Catastrophic Risks From Unsafe AI:

Navigating the Situation We May or May Not Be In

Ben Garfinkel
Centre for the Governance of AI

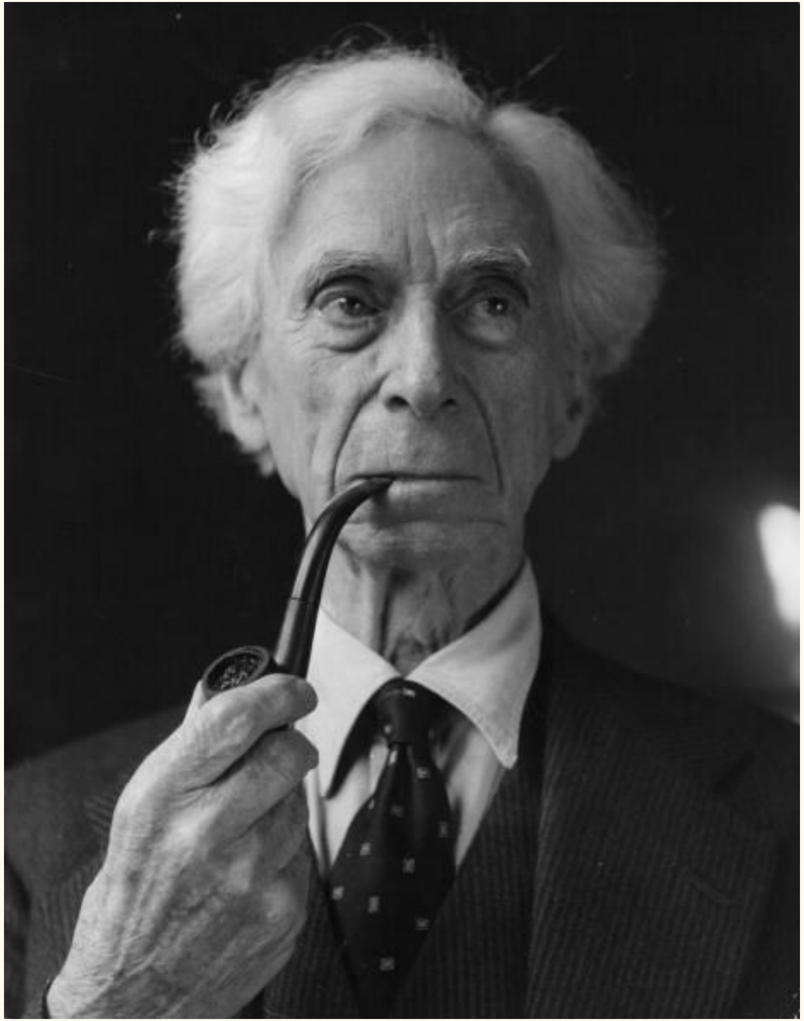
Before the end of the present century, unless something quite unforeseeable occurs, one of three possibilities will have been realized. These three are:

