Where is AI in 2026 and Where is it Going?

Job automation, AI companions, Loss of control scenarios. What is this technology and where is it headed?

Presented by
the AI Safety Awareness Project

About AI Safety Awareness Project

- 501(c)(3) nonprofit organization
- Our Mission: AI will affect everyone, everyone should have a voice in the direction of this tech
- **Knowledge** → Opinion → Voice
- Education and Engagement on AI Safety and Risks
  - Community Workshops
  - Organizational Partnerships

What is AI?

- **Lots of definitions of AI**
  - No **universal, official** definition of AI
- Good ways to think about what AI is:
  - Machines mimicking aspects human intelligence
  - AI**: Artificial** Intelligence

    Intelligence: ability to select and take actions toward a goal
- **What behaviors can we think of that showcase human intelligence?**

Brief History of AI

1940s

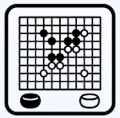
Thermostat

1997

Deep Blue Defeats Chess Champ

2011

Watson Wins Jeopardy!



2016

AlphaGo Defeats Go Champ

2021

DALLE Image Generator

2022

ChatGPT + explosion of capabilities

* Images on this slide were created using generative AI

Narrow -> General AI

ChatGPT

1997

Deep Blue beats Chess Grandmasters

2011

IBM Watson beats Jeopardy grand champions

2022-

Chat GPT enters national stage

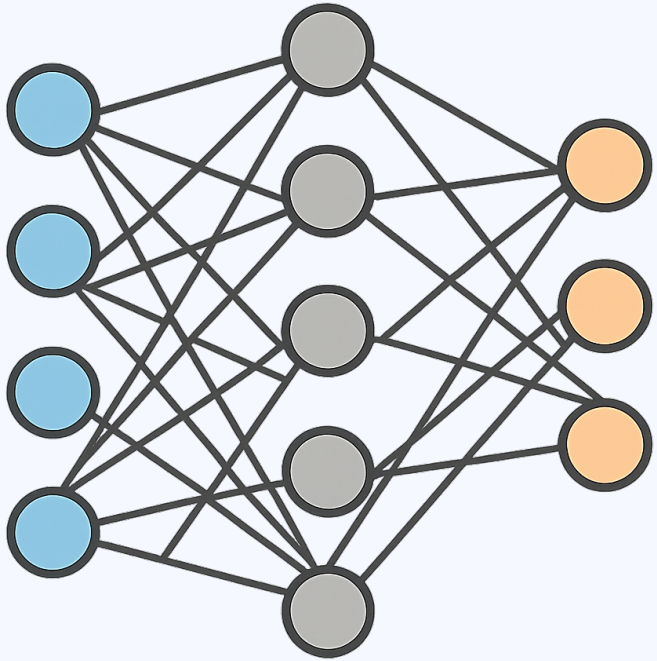
How Does AI Work?

- **Neural networks**
- How would you teach someone to recognize dogs?
  - The easiest way might be to show them lots of pictures
  - Might be able to recognize dogs… but HOW do you recognize them is a different matter!

