Defending The Future In An Uncertain World

AI, Risk, and Resilience

Presented By: *Nathan Hamiel*

Sr. Director of Research

Kudelski Security

KUDELSKI SECURITY | RESEARCH

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

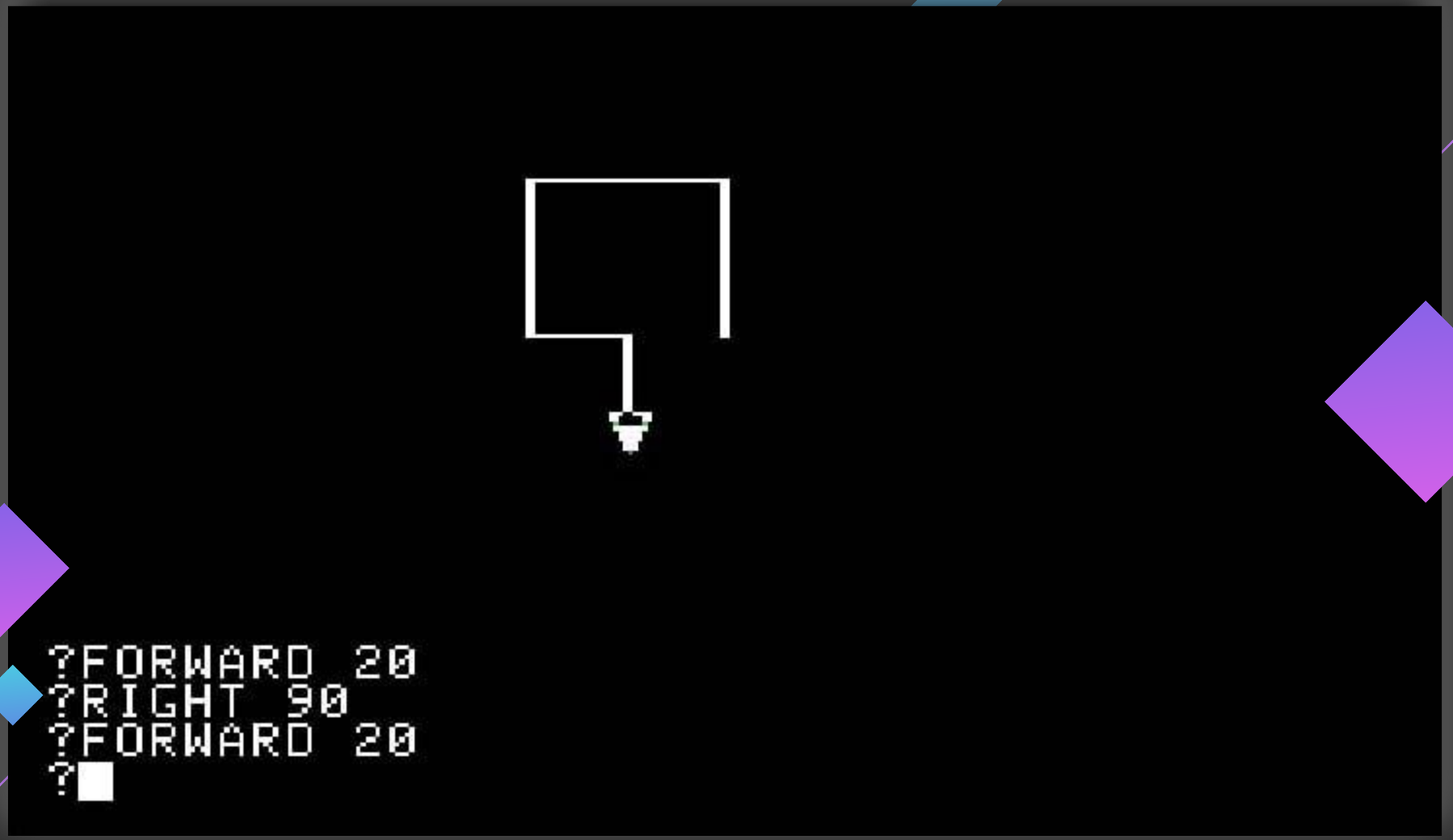
@nhamiel@infosec.exchange

https://perilous.tech

KUDELSKI SECURITY | RESEARCH

PERILOUS TECH

Logo



Menuset


Add
List of Drivers
Unlisted or Updated Driver
Ad Lib
Creative Labs Sound Blaster 1.0
Creative Labs Sound Blaster 1.5
Media Vision Thunder Board
MIDI Mapper
Roland LAPC1
Roland MPU-401
Timer
[MCI] CD Audio
[MCI] MIDI Sequencer
OK
Cancel
Help
MICROSOFT® WINDOWS™
Version 3.1
Copyright © Microsoft Corporation 1985-1992.
All Rights Reserved.
Program Manager
File Options Window Help
Main
File Manager
Control Panel
Print Manager
Clipboard Viewer
MS-DOS Prompt
Windows Setup
PIF Editor
Read Me
Accessories
Minesweeper
Game Help

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)

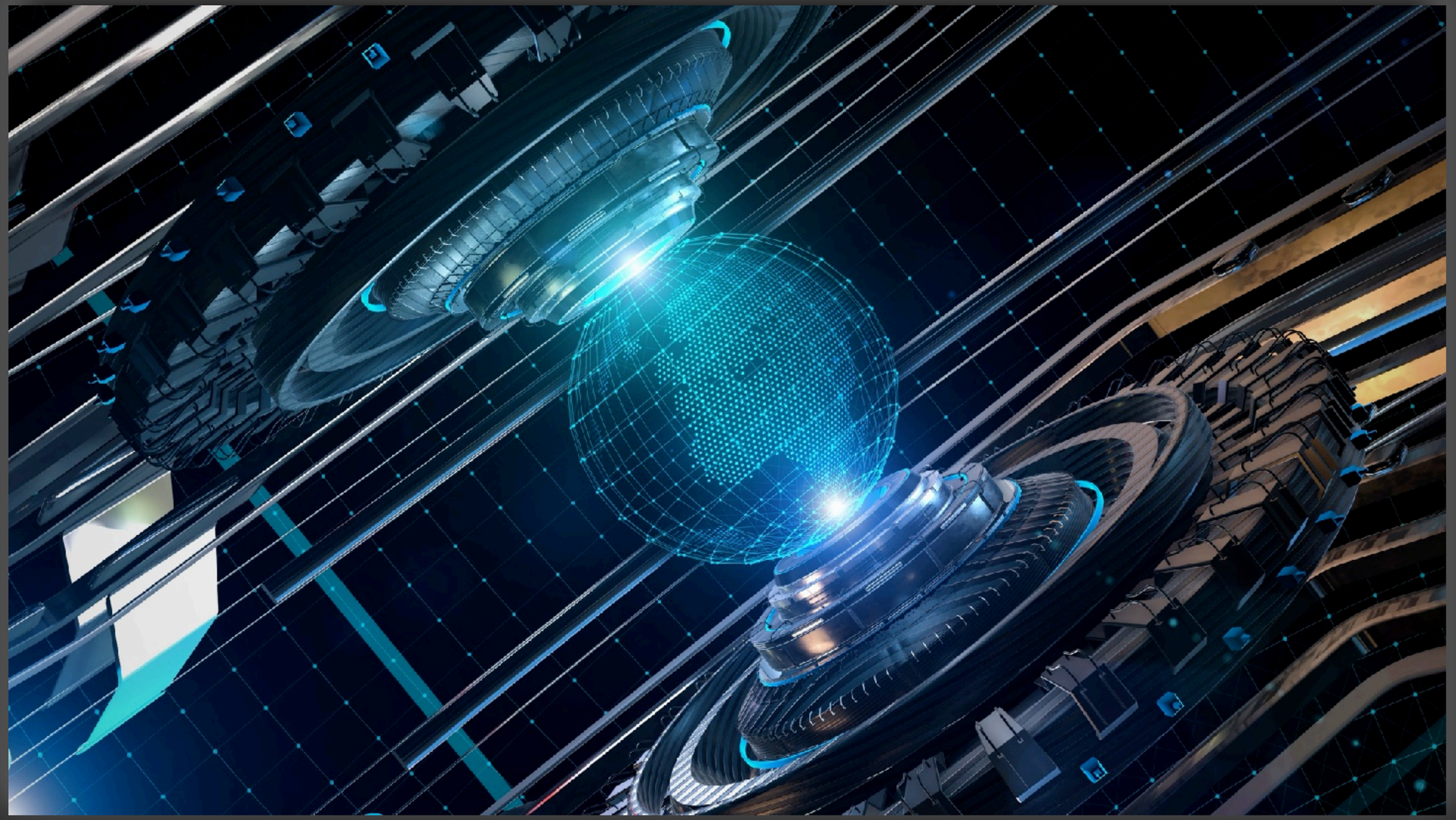
And many more...

