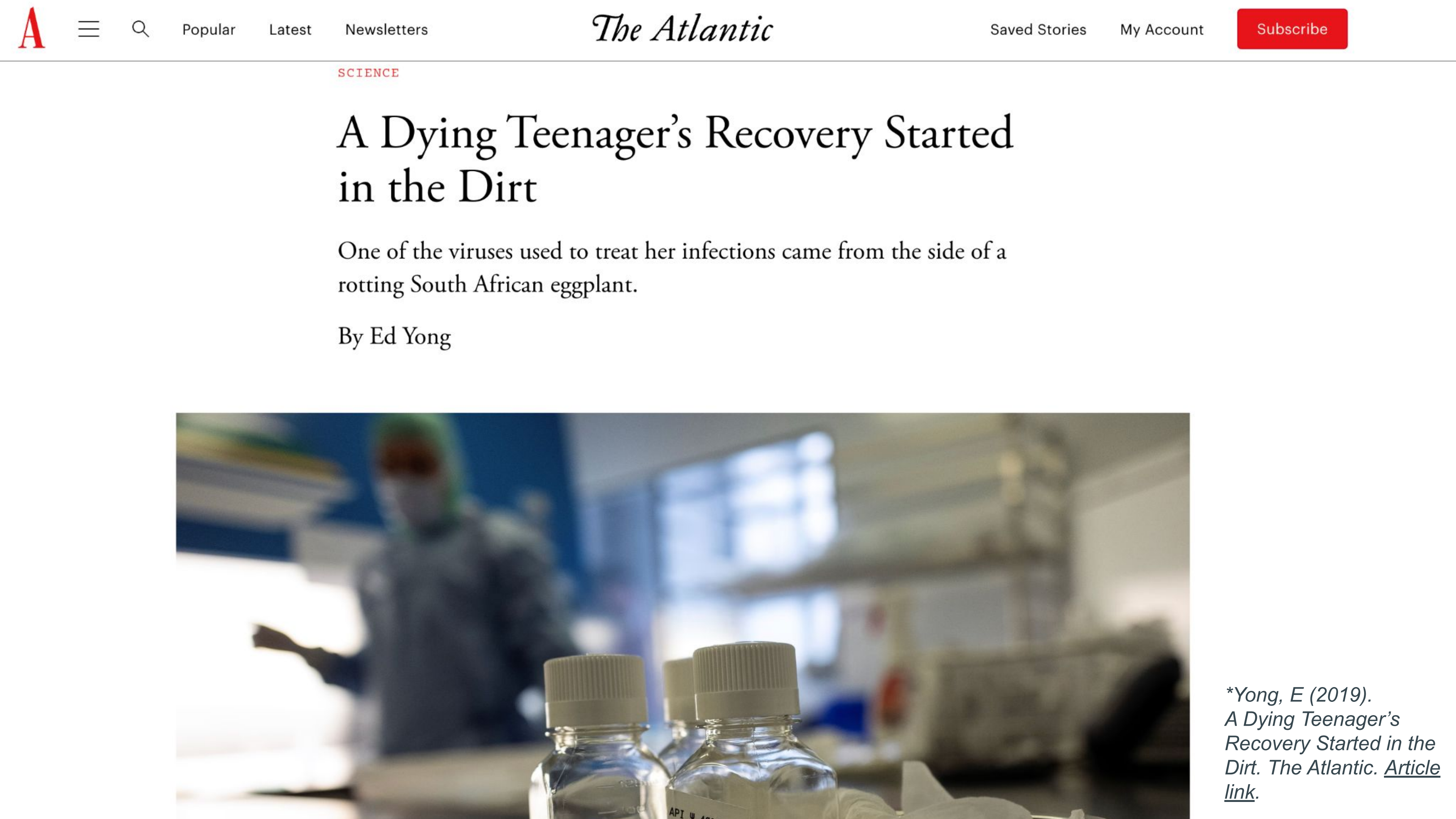
Violet teaming and Strategic Cooperation on AI-Driven Innovation