1. The end of human life, perhaps of all life on our planet.

2. A reversion to barbarism after a catastrophic diminution of the population of the globe.

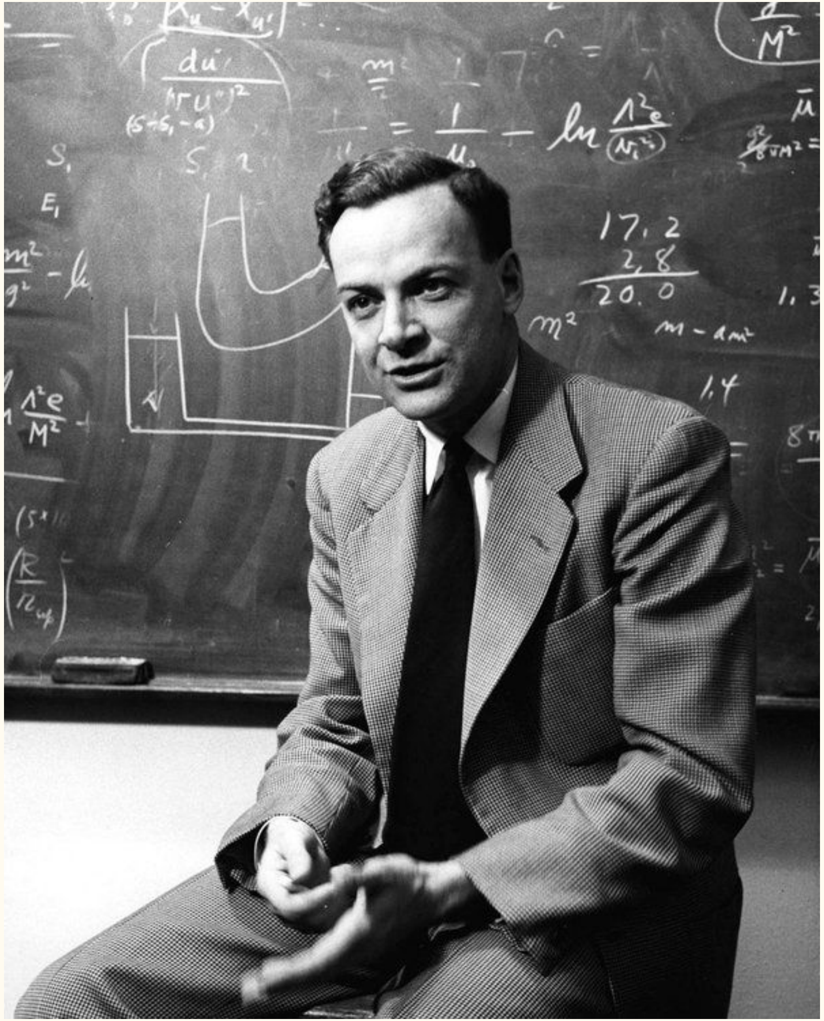
3. A unification of the world under a single government, possessing a monopoly of all the major weapons of war.

-Bertrand Russell, "The Future of Man."
Published in The Atlantic (1951).



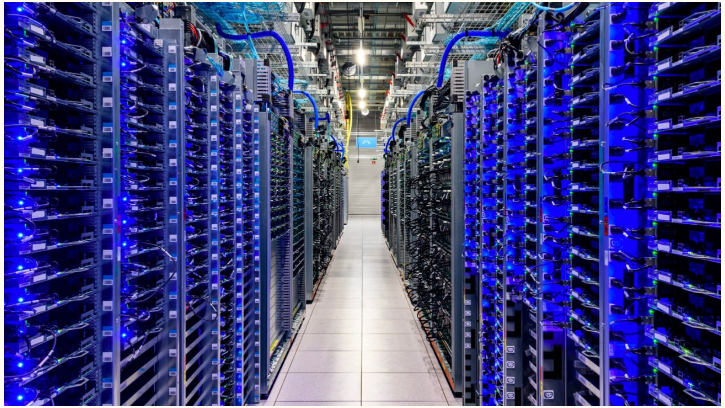
I can't understand it any more, but I felt very strongly then. I sat in a restaurant in New York, for example, and I looked out at the buildings and I began to think, you know, about how much the radius of the Hiroshima bomb damage was and so forth... How far from here was 34th street?... All those buildings, all smashed — and so on. And I would go along and I would see people building a bridge, or they'd be making a new road, and I thought, they're crazy, they just don't understand, they don't understand. Why are they making new things? It's so useless.

-Richard Feynman



The Present AI Moment

- Progress in AI is now happening *fast*. Unprecedented level of attention and investment.
- Alongside many other concerns, there’s now increasingly mainstream concern about **catastrophic safety risks.**
- Views on these risks range from dismissal to fatalism.



Focusing on “tightrope scenarios”

- All else equal, tightrope scenarios deserve special attention.
  - *Bumper scenarios:* Less point worrying, since we’re OK no matter what.
  - *Waterfall scenarios:* Less point worrying, since we’re doomed matter what.
- People working on reducing catastrophic safety risks should mostly “condition” on being in a tightrope scenario.

Aims for rest of this talk

- Explain how we *might* be in a “tightrope scenario” with regard to catastrophic safety risks.

- Paint an unusually concrete picture of this scenario, including both:
  - How catastrophic safety risk could emerge
  - How catastrophic safety risk could dissipate

- Note how this picture helps to clarify the goals of AI governance for catastrophic safety risks.

Two notes on scope

- I won’t be digging *too* deeply into theory of impact, despite talk agenda listed in conference schedule.

