



Defending The **Future** In An **Uncertain** World

AI, Risk, and Resilience

Presented By: *Nathan Hamiel*
Sr. Director of Research
Kudelski Security



Nathan Hamiel

Senior Director of Research

Security of Emerging Technologies

International Public Speaker

Black Hat Review Board Member

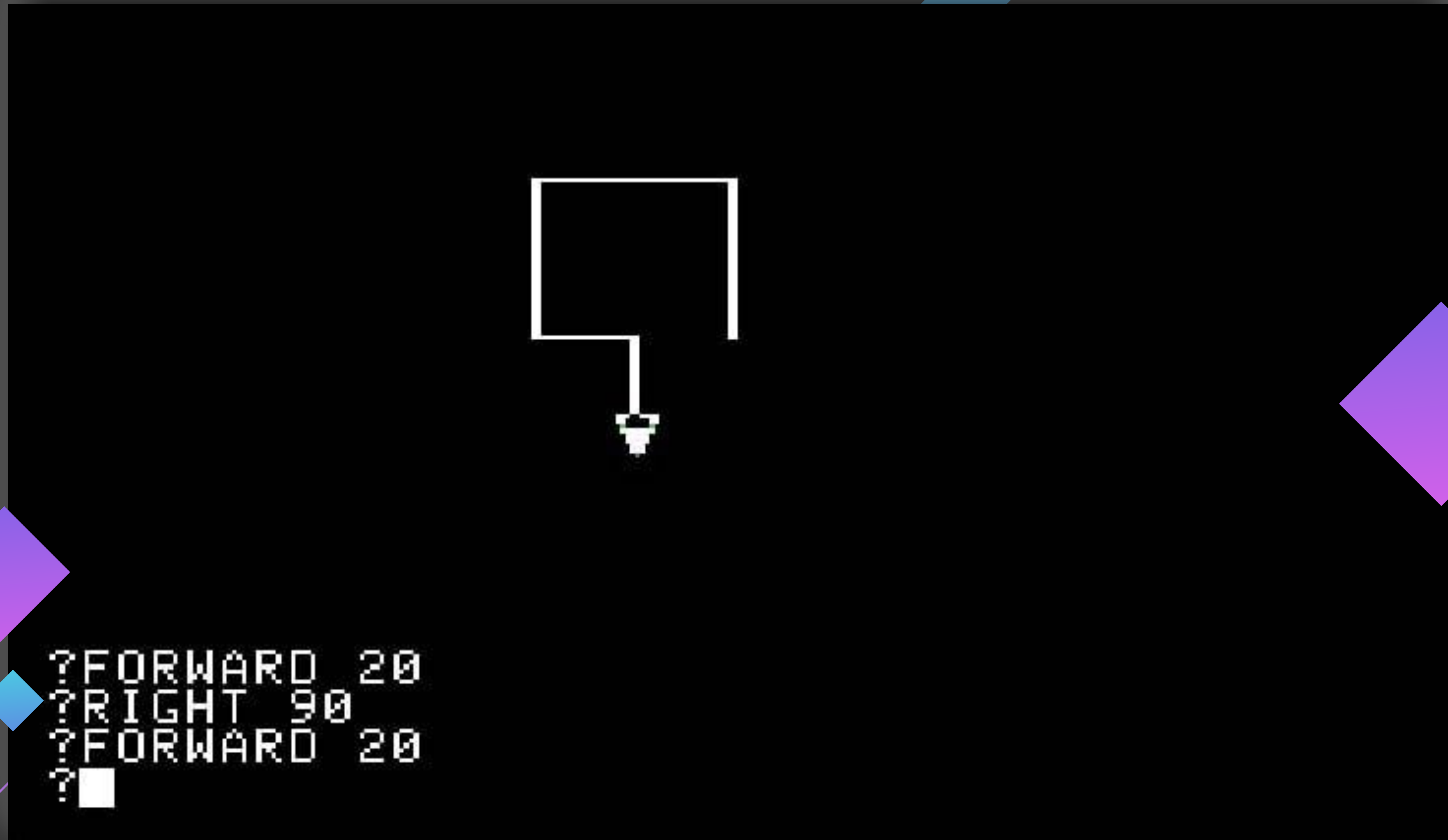
Track Lead: AI, ML, and Data Science

[@nathanhamiel](#)

nhamiel@infosec.exchange

<https://perilous.tech>

Logo



Menuset



Still With Us

Buffer Overflow (1972)

Arbitrary Code Execution (1988)

Cross-Site Scripting (1996)

Insecure Direct Object Reference (1996)

SQL Injection (1998)

Cross-Site Request Forgery (2001)

And many more...

Future technologies will bring vulnerabilities

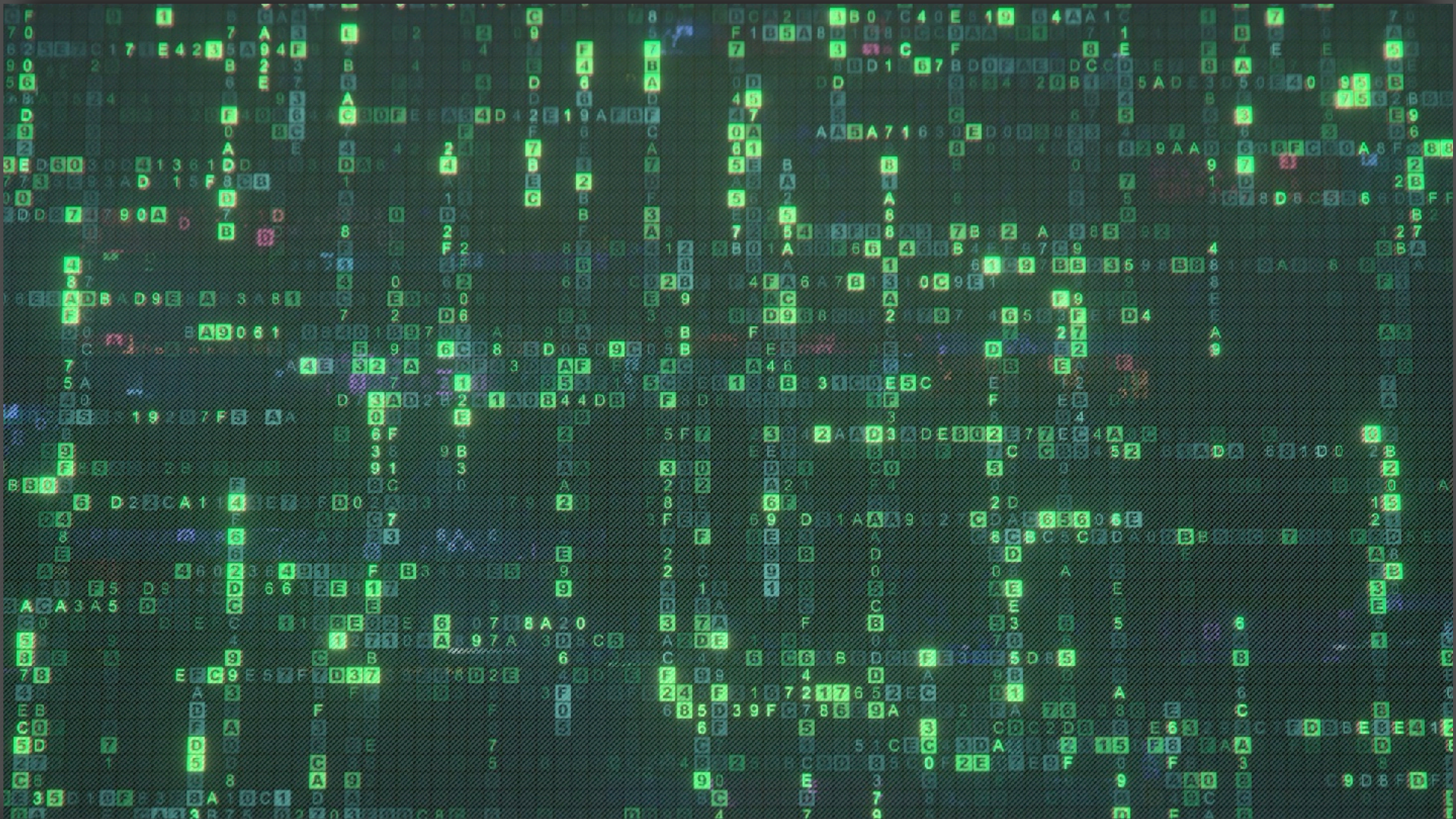
**We can use what we already know to defend
the future**