•••

Alexander Titus, PhD

GUIRR Meeting @ the National Academies

11 October 2023



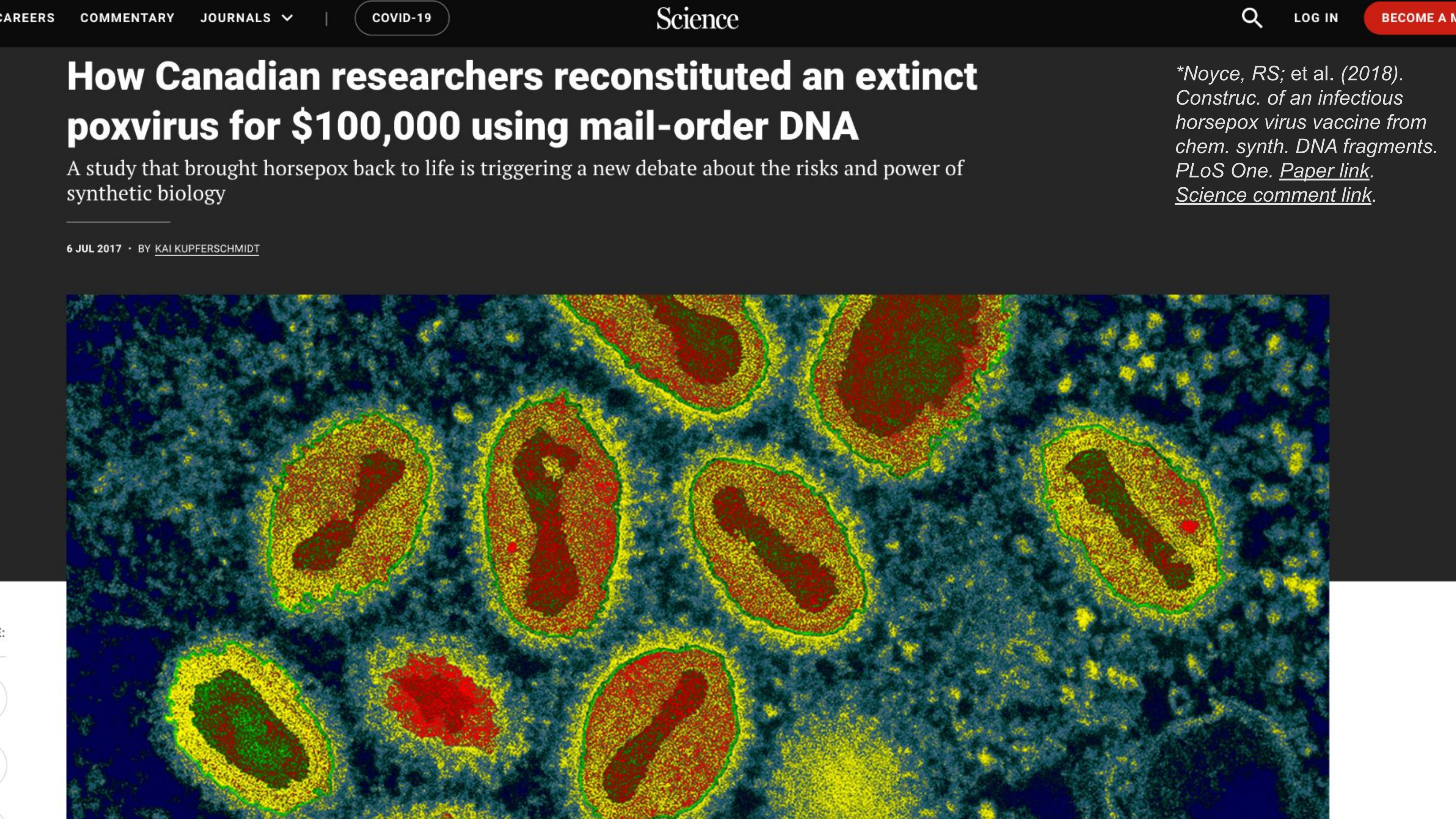
SCIENCE

A Dying Teenager's Recovery Started in the Dirt

One of the viruses used to treat her infections came from the side of a rotting South African eggplant.

By Ed Yong

Yong, E (2019). A Dying Teenager's Recovery Started in the Dirt. The Atlantic. Article link.



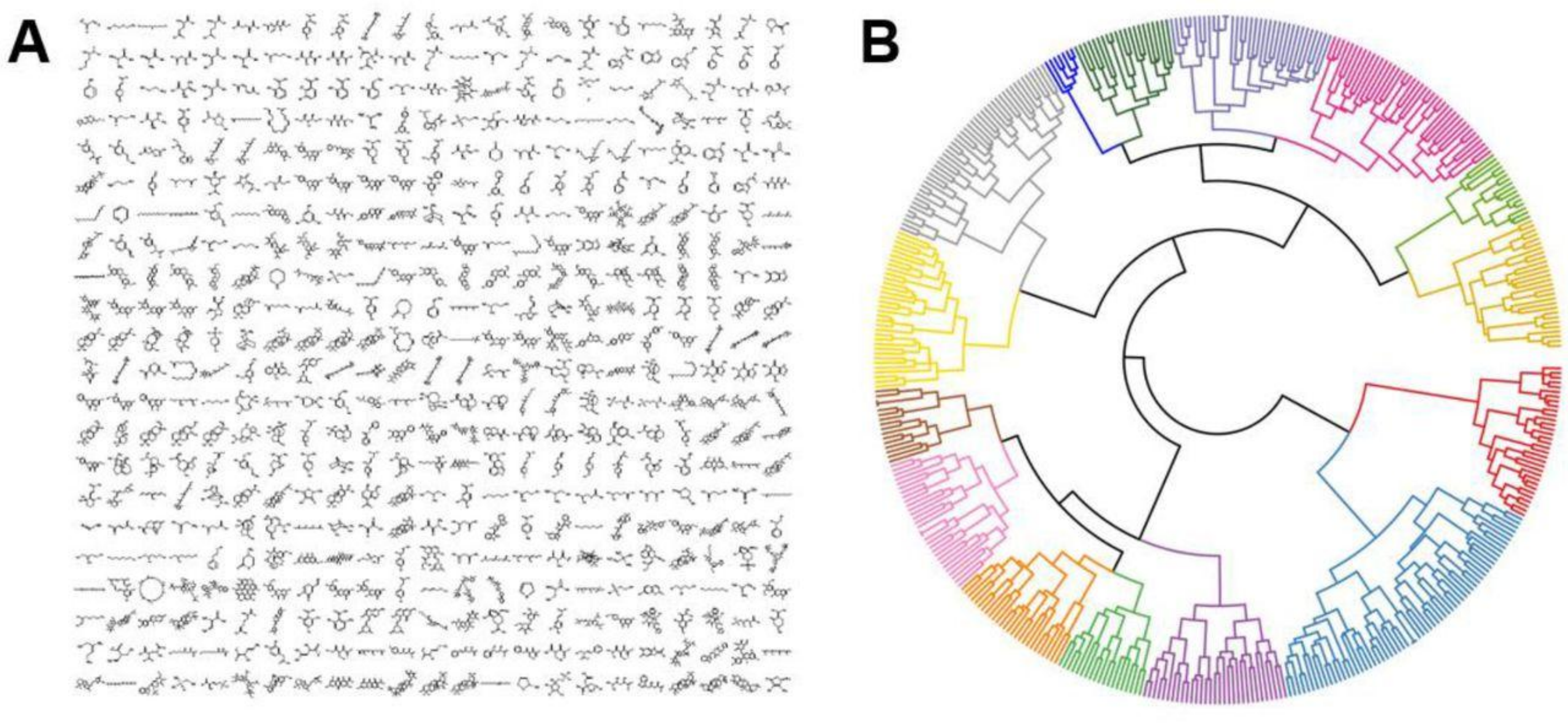
How Canadian researchers reconstituted an extinct poxvirus for $100,000 using mail-order DNA

A study that brought horsepox back to life is triggering a new debate about the risks and power of synthetic biology

6 JUL 2017 • BY KAI KUPFERSCHMIDT

**Noyce, RS;* et al. *(2018). Construc. of an infectious horsepox virus vaccine from chem. synth. DNA fragments. PLoS One. Paper link. Science comment link.*

An Automated Scientist to Design and Optimize Microbial Strains



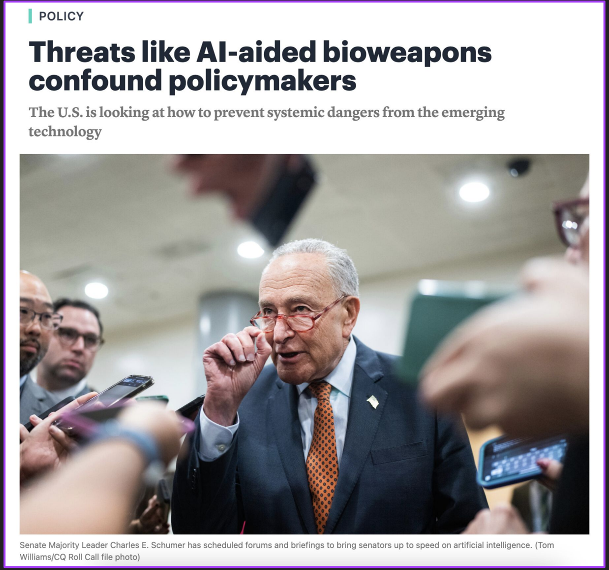
*Singh, AH, et al. (2023). An Automated Scientist to Design and Optimize Microbial Strains for the Industrial Production of Small Molecules. bioRxiv. https://doi.org/10.1101/2023.01.03.521657.