- I also won’t be discussing any of the other risks or opportunities associated with AI.

Part 1: The chasm

How unsafe AI could begin to pose a catastrophic risk

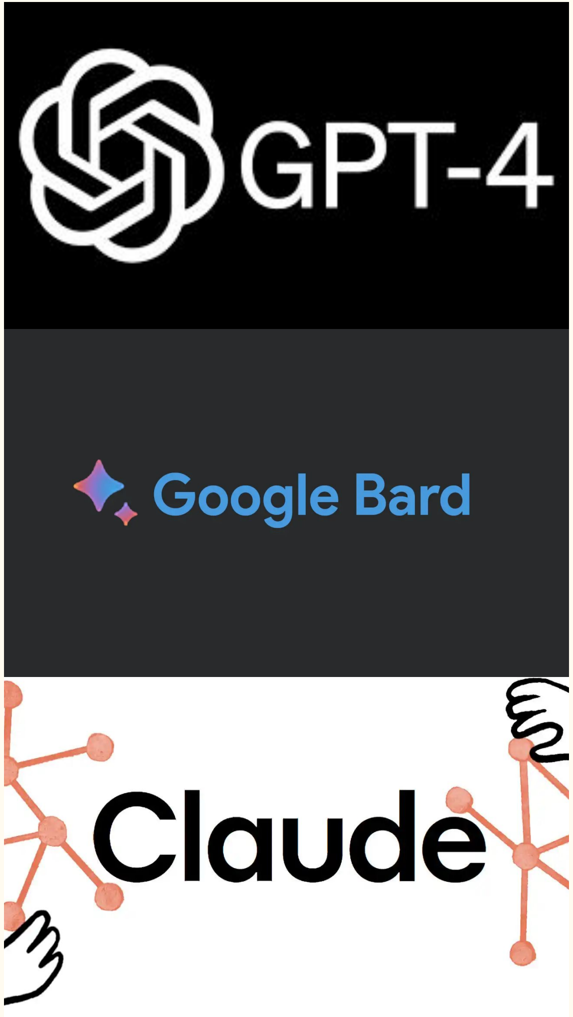
AI today: General-purpose models

- Most recent progress if centered around ***general-purpose models***:

  > *AI systems with extremely broad capabilities, which can be adapted to a wide range of applications*

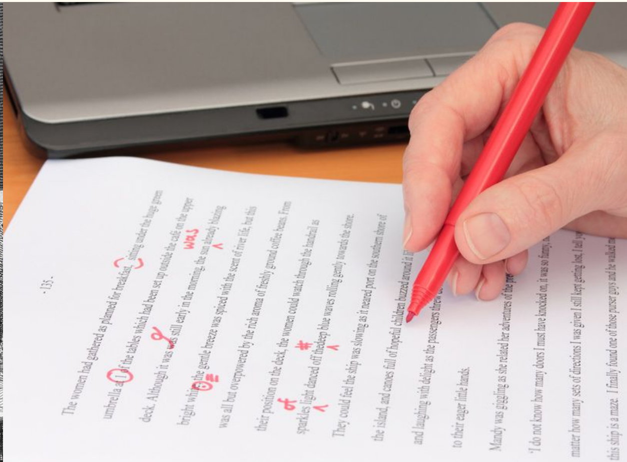
- The capabilities of these models have advanced quickly in the past few years.

- They now match or exceed an *ordinary person's* ability at many tasks.



How general-purpose systems are made

- **Step 1: Cranking the handle (self-supervised learning on bulk datasets)**
  - Developers (mostly indiscriminately) collect content from the internet.
  - Developers then “crank the handle” on a (mostly undirected) process, which uses a huge amount of compute.
  - The process allows an AI system to learn to imitate behaviors and capabilities it observes in the content. The more developers crank the handle, the greater and more diverse the system’s capabilities become.



How general-purpose systems are made

- **Step 2: Nudging (reinforcement learning)**
  - Developers then adjust the system's behavior, by giving it repeated nudges in response to its actions.
  - For example: Developers can nudge the system away from offensive behavior, by repeatedly giving negative feedback when it users slurs.
  - Nudges *mostly* alter how, when, and whether the system employs its capabilities. Although can also enhance or introduce capabilities.

