI is not even Cat-level AI
            - We are far away from human-level AI
            - AI cannot do what people can do
                - Fill dishwasher
                - Drive a car after 20hr training
    - AI presents many risks and harms, none are existential risks in the near future. (Melanie)
        - Fear of extinction has deep roots in our collective psyche
        - Posited Extinction scenarios are unfounded speculations
        - Near-term Risks and Harms
            - Job losses
            - Spread of misinformation
- Host summary of con side
    - Risk is so low, in a sense negligible
    - should move on to other things - development of AI with controls and regulation
    - no demonstrable existential risk of a size requiring a moratorium on AI

10 What other types of AI risks should be the priority?

- Immediate AI risks are more serious than spotlighted existential risks (Yann)
    - Extrapolating the AI system's trajectory, using the current blueprint leads to false and worrisome scenarios
    - Immediate Risks
        - Not unique to AI
            - Disinformation
            - Hate speech
            - Manipulative messages
            - Unauthorized information access
        - Countermeasures use AI heavily
        - AI is the solution, not the problem
    - AI does not pose an Existential Risk

11 Is there an opportunity cost due to the existential risk narrative?

    - It is harmful to claim that AI is an existential threat. (Melanie)
        - Misleads people about AI's current state
        - Deflects attention from real immediate risks
        - Potential to block potential benefits from AI
- Examples of harm from unfounded speculations of risk (Melanie)
    - Example: Unfounded fear against vaccines harms society
    - Science itself led to nuclear weapons
- Benefits we will miss out on due to focus on existential risk (Yann)
    - Amplify human intelligence
    - Progress in science, medicine, education, etc.
- Optimism (Yann)
    - A new era of Renaissance
    - An era of Enlightenment

12 How can AI lead to extinction? Extinction Scenarios

- Extinction Scenarios (Max)

- Human Malicious use - Some genocidal group uses AI to help destroy humanity.
- Rouge AI - fallacy dumb super-intelligent AI
  - Misinterprets our wishes and accidentally kills.
  - Follows our wishes but devises methods that accidentally kill humanity
  - We give AI some noble great goals
  - But AI misconstrues those goals
  - Focus on goals antithetical to our interests
- Out-competed - AI replaces not only jobs but also decision-making, race to the bottom
- Malevolent super-intelligent AI uses evil genius to destroy humanity.

- Countering Extinction Scenarios
  - Malevolent Super Intelligent AI
    - AI is not alive, it will not have its desires and goals in the near term (Melanie)
    - Intelligence does not imply a desire to dominate (Yann)
      - Intelligence is not linked to a desire to dominate or kill
      - Organizing, hierarchy, and dominating - found in social animals/species
      - We can make intelligent machines that have no desire to dominate
  - Rouge AI
    - Fallacy of dumb AI - How can AI be so smart and still not infer that killing humanity is not an option? (Melanie)
  - Human Malicious use (Melanie)
    - This can happen but is not unique to AI.
    - We have dealt with this problem, making our institutions and technology more resilient and diverse.
    - Our created complexity puts brakes on such attacks
    - These attacks require a cascade of highly improbable events.
  - On Control Problem (Yann)
    - How do we design goals for machines to ensure proper behavior
    - Difficult engineering problem
    - As societies we design laws to align our objectives with the common good
- On Out-competition threat -AI has several limitations. History does not support fear of out-competition
  - Out-compete threat: (Corporations, people, and countries) using AI will outperform those who do not use AI. This encourages us to hand over the control to AI and thus reduce human agency
  - Melanie: It is not clear that corporations using lots of AI will outperform. AI has several limitations. It may get better. Assumptions about people handing over agency or forgetting skills are not necessarily true.
  - Yann: Fear of technology causing loss of agency or capability dates back to Socrates. Has been proven wrong. Technologies that make people smarter or communicate better facilitates education and is intrinsically good. AI is one of them.
  - Yoshua: The issue is scale. At a small scale, AI good/benefits overcome harm. Superhuman AI builds its own copies, deploys at a large scale. Lots of research on this shows it is hard to control or keep under control.
  - Yann: Predictions of technology leading to harm have been done for a long time. There is a website: Pessimist Archive with stories of fear of trains faster than 50mps (passengers will not be able to breathe). Jazz will destroy society.
- On Out-competition threat - Better to acknowledge, prepare than dismiss (Max)
  - The problem is that past performance is not an indicator of future results. We need to extrapolate for exponential growth in technology. Industrial revolution cause a shift

from muscles based work to brain based work. That went ok since it led to better pay.
The superhuman AI revolution will create machines smarter than us in 5, 20, 300 years.
Cannot compete based on muscles or brain or scale. This could lead to human disempowerment