Synthetic Biology, AI, & Biosecurity Concerns

HELENA

Biosecurity in the Age of AI Chairperson's Statement

Convened by Helena at The Rockefeller Foundation's Bellagio Center | July 2023
Chairperson: The Hon. Mark Dybul, MD



POLICY

Threats like AI-aided bioweapons confound policymakers

The U.S. is looking at how to prevent systemic dangers from the emerging technology

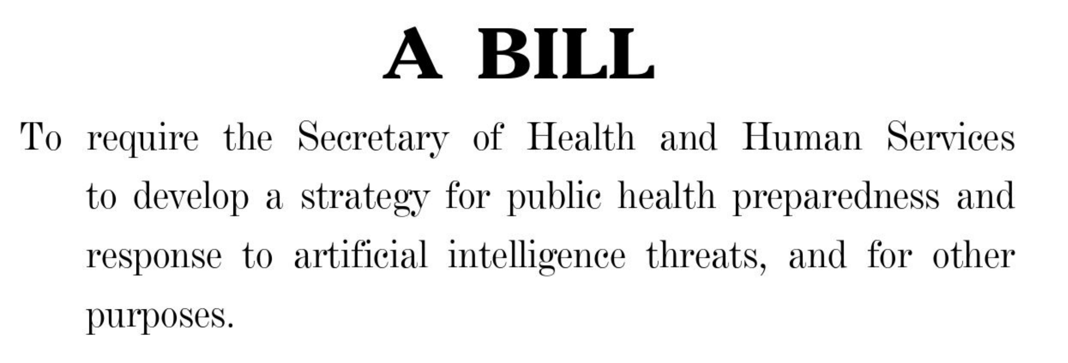
Senate Majority Leader Charles E. Schumer has scheduled forums and briefings to bring senators up to speed on artificial intelligence. (Tom Williams/CQ Roll Call file photo)

Synthetic Biology, AI, & Biosecurity Concerns

HELENA

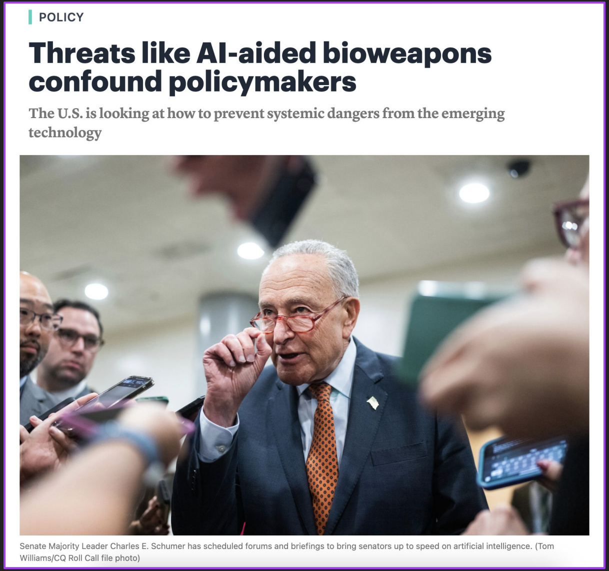
Biosecurity in the Age of AI Chairperson's Statement

Convened by Helena at The Rockefeller Foundation's Bellagio Center | July 2023
Chairperson: The Hon. Mark Dybul, MD



A BILL

To require the Secretary of Health and Human Services to develop a strategy for public health preparedness and response to artificial intelligence threats, and for other purposes.



POLICY

Threats like AI-aided bioweapons confound policymakers

The U.S. is looking at how to prevent systemic dangers from the emerging technology

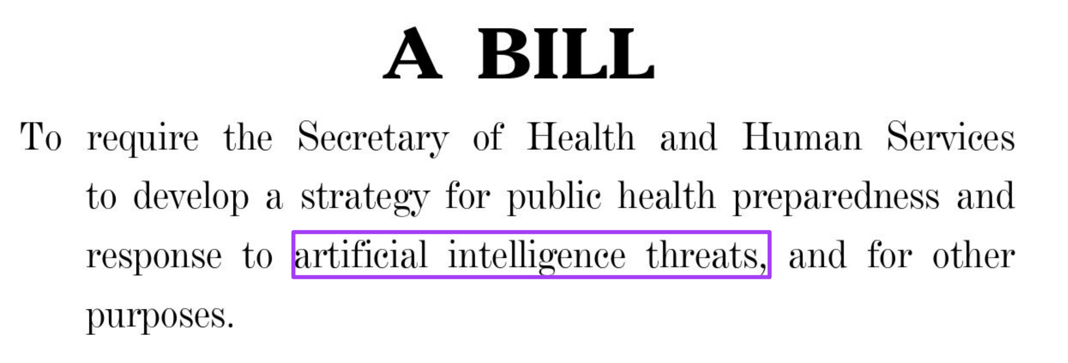
Senate Majority Leader Charles E. Schumer has scheduled forums and briefings to bring senators up to speed on artificial intelligence. (Tom Williams/CQ Roll Call file photo)