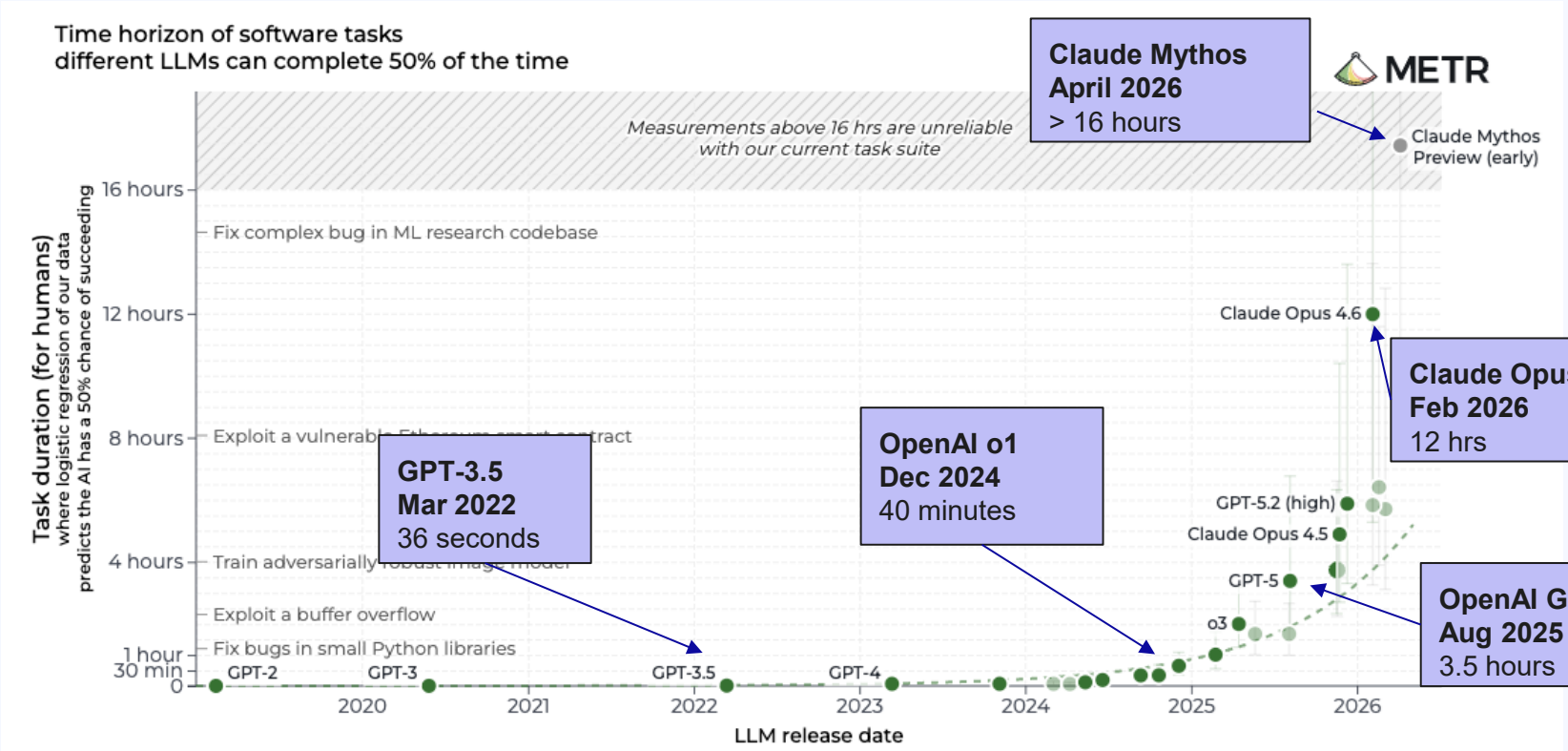
How Does AI Work?

Systems that are built on patterns of information

- NOT Software
  - No step by step rules
  - Behavior emerges from data
- Trained, NOT designed
  - Goal is defined
  - Model figures out how to achieve it



Pace of Recent Progress



AI Video Generation

AI Generated Video:
State of the Art in
March 2023

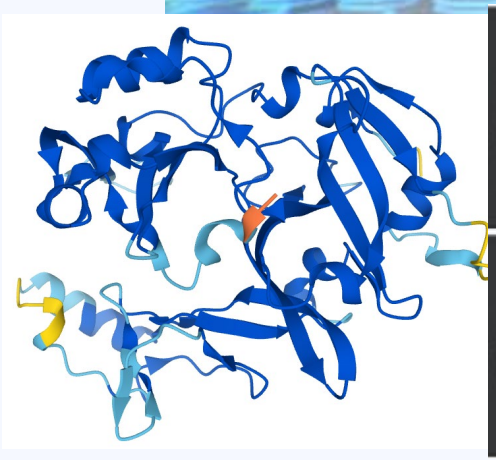
AI Video Generation

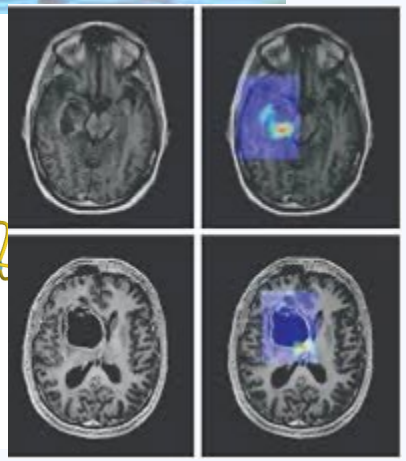
AI Generated Videos
State of the Art in
Feb 2026

Why is AI such a big deal?