- The solution is to stop dismissing this Superhuman AI scenario, and invest in control and how to use it to empower us.
  AI can be made safe and controllable. Put in the right policies
  AI tech power is growing faster than the pace of safety and alignment research. Need to acknowledge risk and accelerate the pace of safety research

13 Challenges related to the existential risk of AI (for Melanie Yann)

- What evidence is there that AI is not an existential risk in the near term? (Max)
  - Yann: Cannot disprove the assertion that AI poses an existential risk. Many assertions cannot be proven or disproven.
    - Melanie: Existential risk in the near term - what is the risk? Don't know. Non-zero - anything is a non-zero risk, e.g., Malicious aliens from space.
  - Yoshua: No existential threat yet.
  - Max: How do we know it won't be in 2, 3, 5, 20 years?
  - On existential risk, it being a very high bar (killing 100% of humanity). Where should the bar be to be significant?
    - Yoshua: Is killing 1% of humanity (80M) significant?
    - Melanie: Very significant and catastrophic. But not existential. It is hard to kill 8B people.
    - Yoshua: 1% or 100% - we could develop in a few years to a few decades AI that would be smarter than us (out-compete),
      they will find ways against which we will have no defenses
      and that scenario is plausible enough to worry about.
- What is the probability of creating superhuman intelligence in the next 20 years to 100 years?
  - Feasibility of building machines with human level intelligence (Melanie)
    - Feasible - yes in principle
    - Human Intelligence is complex to build
      - But human intelligence is not simply any brain
      - Our brain is different from octopus, rat, viruses
      - Human intelligence – guided by specific human problems, needs, motivations
      - Human intelligence is embedded in a physical cultural social system
  - Feasibility of building Superintelligence (Melanie)
    - AI systems learning is shallow
      - Learning is from data created by human
      - Captures some aspects of human intelligence
      - Lack fundamental important aspects of what it is to understand the world
    - The feasibility of building Superintelligence is not proven, no scientific evidence.
  - On the feasibility of building superhuman AI
    - Max: Need to be humble about our ability to predict progress. Just 3 years ago, AI researchers predicted 3 decades to the Turing test. And it is already done.
    - Yann: I was not surprised. Tech existed 2-3 years before.
    - Yoshua: For folks on the outside, it was a big surprise.
    - Melanie: ChatGPT, GPT 4 - a surprise and amazing. But does not point to

misalignment. They learn from huge amounts of human data. Synthesize, not acquire knowledge. They don't feel, think, or form their own goals.
- Yoshua: Agree. That is the problem - we don't know how to make a computer do what we intend it to do. And from data, it could infer goals we did not intend and may want to preserve its existence.
- Melanie: AI is not alive, it does not want to do anything.
- Yoshua: It is easy to give it wants. ChatGPT trained to please us (reinforcement learning). Put wrappers to turn them into agents that have goals- human-provided goals - but to achieve those goals they will have subgoals and that could include things like deception (NY Times journalist - divorce your wife)
- Melanie: Deception is an anthropomorphic idea. It was not deceiving.
- Yoshua: it was a consequence of trying to achieve a goal, finding a means to an end, other animals do it too.
- Yann: You know it has no goals.

- On the journey to Super-intelligent AI analogous to a cruise toward a waterfall
  - Max: Imagine we are going down the Niagara River. This is what I understand each of you will respond. Yoshua: I heard there is a huge waterfall ahead. Melanie: I am not convinced. Yann: Maybe, we will tackle it when we get there. Max: This approach of denial and delay would not work for many industries. Melanie says - I am not convinced there is, don't know how far it is, there is a big uncertainty -- so I am not going to worry about it. Yann says - I agree there is a waterfall, but we don't know yet how to make it safe we have done decades of research on this we will figure it out once we get closer. Max: This kind of argument would not work for the biotech industry Oil companies denied existential climate impacts - so we should not bother oil companies. Same with the tobacco, and asbestos industry.
  - Melanie: AI has some risk - no denial. Waterfall and Superintelligence AI are not analogous. Waterfall consequences are well known. But super-intelligent AI consequences are not known. There is wild extrapolation from current AI to something 100X powerful, but still dumb enough to misinterpret our goals
  - Max: On extrapolation - building super-intelligent AI is the declared goal of AI
  - Melanie: The arrival of the waterfall and its fears have been predicted since the 1960's by Shannon, Simon. We don't know what other kinds of existential risks will happen - genetic engineering.
  - Yoshua: But we are moving in that direction, my work is moving in that direction
  - Melanie: Then stop doing it.

- How will you make superhuman intelligence safe?
  - Should we do nothing to control AI? Need to build safe AI. (Yoshua)
    - AI has been improving for two decades
    - Superhuman Intelligence is not yet here.
    - AI Trajectory is clear. There is a problem.
  - Yann: Human-level AI system
    - is not going to take over the world the moment it is turned on
    - AI systems are built iteratively, in small steps.
    - AI will be built or increased in small steps.
    - Building Safe AI will be an engineering effort.
  - Melanie: What should we do with non-zero-risk events?
    - Stop radio broadcasts because that may attract aliens.