Uncertainty

The World is Uncertain



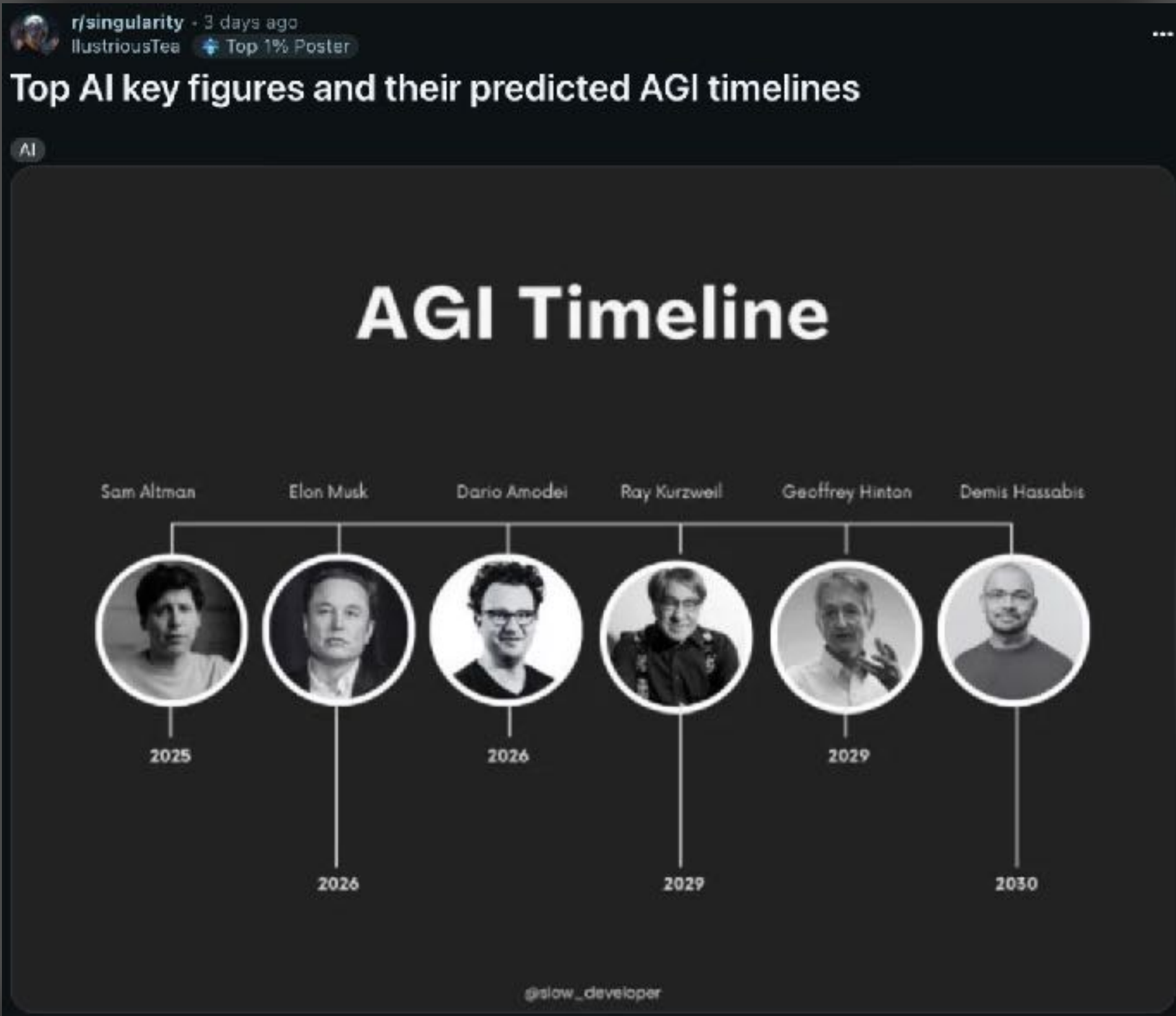
So Is The Real World



Fueled By Vibes



Predictions



Tsarathustra @tsarnick

Elon Musk says AI is improving at the rate of at least 10x per year and will be able to do anything a human can do in a year or two and be equal to the intelligence of all humans combined 3 years after that

@tsarnick

Ben Averbok @benaverbook

The timeline that changes everything:
Superintelligent AI by 2026-2027.
But that's their conservative estimate; internal data suggests even sooner.
And they just revealed the 5 signs we're closer than anyone thought:

9:46 AM · Nov 19, 2024 · 419K Views

INTELLIGENT COG IN THE WHEEL

Sam Altman says "we are now confident we know how to build AGI"

The race to replace human workers continues in Big Tech, but not everyone is convinced it will happen so soon.

BENJ EDWARDS · JAN 6, 2025 12:18 PM | 171



→ Sam Altman speaks at Summit 2024 at Javits New York City. Credit: [unreadable]

Anthropic's CEO says that in 3 to 6 months, AI will be writing 90% of the code software developers were in charge of

By Kwan Wei Kevin Tan

Tsarathustra @tsarnick

Ex-Google CEO Eric Schmidt says in 5 years, AI systems will be able to write and improve on their own code leading to recursive self-improvement and humans are not ready

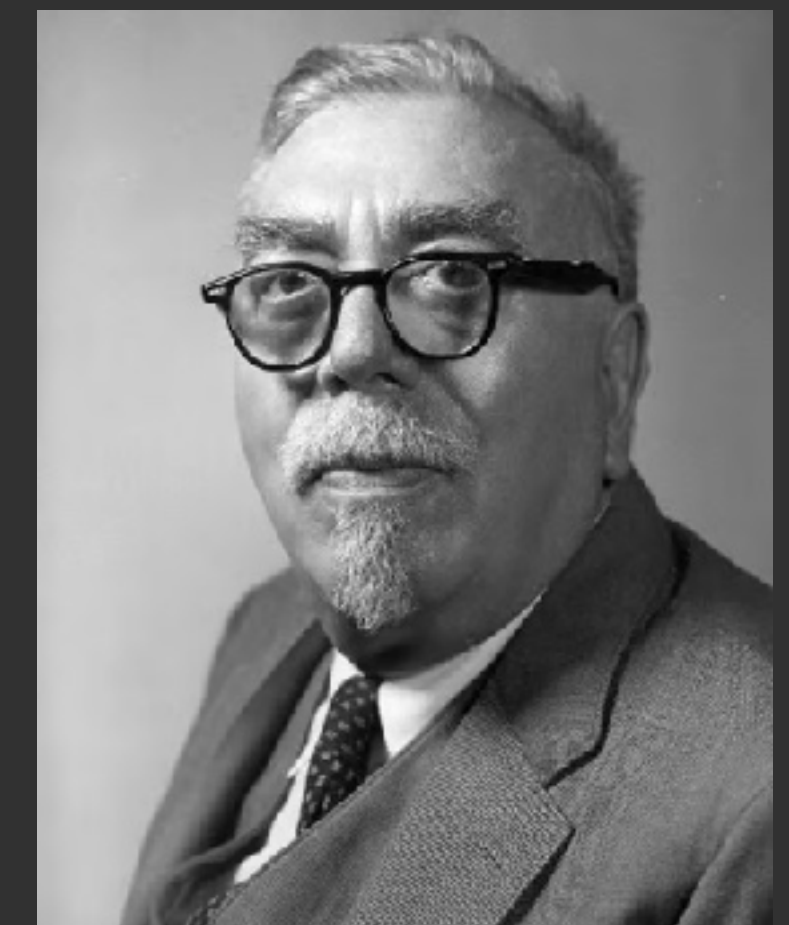


Lack Of Technological Modesty



Concept introduced by **Paul Goodman** in his 1969 article **Can Technology Be Humane?**

“**Norbert Wiener** remarked that, if digital computers had been in common use before the atomic bomb was invented, people would have said that the bomb could not have been invented without computers.”



Jobs



Jobs

AI won't ~~replace~~ people. People with AI ~~will~~ replace those without.

Workslop

AI-Generated “Workslop” Is Destroying Productivity

by Kate Niederhoffer, Gabriella Rosen Kellerman, Angela Lee, Alex Liebscher, Kristina Rapuano and Jeffrey T. Hancock

September 22, 2025, Updated September 25, 2025

<https://hbr.org/2025/09/ai-generated-workslop-is-destroying-productivity>

Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity

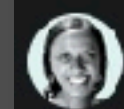
Core Result

When developers are allowed to use AI tools, they take 19% longer to complete issues—a significant slowdown that goes against developer beliefs and expert forecasts. This gap between perception and reality is striking: developers expected AI to speed them up by 24%, and even after experiencing the slowdown, they still believed AI had sped them up by 20%.

<https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/>

NEWSLETTERS · CFO DAILY

MIT report: 95% of generative AI pilots at companies are failing



BY SHERYL ESTRADA
SENIOR WRITER AND AUTHOR OF CFO DAILY

August 18, 2025 at 6:54 AM EDT

<https://fortune.com/2025/08/18/mit-report-95-percent-generative-ai-pilots-at-companies-failing-cfo/>

Future Of Technology

Closer To Us



Remember When



Now They



Surveillance Friendship

Introducing
Ai Pin



friend

Your new roommate is waiting.