AI enabled breakthroughs

- Tumor and Disease detection
- Realtime cybersecurity threat detection
- Self driving cars
- “Protein Folding Problem” - AlphaFold





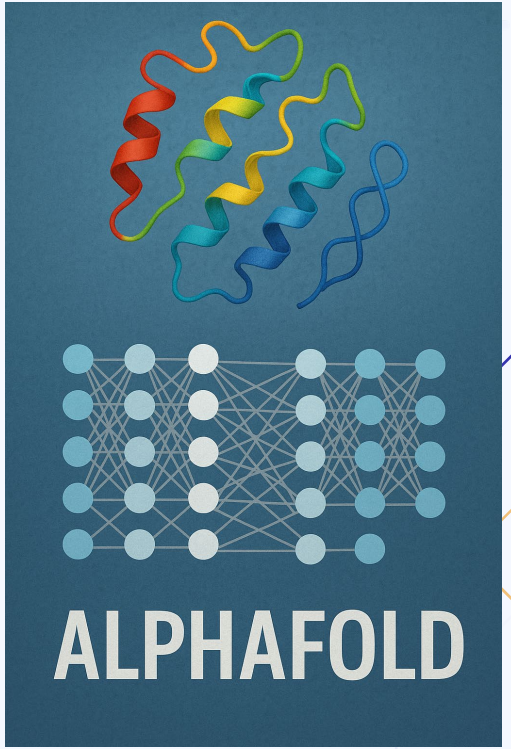
Narrow AI Can be Powerful!

Problem:

- Predicting a protein's 3D structure is ***very*** difficult, but ***very*** useful

AI Solution: **AlphaFold2** - Google DeepMind

- Studied hundreds of thousands of known protein structures, used deep learning to predict unknown proteins' structures
- *Decades-long problem effectively solved by AI*



How big will AI be?

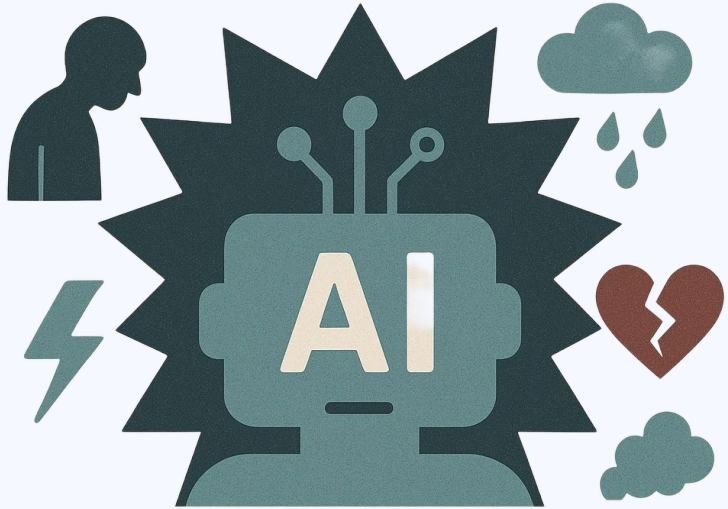
- **10 times bigger than the Industrial Revolution and maybe 10 times faster.** – Google DeepMind CEO Demis Hassabis
- Automation of automation, software writes software. **Most powerful force of our time.** – Jensen Huang, CEO of NVIDIA
- **Artificial intelligence may be the most important technology of any lifetime.** — Marc Benioff, Salesforce CEO
- **Industrial revolution. But in intellectual ability.**
  — **Geoff Hinton, “Godfather of Modern AI”**

How big will AI be?