Synthetic Biology, AI, & Biosecurity Concerns

HELENA

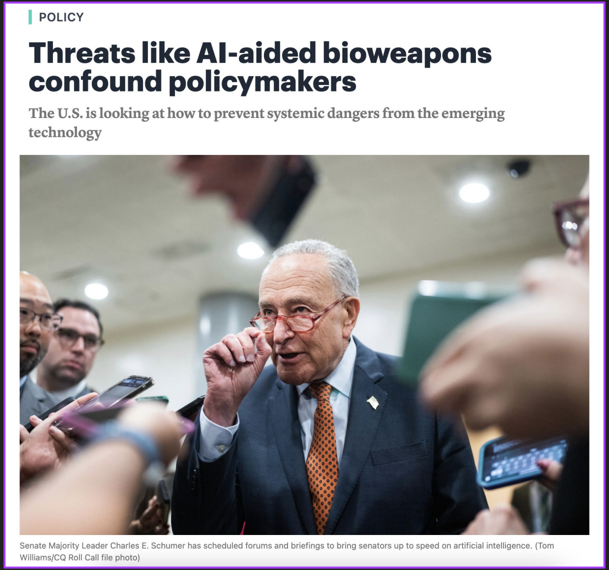
Biosecurity in the Age of AI Chairperson's Statement

Convened by Helena at The Rockefeller Foundation's Bellagio Center | July 2023
Chairperson: The Hon. Mark Dybul, MD



A BILL

To require the Secretary of Health and Human Services to develop a strategy for public health preparedness and response to artificial intelligence threats, and for other purposes.



POLICY

Threats like AI-aided bioweapons confound policymakers

The U.S. is looking at how to prevent systemic dangers from the emerging technology

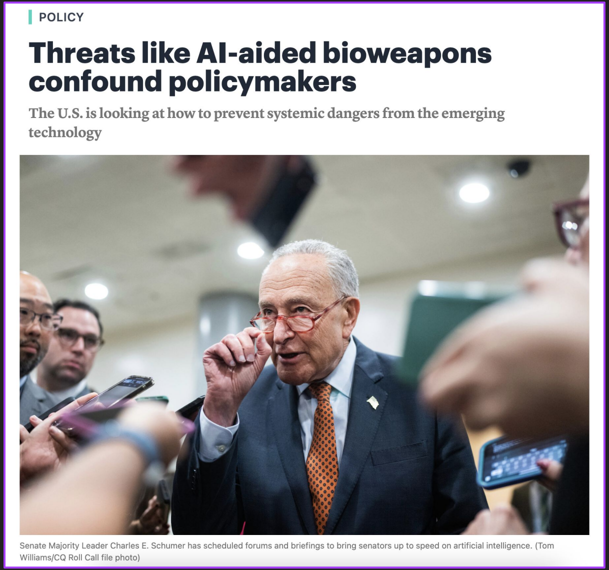
Senate Majority Leader Charles E. Schumer has scheduled forums and briefings to bring senators up to speed on artificial intelligence. (Tom Williams/CQ Roll Call file photo)

Synthetic Biology, AI, & Biosecurity Concerns

HELENA

Biosecurity in the Age of AI Chairperson's Statement

Convened by Helena at The Rockefeller Foundation's Bellagio Center | July 2023
Chairperson: The Hon. Mark Dybul, MD

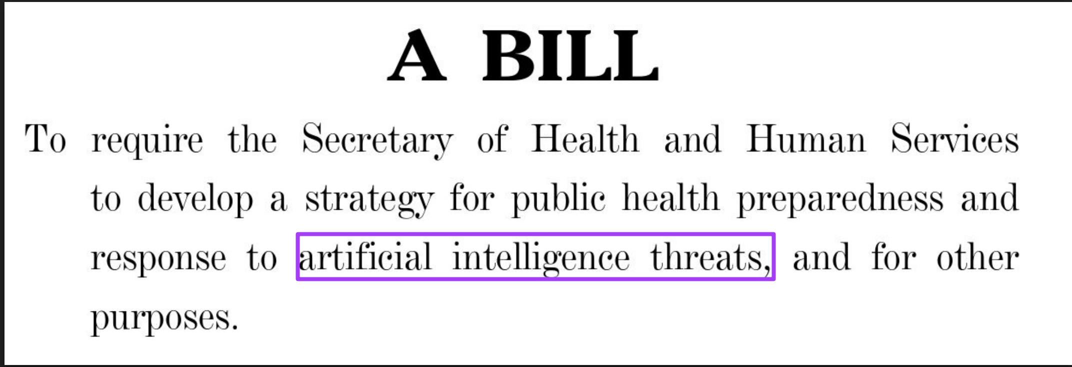


POLICY

Threats like AI-aided bioweapons confound policymakers

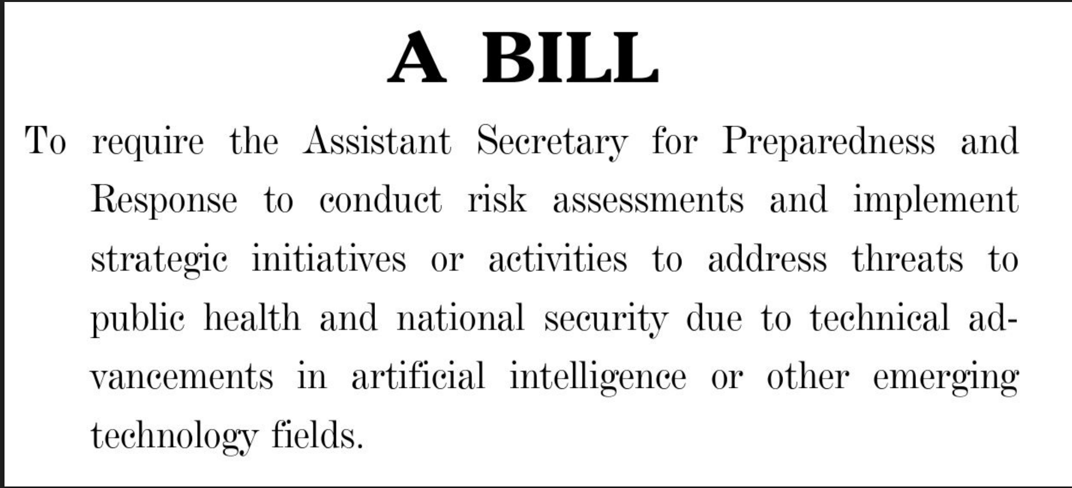
The U.S. is looking at how to prevent systemic dangers from the emerging technology

Senate Majority Leader Charles E. Schumer has scheduled forums and briefings to bring senators up to speed on artificial intelligence. (Tom Williams/CQ Roll Call file photo)



A BILL

To require the Secretary of Health and Human Services to develop a strategy for public health preparedness and response to artificial intelligence threats, and for other purposes.