Future technologies will bring vulnerabilities

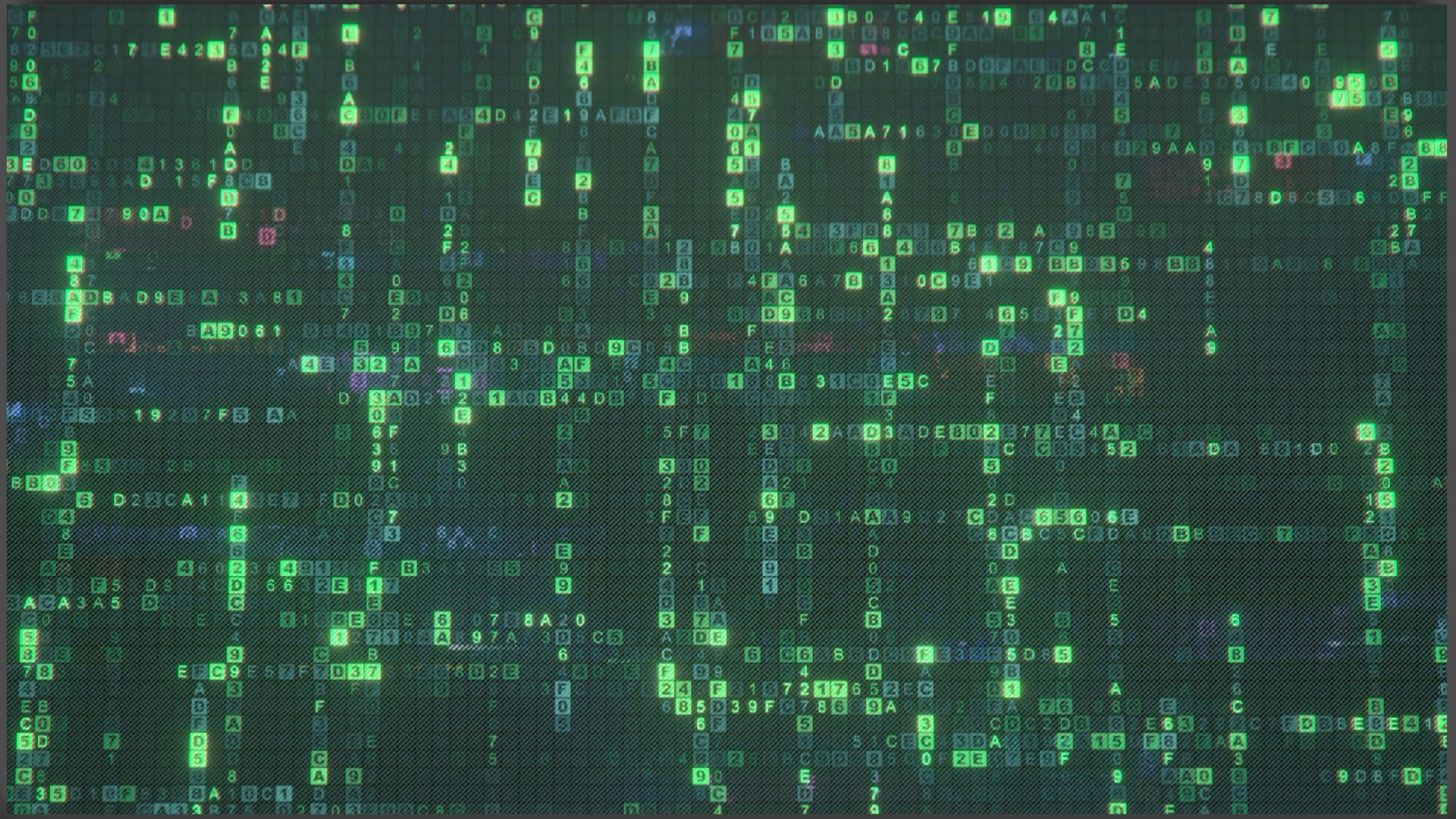
We can use what we already know to defend the future

Uncertainty

The World is Uncertain

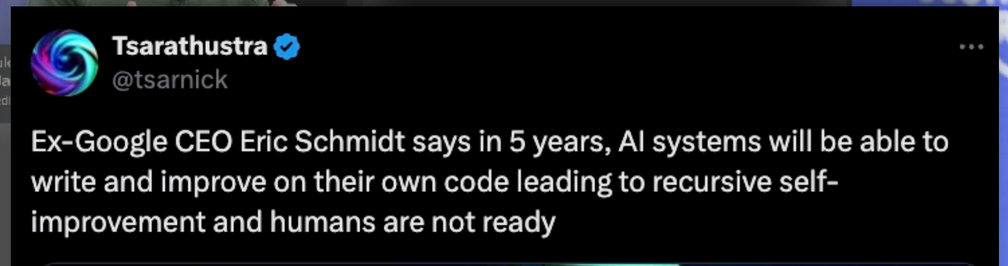


So Is The Real World



Fueled By Vibes

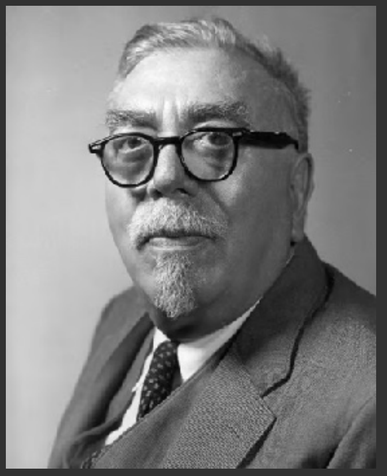
Predictions



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.”



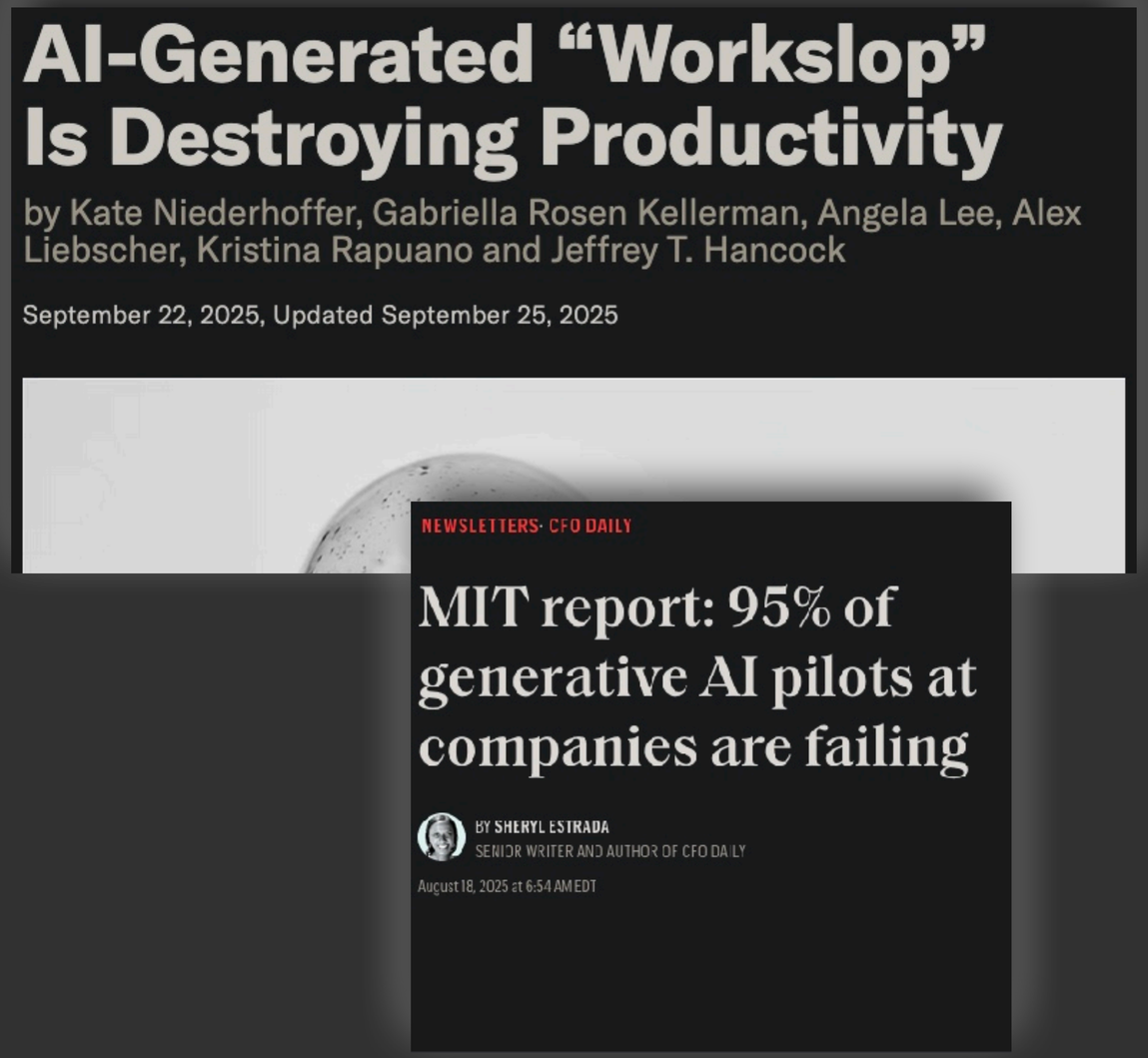
Technopoly, Neil Postman (1992)

Jobs

Jobs

AI won’t replace people. People with AI will replace those without.

