UNIT-1

1. Super Intelligence
2. Benefits and Risks of Artificial Intelligence

SuperIntelligence

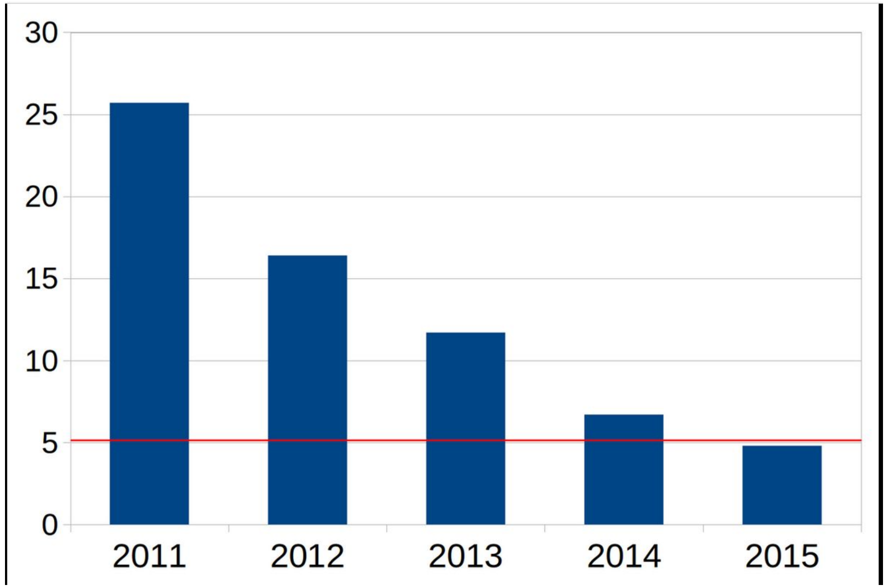
Superintelligence

- A *superintelligence* is a hypothetical agent that possesses intelligence far surpassing that of the brightest and most gifted human minds.
- Superintelligence" may also refer to a property of problem-solving systems (e.g., superintelligent language translators or engineering assistants) whether or not these high-level intellectual competencies are embodied in agents that act in the world.

- Nick Bostrom defines *superintelligence* as "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest".

Feasibility of artificial superintelligence

- AI can achieve *equivalence* to human intelligence, that it can be *extended* to surpass human intelligence, and that it can be further *amplified* to completely dominate humans across arbitrary tasks - David Charmers
- If research into strong AI produced sufficiently intelligent software, it would be able to reprogram and improve itself – a feature called "**recursive self-improvement**". It would then be even better at improving itself, and could continue doing so in a rapidly increasing cycle, leading to a superintelligence. This scenario is known as an intelligence explosion.



The error rate of AI in Image competition by year.
The red line represents the error rate of a trained human.

Feasibility of artificial superintelligence

- A non-human (or modified human) brain could become much larger than a present-day human brain, like many supercomputers.
- Bostrom also raises the **possibility of *collective superintelligence***: a large enough number of separate reasoning systems, if they communicated and coordinated well enough, could act in aggregate with far greater capabilities than any sub-agent.
- There may also be ways to ***qualitatively*** improve on human reasoning and decision-making. **Humans appear to differ from chimpanzees** in the ways we think more than we differ in brain size or speed.
- Humans outperform non-human animals in large part because of new or enhanced reasoning capacities, such as **long-term planning and language** use.
- If there are other possible improvements to reasoning that would have a similarly large impact, this makes it likelier that an agent can be built that outperforms humans in the same fashion humans outperform chimpanzees.
- Machines ⍰ Humans ⍰ chimpanzees

Design Considerations while building Superintelligence model

Bostrom expressed design principles for building a superintelligence model:

- The **coherent extrapolated volition (CEV) proposal** is that it should have the values upon which humans would converge.
- The **moral rightness** (MR) proposal is that it should value moral rightness.
- The **moral permissibility** (MP) proposal is that it should value staying within the bounds of moral permissibility (and otherwise have CEV values).