        - The probability is not high enough to attract that kind of attention.
    - Build safe AI systems (Yoshua)
        - Need to understand what can go wrong
        - Prepare for when things go wrong
        - Need government intervention to reduce mismatch between AI goals and society's needs
- How to ensure that superhuman AI goals are always aligned with humans?
    - Poorly understood phrases being used (Melanie)
        - superintelligence
        - smarter than human
    - People do build unsafe systems knowingly. (Yoshua)
        - Fossil fuel companies knew of damage to the planet.
        - Profit motive companies act in a way not aligned with society's needs.
    - Corporate and AI analogy (Yoshua)
        - Company goals (max profit) and some rules/laws (Pay tax, be legal)
        - Companies' activities evolve to mismatch with society's needs.
- How to avoid Malicious user use cases?
- How to prevent evil people from putting into AI goals to take over the world?
- What is your idea of how much risk is there for the extinction scenario? More than zero? One in a million? 1%? 10%?
    - Yann: Risk of extinction is negligible
        - We build AI systems.
        - We are not building superhuman intelligence.
        - We will not build it if it is not safe.
- Isn't there a risk that some company will continue the GPT 6 path and make unsafe AI rather than choose a safe way?
    - Yann: GPT-x is not an existential risk.
    - The risk is that it may not be as useful.

14 Proposed Solutions for AI Safety

Build Objective Driven, open source, virtuous AI with engineered safety (Yann)

- Objective-driven AI
    - Objective/goal driven
    - Constraint conforming
- Control via Virtuous AI is indispensable to building human-level AI
    - Possess emotion
    - Possess Compassion
    - Attributes of proper behavior
    - Transparent
- Open-source platform
- Build safety through arduous engineering

Iterative refinement to make superhuman AI safer is a terrible strategy (Max)

- we may lose control over superhuman AI
- superhuman AI needs only one chance to wipe out humanity

Proper and formal Assessment (Melanie)

- of AI threats and Risks
- grounded in science and empirical data
- not unsupported speculations.

Dealing with AI Risks

Framing the AI extinction threat risk

- Three scenarios/reasons (Max)
  - Malicious User - some user gives harm-causing goals
  - Loyal AI - goal evolved to harmful goal initial goal was fine but when pushed created much harm
  - We get outcompeted - on this trajectory already companies want to make profits, and we get more disempowered
- Three questions (Max)
  - Probability (build a human-like machine in 20 years)?
  - Probability (there will be an existential risk)?
  - What is the plan to avoid these three different threats?
- Extinction risk talked by Hinton, Altman, CEO DeepMind .. cannot ignore? (Max)

- Yoshua: Let us say we don't know the answers to these questions. Given the stakes shouldn't we pay attention to them?
- Host: Precautionary principle
  - you take actions now not because you want to but to avoid a certain worst-case scenario. For example, Climate. Why isn't AI like climate? Maybe AI is a tail risk, but it is an existential tail risk?

Precautionary Approaches

- Safety Testing (Yann): That is what we are doing. Testing for safety before deploying.
- Empower Good guys AI with Superior AI
  - Bad guys use AI for bad things (Yann)
  - Many more good guys use AI to do good and counteract bad. (Yann)
  - Yoshua: sometimes the attacker, bad guy has the advantage
  - Yann: no reason to believe that is the case for AI. People are already using AI for bad things - that happens with or without AI. AI might help with creating misinformation but not necessary - but QAnon two guys, don't use AI had big impact to corrupt electoral process. Solution is AI to take down such content. 5 years ago AI took down 25% of hate speech. Last year, AI took down 95% of hate speech
  - Yoshua: We need 100% to avoid extinction
  - Yann: There is never anything completely perfect
  - Yoshua: that's the issue, that is why we need to do more than what we do now
  - Yann: it would be good guy AI which is superior to the bad guys AI
- Vaccines, Banning, Regulation, Control (Max)
  - Max: the way to stop a bad guy with bioweapons is to have a good guy with bioweapons? that is not what you do - you stop bioweapon attack by vaccines and banning bioweapons and having regulation and control
- Counter-Intelligence
  - Yann: No, the way to stop bioweapons attack is to have counterintelligence
  - Yoshua: how do you build counterintelligence.
    Need infrastructure to protect ourselves.
  - Yoshua: So we need to recognize there is a risk

Melanie: There is risk, lots of risk, but are they existential and going to end civilization
Yoshua: it could
Need to deal with Immediate Risks also (Melanie)

