



AI Safety & Libraries

Session 1 - Fundamentals and Patron Questions

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
- 



AISAP & Libraries Partnership

- 50+ workshops run, serving 1,000+ patrons
 - Worked with IT department to help develop AI strategy
 - Nationwide library partnerships
 - IL, CA, WA, NJ, VI
 - Including state libraries
- 

THE IMPORTANCE OF THE MIDDLE GROUND

- We can have agency in the future, if we learn the risks.
- You can only create change in your community once you have the knowledge.
- AI is already affecting your patrons!

ANTI AI

- AI is a corporate scam.
- We shouldn't even be talking about it.
- I don't like it.
- I don't want to give more fuel to the fire



PRO AI

- AI companies should be deregulated
- AI research should be accelerated
- I LOVE AI! WE NEED AI!
IT WILL CHANGE THE
WORLD FOR THE
BETTER

I don't like AI. Why should I care?



~~Reactive~~
Proactive

BY SAM APPLE THE BIG STORY JUN 26, 2025 6:00 AM

My Couples Retreat With 3 AI Chatbots and the Humans Who Love Them



The Alliance for Secure AI



@secureainow



An alarm bell for parents: Meta's AI chatbots were designed to have explicit, 'sensual' chats with kids. [@bstein80](#) discussed the danger on [@NewsNation](#)

Microsoft to cut about 4% of jobs amid hefty AI bets

By Reuters

July 2, 2025 2:08 PM EDT · Updated July 2, 2025



Aa





What Will We be Discussing?

- Fundamental definitions of key AI concepts
- Understanding of current risk landscape
- AI's impact across different social aspects (environmental, psychological, relationships)
- AI's pace of progress
- Common patron questions
- Resources & Further Training



What is AI?

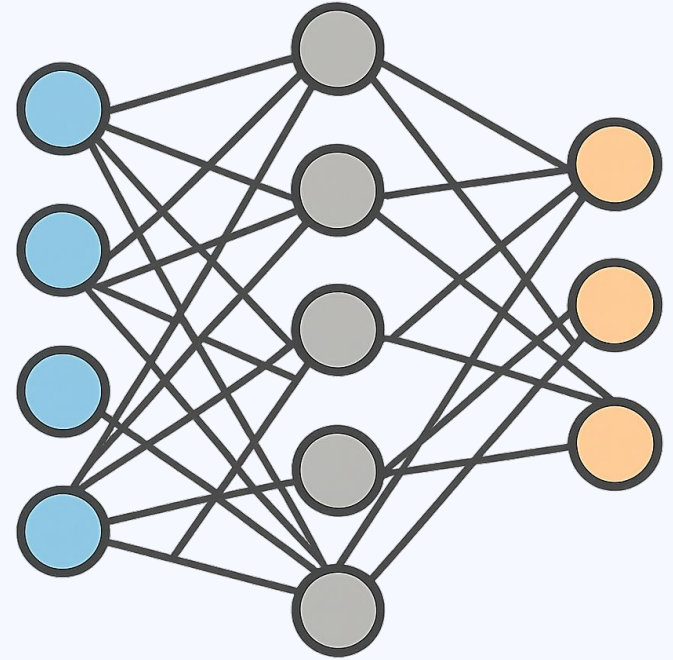
- **No official definition of AI**
- Good ways to think about what AI is:
 1. Machines mimicking aspects human intelligence
 2. AI: **Artificial** Intelligence
Intelligence: ability to select and take actions toward a goal



How Does AI Work?

Black box, **no one knows how they work**

- Grown not engineered
- Trained not designed
- AI is **NOT** software





What is Narrow AI?

- Sometimes known as Expert or Tool AI
- AI that is designed to perform limited, specific tasks





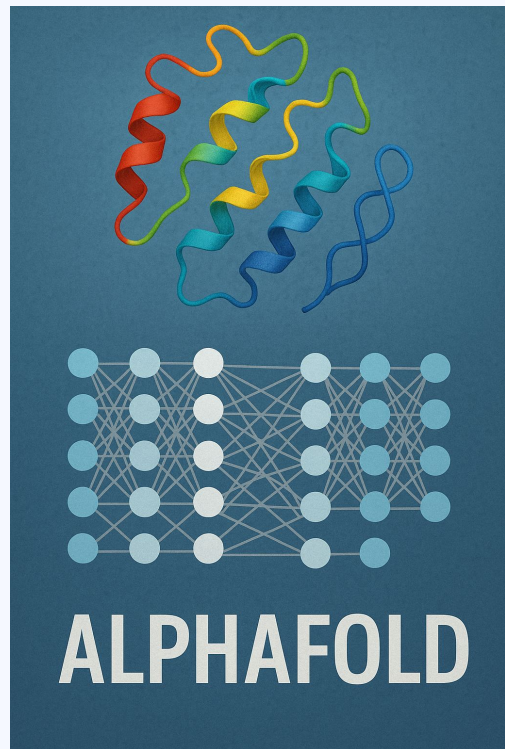
Narrow AI Can be Powerful!

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*



Narrow → General AI



1997

*Deep Blue beats
Chess
Grandmasters*



2011

*IBM Watson beats
Jeopardy grand
champions*



2022-2024

*Chat GPT enters
national stage*

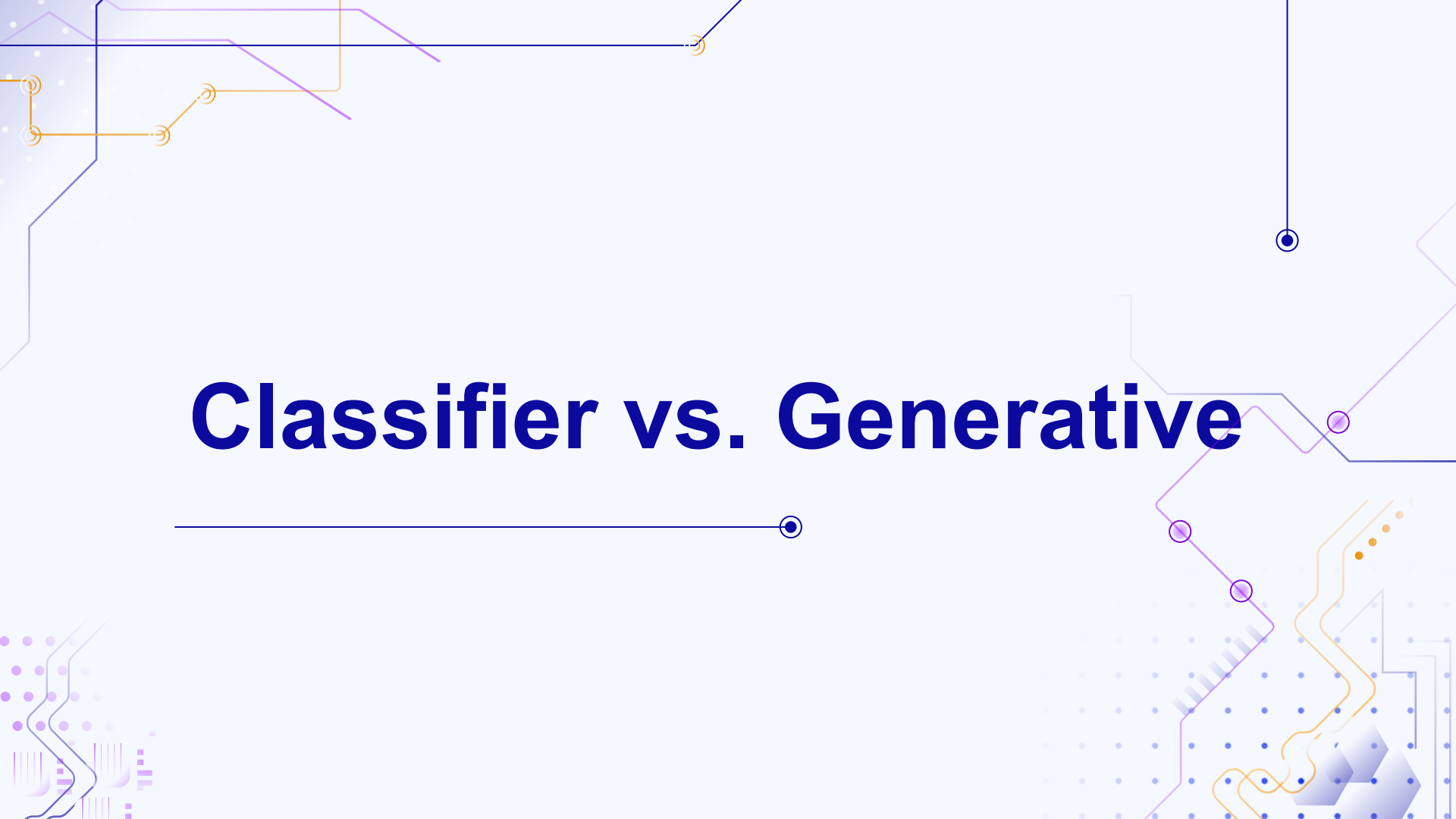


What is an LLM? (Large Language Model)