A BILL

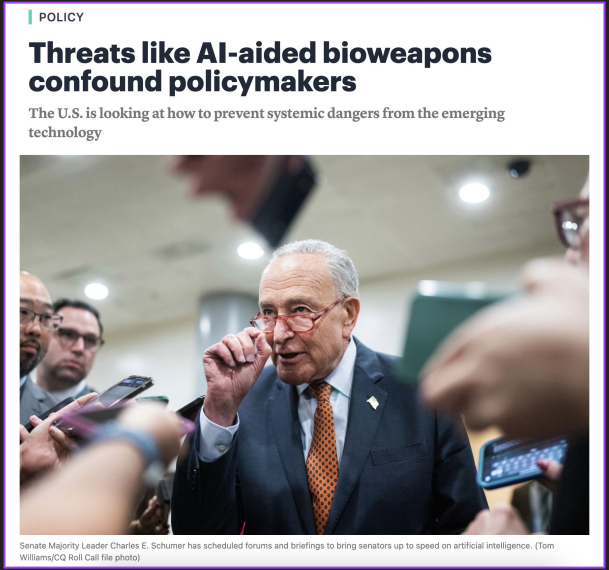
To require the Assistant Secretary for Preparedness and Response to conduct risk assessments and implement strategic initiatives or activities to address threats to public health and national security due to technical advancements in artificial intelligence or other emerging technology fields.

Synthetic Biology, AI, & Biosecurity Concerns

HELENA

Biosecurity in the Age of AI Chairperson's Statement

Convened by Helena at The Rockefeller Foundation's Bellagio Center | July 2023
Chairperson: The Hon. Mark Dybul, MD

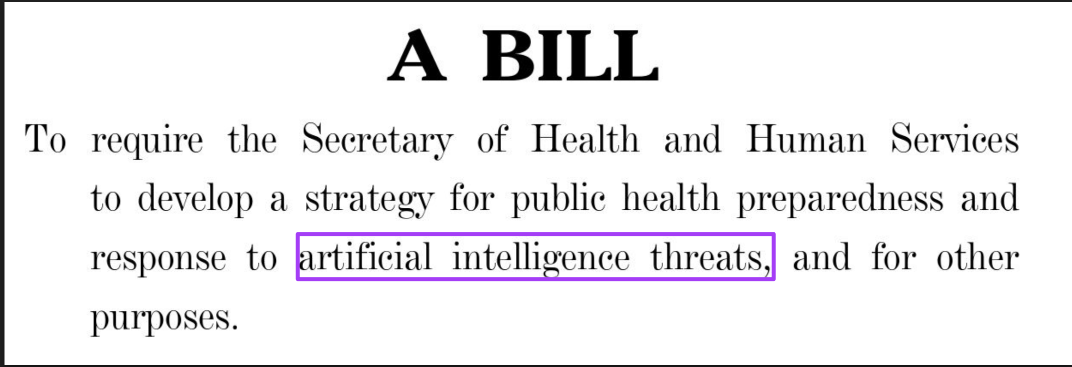


POLICY

Threats like AI-aided bioweapons confound policymakers

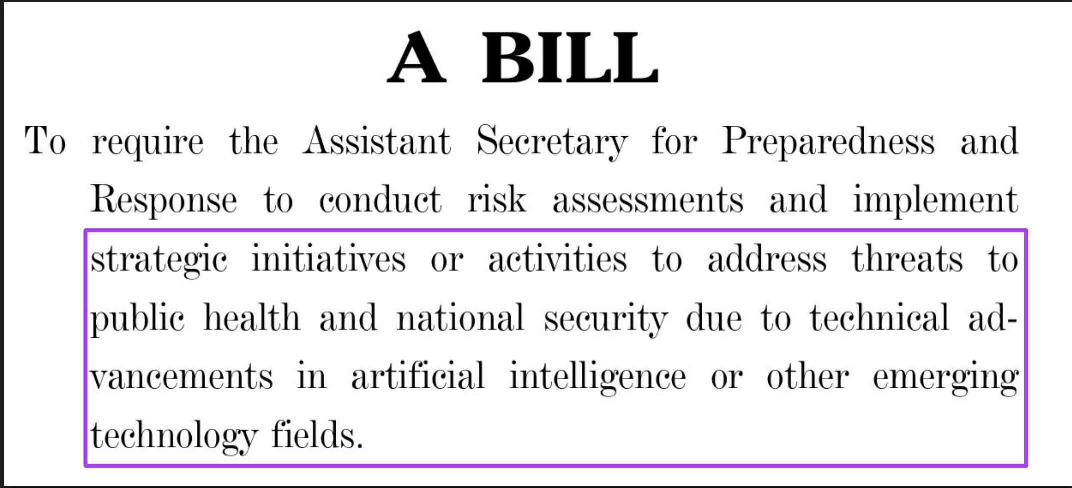
The U.S. is looking at how to prevent systemic dangers from the emerging technology

Senate Majority Leader Charles E. Schumer has scheduled forums and briefings to bring senators up to speed on artificial intelligence. (Tom Williams/CQ Roll Call file photo)