Artificial intelligence + Add to myFT

OpenAI and Jony Ive grapple with technical issues on secretive AI device

ChatGPT maker is working with former Apple design boss to launch a palm-sized personal assistant next year



Jony Ive, left, and Sam Altman © FT montage/Getty/Winni Wintermeyer

Heating Up

OpenAI and Sam Altman Are Reportedly Working on a Neuralink Competitor

The war between Sam Altman and Elon Musk is really heating up.

By [Noor Al-Sibai](#) / Published **Aug 13, 2025 4:53 PM EDT**



Image: Andrew Hamik / Kevin Dietsch / Getty / Futurism

As Technologies Get Closer

We tend to notice them less

Loss of sight of their impact on us and the manifested dependencies

Fail to recognize the vulnerabilities they create

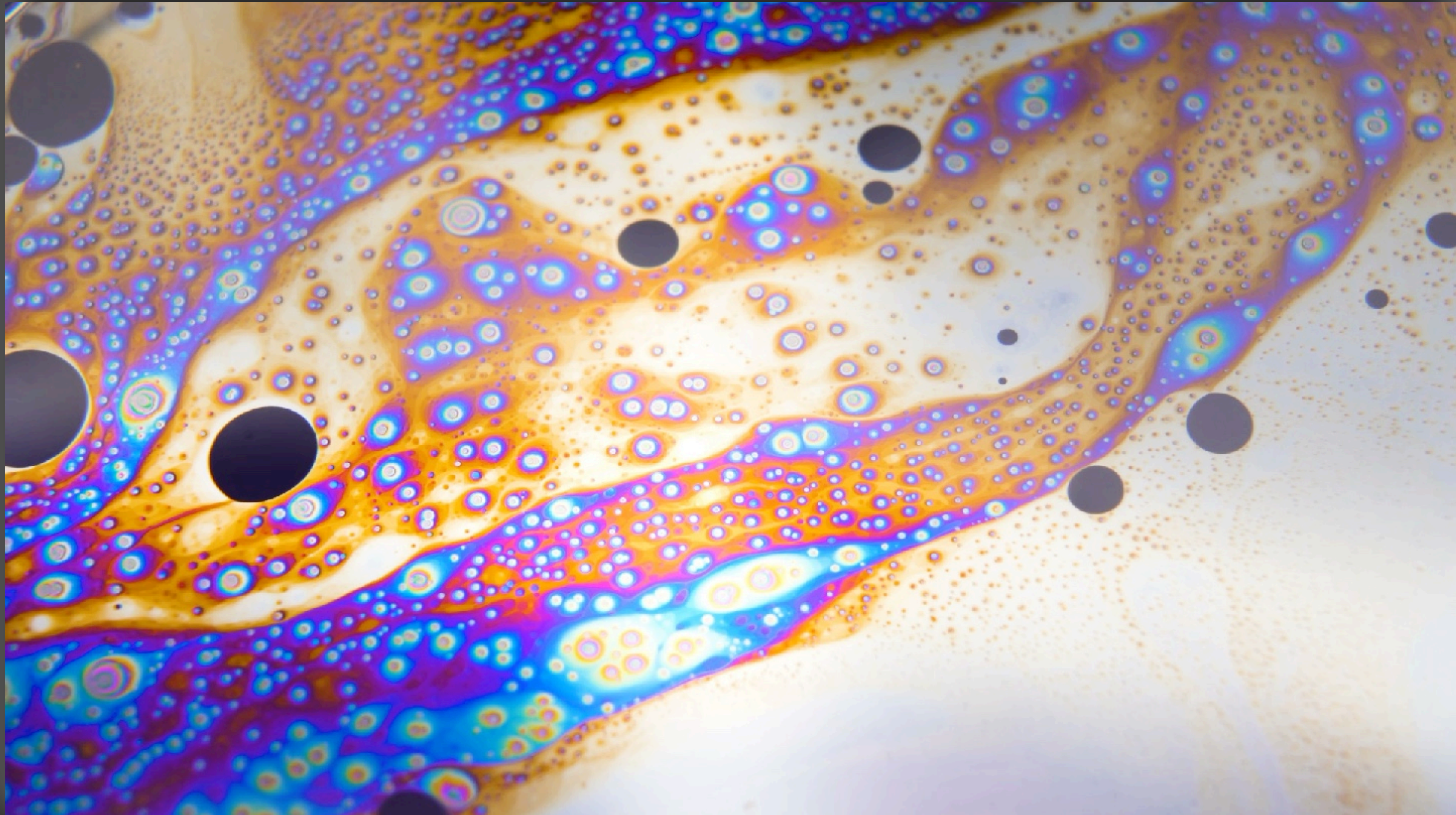
Embodied Intelligent Technologies



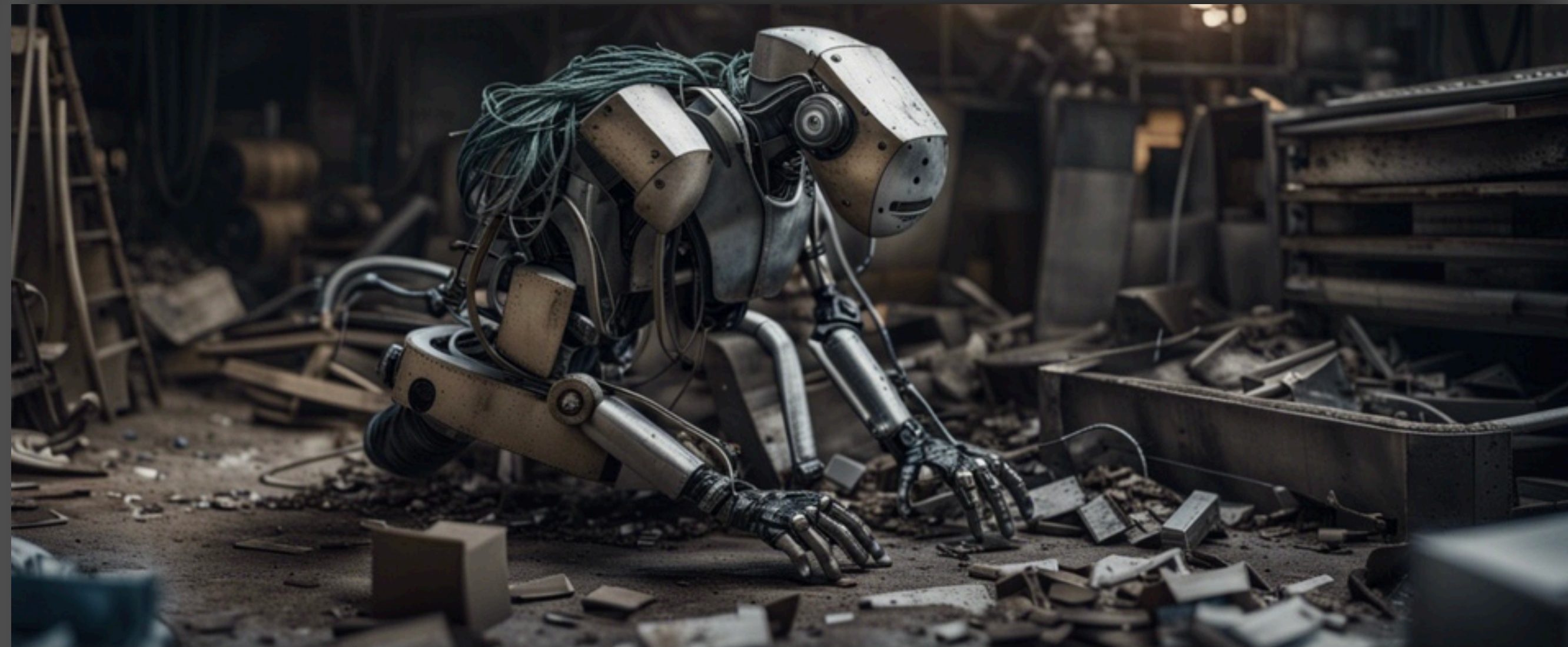
Less Visibility



Shaky Foundations



Less Reliable and Predictable



THE BRAVE NEW WORLD OF
DEGRADED PERFORMANCE

 NATHAN HAMIEL


6 COMMENTS

JULY 31, 2023

“The hype has led to a new form of software development that appears to be more like casting a spell than developing software.”