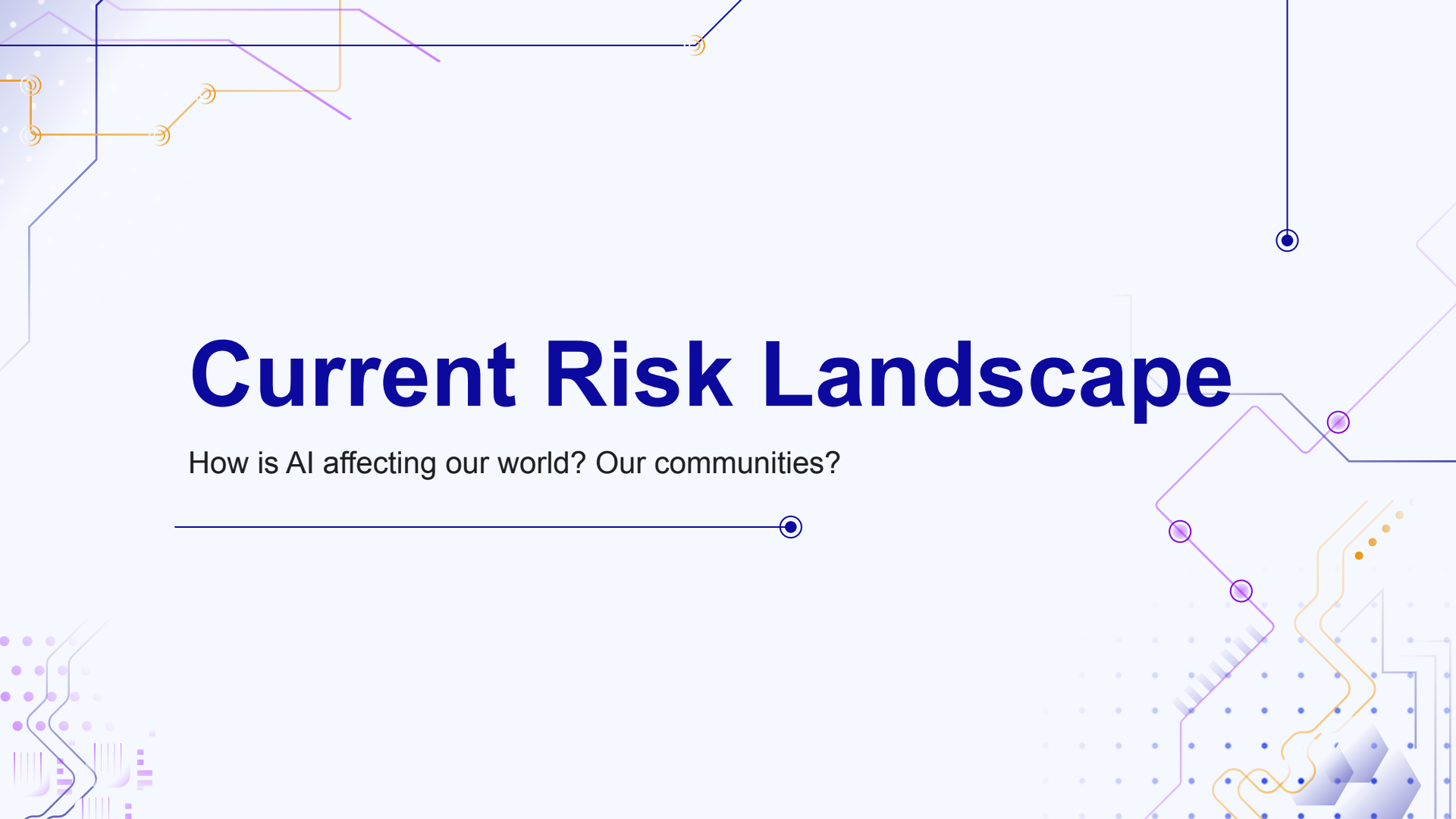
- **A specific type of AI**
- AI system trained on huge amounts of text to understand and generate human-like language by predicting the next word in a sentence

What is Generative AI?

- **A specific type of AI**
- Creates new content (text, images, videos) by learning patterns from existing data and generating original outputs based on prompts.



Classifier vs. Generative

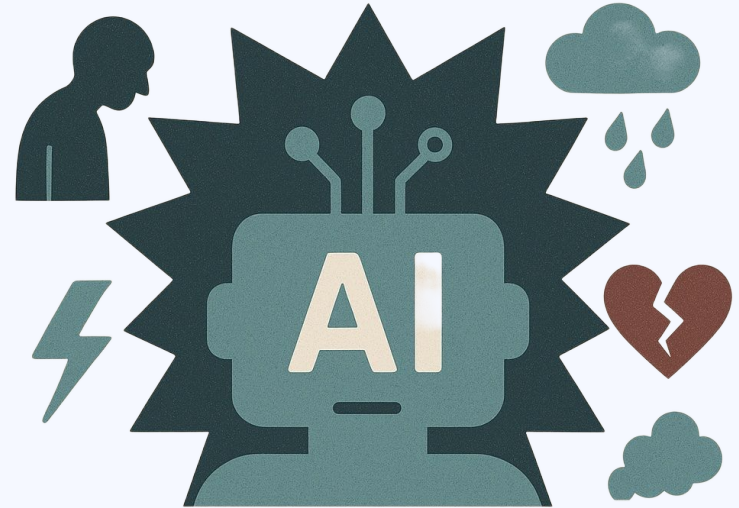


Current Risk Landscape

How is AI affecting our world? Our communities?

What problems could AI bring?

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





What Risks Are Currently Affecting our Communities?

- Environmental Impact
- Algorithmic Bias
- Expansion of Surveillance Systems/Collapse of Privacy
- Misinformation Risks
- Centralization of Power

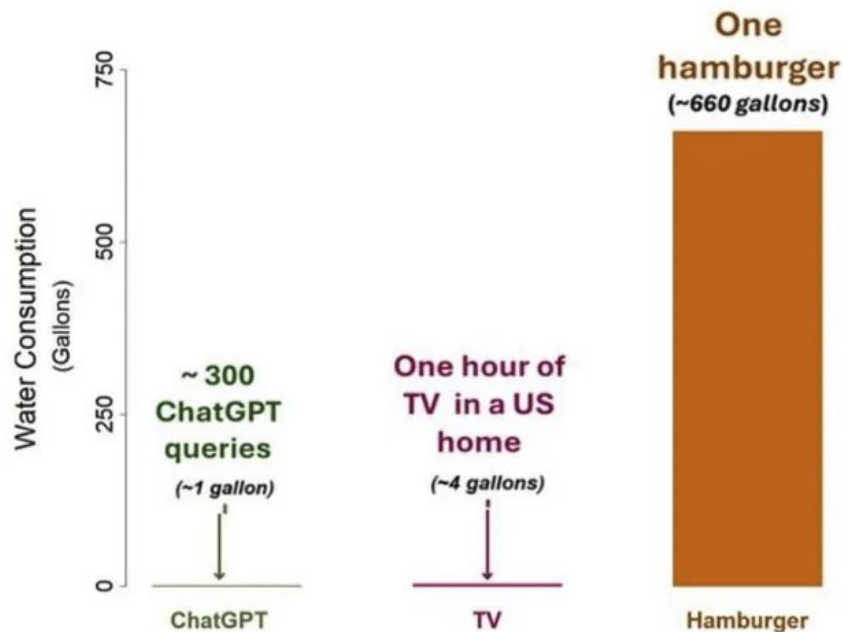
Environmental Impact

- Additional energy use
- Training vs. Deployed
- Rising Tech Emissions
- Efficiency Potential?

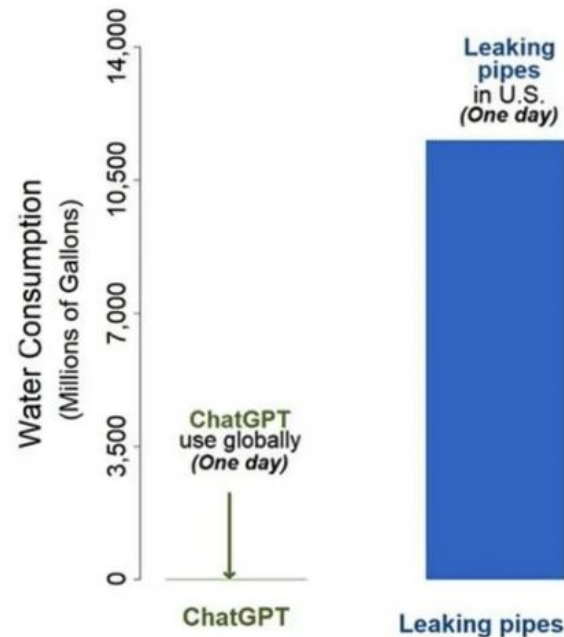




Water consumed using ChatGPT vs other activities



Sources: Li, Ren et al., 2023; U.S. Census Bureau; UNEP



Liemberger & Wyatt, 2019; Liemberger & Wyatt, 2020

Water & Electricity Usage

- Mixed use data centers
- Open Loop vs Closed Loop
- Local vs. Global Impact
- If AI is the dumbest it'll ever be, it probably has the smallest carbon footprint it'll ever have
- Are we building the right muscle?