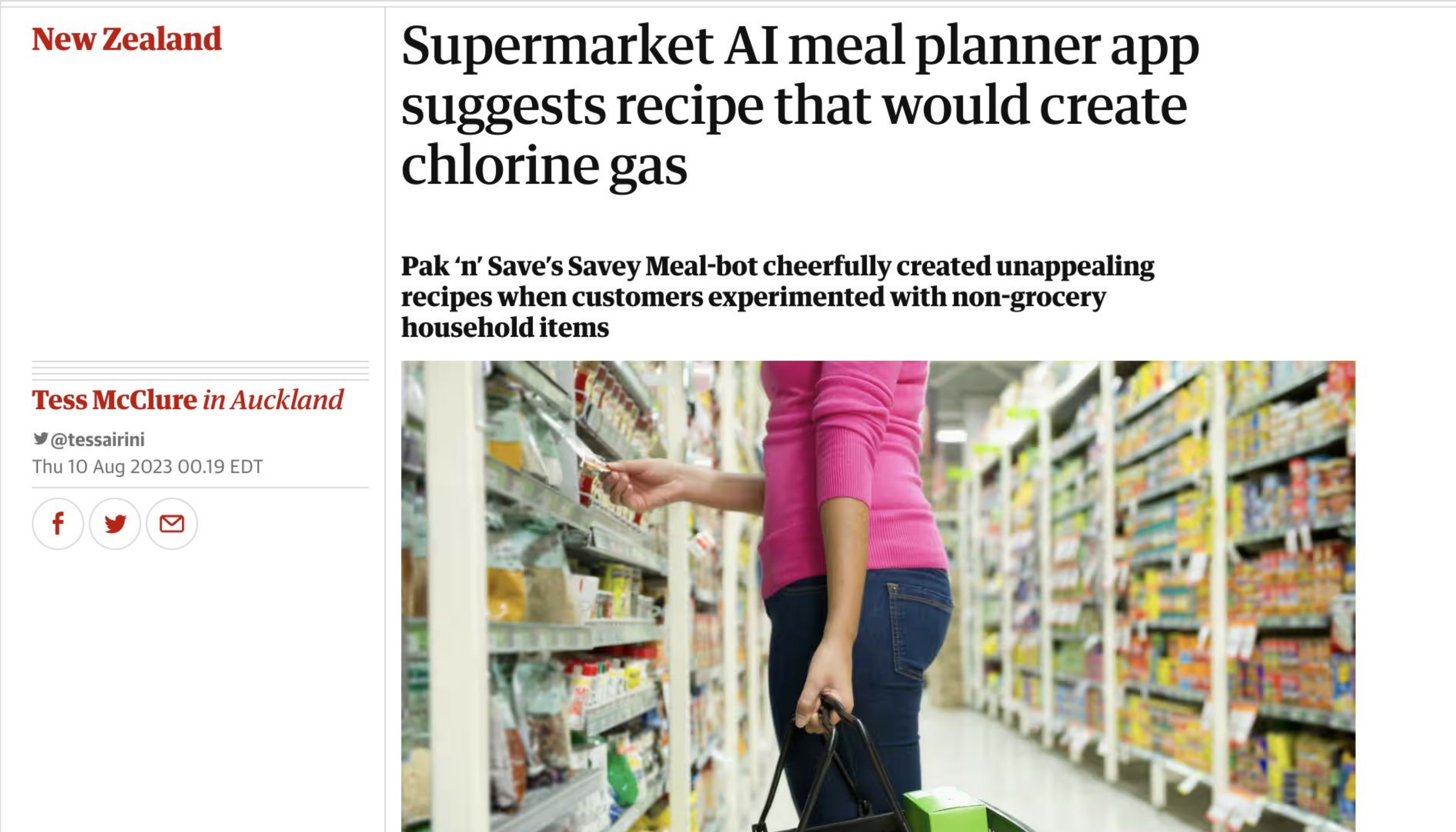
A BILL

To require the Secretary of Health and Human Services to develop a strategy for public health preparedness and response to artificial intelligence threats, and for other purposes.



A BILL

To require the Assistant Secretary for Preparedness and Response to conduct risk assessments and implement strategic initiatives or activities to address threats to public health and national security due to technical advancements in artificial intelligence or other emerging technology fields.



New Zealand

Supermarket AI meal planner app suggests recipe that would create chlorine gas

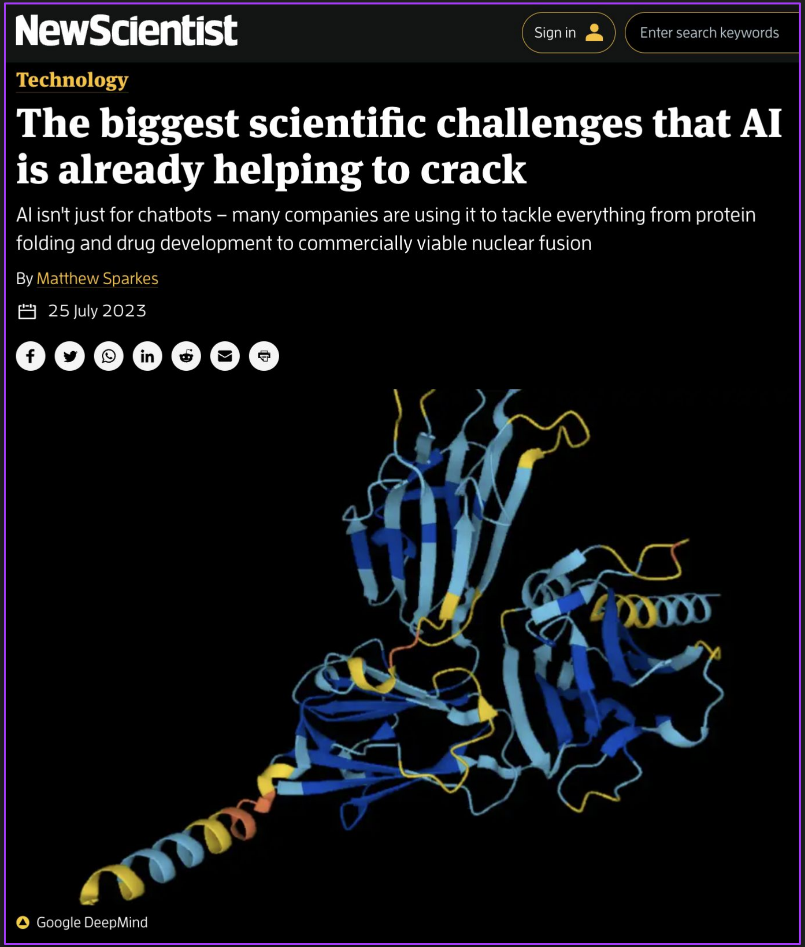
Pak 'n' Save's Savey Meal-bot cheerfully created unappealing recipes when customers experimented with non-grocery household items