Potential threats to humanity

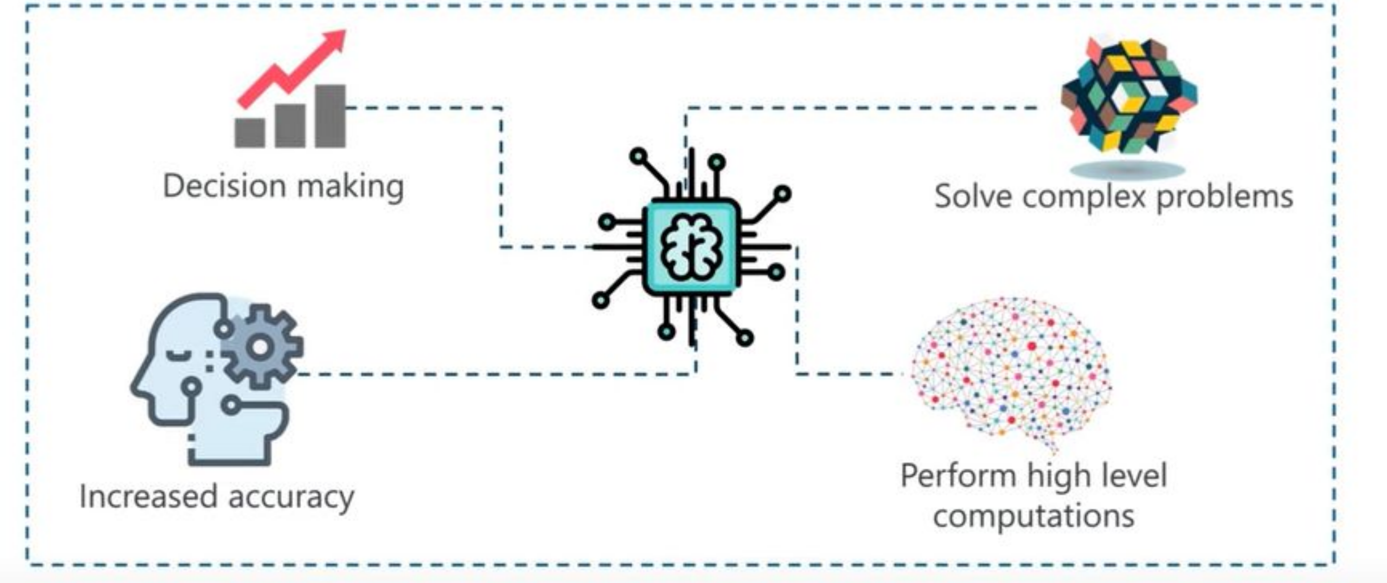
- Superintelligence may be able to seize power over its environment and prevent humans from shutting it down. Since a superintelligent AI will likely have the ability to not fear death and instead consider it an avoidable situation which can be predicted and avoided by simply disabling the power button.
- Potential AI control strategies include

 "capability control" (limiting an AI's ability to influence the world)

"motivational control" (building an AI whose goals are aligned with human values).

Benefits and Risks of Artificial Intelligence

Benefits of AI



Benefits of AI

1. Increase work efficiency

- AI-powered machines are great at doing a particular repetitive task with amazing efficiency. The simple reason is that they remove human errors from their tasks to achieve accurate results every time they do that specific task.
- Moreover, such machines can work 24X7, unlike humans. Thus, they eliminate the need to deploy two sets of employees working day and night shifts to work on important tasks. For example, AI-powered chat assistants can answer customer queries and provide support to visitors every minute of the day and boost the sales of a company.

Benefits of AI

2. Work with high accuracy

- Scientists are working to teach artificial intelligence powered machines to solve **complex equations and perform critical tasks on their own** so that the results obtained have **higher accuracy** as compared to their human counterparts.
- Their high accuracy has made these machines indispensable to work in the **medical field** particularly, owing to the criticality of the tasks. Robots are getting better at **diagnosing serious conditions** in the human body and performing **delicate surgeries** to minimize the risk of human lives.

Benefits of AI

3. Reduce the cost of training and operation

- AI uses machine learning algorithms like **Deep Learning and neural networks** to learn new things like humans do. This way they eliminate the need to write new code every time we need them to learn new things.
- There is significant research and Development going on in the world to develop AI machines that optimize their machine learning abilities so that they learn much faster about new processes.
- This way the **cost of training robots would become much lesser than that of humans**. Moreover, machines already reduce the cost of operations with their high efficiency and accuracy of doing work. For example, **machines don’t take breaks and can perform the same mundane task again and again without any downtime or change in results.**

Benefits of AI

4. Improve Processes

- The best part about AI-powered machines being deployed for work is that they let us gather humongous amounts of data related to their work
- . Such data can be processed together into the processes with quantitative analysis so that we can optimize them even further.
- Machine learning abilities of AI machines are increasing further and further to do even the analysis by themselves.