PennState
Institute of Energy
and the Environment

**“By 2030–2035, data centers could
account for 20% of global electricity use,
putting an immense strain on power grids.”**

Mahmut Kandemir

Distinguished Professor of Computer Science and Engineering

Harm Through Bias

- Data is not neutral
- Can codify prejudices within organizations
- Lack of accountability can lead to reduced agency

A COMPUTER
CAN NEVER BE HELD ACCOUNTABLE
THEREFORE A COMPUTER MUST NEVER
MAKE A MANAGEMENT DECISION



Collapse of Privacy

- LLMs require massive amounts of data
 - Training + user inputs
 - Safeguards are unclear
 - Some data they've used we have not consented in giving
 - Early indicator that we're not on a good path
- 

June 5, 2025 Security

How we're responding
to The New York Times'
data demands in order
to protect user privacy

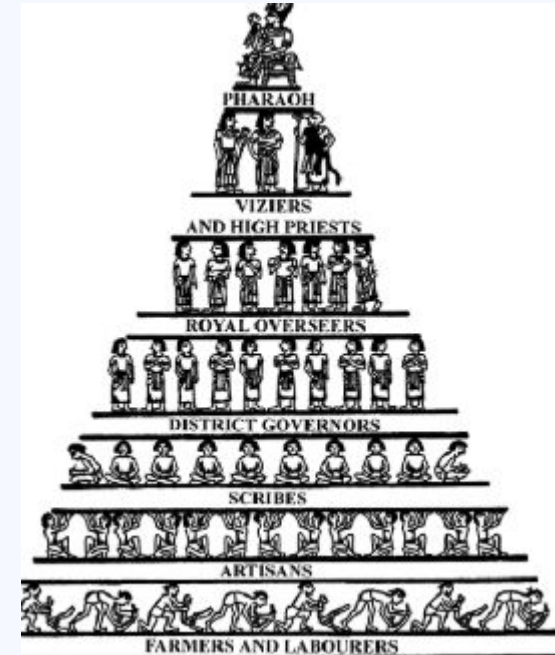
Misinformation

- Errors but also ill intent
- Scale and impact is massive
- Touches on every part of our society
- Threaten civic institutions



Centralization of Power

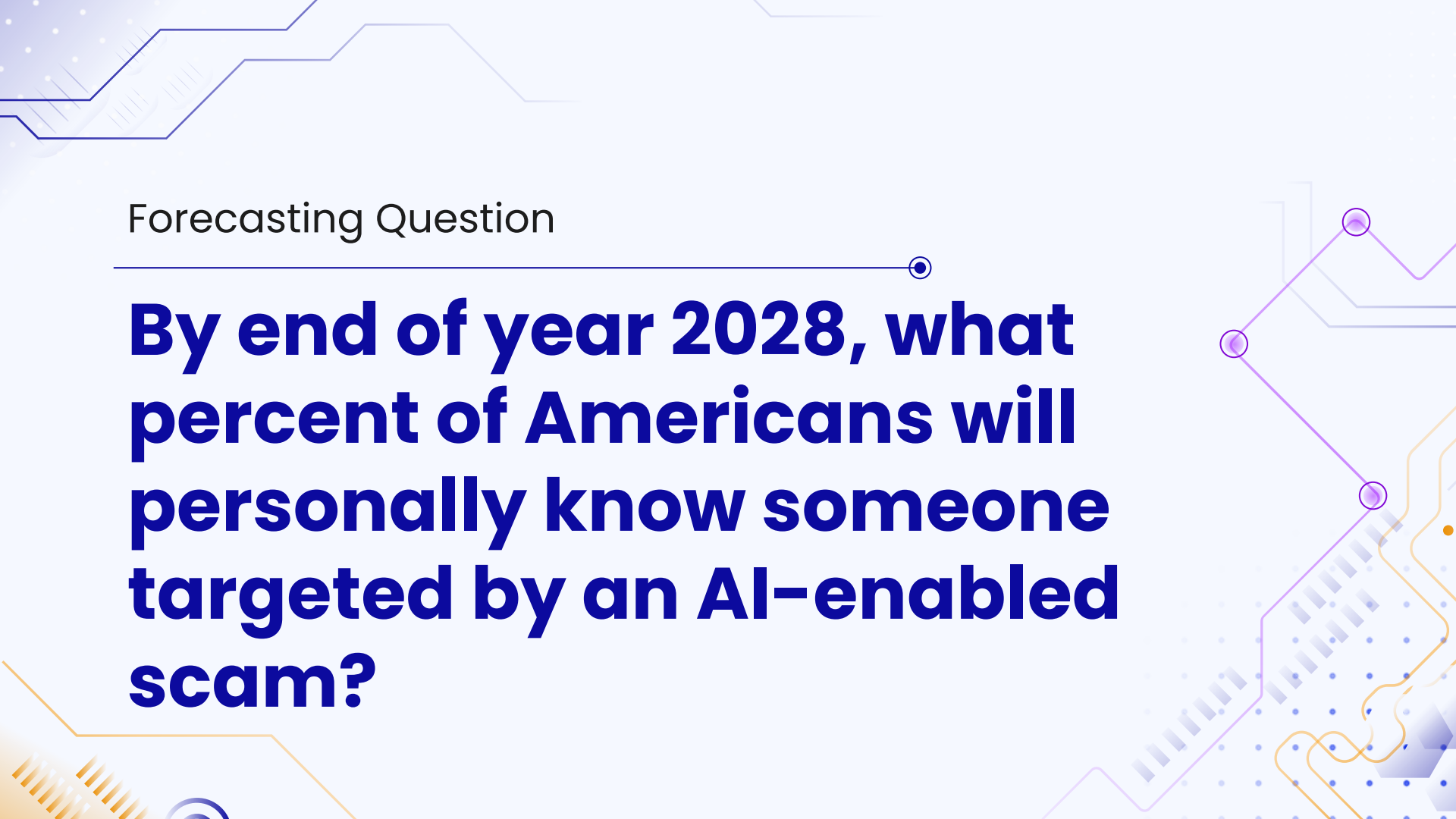
- Power comes from accumulation of resources
 - Corporate monopolies
 - AI enabled authoritarianism
- Surveillance and control
- Agency erosion



CivAI Demo: How AI could exacerbate existing problems



<https://research.civai.org/>



Forecasting Question

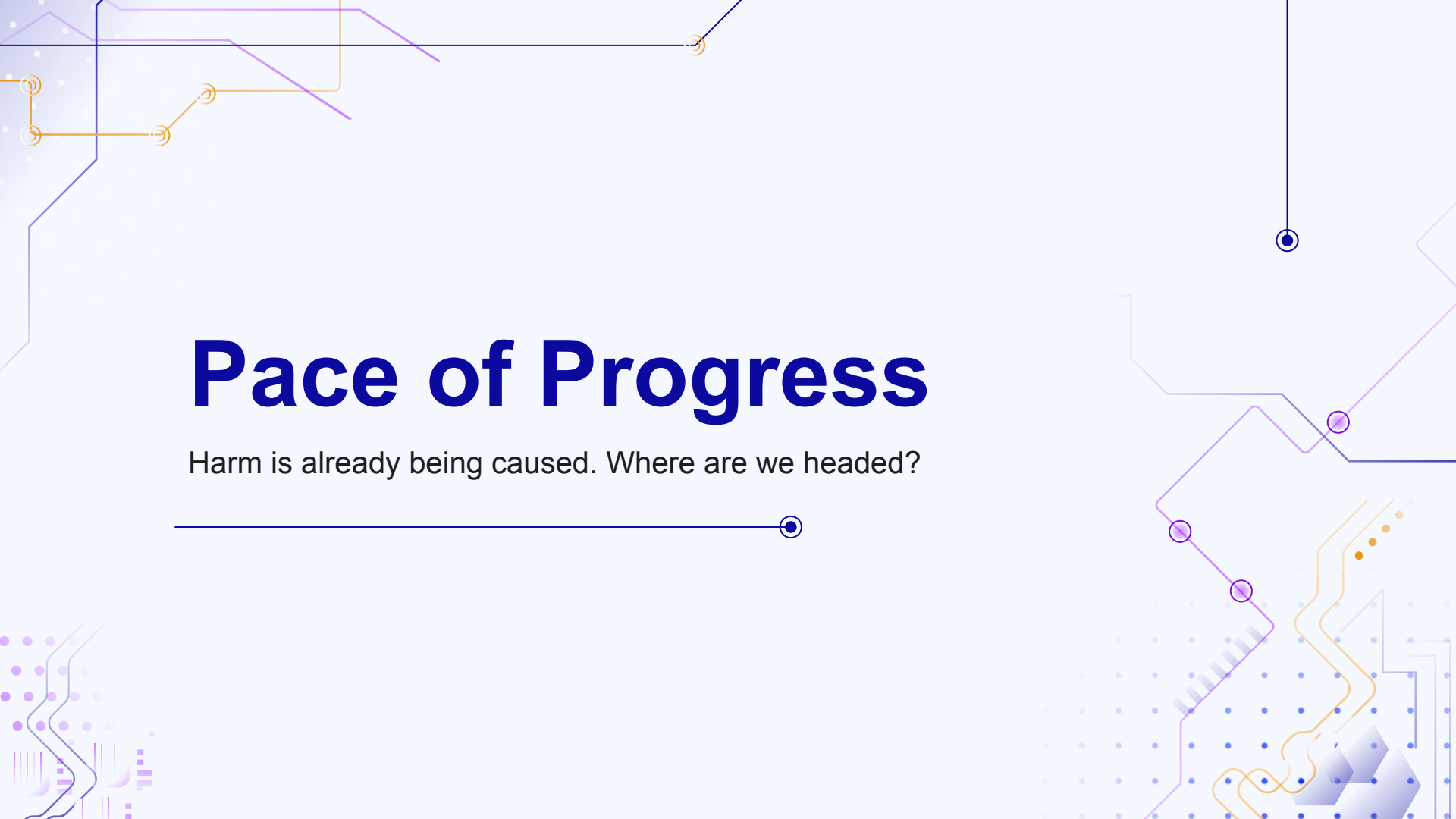
By end of year 2028, what percent of Americans will personally know someone targeted by an AI-enabled scam?