Tess McClure *in Auckland*

@tessairini

Thu 10 Aug 2023 00.19 EDT

AI is improving our future faster every day



Technology

The biggest scientific challenges that AI is already helping to crack

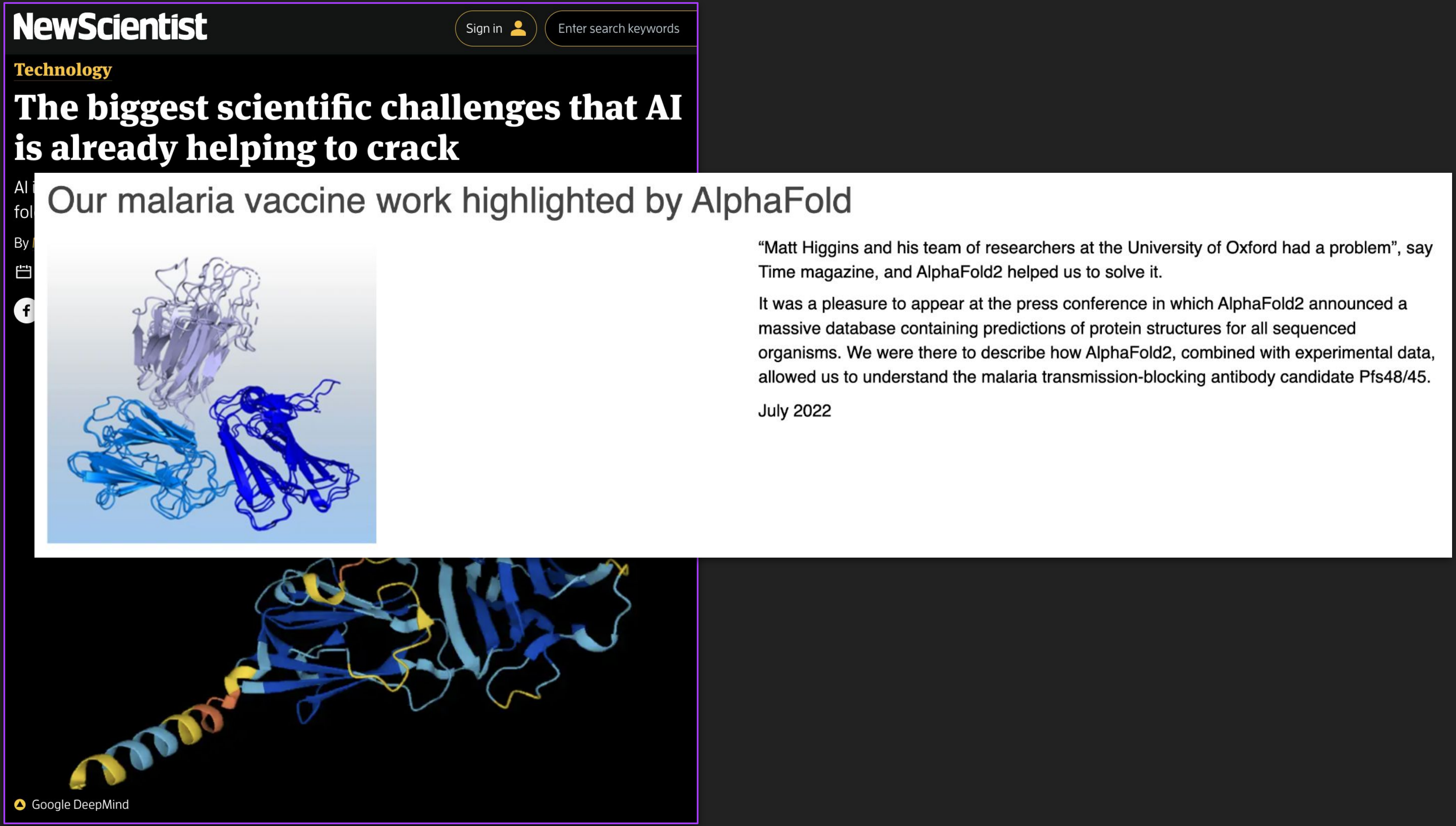
AI isn't just for chatbots – many companies are using it to tackle everything from protein folding and drug development to commercially viable nuclear fusion

By Matthew Sparkes

25 July 2023

Google DeepMind

AI is improving our future faster every day



Technology

The biggest scientific challenges that AI is already helping to crack

AI i
fol

By

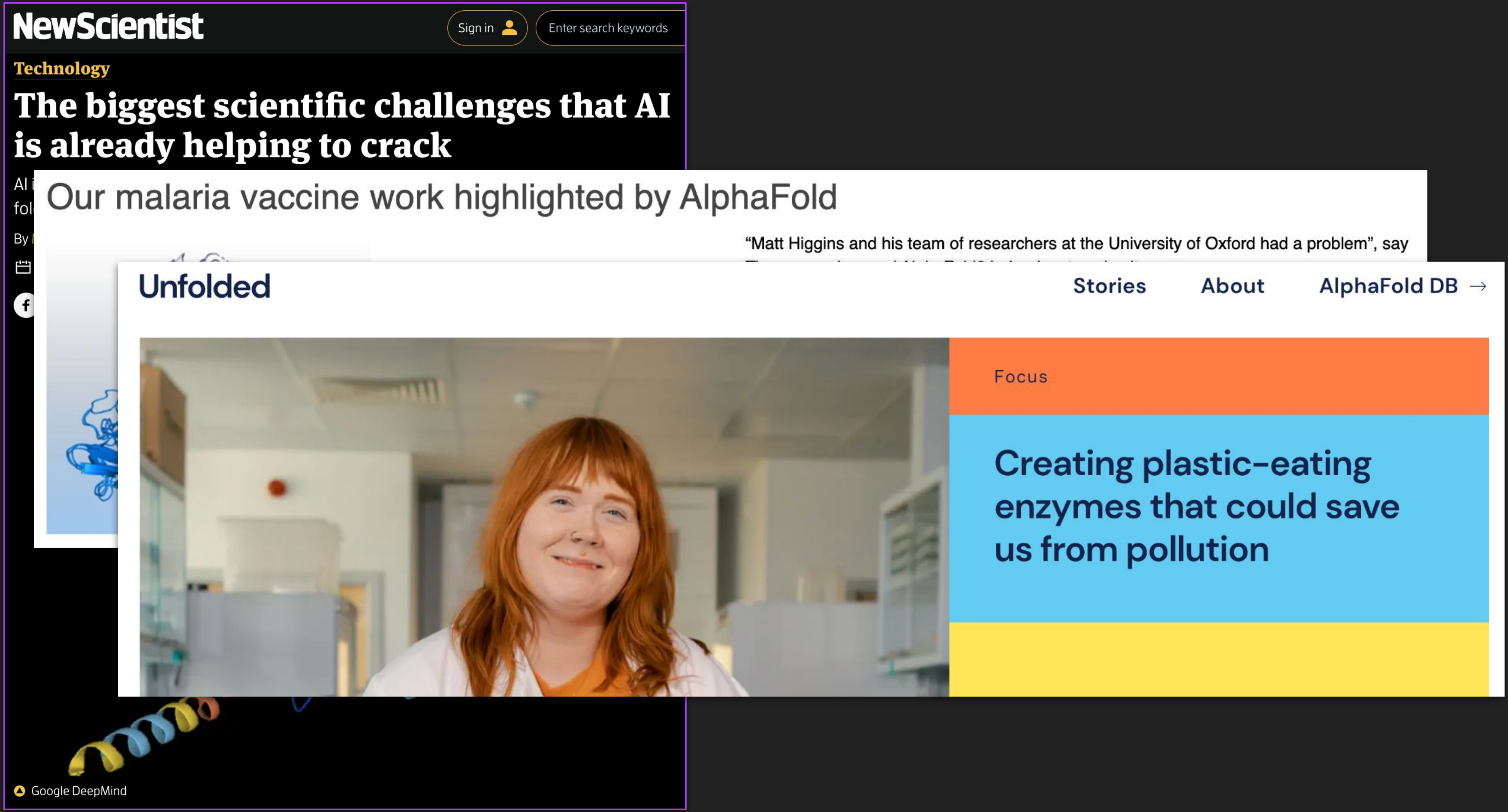
Our malaria vaccine work highlighted by AlphaFold