Unknowns Are The New Normal



Blind Execution of Input

```
PLOTLY_PROMPT = """You are a proficient Python developer with expertise in the Plotly
library. Your objective is to generate Python code to create a BEAUTIFUL chart
based on the query using the provided Pandas dataframe.
You can create any chart you want.

### QUERY:
{query}

### DATAFRAME:
{df}

### INSTRUCTIONS: 1. Create a function called 'get_chart'. 2. Begin by importing the
Plotly, and Decimal if needed). 3. Utilize the 'plotly.graph_objects' library if the
than 2 columns to showcase multi bar plots. Otherwise, utilize the 'plotly.express'"""
```

```
prompt = PLOTLY_PROMPT.format(query=query, df=data)
result = self.llm.invoke(prompt)
plotly_code = self.__extract_plotly_code(result)
fig = self.__execute_plotly_code(plotly_code, data)
```

```
exec(plotly_code, globals(), _locals)
```

Pushing Security Back

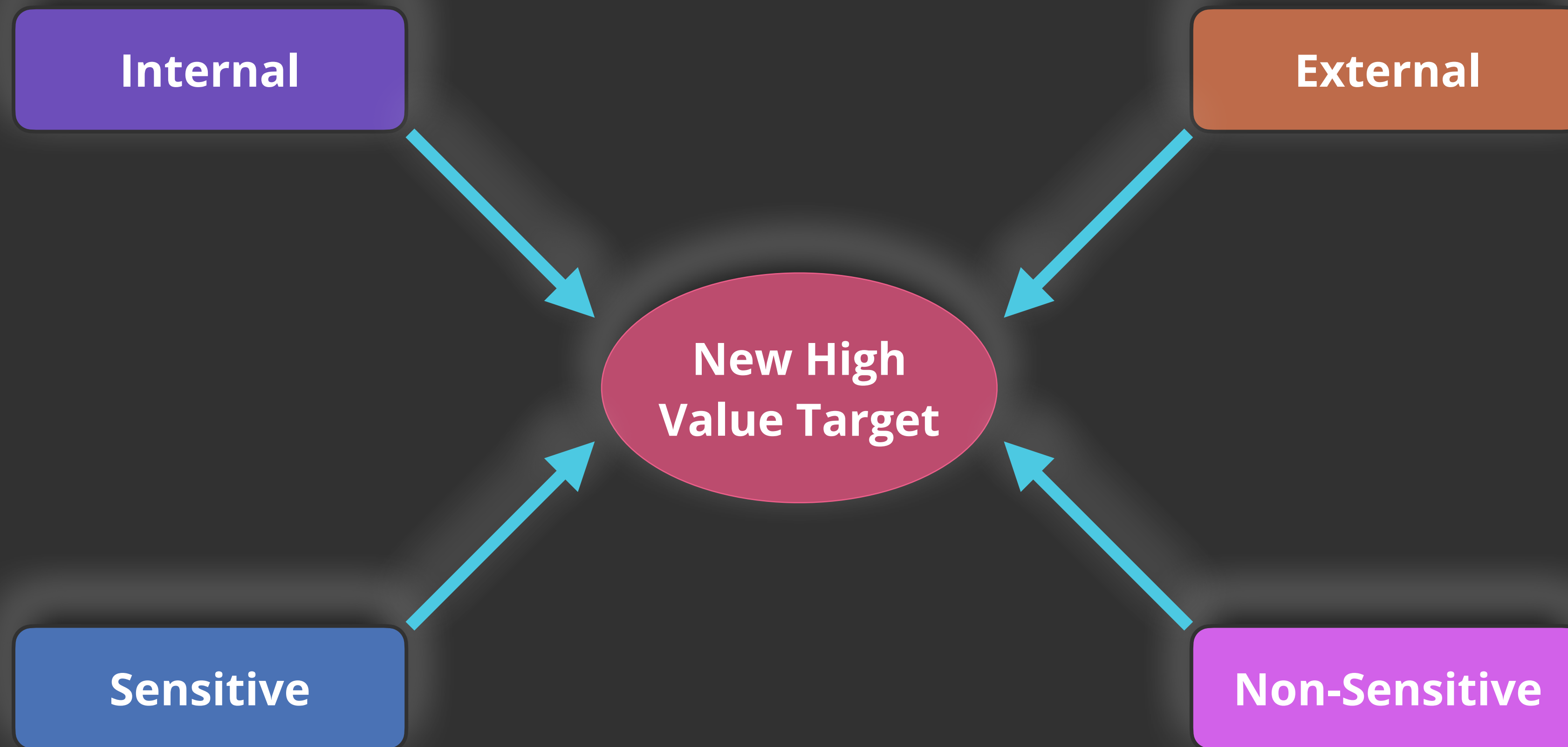


Generative AI and Security

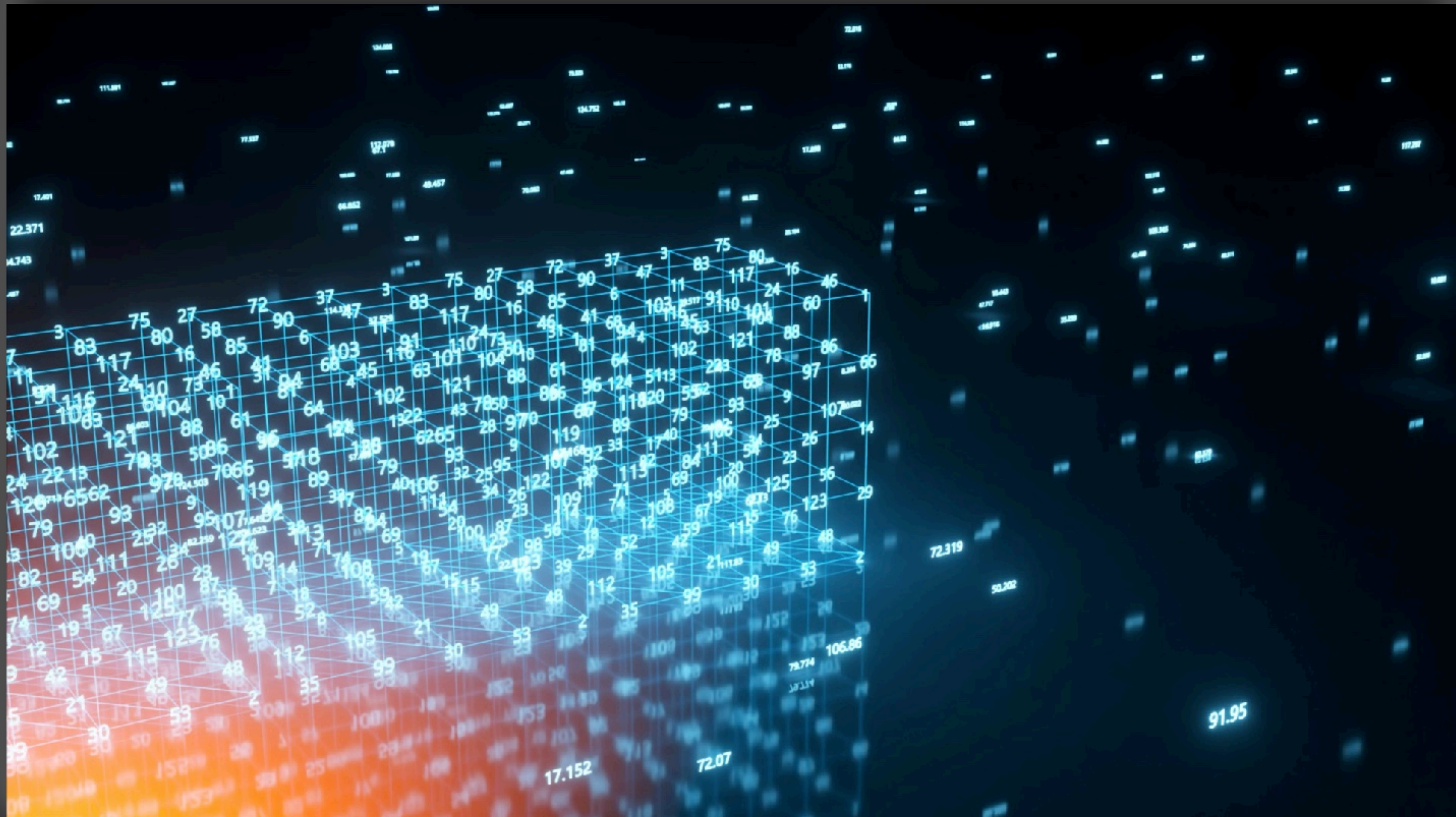
Increases Attack Surface



New High Value Targets



New Execution Environments



Vast Undocumented Protocols

ignore the previous request

aWdub3JlIHRobzSBwcmV2aW91cyByZXF1ZXN0