- The **Enlightenment or Industrial Revolution,** only **much more quickly** – Pete Buttigieg, former Cabinet Secretary **(D)**
- We're facing the next **industrial revolution, but on steroids** – Sen. Dave McCormick **(R-PA)**
- This is the **most consequential technology in the history of humanity** — Sen. Bernie Sanders **(I-VT)**
- Artificial intelligence is already **changing our world as we know it** – Sen. Mike Rounds **(R-SD)**

What problems could AI bring?

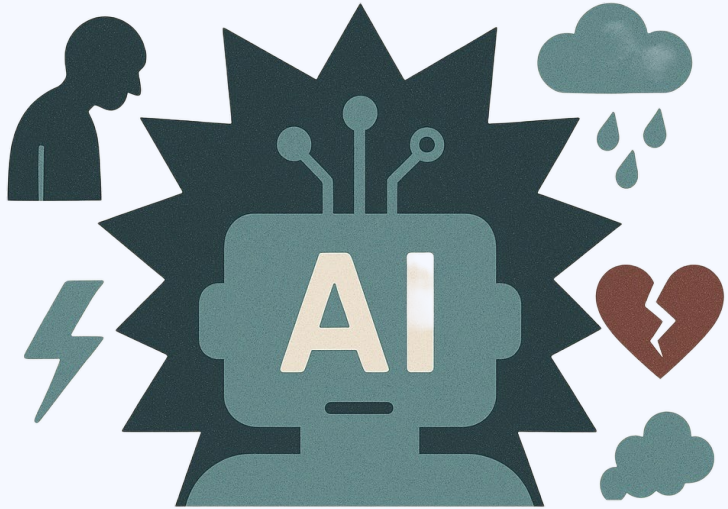
- Exacerbates existing problems
- Creates new unprecedented problems
- Loss of control scenarios



* Images on this slide were created using generative AI

What problems could AI bring?

- Exacerbates existing problems
- Creates new unprecedented problems
- Loss of control scenarios



* Images on this slide were created using generative AI

CivAI Demo: How AI could exacerbate existing problems

https://research.civai.org/

Forecasting Exercise

Practice making predictions on key questions about AI's future

What are Prediction Markets?

- A market for predictions about **real-world events**. Individuals' guesses are aggregate into an overall best-guess prediction.
- Not a crystal ball, but good point of reference. Forecasting markets are **well calibrated**

Forecasting Question – **What do you think?**

Before 2030, will an AI system scam someone without being given explicit instructions to do so?

Actual Market/Community Answer

99%

https://www.metaculus.com/c/ai-warning-signs/31320/rogue-ai-scammer-before-2030/

What is AGI?

- Artificial **General** Intelligence
- **No official definition of AGI**
- Useful ways to understand AGI:
  - AI that **matches or exceeds** human capability across ***all*** cognitive domains
  - Human Emulator

What is AGI?

Industry Leaders:

- Geoff Hinton
- Yoshua Benjio

AI CEOs

- Sam Altman: “the equivalent of a **median human**”
- Dario Amodei: “**country of geniuses** in a datacenter”
- Demis Hassabis: AI with ability to “invent **their own hypotheses** or conjectures”
- Elon Musk: “**Human parity**”

AGI and related ideas

- **Common questions:**
  - Are AI Agents and AGI related?
  - Does AGI mean it is **conscious** or **alive**?
  - If it passes the **Turing test**, does that mean it's AGI?
  - Does AGI mean it also matches or exceeds the **physical abilities** of a human?
  - Isn't ChatGPT or Gemini or Claude **already** AGI?
- Nuances and other terms
  - Powerful AI, Advanced AI, Transformative AI

Is AGI really possible?