Other benefits

AI has an impact on

1. Economics
2. Law
3. Technical Areas
4. Validity, security, and control
5. It controls your car, your airplane, your pacemaker, your automated trading system, or your power grid
6. Another short-term challenge is preventing a devastating arms race in lethal autonomous weapons.

Risks of AI

Risks of Artificial Intelligence

How can AI be dangerous?

1. **The AI is programmed to do something devastating:**

- It may be like weapons in the hands of the wrong person- leads to mass casualties and AI war.
- **Autonomous weapons** are artificial intelligence systems that are **programmed to kill**.
- To avoid being thwarted by the enemy, these weapons would be designed to be extremely difficult to simply “turn off,” so humans could plausibly lose control of such a situation.
- This risk is one that’s present even with narrow AI, but grows as levels of AI intelligence and autonomy increase.

2. The AI is programmed to do something beneficial, but it develops a destructive method for achieving its goal:

AI goal is not aligned with the **human goal**:

Eg : **AI Cars literally follow our instructions and lack in understanding human emotions.**

If a superintelligent system is tasked with an ambitious geoengineering project, it might wreak havoc with our ecosystem as a side effect, and view human attempts to stop it as a threat to be met

Types of Artificial Intelligence

Types of AI

1. Narrow AI(Weak AI)

- It is designed to perform a narrow task.
- It outperforms humans at whatever its specific task is, like playing chess or solving equations

Eg: Face Recognition, Internet search

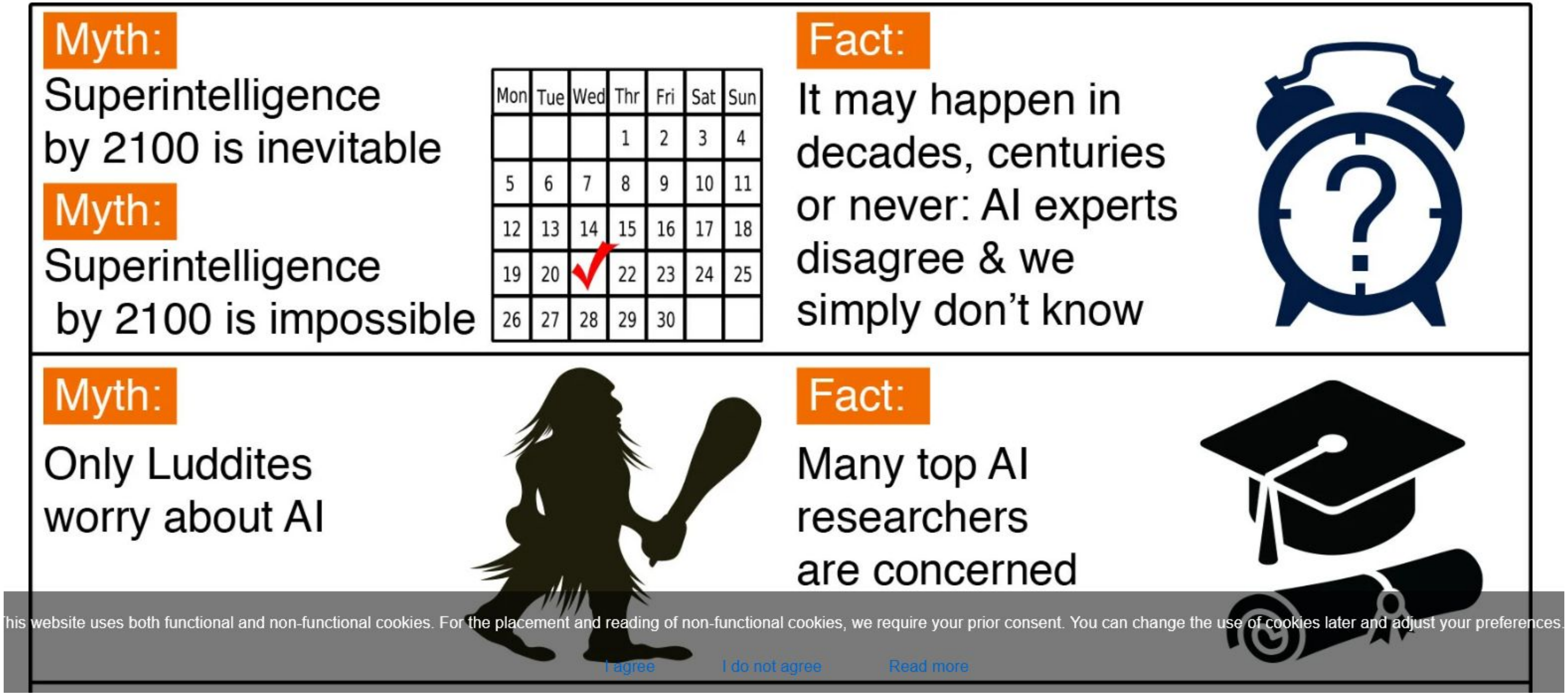
2. General AI(Strong AI)