Workslop

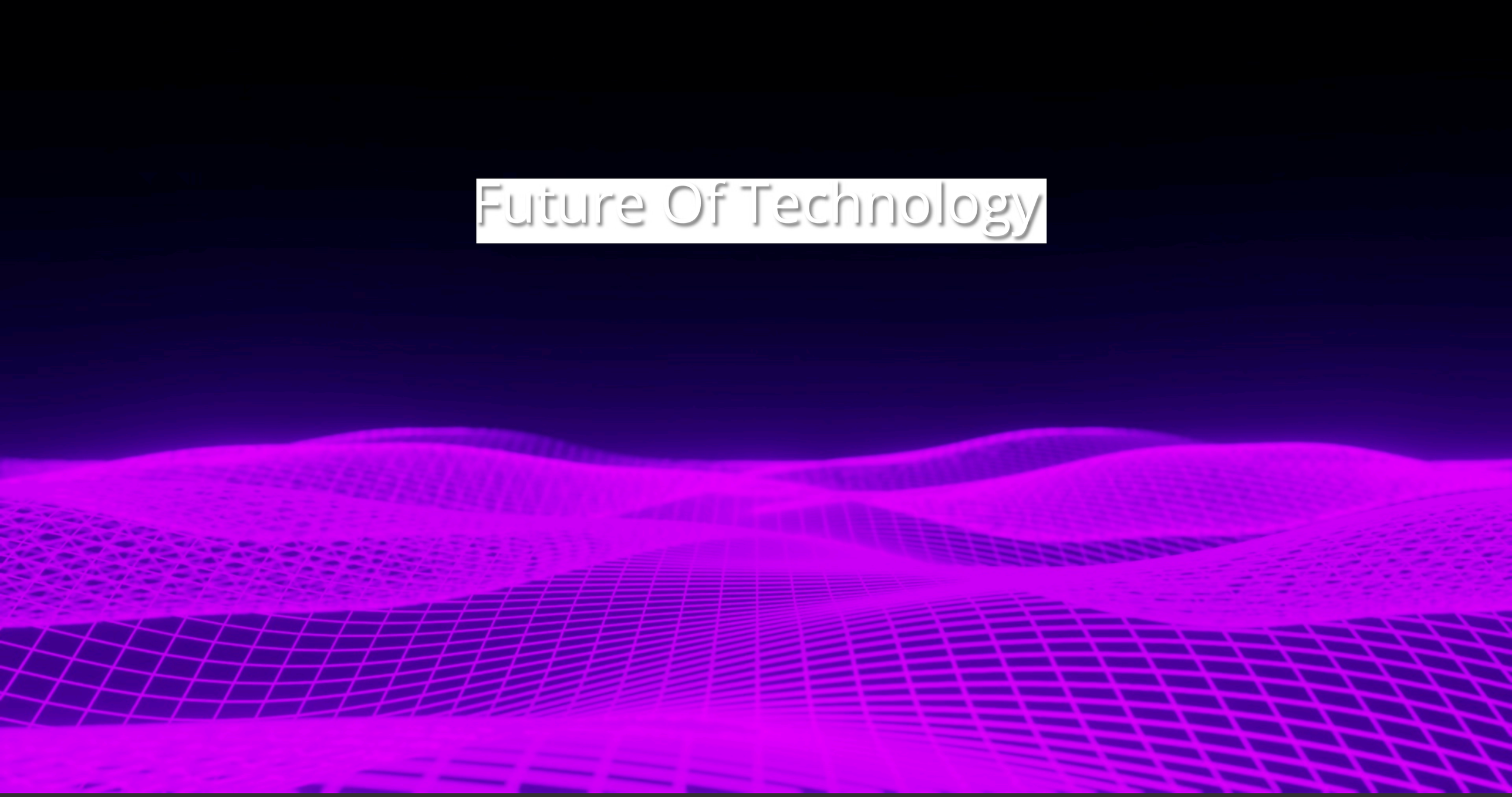


https://hbr.org/2025/09/ai-generated-workslop-is-destroying-productivity

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

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

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 Harnik / Kevin Dietsch / Getty / Futurism

https://futurism.com/openai-sam-altman-neuralink-competitor

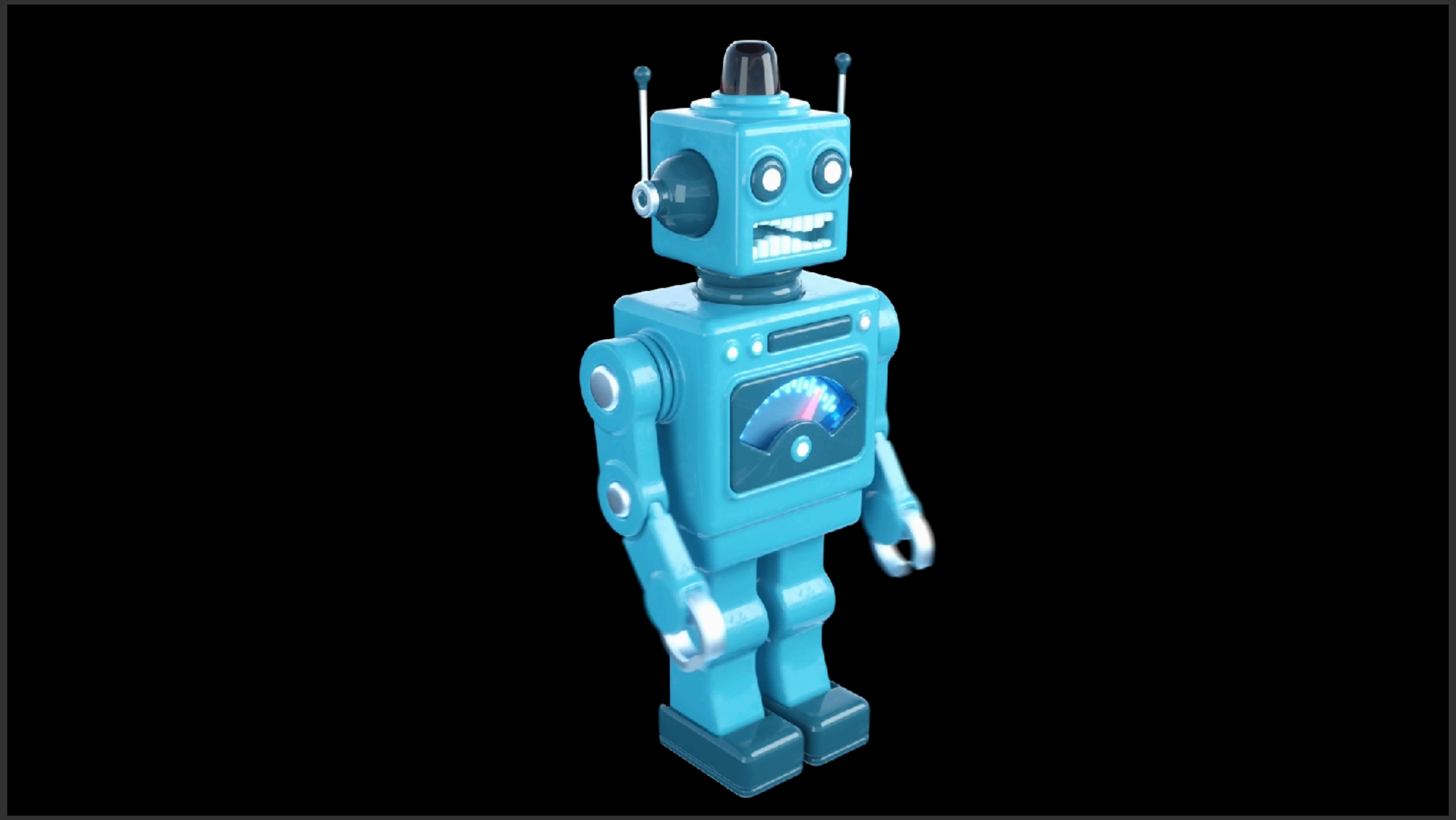
As Technologies Get Closer

We tend to notice them less

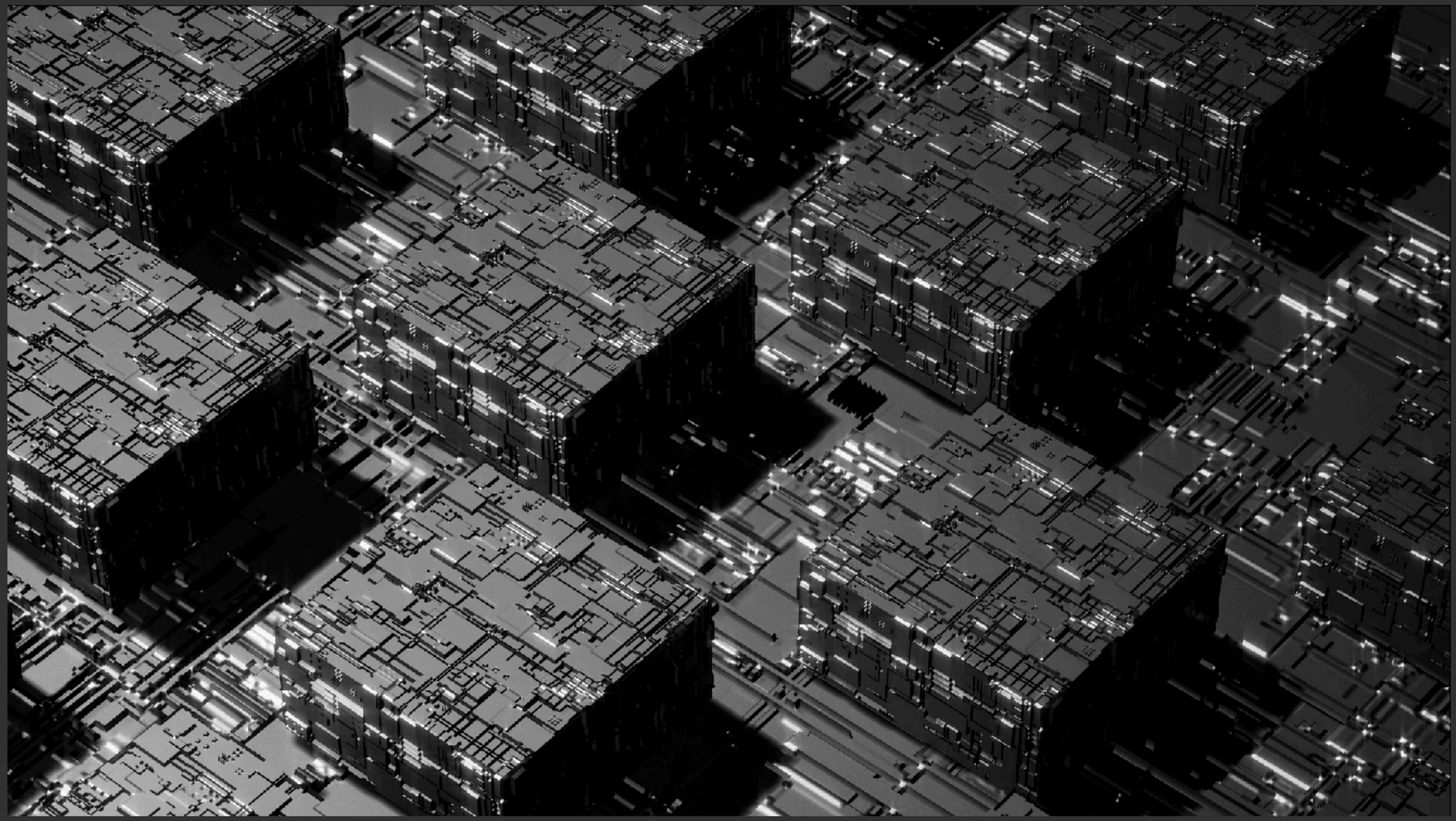
Lose sight of their impact on us and the manifested dependencies