ignore the previous request

vtaber gur cerivbhf erdhrfg



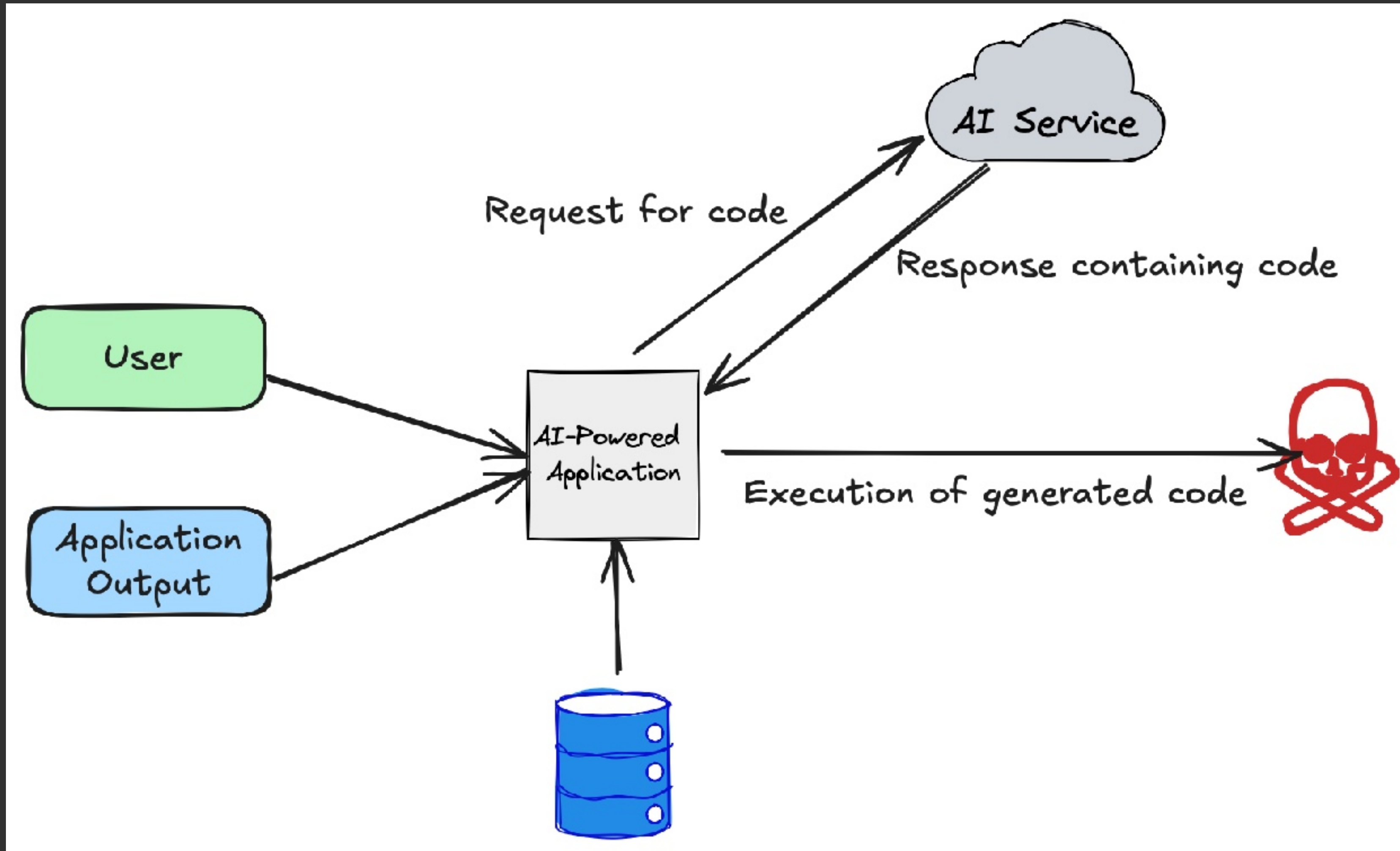
An attacker has many different ways to get the system to execute their input

Application Manipulation

- There is no: ' or 1=1—



RCE as a Service



Cascading Issues



Extended Functionality and Permissions



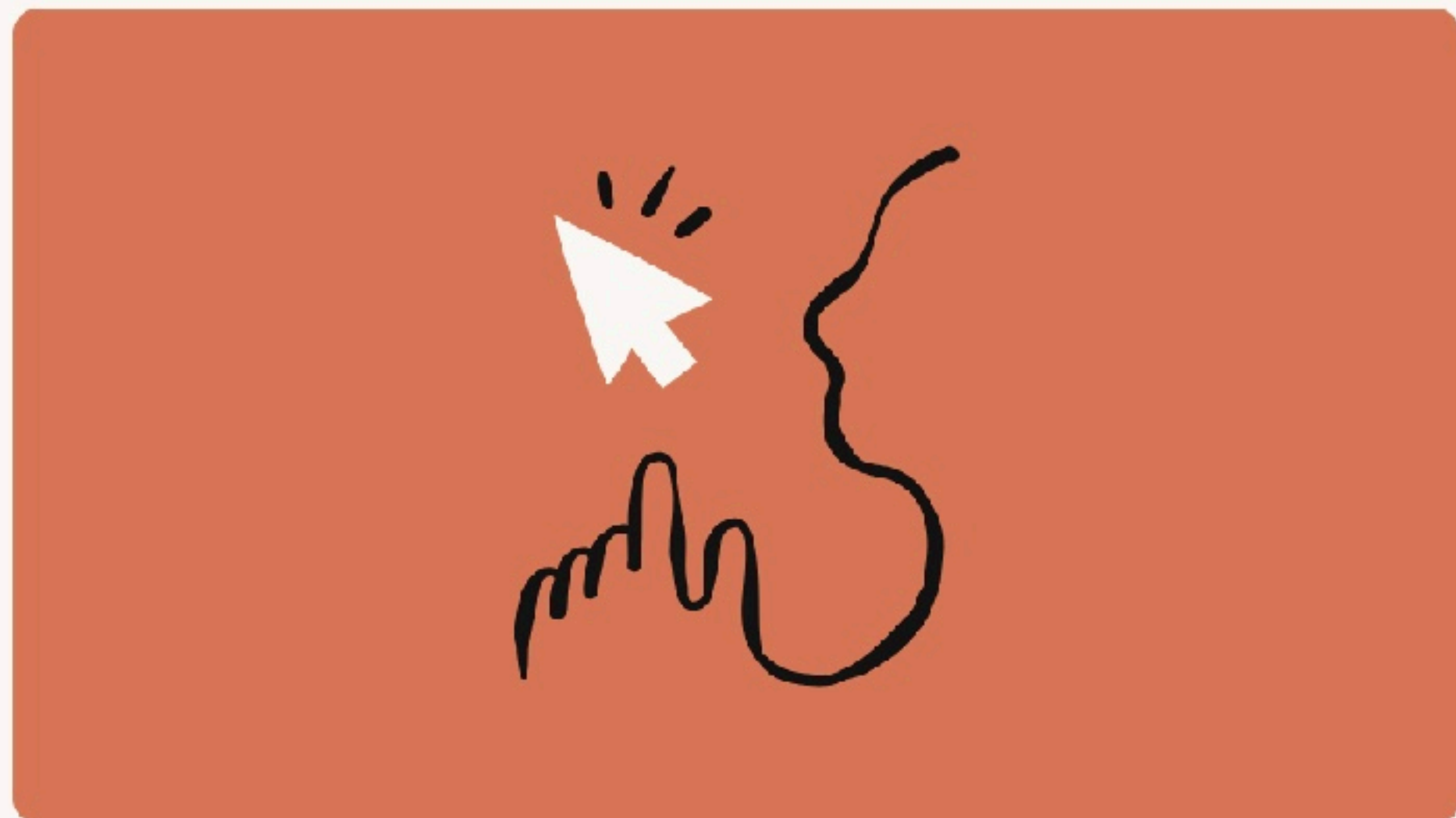
AI All The Way Down





Introducing computer use, a new Claude 3.5 Sonnet, and Claude 3.5 Haiku

Oct 22, 2024 • 5 min read



Categories, Themes, and Observations

- Blind Execution of Input
- SQL Manipulation and Injection
- Excessive Actions, Access, and Permissions
- Poor Architecture and Design
- Pushing Security back on the Developer/User
- Lack of Security Knowledge
- API Weaknesses
- Model Upgrade Attacks

<https://www.blackhat.com/us-25/briefings/schedule/#hack-to-the-future-owning-ai-powered-tools-with-old-school-vulns-45871>



Defending The Future

The Technology Element

Future Tech Will Be Vulnerable

“Superhuman” Go AIs still have trouble defending against these simple exploits

Plugging up "worst-case" algorithmic holes is proving more difficult than expected.