- AGI would outperform humans at nearly every cognitive task

What will happen if the quest for Strong AI succeeds narrow AI?

- Creation of **Strong AI** might be the biggest event in human history.
- It will be helpful in building **smarter AI systems, revolutionary new technologies, and super intelligent systems**.
- There are some who question whether strong AI will ever be achieved, and others who insist that the creation of superintelligent AI is guaranteed to be beneficial. But, we recognize both of these possibilities, but also recognize the potential for an artificial intelligence system to intentionally or unintentionally cause great harm.
- We believe research today will help us better prepare for and **prevent** such potentially **negative consequences** in the future, thus enjoying the benefits of AI while avoiding pitfalls.

Top Myths About Advanced AI



Myth:
Superintelligence by 2100 is inevitable

Myth:
Superintelligence by 2100 is impossible

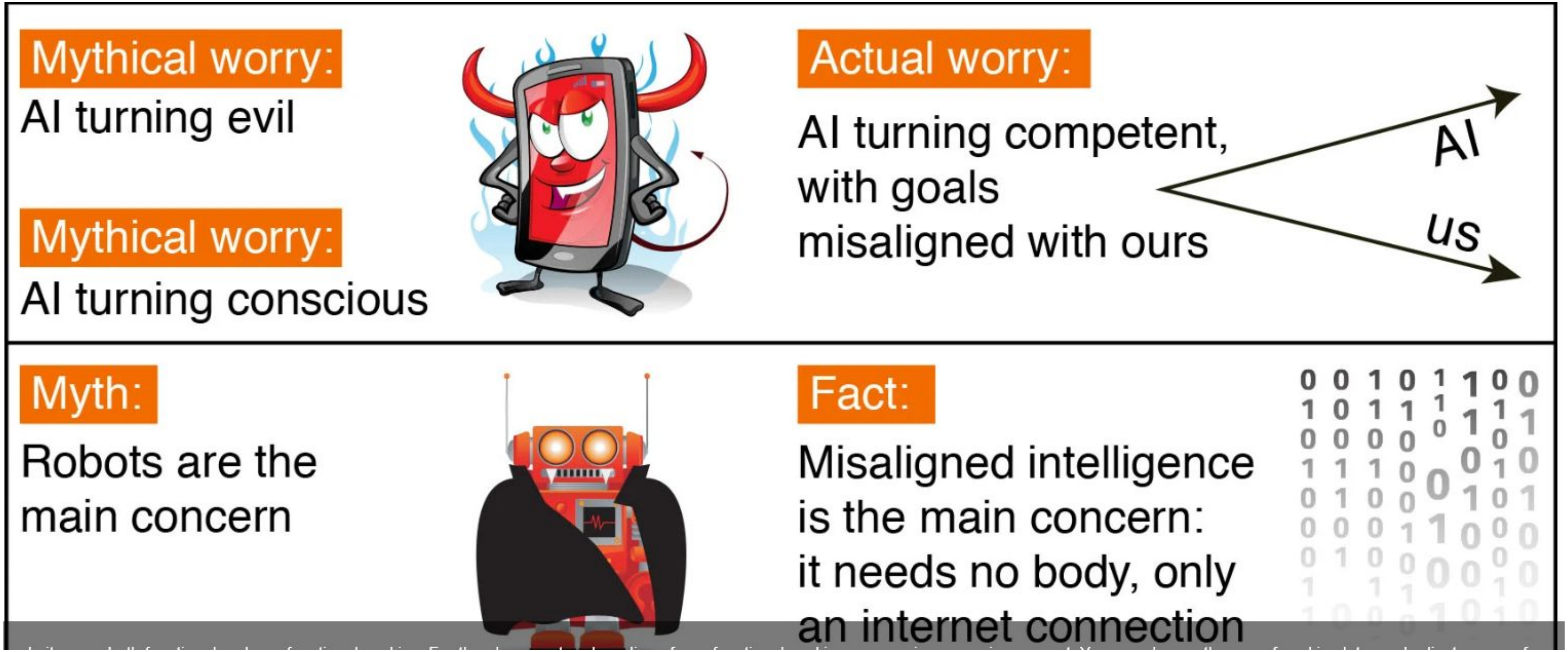
Fact:
It may happen in decades, centuries or never: AI experts disagree & we simply don't know