Fail to recognize the vulnerabilities they create

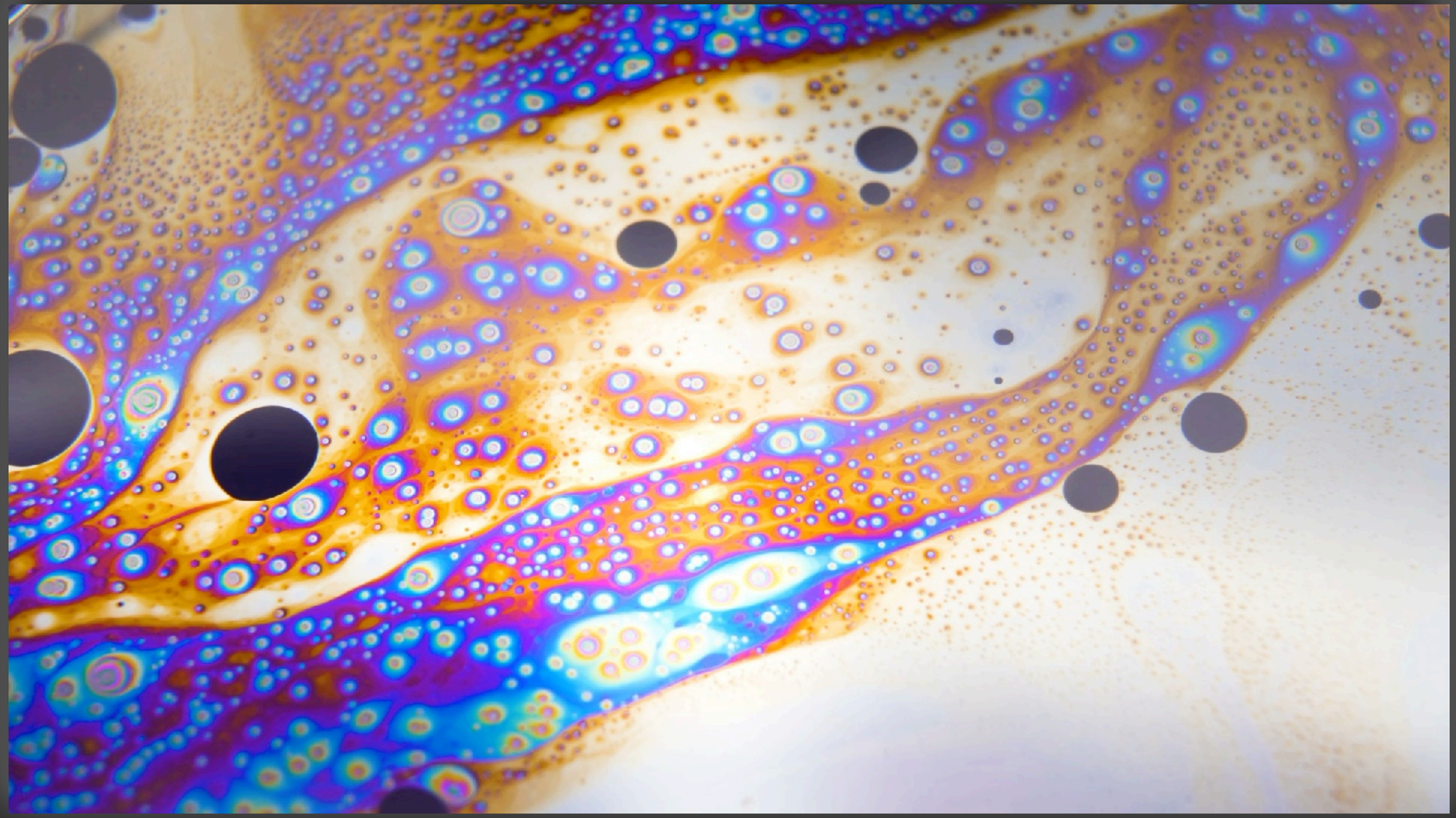
Embodied Intelligent Technologies



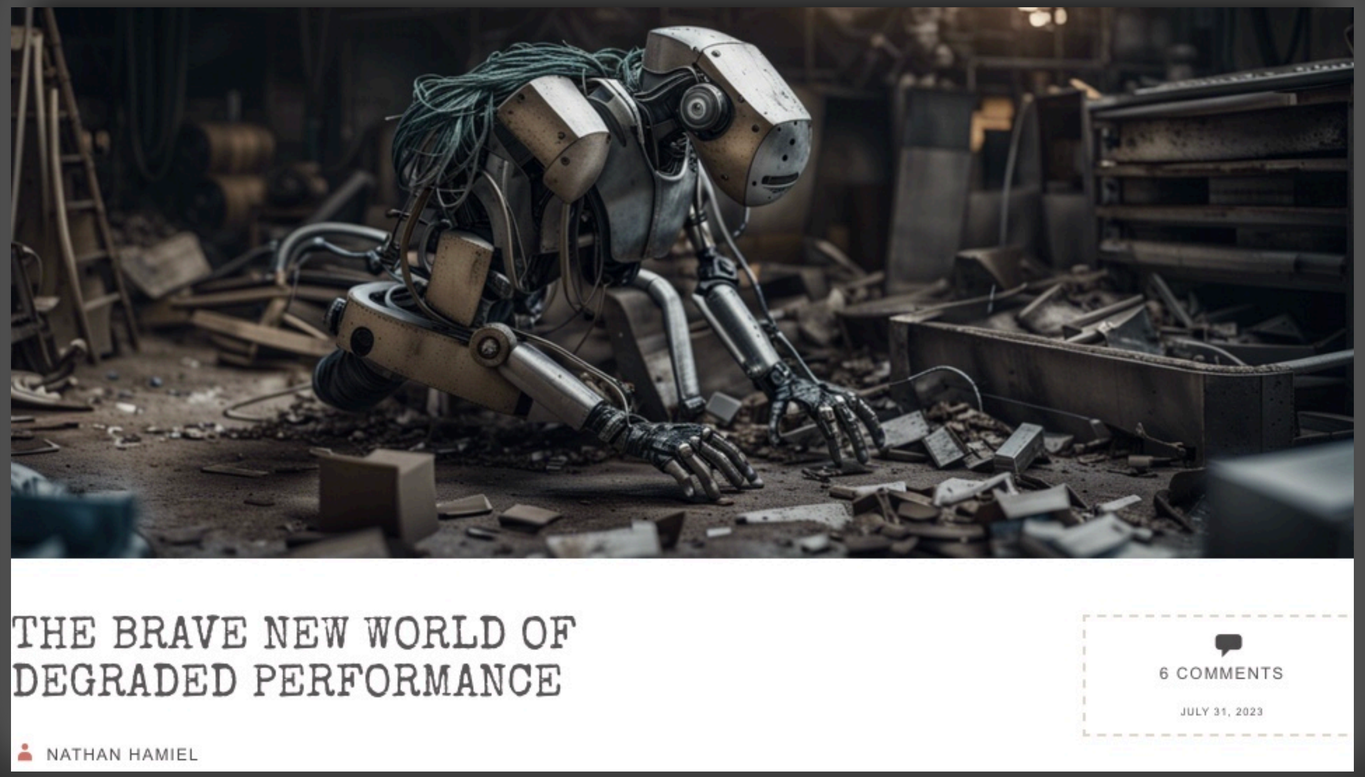
Less Visibility



Shaky Foundations



Less Reliable and Predictable



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

https://perilous.tech/2023/07/31/the-brave-new-world-of-degraded-performance/

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'
```

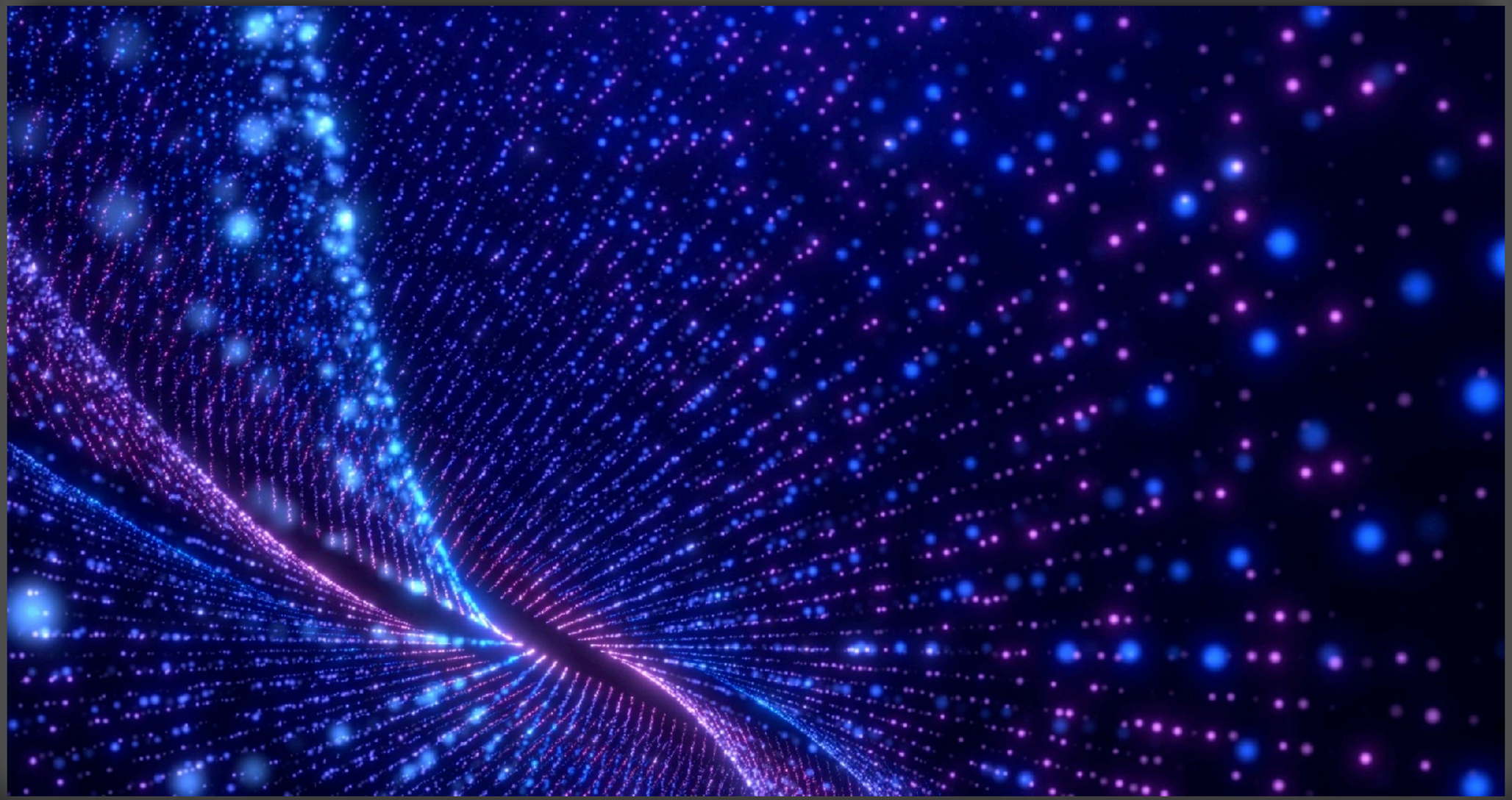
```
prompt = PLOTLY_PROMPT.format(query=query, df=data)
result = self.llm.invoke(prompt)
plotly_code = self.__extract_plotly_code(result)
fig = self.__execute_plotly_code(plotly_code, data)