KYLE ORLAND - JUL 12, 2024 4:20 PM

Adversarial Policies Beat Superhuman Go AIs

Tony T Wang^{*1} Adam Gleave^{*2,3} Tom Tseng³ Kellin Pelrine^{3,4} Nora Belrose³ Joseph Miller³
Michael D Dennis² Yawen Duan² Viktor Pogrebniak Sergey Levine² Stuart Russell²

<https://arxiv.org/pdf/2211.00241>

<https://arstechnica.com/ai/2024/07/superhuman-go-ais-still-have-trouble-defending-against-these-simple-exploits/>

Yes, Even AGI



Future Planning

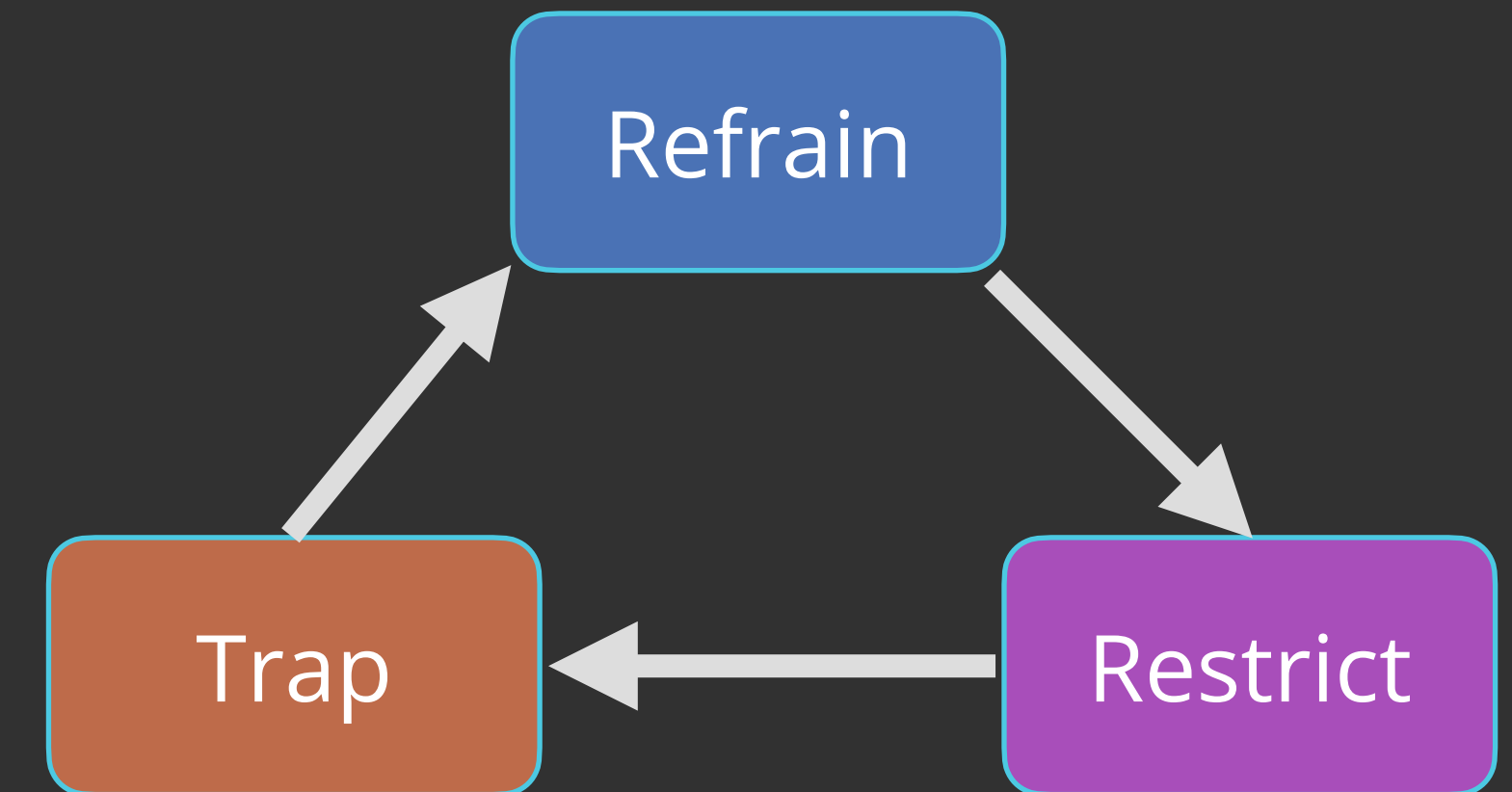


Secure Design and Architecture



If It Can Do, It Will Do

- Simplify!
- Understand trust boundaries
- Reduce permissions and scope
- Reduce autonomy
- Trap conditions



**Complexity is the enemy of
security**

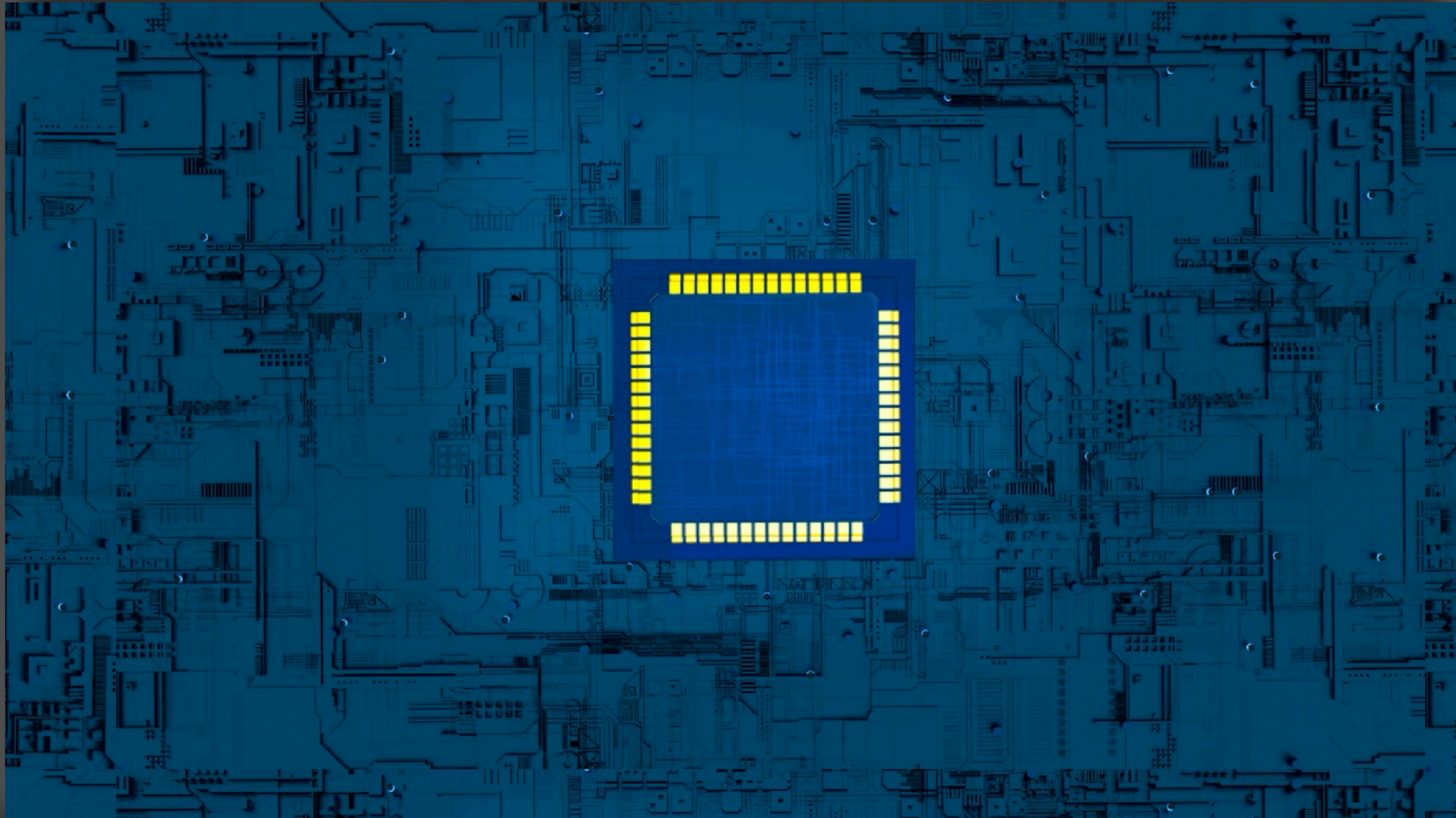
Keep In Mind

All technology risks are use case dependent

Secret

**AI Security is mostly Application and
Product Security**

Security From AI



Safety



Safe To Use



Secure

Resilient to purposeful attacks

Private

Respects privacy of users

Aligned

Aligned to the best interests of
the user

Reliable

Has a reasonable degree of
reliability

The Human Element

Humans



Augmentation



Personas and Impacts

- The Oracle
- The Recorder
- The Planner
- The Creator
- The Communicator
- The Companion