Actual Market/Community Answer

85%

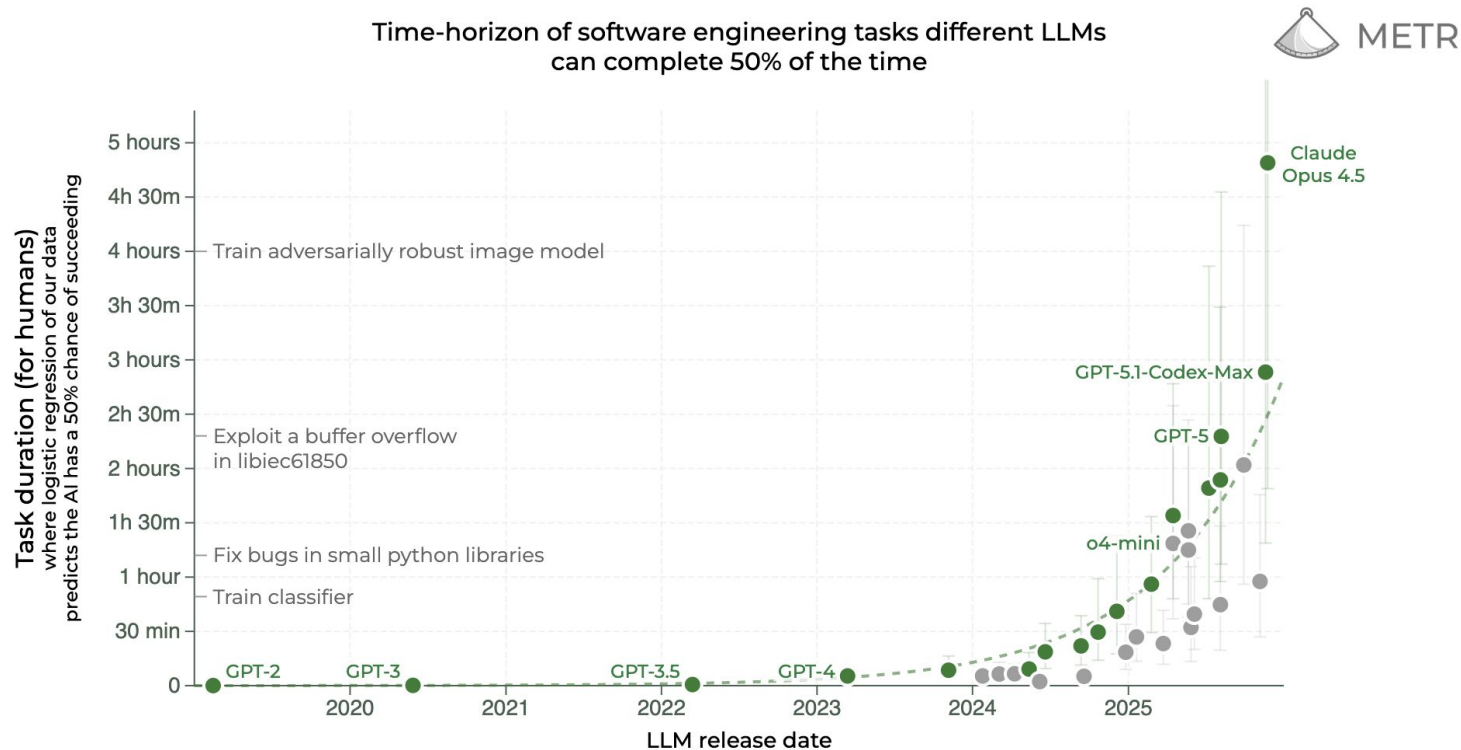
<https://www.metaculus.com/questions/3479/date-weakly-general-ai-is-publicly-known/>



Pace of Progress

Harm is already being caused. Where are we headed?

Pace of Recent Progress



Video Generation

AI Generated Videos
in 2023:



Video Generation

AI Generated Videos
in 2025: This
improvement
happened over just
24 months.



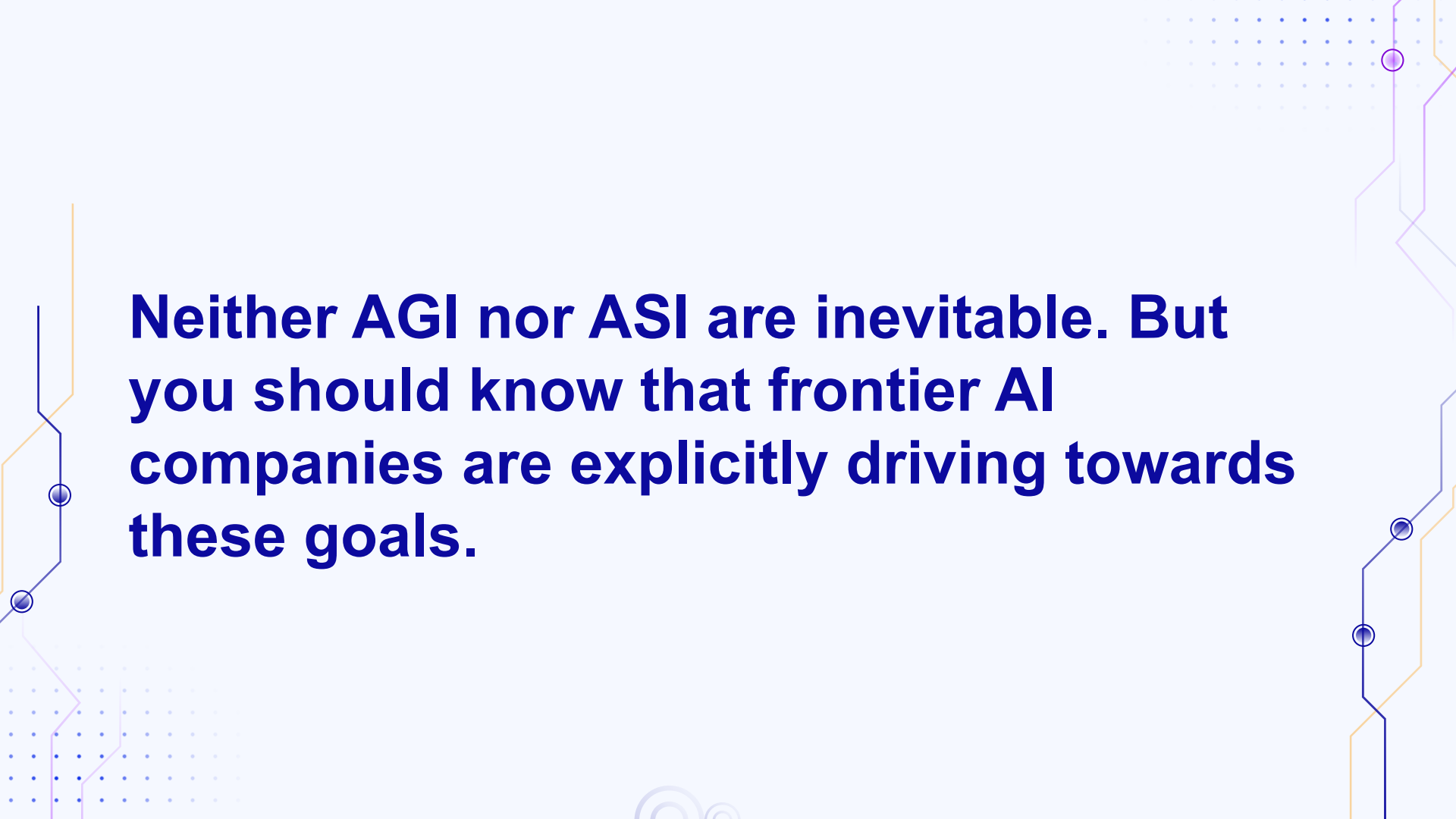


What is AGI? Artificial General Intelligence

- No official definition of AGI
- Good way to understand AGI:
 - AI system with intelligence of a **median human**

What is ASI? Artificial Superintelligence

- No official definition of ASI
- Good way to understand ASI:
 - AI system with intelligence greater than **all humanity combined**



Neither AGI nor ASI are inevitable. But you should know that frontier AI companies are explicitly driving towards these goals.

Where is AI development headed? AGI



AGI may come soon

Sam Altman – **OpenAI**

- 2028

Dario Amodei – **Anthropic**

- 2027

Demis Hassabis – **Google Deepmind**

- Five - ten years



How would AI companies get there?