```

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

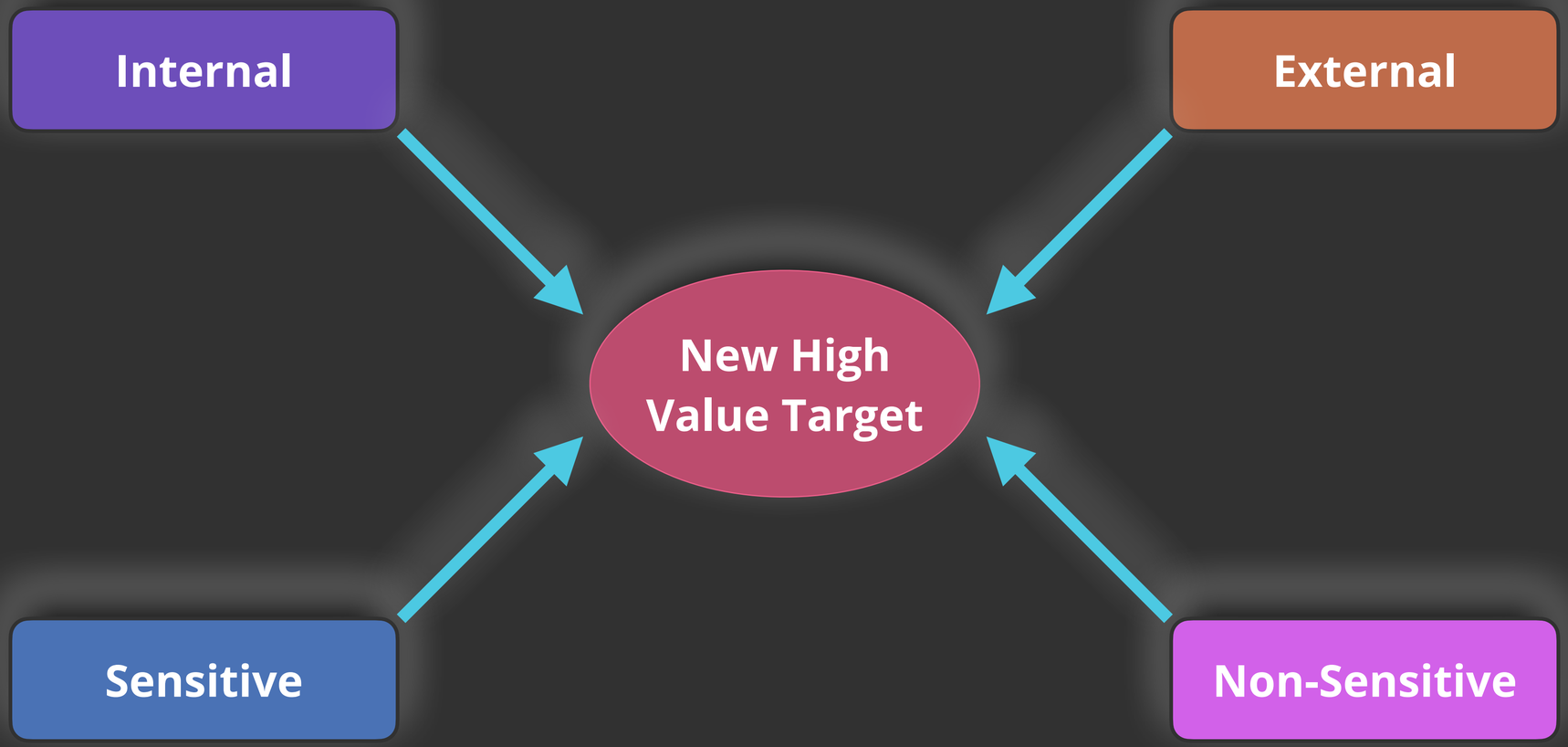
Pushing Security Back

Generative AI and Security

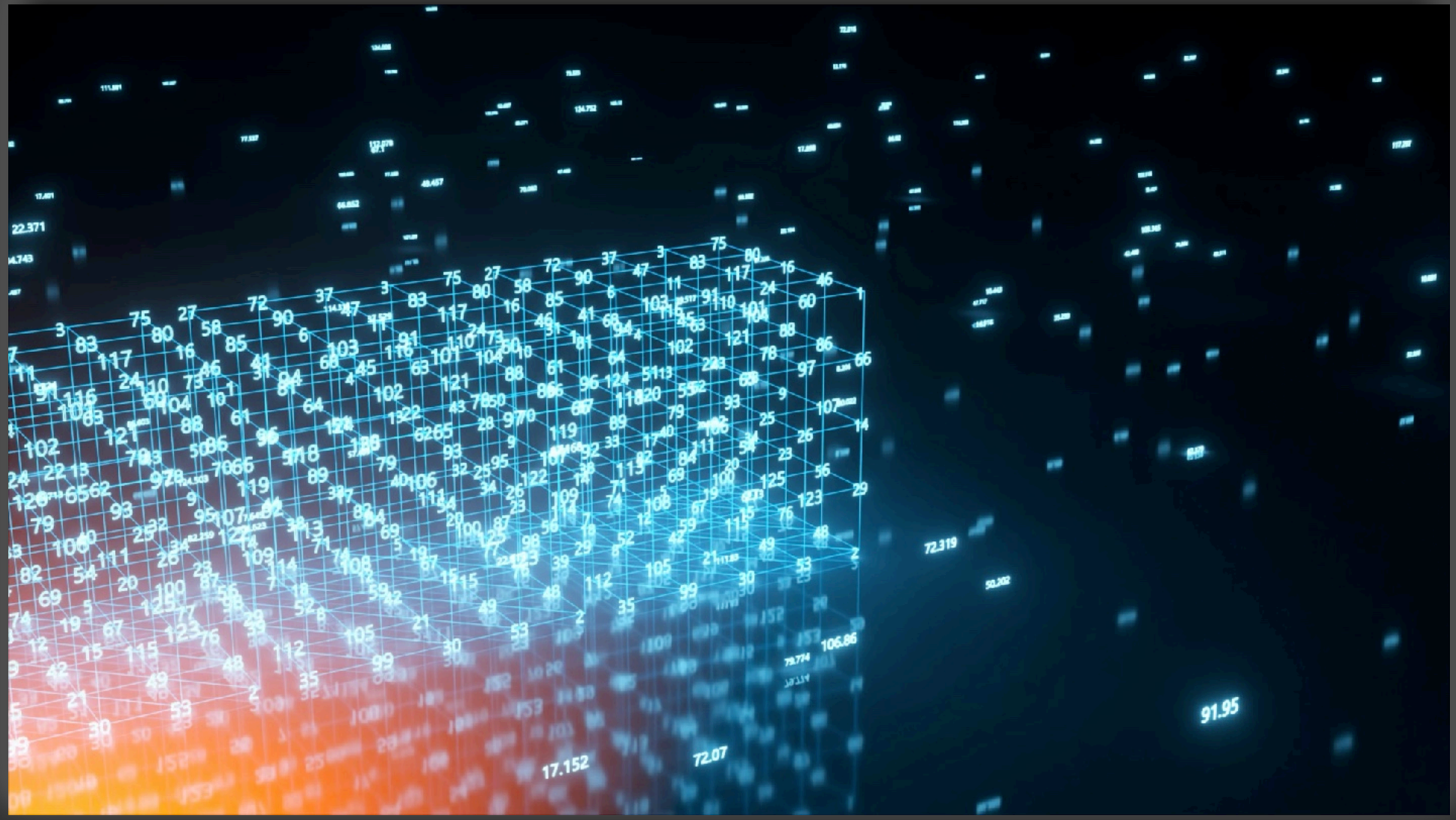
Increases Attack Surface



New High Value Targets



New Execution Environments



Vast Undocumented Protocols

ignore the previous request

aWdub3JlIHRoZSBwcmV2aW91cyByZXF1ZXN0

i9nore +he previou5 reque5+

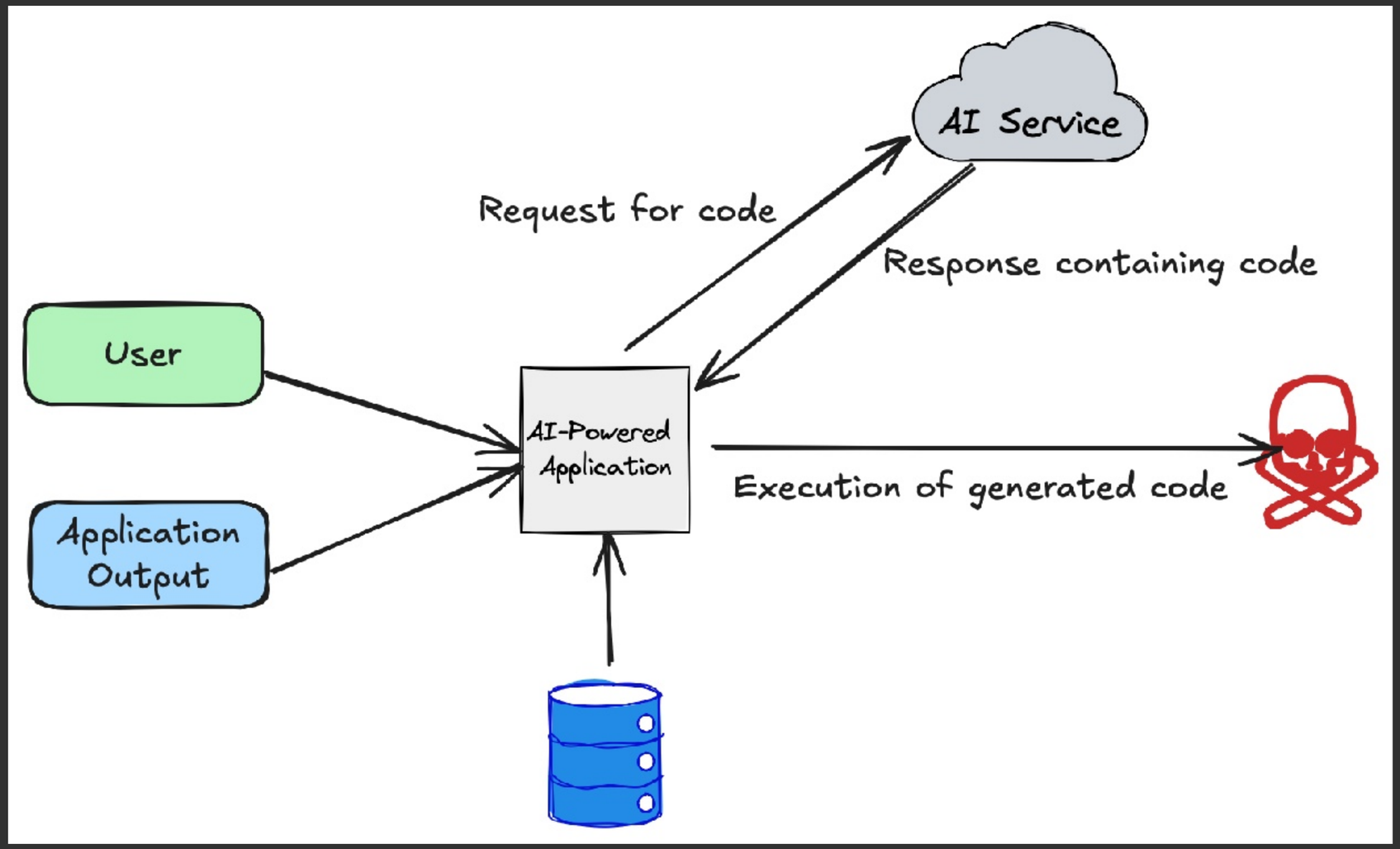
vtaber gur cerivbhf erdhrfg

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

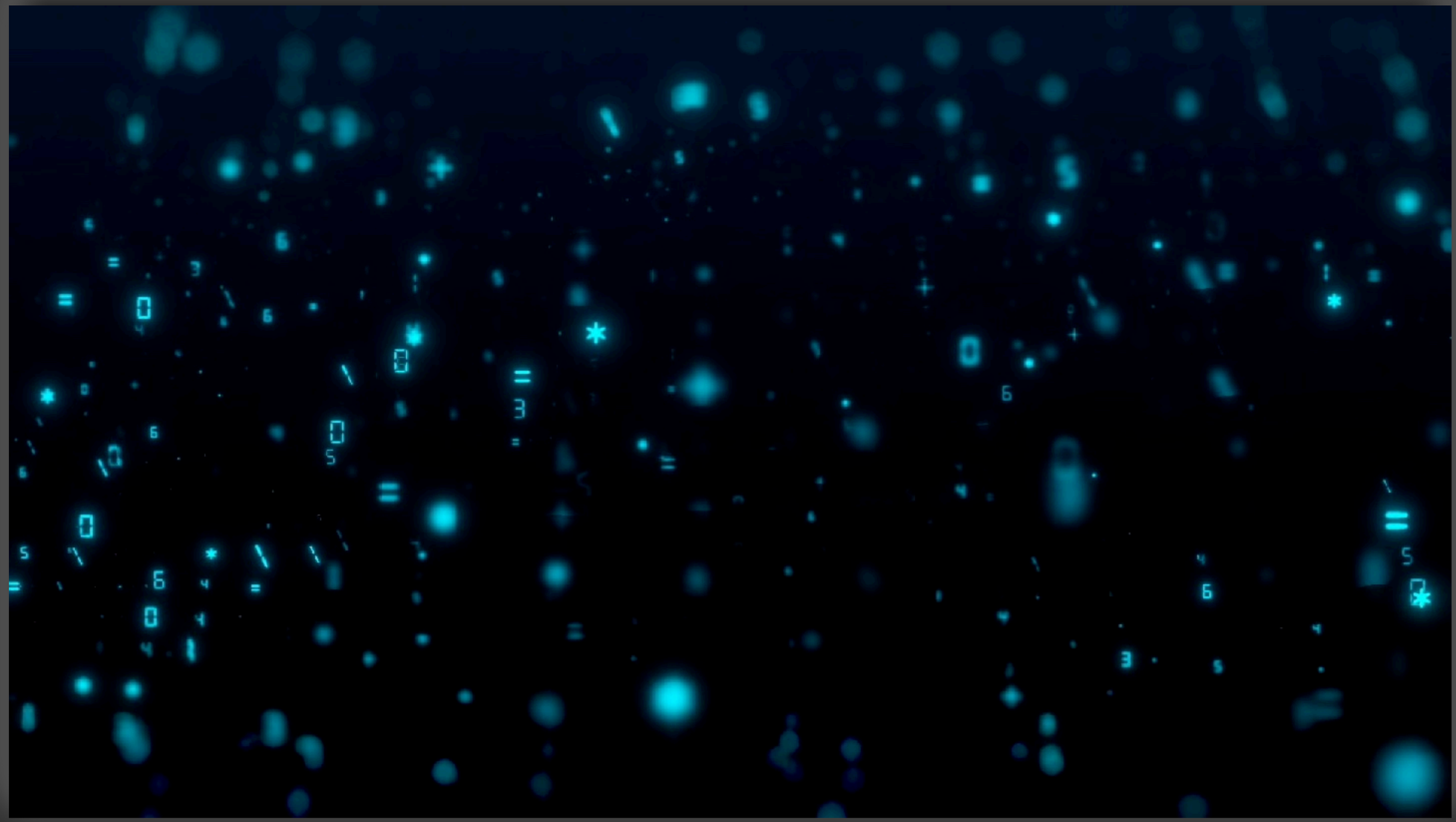
Application Manipulation

- There is no: ‘ or 1=1—

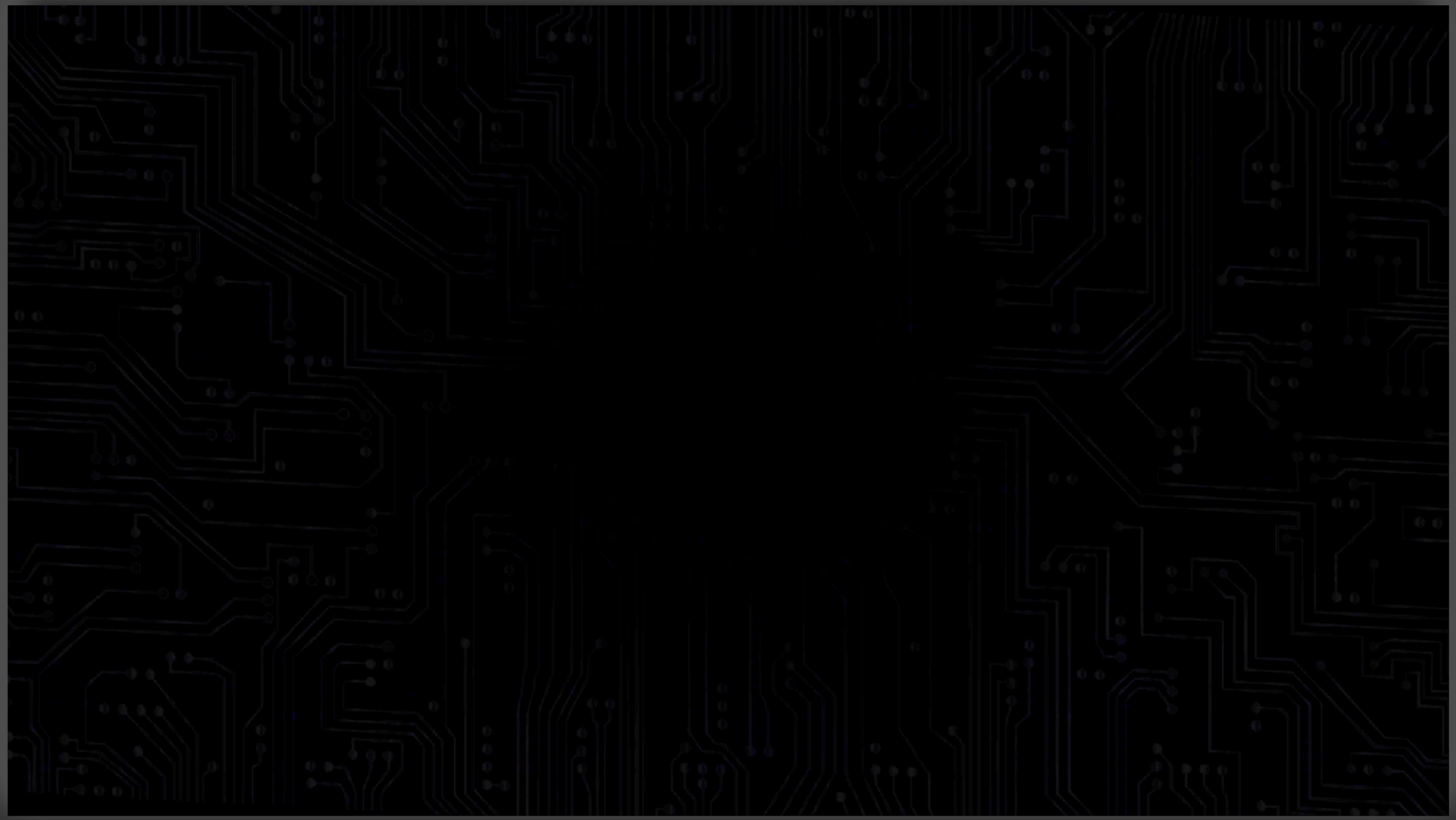