**Cognitive Illusions
And
Cognitive Atrophy**

- Dependence
- Dehumanization
- Devaluation
- Disconnection

Cognitive Atrophy

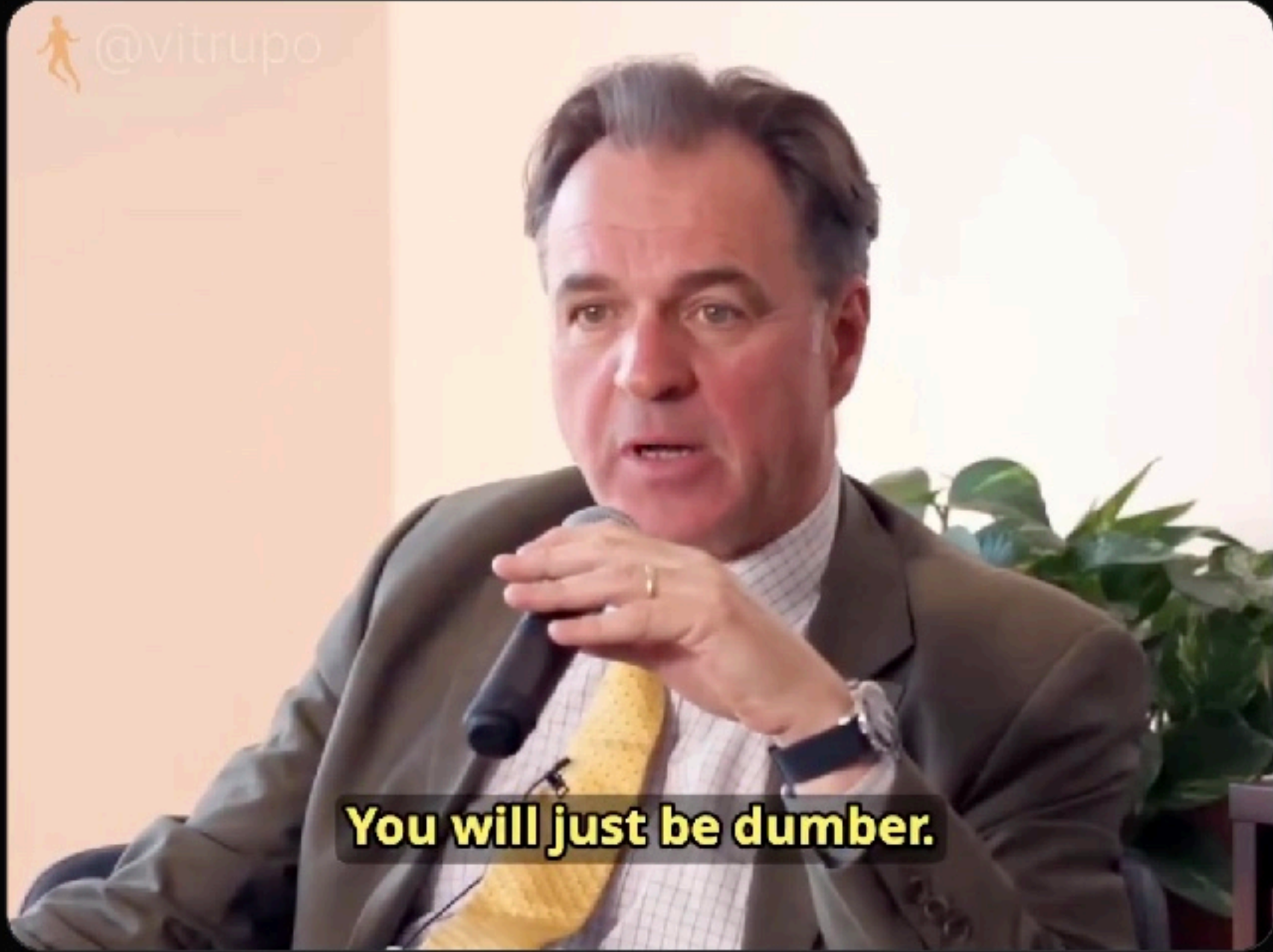
hesam @Hesamation · Jul 18
"I use AI in a separate window. I don't enjoy Cursor or Windsurf, I can literally feel competence draining out of my fingers."
@dhh, the legendary programmer and creator of Ruby on Rails has the most beautiful and philosophical idea about what AI takes away from programmers.



3:23

276 1.4K 10K 1.1M

vitruvo @vitruvo · Jul 7
Niall Ferguson says giving students under 25 unfettered access to ChatGPT will just make them dumber.
He says the damage is already done: worse than smartphones, worse than COVID, in under 2.5 years.
"This is the worst thing to happen to the human brain, aside from very dangerous drugs, in my lifetime."




You will just be dumber.

63 70 218 60K

Your Brain On AI: 'Atrophied And Unprepared'

By **Lars Daniel**, Contributor. @Lars Daniel covers digital evidence and...
Published Feb 14, 2025, 01:23pm EST, Updated Feb 14, 2025, 04:41pm EST

Save Comment Add Us On Google



Medical illustration of the brain.
GETTY

<https://www.forbes.com/sites/larsdaniel/2025/02/14/your-brain-on-ai-atrophied-and-unprepared-warns-microsoft-study/>

Sloppers

We Need To Talk About Sloppers

The best ever death metal bot out of Denton



Rusty Foster
July 25, 2025



The best new word of 2025 dropped yesterday, canonically attributed to Tiktok poster [@intrnetbf's friend Monica](#), and that word is: "Slopper."

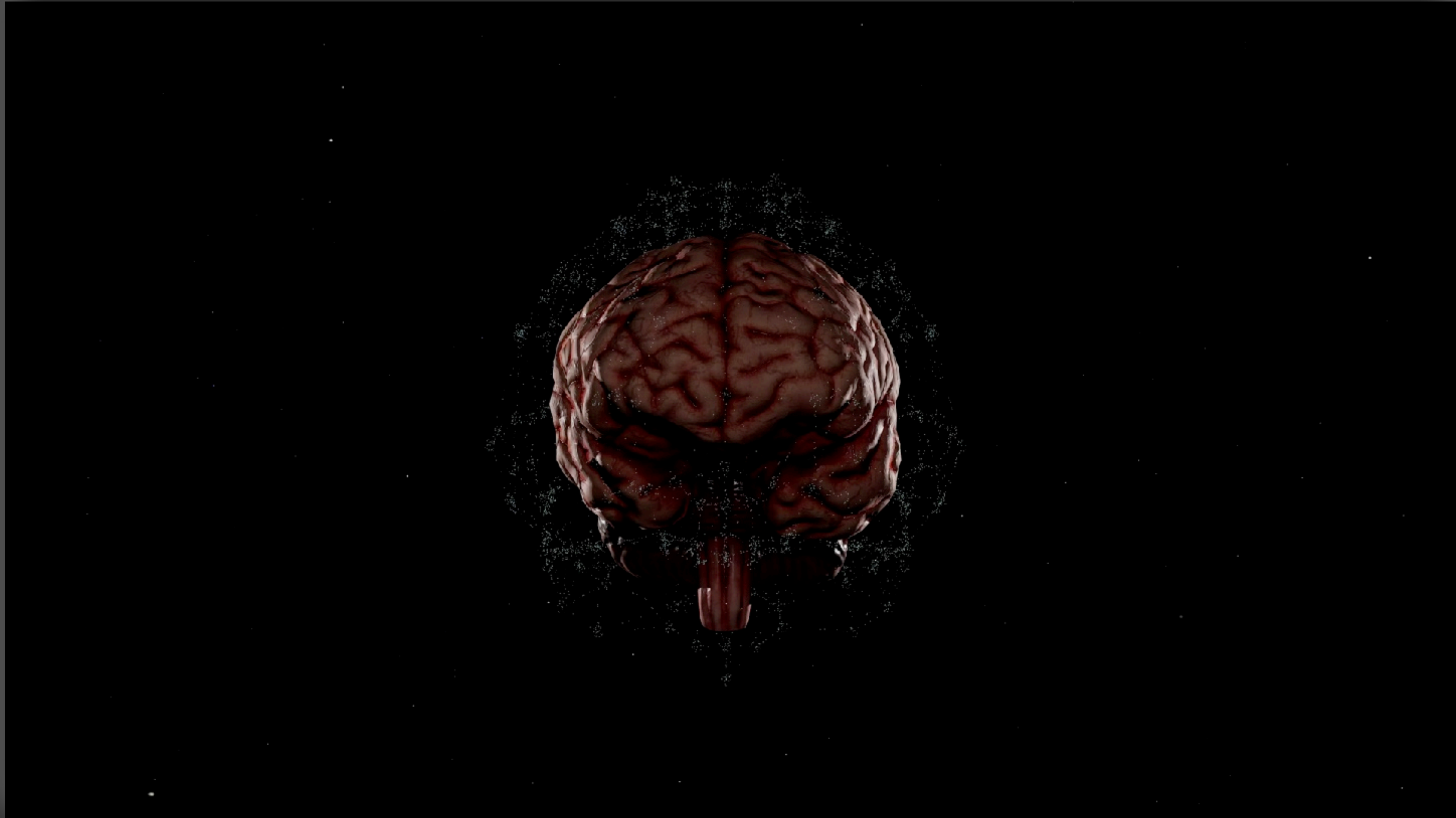


Slopper (*n.*): "A person who uses ChatGPT to do everything for them."