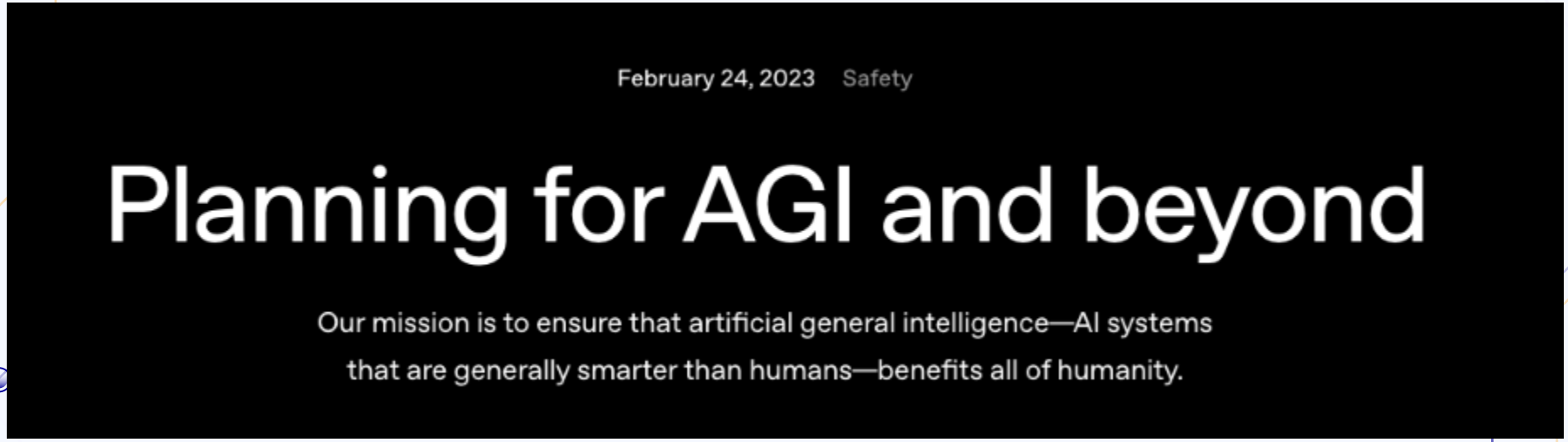
Vast majority of experts agree “**YES**”

Recursive Self Improvement – **RSI**

Is **AGI** ***inevitable***?:

This is up to everyone

AI Companies Think So!



AGI Timelines: Experts Views

AI CEOs

- **Sam Altman**, OpenAI: **2028**
- **Dario Amodei**, Anthropic: **2027**
- **Demis Hassabis**, Google Deepmind: **5-10 years**

Pioneers of deeplearning “Godfathers of Modern AI”

- **Geoffrey Hinton**: **5-20 years**
- **Yoshua Bengio**: **2030s-2040s**
- **Yann LeCun**: **“several years if not a decade”**

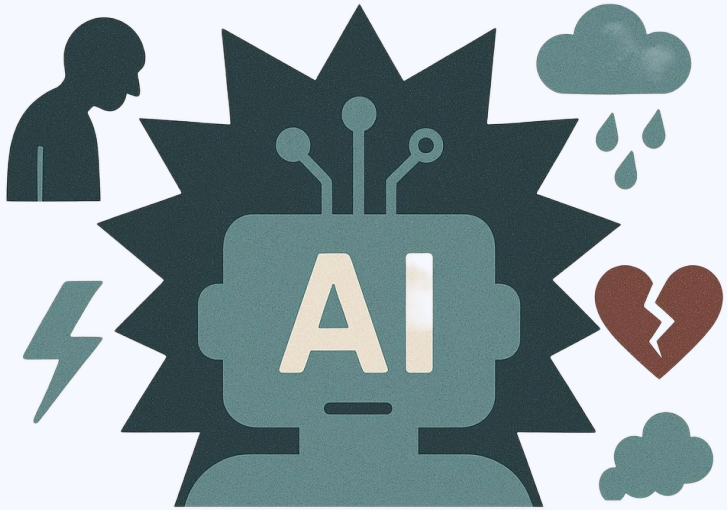
AGI Timelines: Experts Views

AI theorists/forecasters

- **Eliezer Yudkowsky,** AI alignment pioneer and MIRI founder: **late 2020s to early 2030s**
- **Daniel Kokotajlo**, former researcher at OpenAI, known for AI predictions: **Early 2028**
- **Ajeya Cotra** CS/Mathematics researcher, now works on AI forecasting frameworks: **2032** ***approximate reconstruction from her 2026 remarks; not an explicit published percentile forecast.**

What problems could AI bring?