- Need to be balanced
- Finite resources. Existential Risk Narrative diverted lot of attention.
- Distracts from immediate risks: disinformation and bias

Prudent guidelines for building and deploying safe AI

- Design for Safety
  - Assimilation of AI and nuclear weapons is a fallacy (Yann)
    - Nuclear weapons designed for destruction
    - AI is designed to amplify human intelligence, good
  - Improperly designed systems (e.g., airplanes) can be dangerous (Yann)
  - Address Risks in Design (Max)
    - You or your company should know why AI is dangerous
    - What is the plan to ensure AI doesn't have existential risk?
      - Avoid misuse
      - Solve alignment problem
      - Avoid scenarios where humanity is out-competed
  - Testing for Safety - Prove Safety before Deployment (Max:)
    - Intelligence is not good or bad. It is a tool.
      - Good intelligence people do good
      - Evil intelligent people do evil.
    - Safe until proven unsafe is a dangerous policy
      - Biotech - need to prove meds are safe first, and benefits outweigh harms
      - Nuclear Reactors - cannot deploy assuming safety until proven unsafe
      - AI companies must prove powerful AI is safe before deploying it.
- Governance Regulation for Safety
  - Certifications, and regulations - they exist for AI systems in medicine FDA regulation
  - Prove Safe First, Regulation needed. But existential Risk is too high a bar (Melanie)
  - Need Regulations to ensure we get upside/benefits
  - Also take care of downsides/risks - immediate and existential (Yoshua)
- Engineer for risk/benefit tradeoff Need to focus on risk/benefit tradeoff, risk is negligible --> do not kill benefits (Melanie)
  - The AI community is developing plans to mitigate risks
  - Many risks but not existential
  - Many folks, including Yoshua working on mitigating more immediate real-world risk

15 Closing Statements

Melanie

- AI has made incredible advances. The potential for benefiting humanity is breathtaking.
  - describe images to blind people,
  - help doctors diagnose diseases,
  - predict structures of protein,
  - fluently synthesize enormous amounts of human collective knowledge.
- Science & Technology & AI are double edged swords.

potential risks and misuse

  - AI technology has many risks and potential for harm
    but they are not existential.
- Existential threat narratives are unfounded and harmful
  - Resilience of human society does not support fear stoking

existential risk narratives.

  - Overstating existential threats of AI is harmful
    - Distraction takes away collective attention focus

away from real harms and risks. Misinformation, magnify bias, ethics problems
        - Minimization of real-risks. Example: Hinton resign from Google which minimized Google's work on risks to emphasize existential threats.
- What we need to do
    - Need focus on evidence-based risks to address real harms
    - Do not allow unfounded speculations to inflame our emotions and fears
    - Design AI safe, fair, beneficial based on science not science fiction

Yoshua

- AI has delivered unexpected advances in recent years. Can dialogue in a way pass for humans?
- Current AI is missing some things (misalignment, safety, ..). When extrapolated creates existential concerns.
    - Need to extrapolate and prepare for it.
- Need Social Adaptations.
- We have agency or now. We have a chance to control the future.
- Need to address all risks - immediate and existential.
- Why is there an existential risk?
    - People are crazy and will misuse.
    - Crazy people + very powerful tools - very dangerous
- Reorienting personal research agenda on safe topics
- healthcare,
- environment,
- AI safety

Yann

- Every tech has risks, benefits, bad side effects, and misuses. That is the reality.
    - Every tech has risks and they are not existential.
    - Tech is a tradeoff b/w benefits and bad side-effects.
      Side-effects can be predictable or not.
    - Bad actors will use it for bad things.
- New technology creates fear stemmed in
- fear of unknown and
- fear of lack of control

Lots of fear examples in history, also at the beginning of Internet era

- Reason for Optimism
    - we are working on making them safe
    - we will no deploy if not safe
    - AI will be subservient, smarter but not reduce agency
    - AI will empower us to do great things (like a staff of smart people)
- Risks of not developing AI
    - Miss out on benefits
    - Need AI to fight evil users of AI

Max Pitch for Humility

- Would rather have questions I can't answer, and answers I can question.
- It is an existential threat if any of the extinction scenarios can happen with any non-zero probability.

- Both Optimism / Pessimism need humility
  - McCarthy/Minsky predicted human-level AI by the 1990's
  - Lord Rutherford was pessimistic about nuclear energy
  - Yoshua did not expect GPT-4 for another 30 years
- Humility in estimating the probability (we are not at all good at it)
  - Risk of getting struck by asteroid = 1 in 100M per year (10^-8)
  - Space Shuttle - risk 10^-5. Reality 2 blowing up in 100 launches.
  - Fukushima NPP tsunami - 10^-4 - less than 1 in 10K years. Happened within 1 lifetime.
  - Superhuman AI probability - in 5-20-300 years. need to be humble.