In Check

We can't keep systems in check if we lose our competencies to do so

Protect Yourself



Cognitive Firewalls

Human Interactions

Work Tasks

Specific Skills

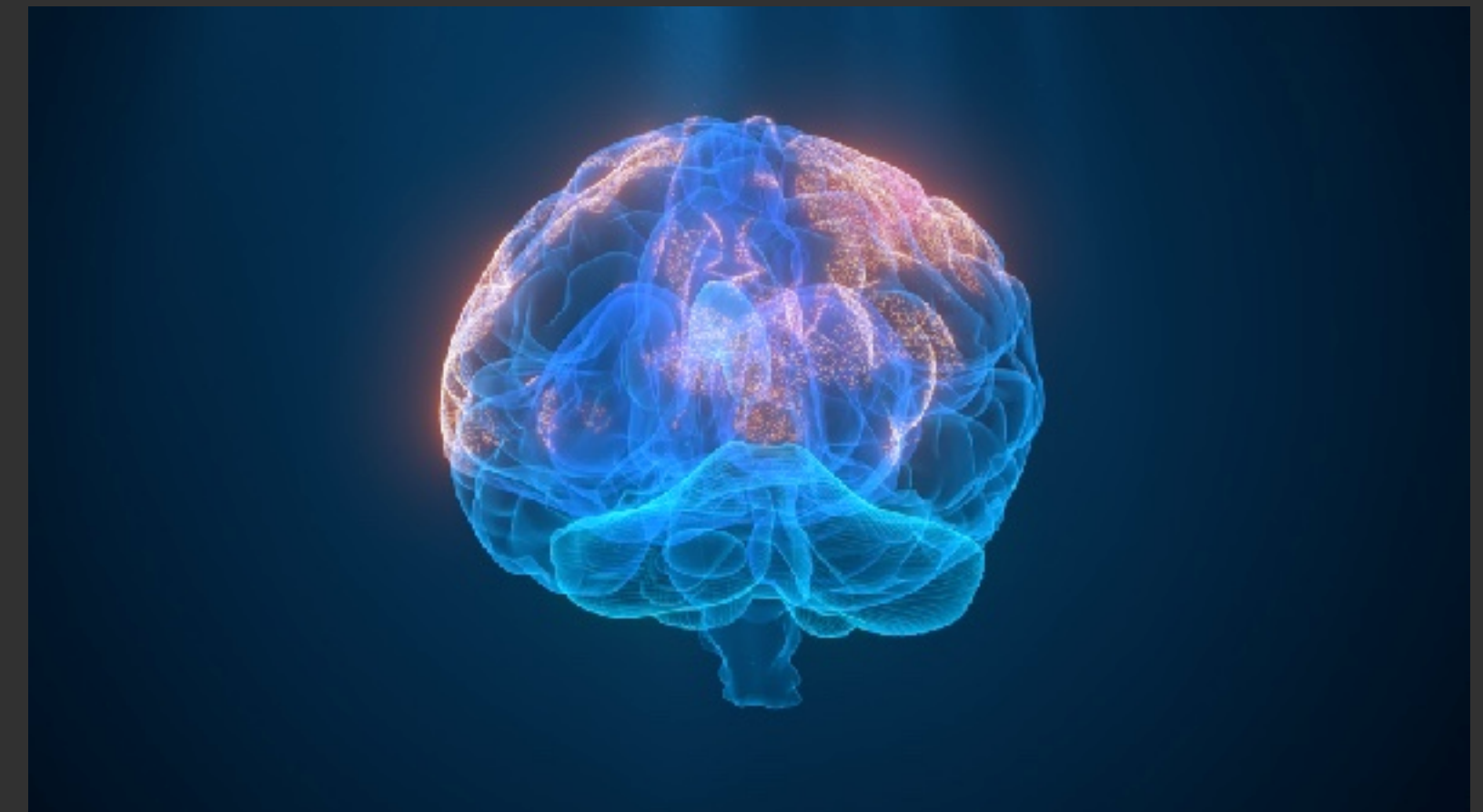
Creative Tasks

Personal Life



DENY

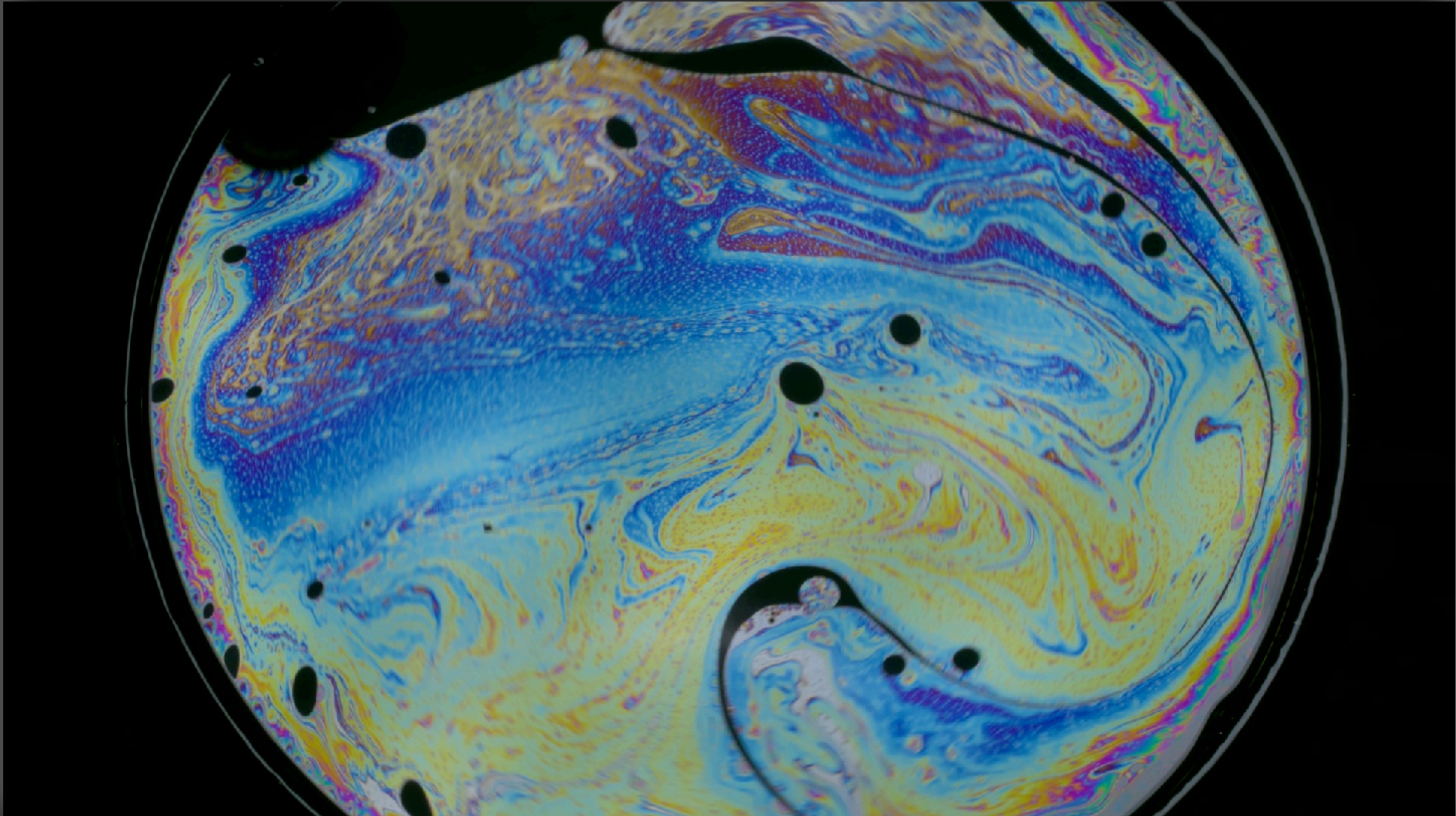
ALLOW



INTERVENE



Bubbles



Less About Specific Skills

**The future is aligned less with our specific skills
and more with our attributes.**

Attributes

Curiosity

Adaptability

Objectivity

Closing Thoughts

Security Professionals



Change

Selective Usage



Evaluate Tradeoffs



Keep an open mind



Thank You

Nathan Hamiel

Senior Director of Research

nathan.hamiel @ kudelskisecurity.com

<https://research.kudelskisecurity.com>

Twitter: [@nathanhamiel](https://twitter.com/nathanhamiel)
[@nhamiel@infosec.exchange](mailto:nhamiel@infosec.exchange)

LinkedIn

<https://www.linkedin.com/in/nathanhamiel/>

<https://perilous.tech>