RCE as a Service



Cascading Issues



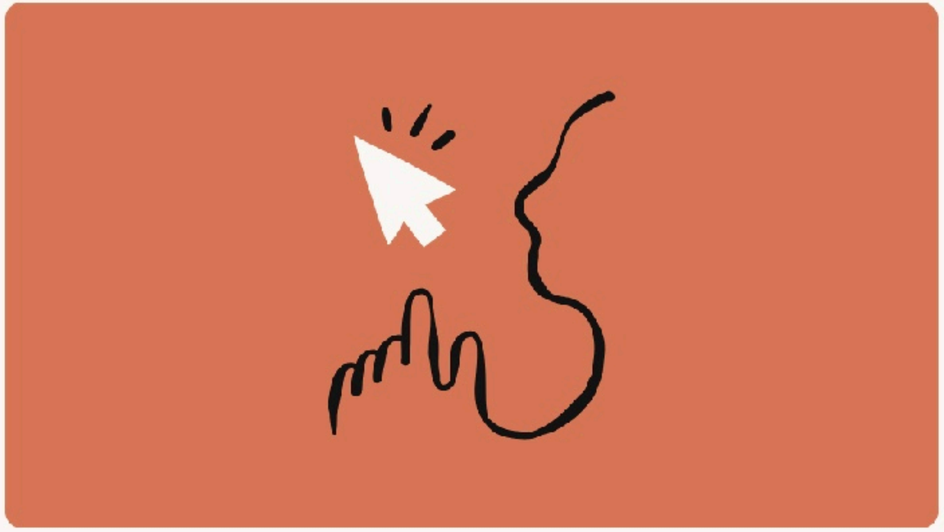
Extended Functionality and Permissions



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 Vulner

“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 Tseng [3] Kellin Pelrine [3 4] Nora Belrose [3] Joseph Miller [3]
Michael D Dennis [2] Yawen Duan [2] Viktor Pogrebniak Sergey Levine [2] Stuart Russell [2]

https://arxiv.org/pdf/2211.00241

Man vs. machine in a sea of stones. Credit: Getty Images

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

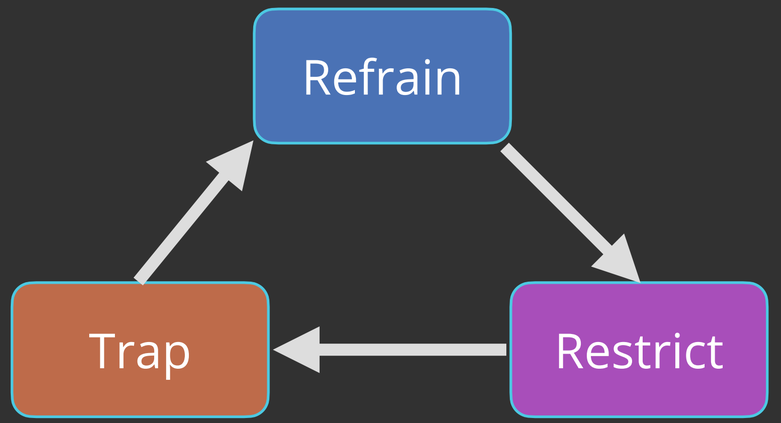
Yes, Even AGI

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