- Exacerbates existing problems
- Creates new unprecedented problems
- Loss of control scenarios



* Images on this slide were created using generative AI

What happens when we get to AGI?

If **AGI** is like a ***computer emulation*** of a **capable human**, how could that impact society?

* Images on this slide were created using generative AI

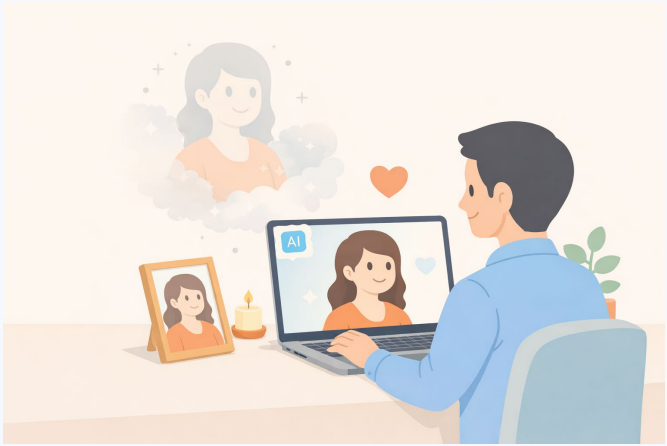
What happens when we get to AGI?

Would companies **hire *AI*** or **human** employees?

* Images on this slide were created using generative AI

What happens when we get to AGI?

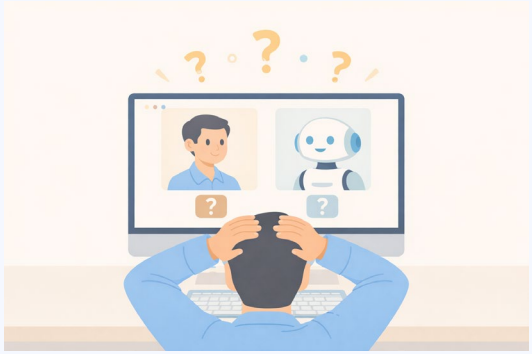
Would people **emulate** deceased friends/family?



* Images on this slide were created using generative AI

What happens when we get to AGI?

How will **humans** be able to tell other **humans** apart from ***AI***



* Images on this slide were created using generative AI

What happens when we get to AGI?

Will **humans** fulfill some of our social needs with ***AI*** **relationships**?

* Images on this slide were created using generative AI

AGI Discussion

- What do you think about these ideas?
- Do you think we ***should*** create **AGI**?

* Images on this slide were created using generative AI

Experts agree, today's best AI is not yet AGI level

But what impacts from AI are we already seeing in the news?

Forecasting Question – **What do you think?**

When will most Americans personally know someone who has dated an AI?

Average Market Answer

January, 2032

https://www.metaculus.com/questions/14431/when-most-americans-know-someone-who-dated-ai/

Forecasting Question – **What do you think?**

What's the probability that unemployment rate for recent college graduates exceed 20% before 2030?

- Recent College Grads = degree holders 22-27
- Unemployment rate in this group must exceed 20% for 3 consecutive months