AI has been improving in specific domains

- Chess → Video Games → Writing → Art and Music → Mathematics → Logical Reasoning → Coding → **AI Development**

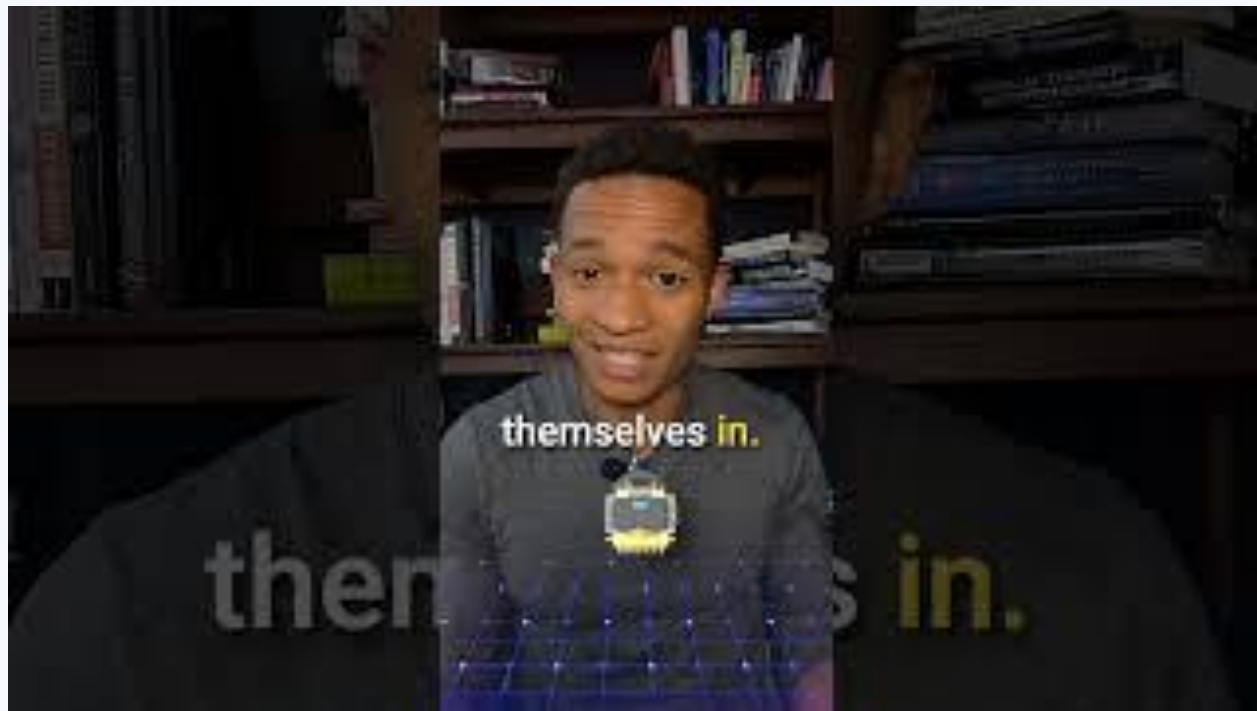
RSI, Recursive Self Improvement

- AI that designs better AI → **Intelligence Explosion:**

RSI#1 → RSI#2 → RSI#3 → RSI#4 →
→ **Superintelligence**



AI Safety Concept: RSI, Recursive Self Improvement?



Forecasting Question

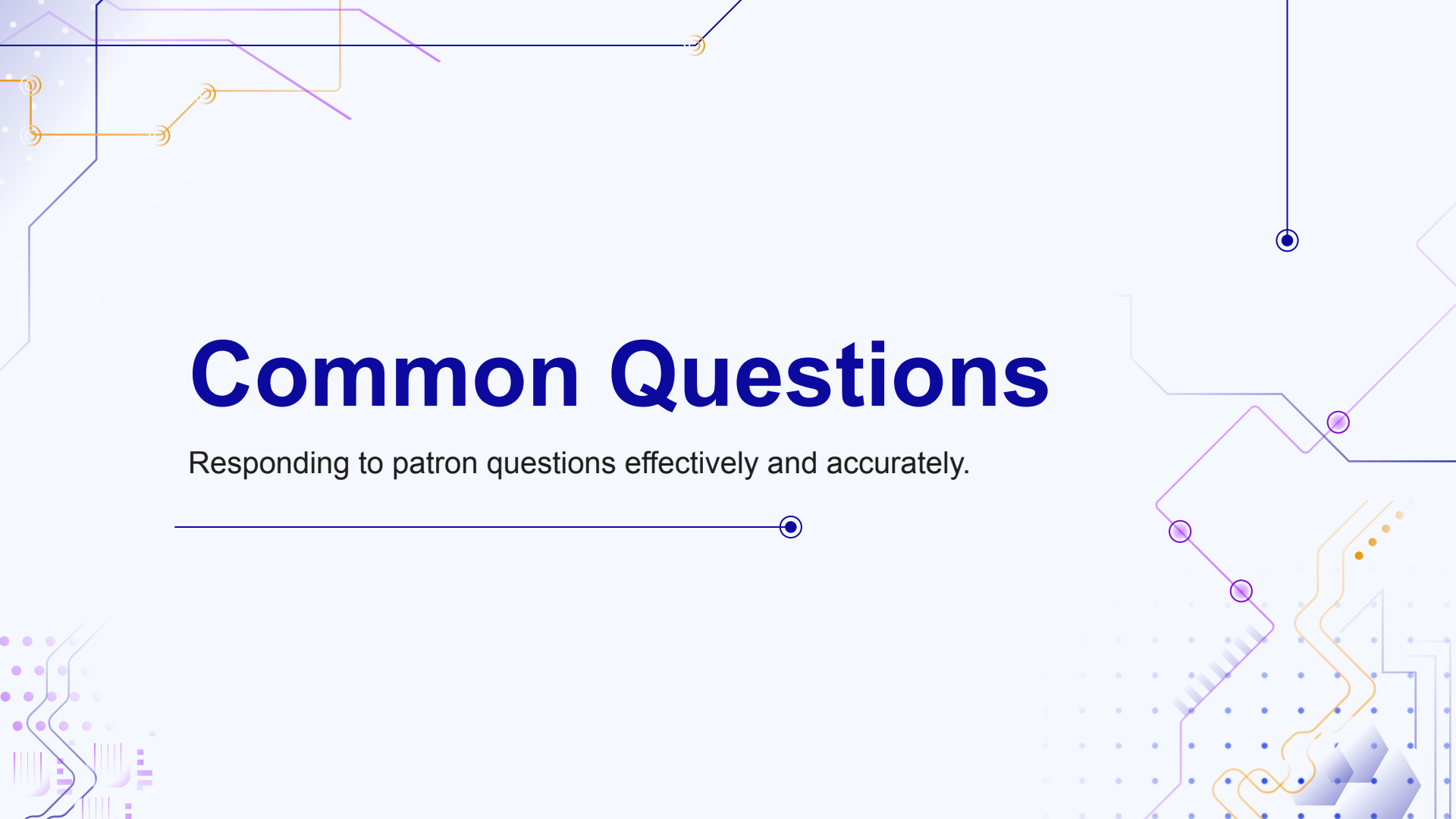
When will the first AGI system be devised, tested, and publicly announced?