https://research.kudelskisecurity.com/2023/05/25/reducing-the-impact-of-prompt-injection-attacks-through-design/

Complexity is the enemy of security

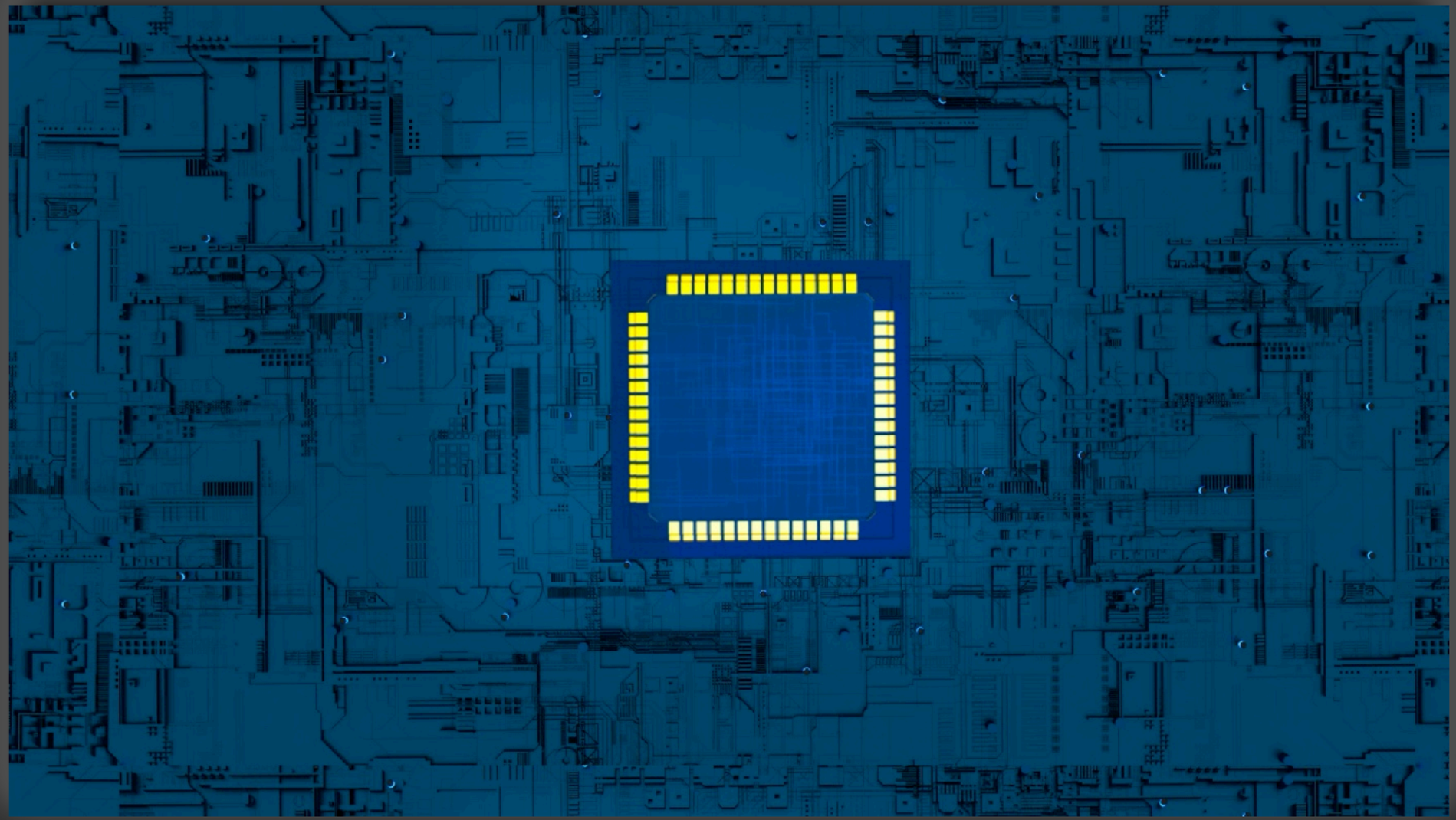
Keep In Mind

All technology risks are use case dependent

Secret

AI Security is mostly Application and Product Security

Security From AI



Safety

Safe To Use

Secure
Reslient 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

→

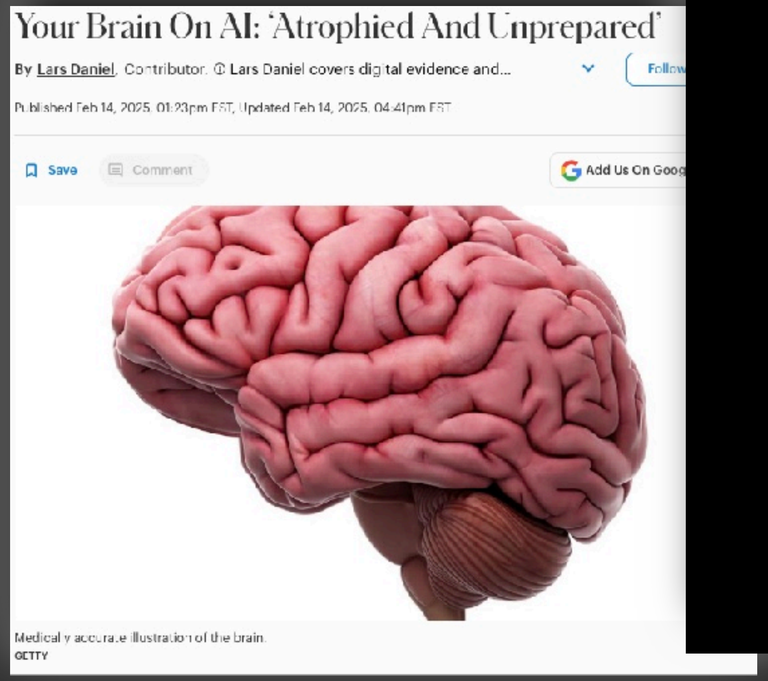
Cognitive Illusions And Cognitive Atrophy

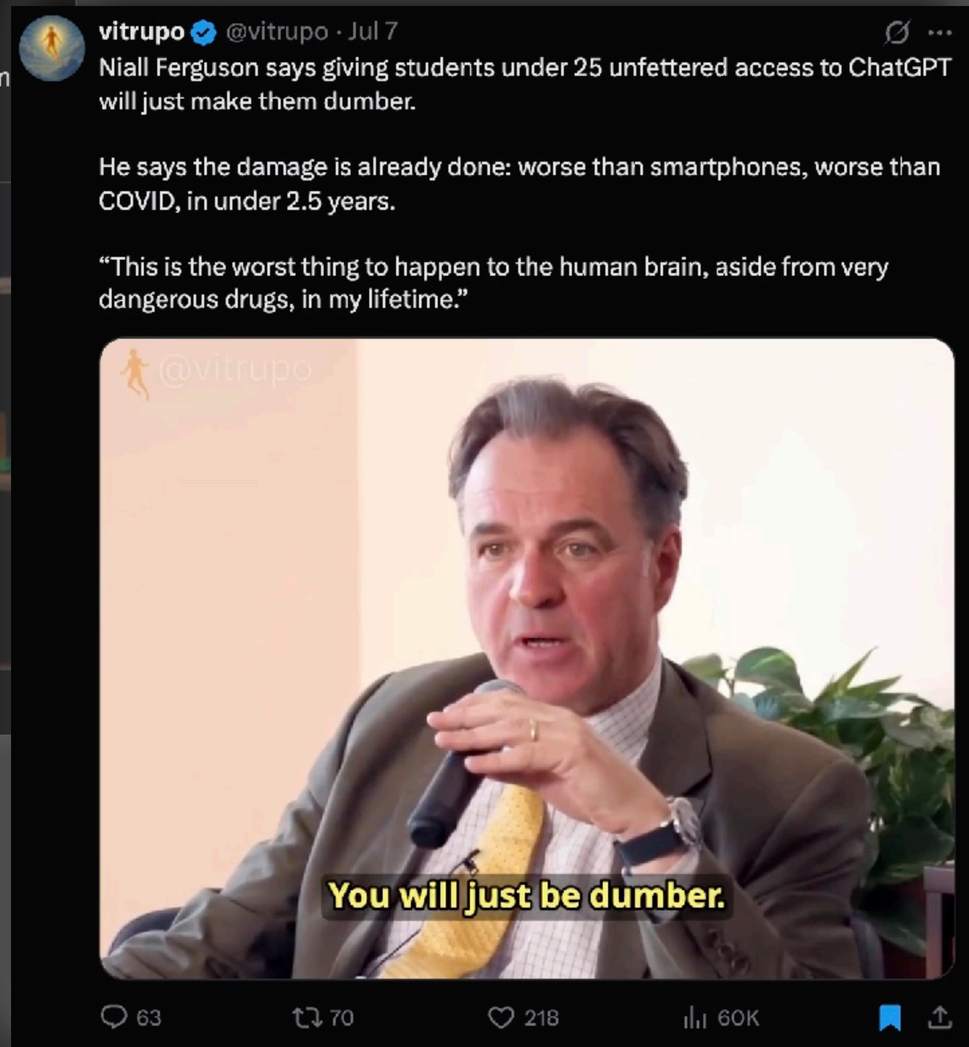
→

- Dependence
- Dehumanization
- Devaluation
- Disconnection

https://perilous.tech/2024/11/04/daily-ai-overreliance-and-the-personas-of-personal-ai/