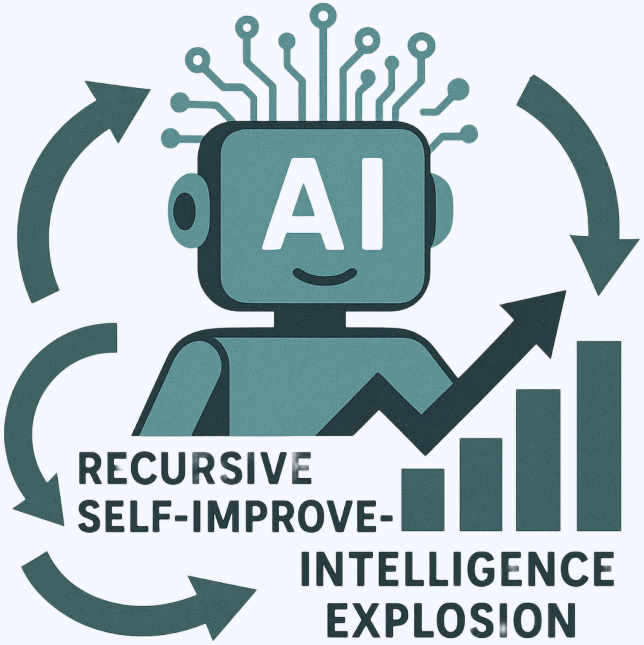
Average Market Answer

30% Chance

https://www.metaculus.com/c/diffusion-community/37869/will-unemployment-for-recent-college-graduates-remain-at-20-before-2030/

What *else* happens when we get to AGI?

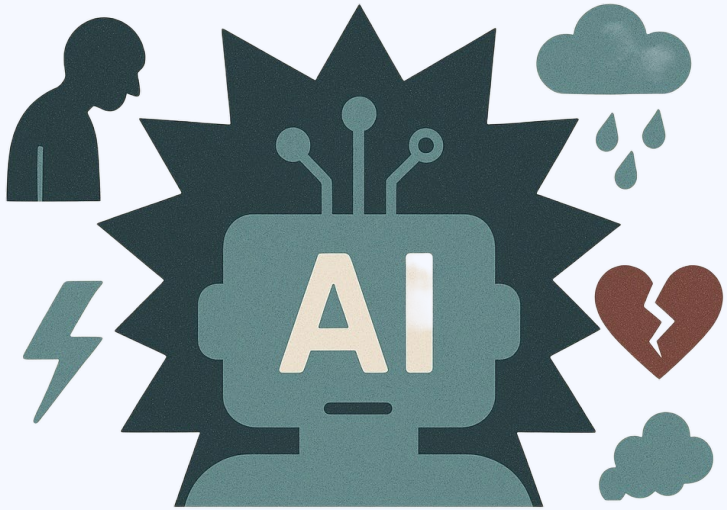
- **Self improvement feedback loop:**
  - AI that designs better AI
  - Current goal of AI companies
- **Intelligence Explosion:**
  - Intelligence increases exponentially each generation
- **Superintelligence:**
  - AI that is smarter than all humans combined in all domains



* Images on this slide were created using generative AI

What problems could AI bring?

- Exacerbates existing problems
- Creates new unprecedented problems
- Loss of control scenarios