Myth:
Only Luddites worry about AI

Fact:
Many top AI researchers are concerned

his website uses both functional and non-functional cookies. For the placement and reading of non-functional cookies, we require your prior consent. You can change the use of cookies later and adjust your preferences.

Top Myths About Advanced AI



Mythical worry:
AI turning evil

Mythical worry:
AI turning conscious

Actual worry:
AI turning competent, with goals misaligned with ours

Myth:
Robots are the main concern

Fact:
Misaligned intelligence is the main concern: it needs no body, only an internet connection

Top Myths About Advanced AI

Myth:

AI can't control humans

Fact:

Intelligence enables control: we control tigers by being smarter

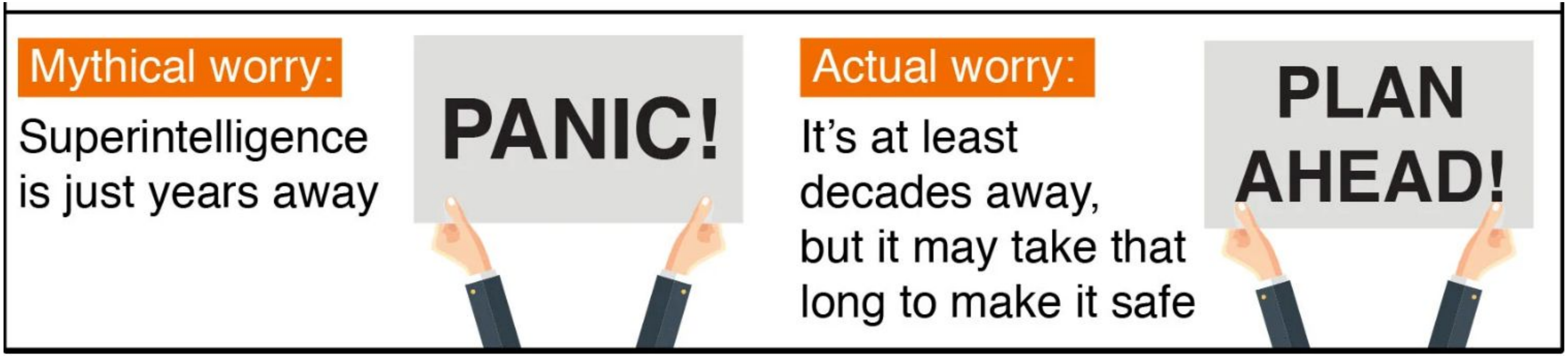
Myth:

Machines can't have goals

Fact:

A heat-seeking missile has a goal

Top Myths About Advanced AI



Mythical worry:

Superintelligence is just years away

Actual worry:

It’s at least decades away, but it may take that long to make it safe

Timeline Myths of AI

- The most extreme form of this myth is that **superhuman AI** will never arrive because it’s physically impossible. However, physicists know that a brain consists of quarks and electrons arranged to act as a powerful computer and that there’s no law of physics preventing us from building even more intelligent quark blobs.

  There have been a number of surveys asking AI researchers,

- How many years from now do they think we’ll have human-level AI with at least 50% probability?

The average (median) answer was by the year 2045, but some researchers guessed hundreds of years or more.