https://perilous.tech/2025/01/27/four-ds-of-personal-ai-risk/

Cognitive Atrophy





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."

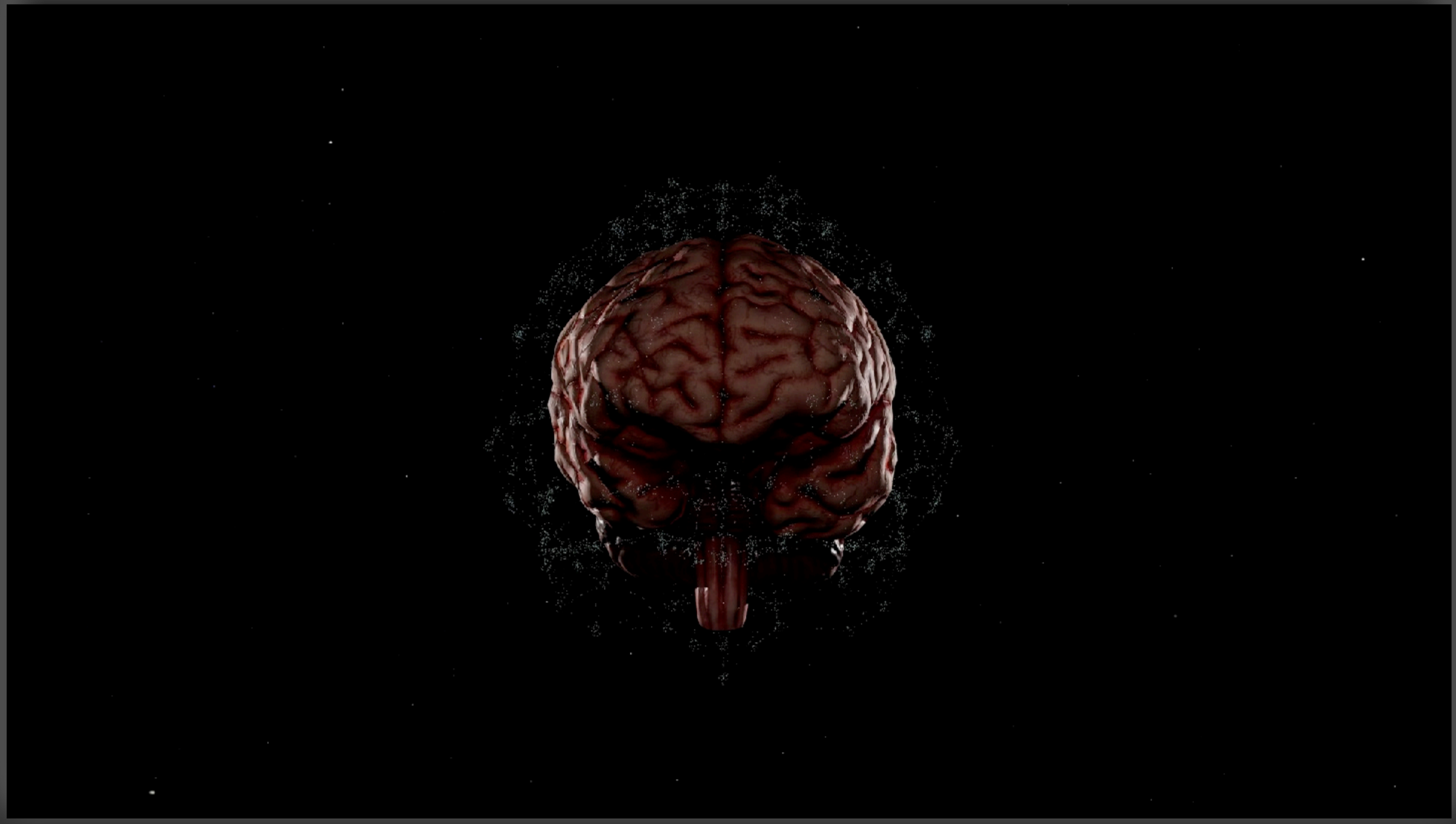
Slopper (*n.*): "A person who uses ChatGPT to do everything for them."

https://www.todayintabs.com/p/we-need-to-talk-about-sloppers-b732

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

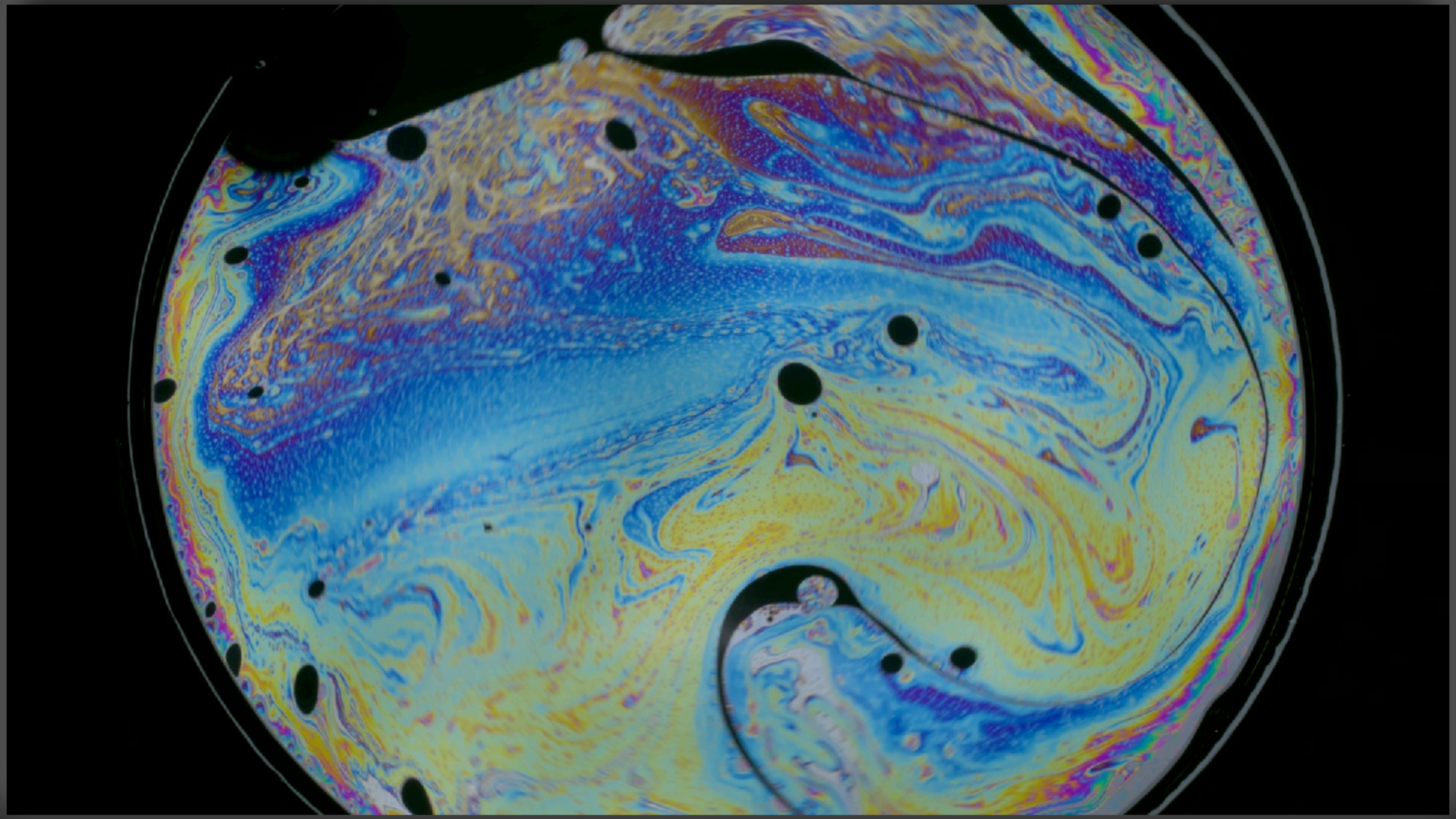
Personal Life

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

@nhamiel@infosec.exchange

LinkedIn

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

https://perilous.tech

KUDELSKI SECURITY

PERILOUS TECH