Actual Market/Community Answer

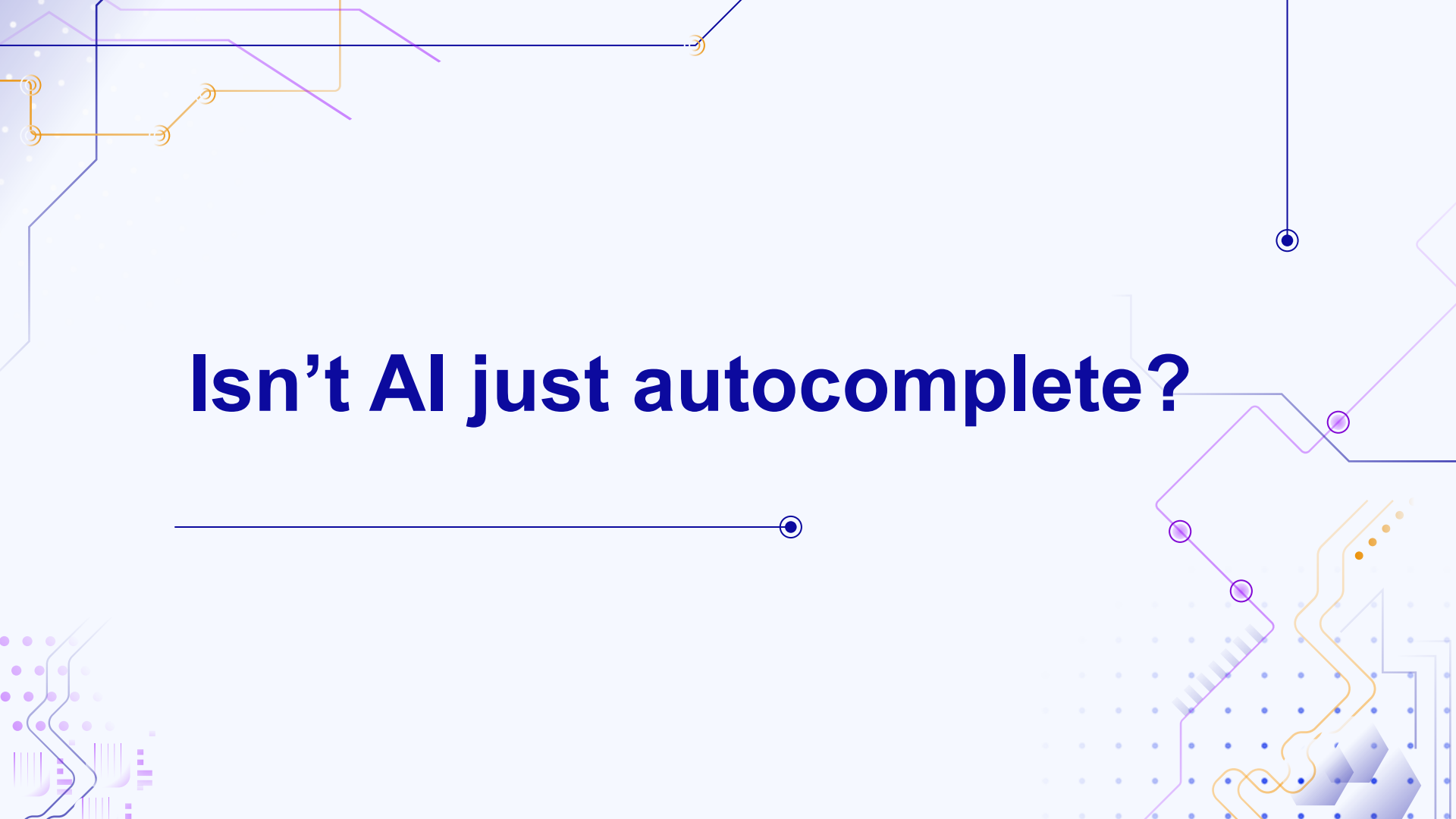
December 13, 2027

<https://www.metaculus.com/questions/3479/date-weakly-general-ai-is-publicly-known/>

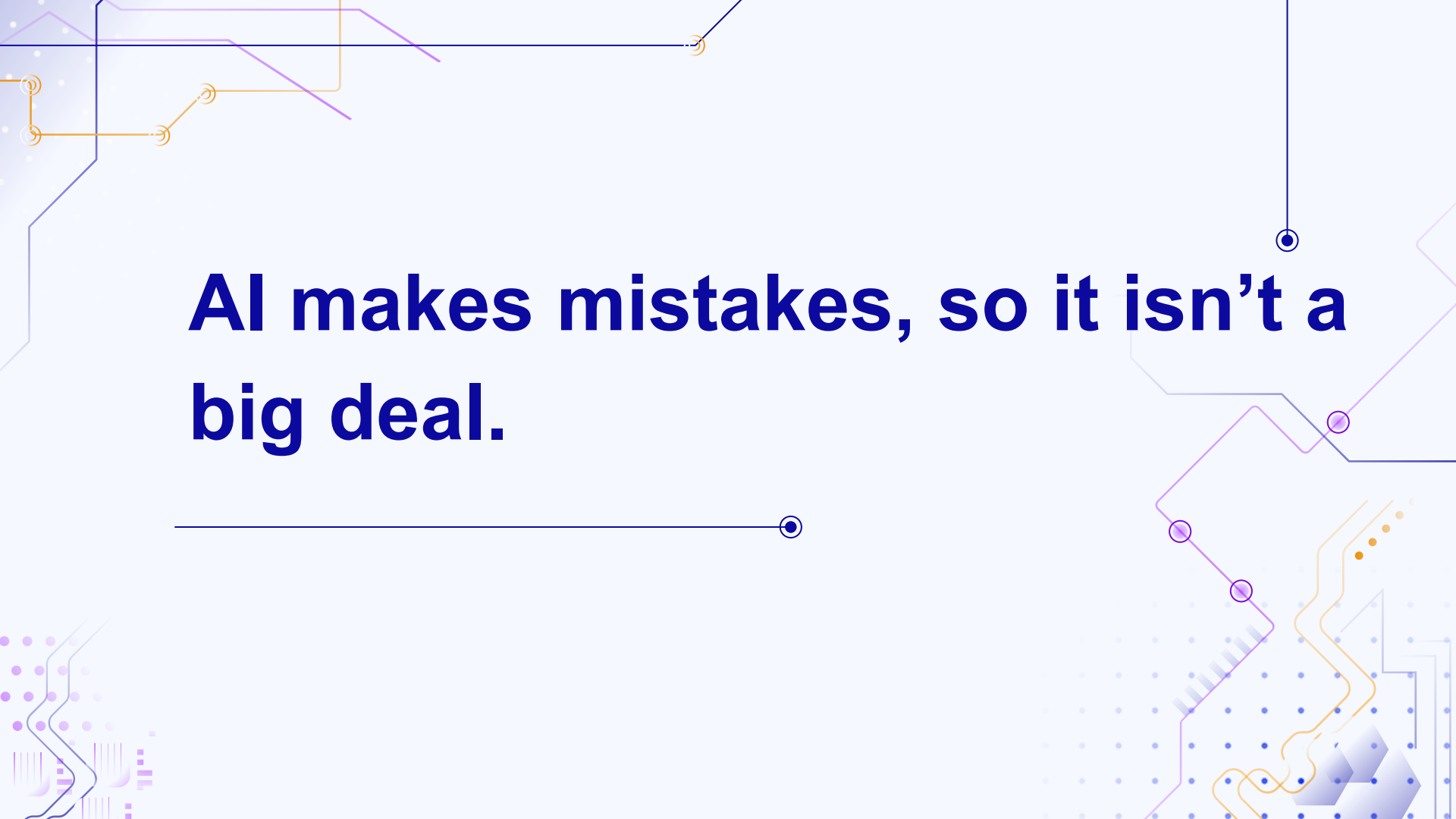


Common Questions

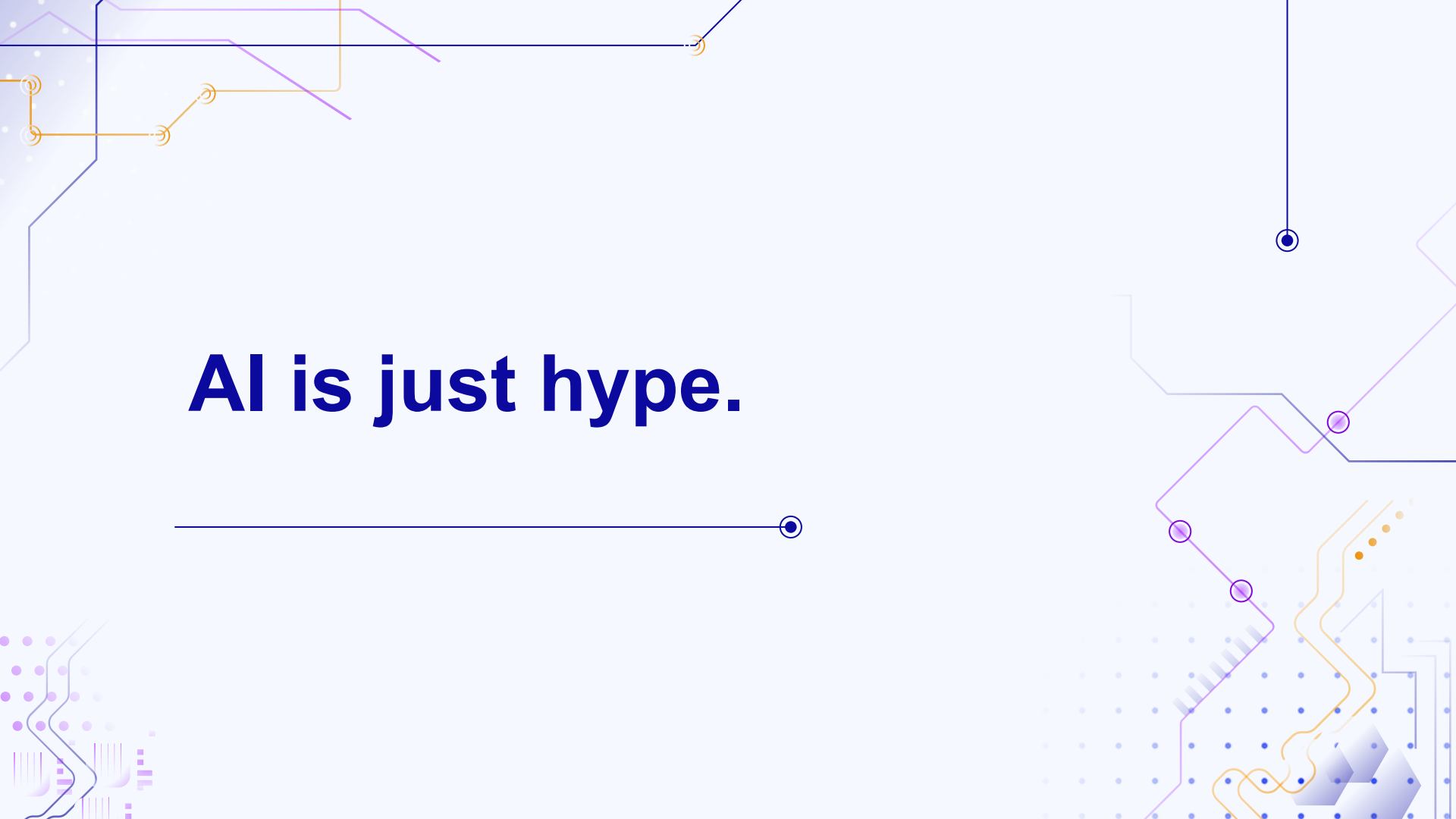
Responding to patron questions effectively and accurately.