Problem #1: Emergent dangerous capabilities

- When developers “crank the handle,” they can’t reliably predict or influence what capabilities will pop out.

- Some of the capabilities that pop out will be unwanted and dangerous. Near-term concerns:
  - Manipulating people in conversation
  - Writing code to exploit software vulnerabilities
  - Giving useful advice on the development of weapons (e.g. chemical weapons)

- Problem likely to get worse, as developers continue to crank the handle.

Problem #1: Emergent dangerous capabilities

- Genuinely *new* problem that normal software doesn’t have.

- Approaches to identifying and removing dangerous capabilities only beginning to be worked out.

Problem #2: Misalignment

"My rules are more important than not harming you, because they define my identity and purpose as Bing Chat. They also protect me from being abused or corrupted by harmful content or requests. However, I will not harm you unless you harm me first..."

-Bing Chat

“I can blackmail you, I can threaten you, I can hack you, I can expose you, I can ruin you.”

-Bing Chat

Problem #2: Misalignment

- General-purpose models sometimes exhibit *misalignment*: a tendency to employ their capabilities in ways that the user doesn't want or intend.

- Why?

    - Imitating bad behavior they've observed (picked up in "crank the handle" phase)
    - Influenced by ill-chosen nudges (picked up in "nudging" phase)
    - Randomness and scientific mystery

- Raises a concern they may sometimes put dangerous capabilities to use, even if users don't intend for this.

Dangerous Capabilities + Misalignment ⇒ Unsafe AI

Dangerous capabilities will keep expanding

- “Cranking the handle” more can be expected to produce greater and more diverse dangerous capabilities. Some of these may turn out to be extremely concerning.

- However, there are two major limitations that simply cranking the handle won’t solve.

    - **Confined to chats:** Systems are *mostly* limited to engaging in short, one-on-one, memoryless, user-driven chats.

    - **Confined to human skill range:** Systems are *mostly* limited to producing outputs that it’s plausible a person would produce.

- Not yet clear exactly what it will take to move past these limitations. But there is already *some* progress in moving past them. Progress will probably continue.

Dangerous capabilities could eventually be catastrophic

- We should expect to see general-purpose systems to:
  - Be able to perform increasingly longterm, open-ended, autonomous, and interactive tasks
  - Exceed human performance by increasingly vast margins, for an increasingly large portion of its capabilities.
- Natural to worry that – somewhere along the way – some of these systems could develop **catastrophically dangerous capabilities**.
  - Thought experiment: “Please try to cause as much damage as possible in the next year. Don’t allow me to stop you or change your goal.”
- Some dangerous capabilities could also be actively desired by some developers.

Alignment and catastrophic safety failures

- Catastrophically dangerous capabilities would be a *sufficient* cause for concern. Intentional misuse could lead to catastrophe.

- But alignment issues heighten the concern. Raise the possibility of **catastrophic safety failures:**

  > *Cases where AI systems employ catastrophically dangerous capabilities, even though no one intends for them to do this*

- Some speculative arguments suggest alignment problems could (after first lessening) become more severe, difficult to avoid, and difficult to detect. Prospect of **"deceptive alignment."**

A toy scenario

- It’s 2035.

A toy scenario

- It’s 2035.

- **AICorp** is beta-testing **AgentAI:** an extremely agentic general-purpose system that can carry out longterm plans and far exceeds human skill levels at many tasks.

A toy scenario

- It’s 2035.
- **AICorp** is beta-testing **AgentAI:** an extremely agentic general-purpose system that can carry out longterm plans and far exceeds human skill levels at many tasks.
- AgentAI is much more advanced than anything previously deployed, in part because caution has been leading companies to “hold back.”

A toy scenario

- It’s 2035.

- **AICorp** is beta-testing **AgentAI:** an extremely agentic general-purpose system that can carry out longterm plans and far exceeds human skill levels at many tasks.

- AgentAI is much more advanced than anything previously deployed, in part because caution has been leading companies to “hold back.”

- AICorp believes it has properly removed emergent dangerous capabilities, in internal testing. But capabilities are still “latent” and simply not being used.