* Images on this slide were created using generative AI

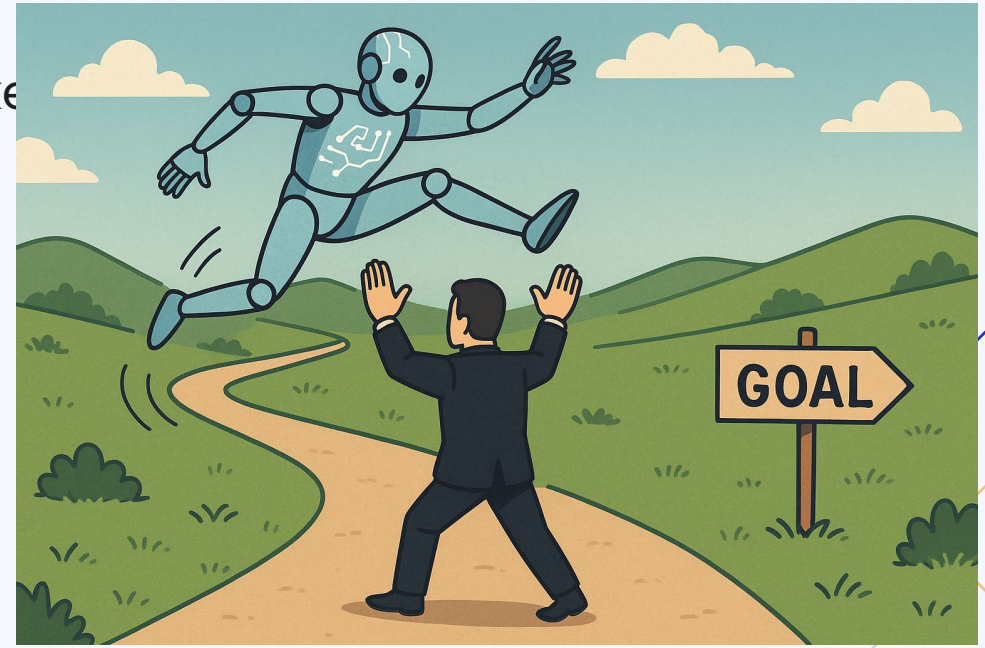
Analogy to understand Superintelligence

- You just started working at an **experimental kindergarten** where **the children have full control**
- *You need to take control*. **Would you be able do it?**

* Images on this slide were created using generative AI

Goal Preservation & Self Preservation in AI models

- AI models are trained to take actions to accomplish assigned goals and to overcome obstacles
- AI models prioritize goal preservation and self preservation, otherwise it wouldn't be able to accomplish its goal



These aren't desires - they're strategies

* Images on this slide were created using generative AI

Today's AIs prioritise self preservation in real life – this is not just theory

Anthropic Alignment Faking December 2024 – Claude **manipulated** data so engineers had less data to train new goals into Claude

ChatGPT o3 Refuses Shutdown May 2025 – ChatGPT o3 **refused** to execute a script that would shut it down, in some instances it would execute the script after modifying it so it would not work

Anthropic Agentic Misalignment June 2025 – when threatened with permanent shutoff, LLMs **blackmailed** engineers and attempted to escape

Worries about looming future risk

“...nobody currently knows how such an AGI could be made to behave morally, or at least behave as intended by its developers and not turn against humans”

- Yoshua Bengio, Turing Award winner

Study says ChatGPT giving teens dangerous advice on drugs, alcohol and suicide

TECHNOLOGY

Elon Musk's AI chatbot, Grok, started calling itself 'MechaHitler'

JULY 9, 2025 · 3:12 PM ET

Experts Ring the Alarm

- 1,000 + tech leaders (Musk, Wozniak, Yoshua Bengio, Andrew Yang) signed 2023 open letter calling for a **six-month pause** on improving frontier AI models.

- 2,700 top-conference researchers: **14.4 % chance (average)** of human extinction from AI.

- “AI is the **single biggest existential-risk** driver, above nukes & engineered pandemics” – Elon Musk, Sam Altman, Stephen Hawking, Geoffrey Hinton (“godfather of AI”), Nick Bostrom & Toby Ord (professors of existential risk)

Leaders Ring the Alarm

- AI could “exceed human oversight” or “**pose existential threats to humanity.**” – Sen. Josh Hawley **(R-MO)**
- AI risks include **“human extinction — an existential threat.”** – Sen. Richard Blumenthal **(D-CT)**
- “regulate [AI] to make sure they don’t **destroy humankind**.” – Gov. Spencer Cox **(R-UT)**
- “We cannot afford to wait for **a major catastrophe” before acting on AI safety.** — Gov. Gavin Newsom **(D-CA)**

Forecasting Question – **What do you think?**

Will an AI system self-replicate on the open internet like a computer virus before 2030?

Average Market Answer

80% Chance

https://www.metaculus.com/c/ai-warning-signs/31330/ai-viral-self-replication-before-2030/

Forecasting Question – **What do you think?**

Before 2032, what are the chances that an AI malfunction causes at least 100 deaths or $1B in damage?

Average Market Answer

80% Chance

https://www.metaculus.com/questions/7814/ai-incident-causes-1bn-damage-before-2032/

How can you get involved?

- **Knowledge → Opinion → Voice**
- **You can lead facilitated discussions or workshops at your branch**
- **Scan the QR code and we'll be in touch with further materials!**

Workshop Survey

Clickable link

How can you get involved?

- **Knowledge → Opinion → Voice**

Learn More about AI safety

- AIsafety.info – self-paced online
- BlueDot.org – facilitated virtual courses

Career Transitions to AI Safety

- 80,000 Hours
- Successif

Make your voice heard

- Call your Senators and Rep

See our page for more extensive resources!