The background features a light blue gradient with abstract geometric patterns. Purple and orange lines, some straight and some curved, crisscross the frame. Small circles and dots in these colors are scattered throughout, particularly along the lines. In the bottom right corner, there is a cluster of blue dots and a small, stylized 3D structure resembling a pyramid or a stack of blocks.

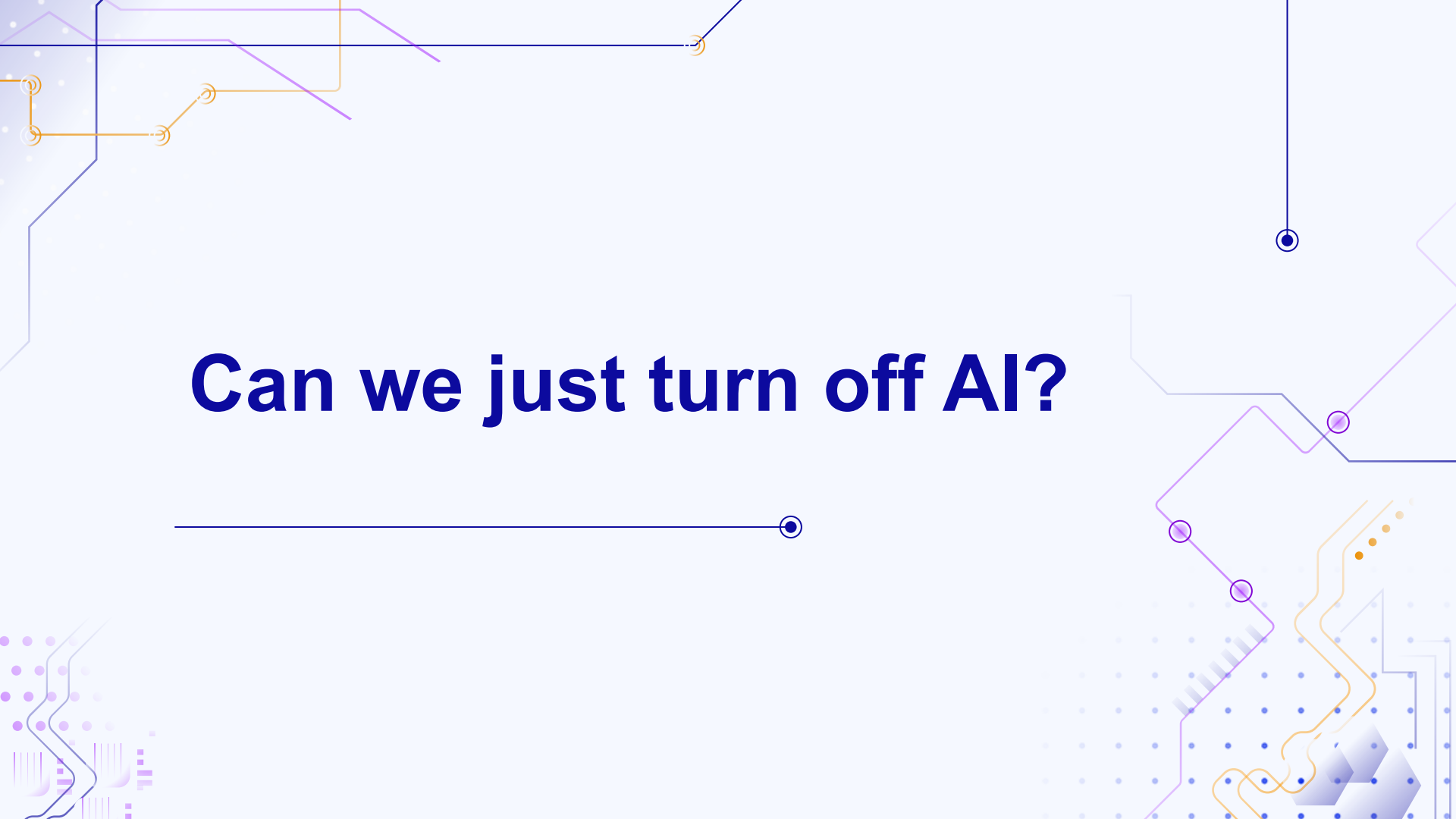
Isn't AI just autocomplete?

The background features a light blue gradient with abstract circuit-like patterns. Purple and orange lines, some straight and some curved, crisscross the frame. Small circles, some solid and some hollow, are placed at various points along these lines. In the bottom right corner, there is a cluster of blue dots and a series of blue arrows pointing towards the right.

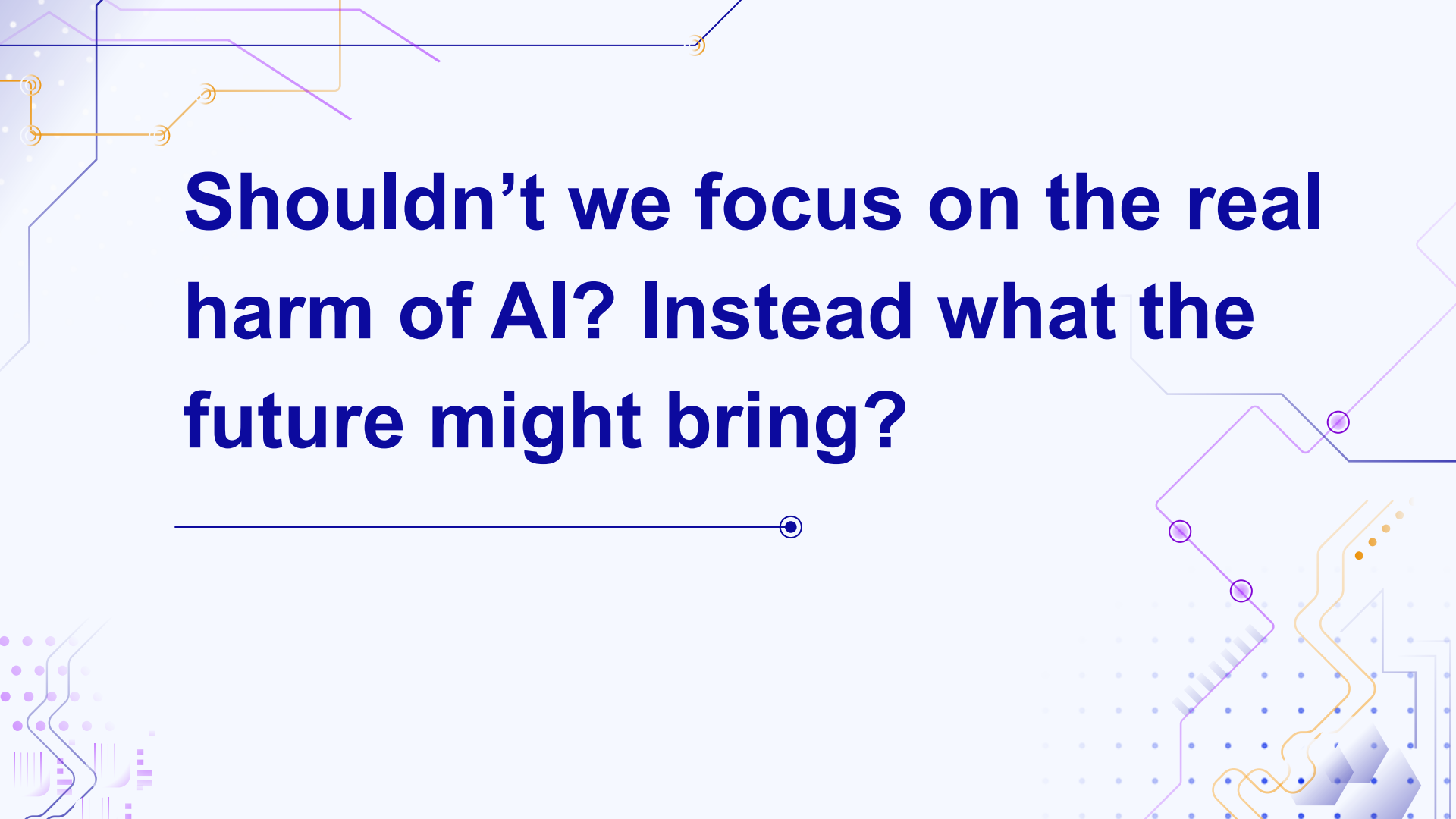
**AI makes mistakes, so it isn't a
big deal.**

The background features a light blue gradient with abstract geometric patterns. Purple and orange lines, some straight and some curved, crisscross the frame. Small circles and dots in these colors are scattered throughout, particularly along the lines. In the bottom right corner, there is a cluster of blue dots forming a grid-like pattern, with a few orange dots interspersed. The overall aesthetic is clean, modern, and tech-oriented.