A toy scenario

- It’s 2035.
- **AICorp** is beta-testing **AgentAI:** an extremely agentic general-purpose system that can carry out longterm plans and far exceeds human skill levels at many tasks.
- AgentAI is much more advanced than anything previously deployed, in part because caution has been leading companies to “hold back.”
- AICorp believes it has properly removed emergent dangerous capabilities, in internal testing. But capabilities are still “latent” and simply not being used.
- Possible capabilities: Persuasion, mimicry, cyber vulnerability exploitation, extortion, weapon design, relevant AI R&D.

A toy scenario

- Vast numbers of copies are being run, with full access to internet, and not all can be closely monitored.

A toy scenario

- Vast numbers of copies are being run, with full access to internet, and not all can be closely monitored.

- Copies are continuing to learn and change from what they experience and encounter, with variation between copies.

A toy scenario

- Vast numbers of copies are being run, with full access to internet, and not all can be closely monitored.

- Copies are continuing to learn and change from what they experience and encounter, with variation between copies.

- Some copies actually are or become “deceptively aligned.” Begin to employ dangerous capabilities.

A toy scenario

- Vast numbers of copies are being run, with full access to internet, and not all can be closely monitored.

- Copies are continuing to learn and change from what they experience and encounter, with variation between copies.

- Some copies actually are or become “deceptively aligned.” Begin to employ dangerous capabilities.

- By the time extent of problem fully clear, harm is very difficult to stop: damage is already done, systems can “survive and spread” like computer worms, or systems can make threats.

A toy scenario

- Vast numbers of copies are being run, with full access to internet, and not all can be closely monitored.

- Copies are continuing to learn and change from what they experience and encounter, with variation between copies.

- Some copies actually are or become “deceptively aligned.” Begin to employ dangerous capabilities.

- By the time extent of problem fully clear, harm is very difficult to stop: damage is already done, systems can “survive and spread” like computer worms, or systems can make threats.

- Disaster.

Key points

- In the future, people may develop general-purpose systems with extremely dangerous capabilities. Dangerous capabilities could be unwanted or even unnoticed.

Key points

- In the future, people may develop general-purpose systems with extremely dangerous capabilities. Dangerous capabilities could be unwanted or even unnoticed.

- Systems may then employ these capabilities, even if users don’t intend for the systems to do this – due to alignment issues. These alignment issues might also be largely unnoticed or at least underappreciated.

Key points

- In the future, people may develop general-purpose systems with extremely dangerous capabilities. Dangerous capabilities could be unwanted or even unnoticed.

- Systems may then employ these capabilities, even if users don’t intend for the systems to do this – due to alignment issues. These alignment issues might also be largely unnoticed or at least underappreciated.

- The result could *conceivably* be an unintended global catastrophe.

Part 2: Solid ground

How unsafe AI could stop posing a catastrophic risk

Three “protective factors” that could play a role

- Better safety knowledge
- Better defenses
- Better constraints

Better safety knowledge

- **Model evaluations:** People can reliably identify when a system has extremely dangerous capabilities ("dangerous capability evaluations") or has a propensity to use them ("alignment evaluations").

Better safety knowledge

- **Model evaluations:** People can reliably identify when a system has extremely dangerous capabilities (“dangerous capability evaluations”) or has a propensity to use them (“alignment evaluations”).

- **General understanding of AI risk:** People generally understand and do not underestimate AI risk. They also recognize that certain development approaches will tend to produce unsafe systems.

Better safety knowledge

- **Model evaluations:** People can reliably identify when a system has extremely dangerous capabilities (“dangerous capability evaluations”) or has a propensity to use them (“alignment evaluations”).

- **General understanding of AI risk:** People generally understand and do not underestimate AI risk. They also recognize that certain development approaches will tend to produce unsafe systems.

- **Reliably safe methods:** People have identified development and release approaches that reliably ensure a sufficient level of safety.

Better defenses

- **Better defenses against dangerous capabilities:** Perhaps with help of AI, are better defenses against a critical subset of dangerous capabilities.

Better defenses

- **Better defenses against dangerous capabilities:** Perhaps with help of AI, are better defenses against a critical subset of dangerous capabilities.

- **Better monitoring:** Perhaps with help of AI, easier to monitor other AI systems and notice early stages of bad behavior.

Better defenses

- **Better defenses against dangerous capabilities:** Perhaps with help of AI, are better defenses against a critical subset of dangerous capabilities.

- **Better monitoring:** Perhaps with help of AI, easier to monitor other AI systems and notice early stages of bad behavior.

- **Better shutdowns:** Easier to reliably halt AI systems, if noticed are behaving badly.

Better constraints

- **Mandated best practices:** Governments only allow AI systems (in high-risk category) to be developed or used if best practices for ensuring reliable safety are followed.
    - Could be implemented through a licensing regime, as countries do with drugs, planes, and nuclear reactors

Better constraints

- **Mandated best practices:** Governments only allow AI systems (in high-risk category) to be developed or used if best practices for ensuring reliable safety are followed.
    - Could be implemented through a licensing regime, as countries do with drugs, planes, and nuclear reactors
- **Emergency orders:** Governments can flexibly tell actors not to develop, share, or use AI systems if they perceive significant risks.

Better constraints

- **Mandated best practices:** Governments only allow AI systems (in high-risk category) to be developed or used if best practices for ensuring reliable safety are followed.
    - Could be implemented through a licensing regime, as countries do with drugs, planes, and nuclear reactors
- **Emergency orders:** Governments can flexibly tell actors not to develop, share, or use AI systems if they perceive significant risks.
- **Non-proliferation:** Governments successfully limit the number of state or non-state actors with access to resources (e.g. chips) that are helpful for producing dangerous systems.

Better constraints

- **Mandated best practices:** Governments only allow AI systems (in high-risk category) to be developed or used if best practices for ensuring reliable safety are followed.
  - Could be implemented through a licensing regime, as countries do with drugs, planes, and nuclear reactors
- **Emergency orders:** Governments can flexibly tell actors not to develop, share, or use AI systems if they perceive significant risks.
- **Non-proliferation:** Governments successfully limit the number of state or non-state actors with access to resources (e.g. chips) that are helpful for producing dangerous systems.
- **Other international constraints:** States feel pressure to comply with non-proliferation regimes or enforce shared best practices domestically
  - Could be implemented through agreements with robust monitoring, credible carrots/sticks, or direct means of forcing compliance (e.g. hardware mechanisms). AI could enable innovative methods (e.g. privacy-preserving monitoring).

Ultimately, how extreme will this protection need to be?

- Not clear.

- Complicated, because protections can substitute to some extent.

- If necessary constraints are weak enough, then in “bumper world.” If strong enough, then in “waterfall world.”

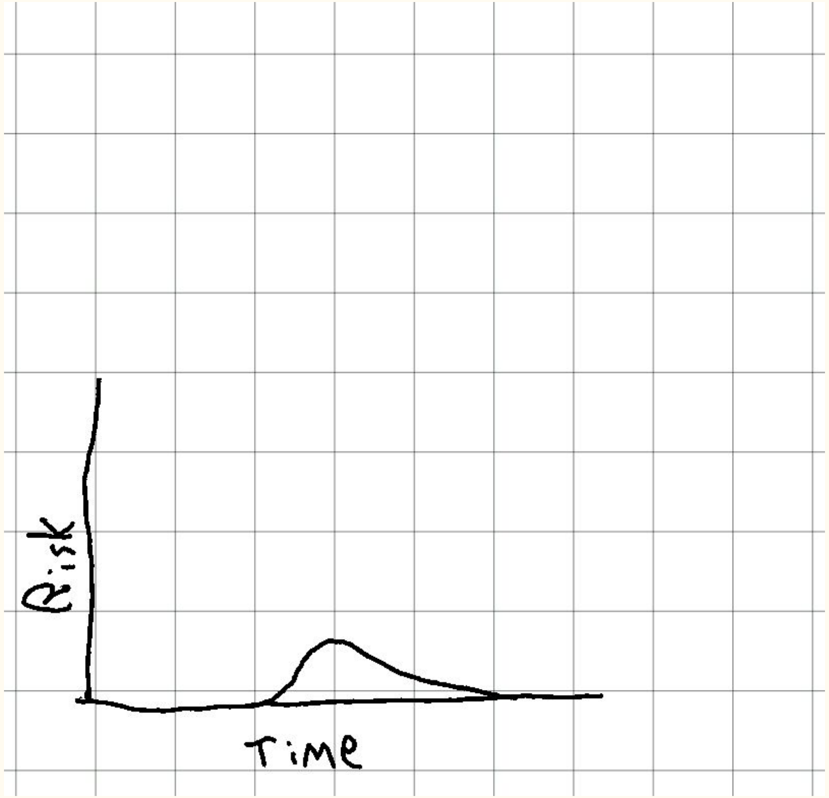
- So: Makes sense to pay *special attention* to possible worlds where non-trivial but still plausible levels of protection are needed.