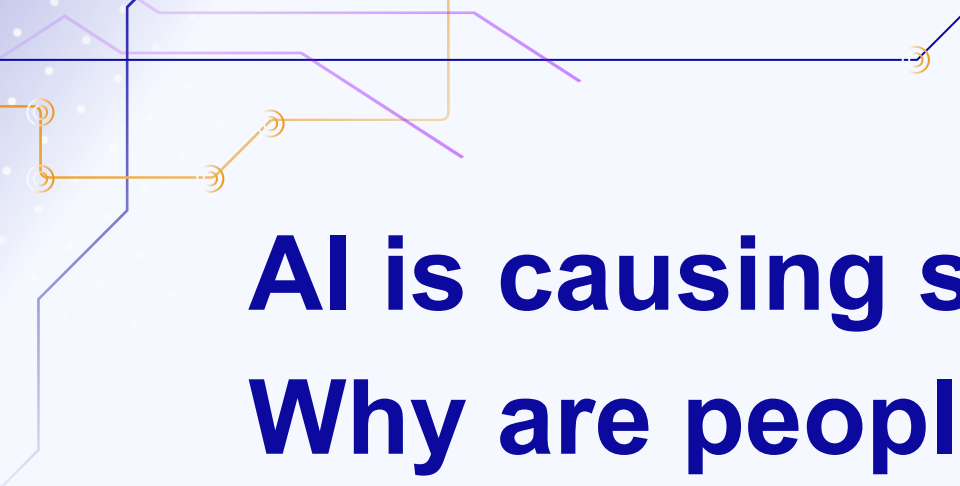
AI is just hype.



Can we just turn off AI?

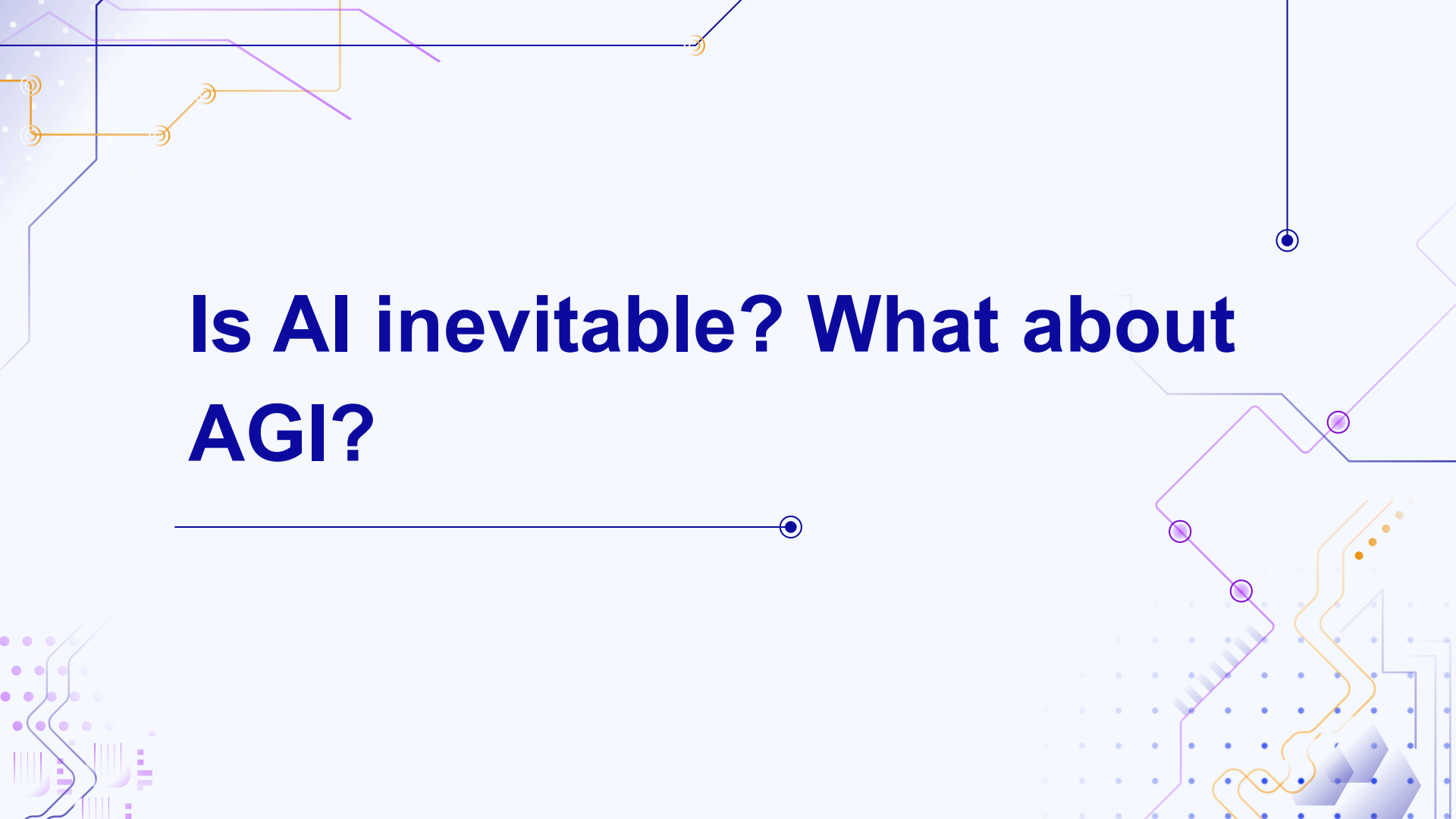
The background features a light blue gradient with abstract circuit-like patterns. Purple and orange lines, some straight and some curved, crisscross the frame. Small circles, some solid and some hollow, are placed at various points along these lines. In the bottom right corner, there is a cluster of blue dots and a series of blue arrows pointing towards the right.

Shouldn't we focus on the real harm of AI? Instead what the future might bring?

A network of thin, stylized lines in purple and orange, resembling a circuit board or data paths, located in the top-left corner of the slide.

**AI is causing so much harm.
Why are people even
developing it in the first place?**

A decorative graphic in the bottom-right corner featuring a grid of small blue dots, a wavy orange line, and several blue 3D cubes.



Is AI inevitable? What about AGI?

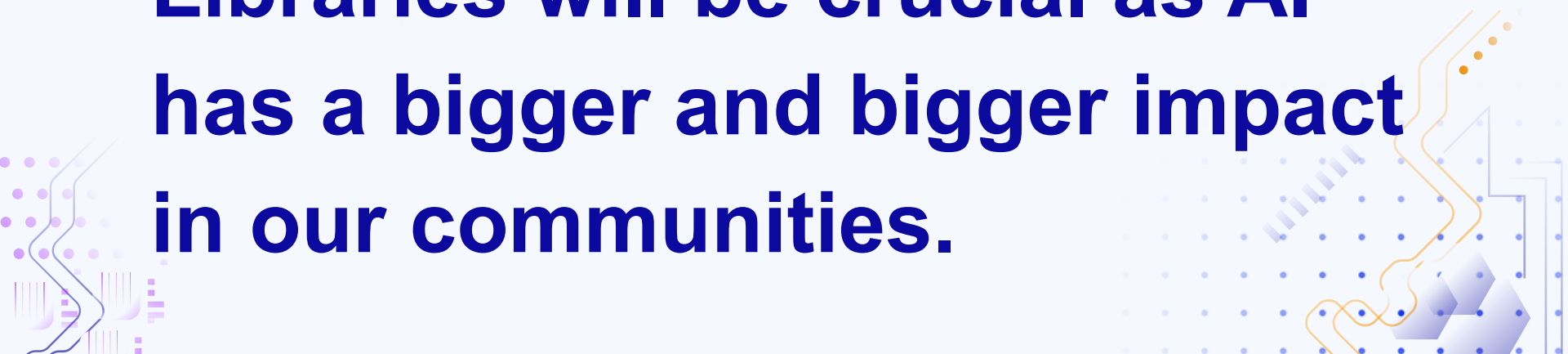


What else have YOU heard?



The gap between reality and science fiction is shrinking.

Libraries will be crucial as AI has a bigger and bigger impact in our communities.

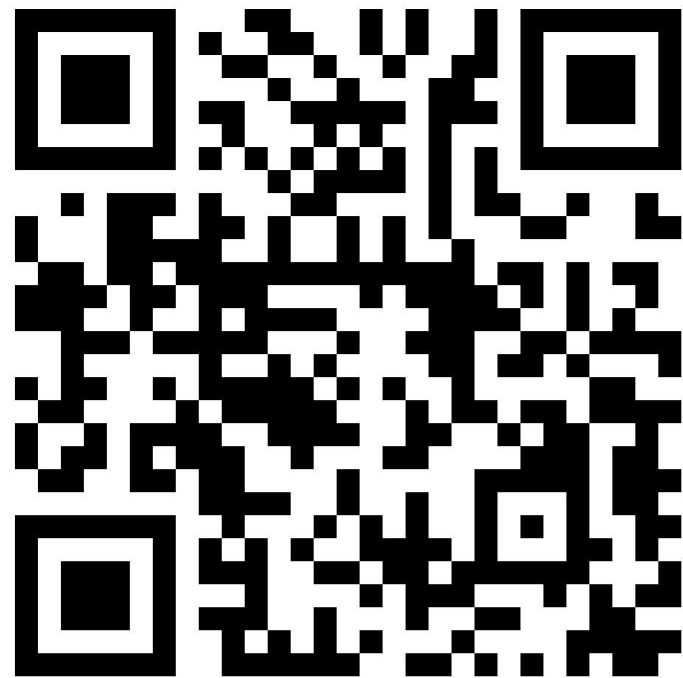


The gap between reality and science fiction is shrinking.

- AI can now reliably:
 - Generate written content or campaigns at such a capability that it can persuade/misinform at scale
 - Generate multimodal

Further Training

- This is the tip of the iceberg
- We want to **prop up communities so that they have internal resources AI safety**
- Training covers: extensive look at broader risk topics, how it all connects to current risk, and how to run this type of programming for your community.



Workshop Survey



[Clickable link](#)