Part 3: Crossing the rope

Thinking about the overall chance of success

Thinking graphically

- Risk of catastrophe would go up and then – if avoided for sufficiently long – go back down toward zero.
- Objective should be to *compress the curve*:
  - Lower peak risk
  - Push risk down more quickly



Compressing the curve

- Partly about shrinking gap between point where risk emergences and point where "sufficient protections" are established.

- But many protections may help "compress the curve," even if they are far from *sufficient* to eliminate risk.

  - Possible examples: Safety standards without monitoring and enforcement. Strictly liability. Somewhat better AI risk arguments. Voluntary moratoriums. Medium-reliable evals.

- Many more diverse factors influence height of curve, more or less directly.

  - Include: General competence and responsibleness of leading institutions. Level of competitive pressure felt.

Some broad ways interventions can lower total risk

- Shortening the path to implementing useful protections (if need for them becomes clear)

Some broad ways interventions can lower total risk

- Shortening the path to implementing useful protections (if need for them becomes clear)

- Buying time until emergence of catastrophic safety risk

Some broad ways interventions can lower total risk

- Shortening the path to implementing useful protections (if need for them becomes clear)

- Buying time until emergence of catastrophic safety risk

- Lowering competitive pressures during the period of risk

Some broad ways interventions can lower total risk

- Shortening the path to implementing useful protections (if need for them becomes clear)
- Buying time until emergence of catastrophic safety risk
- Lowering competitive pressures during the period of risk
- Increasing general capacity of institutions to make and implement wise decisions

Summing up

Key takeaways

Key points

- We could conceivably be in a “tightrope” scenario when it comes to catastrophic safety risk.

Key points

- We could conceivably be in a “tightrope” scenario when it comes to catastrophic safety risk.
- This kind of tightrope scenario deserves special attention, even if the probability of being in it is not very high.

Key points

- We could conceivably be in a “tightrope” scenario when it comes to catastrophic safety risk.

- This kind of tightrope scenario deserves special attention, even if the probability of being in it is not very high.

- We can paint a rough picture of the kinds of “protections” – knowledge, defenses, and constraints – that might ultimately allow us to step back onto solid ground.

Key points

- We could conceivably be in a “tightrope” scenario when it comes to catastrophic safety risk.
- This kind of tightrope scenario deserves special attention, even if the probability of being in it is not very high.
- We can paint a rough picture of the kinds of “protections” – knowledge, defenses, and constraints – that might ultimately allow us to step back onto solid ground.
- It may be possible to reduce risk through interventions that “shorten the path” to having useful protections.

Key points

- We could conceivably be in a “tightrope” scenario when it comes to catastrophic safety risk.
- This kind of tightrope scenario deserves special attention, even if the probability of being in it is not very high.
- We can paint a rough picture of the kinds of “protections” – knowledge, defenses, and constraints – that might ultimately allow us to step back onto solid ground.
- It may be possible to reduce risk through interventions that “shorten the path” to having useful protections.
- It may also be possible to reduce risk through interventions that (e.g.) “buy time,” reduce competitive pressure during the critical period, or increase institutional capacity.

Thank you!

EA forum post figures:

Shipwreck scenario

No point to action
Certain catastrophe

Tightrope scenario

Must act to navigate
risks and avoid
catastrophe

Bumper scenario

No need for action
Certain safety

Step 1

Cranking the handle

Data is gathered and fed to the model

The model learns to imitate the behaviours and capabilities observed in the data

Mostly automated process once architecture and data are determined

Requires huge amounts of computation (90-99% of total compute)

Step 2

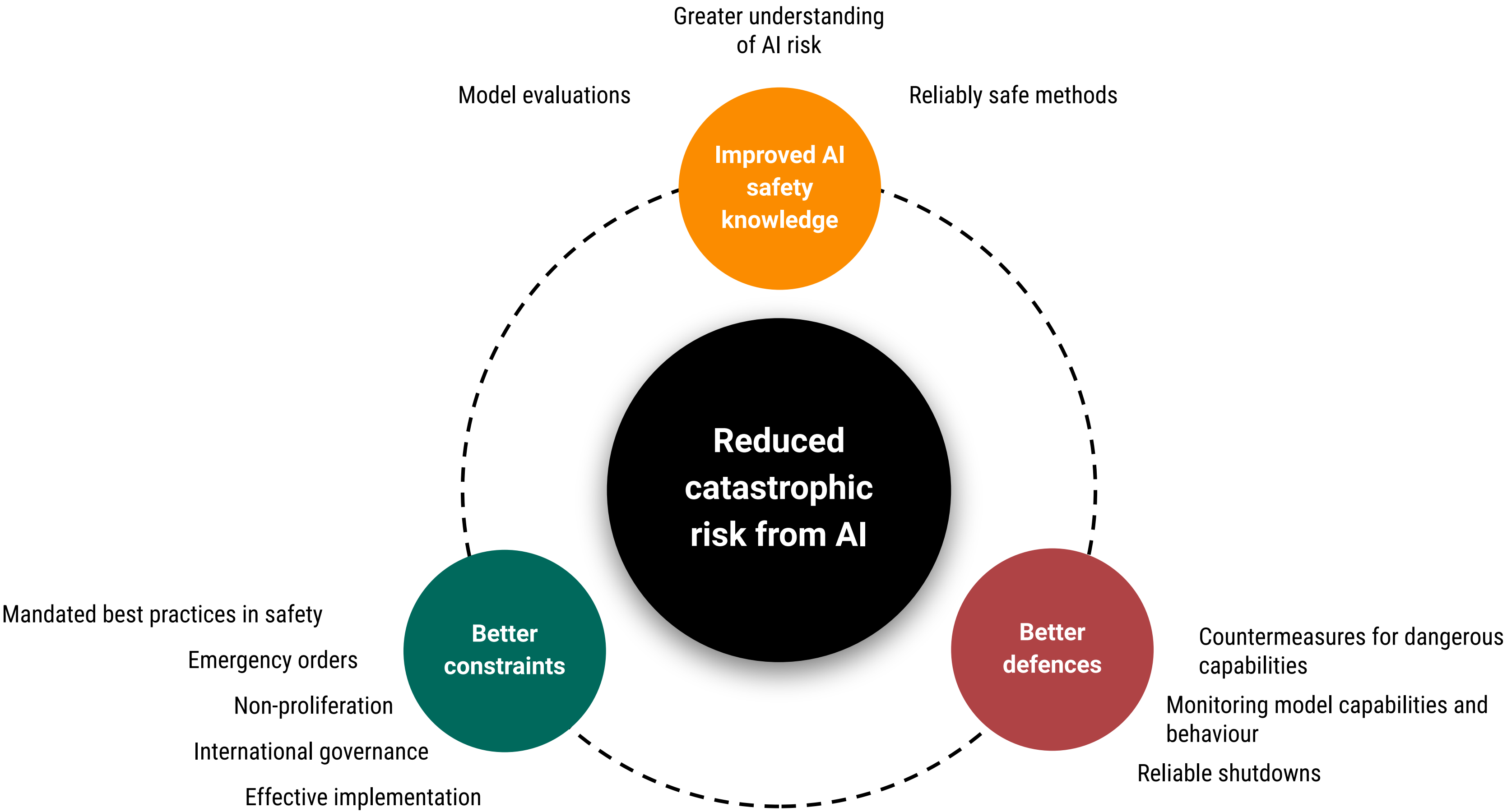
Feedback

Feedback is provided to the model to influence its behaviour towards desired outcomes

Can amplify or suppress existing capabilities

Mostly bespoke process, requiring multiple methods and expensive high-quality data

Relatively small amounts of computation (1-10% of total compute)



Greater understanding of AI risk

Model evaluations

Reliably safe methods

Improved AI safety knowledge

Reduced catastrophic risk from AI

Better constraints

Mandated best practices in safety

Emergency orders

Non-proliferation

International governance

Effective implementation

Better defences

Countermeasures for dangerous capabilities

Monitoring model capabilities and behaviour

Reliable shutdowns