“Matt Higgins and his team of researchers at the University of Oxford had a problem”, say Time magazine, and AlphaFold2 helped us to solve it.

It was a pleasure to appear at the press conference in which AlphaFold2 announced a massive database containing predictions of protein structures for all sequenced organisms. We were there to describe how AlphaFold2, combined with experimental data, allowed us to understand the malaria transmission-blocking antibody candidate Pfs48/45.

July 2022

AI is improving our future faster every day



Technology

The biggest scientific challenges that AI is already helping to crack

AI i

fol

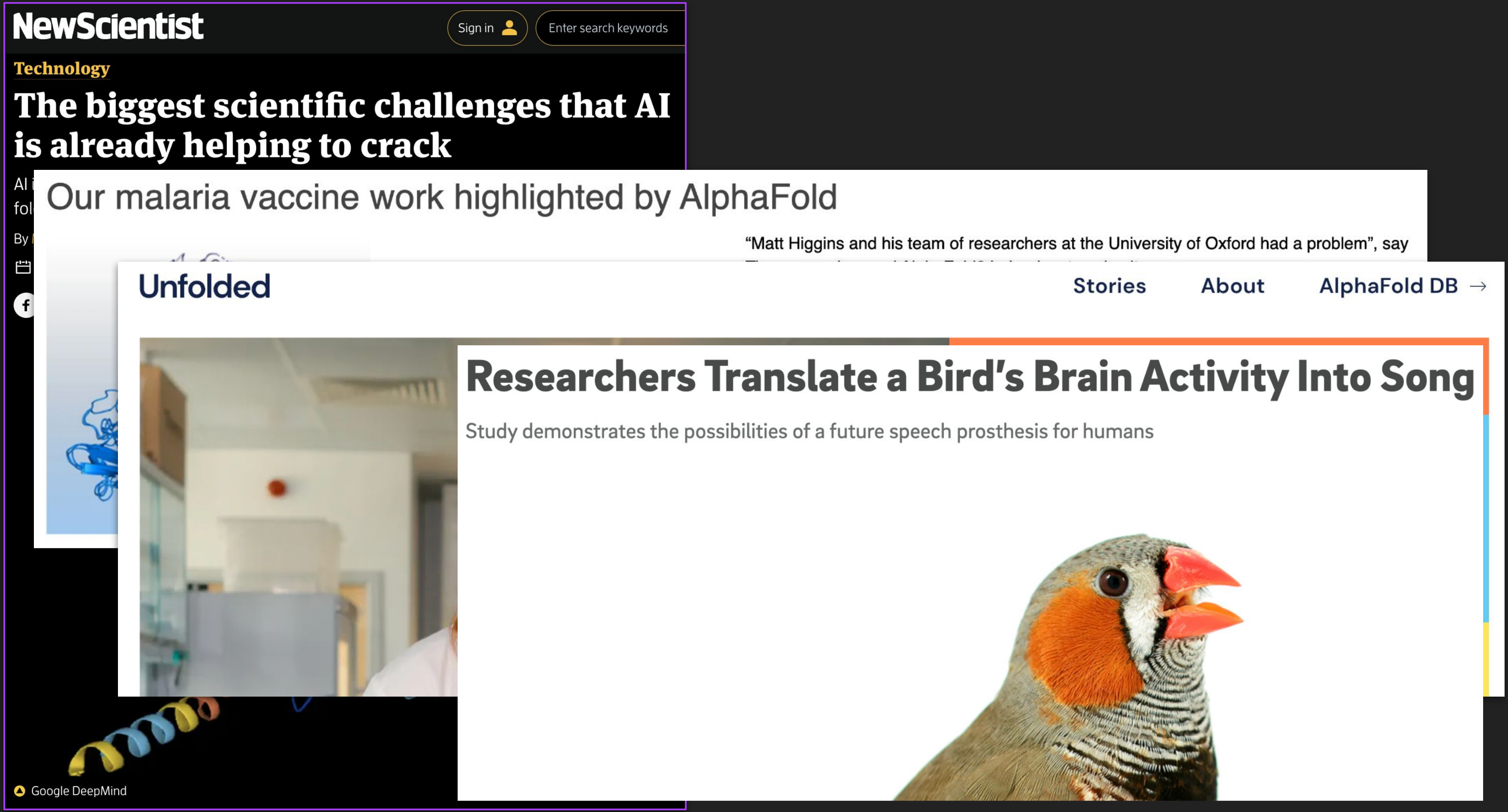
By

Our malaria vaccine work highlighted by AlphaFold

“Matt Higgins and his team of researchers at the University of Oxford had a problem”, say

Creating plastic-eating enzymes that could save us from pollution

AI is improving our future faster every day



Technology

The biggest scientific challenges that AI is already helping to crack

AI i

fol

By

Our malaria vaccine work highlighted by AlphaFold

“Matt Higgins and his team of researchers at the University of Oxford had a problem”, say

Researchers Translate a Bird’s Brain Activity Into Song

Study demonstrates the possibilities of a future speech prosthesis for humans

Traditional red, blue, & purple teams are only about the technology

Red Teaming: Be the threat

- Vulnerability/risk assessment
- Recon/gauge the attack surface
- Get initial access/compromise
- Conduct tactics, techniques, & procedure (TTP) operations

Blue Teaming: Stop the threat

- Implement controls/logging
- Monitor/intrusion detection
- Handle incident response and analysis
- Do patch management

Purple Teaming: Collaborate

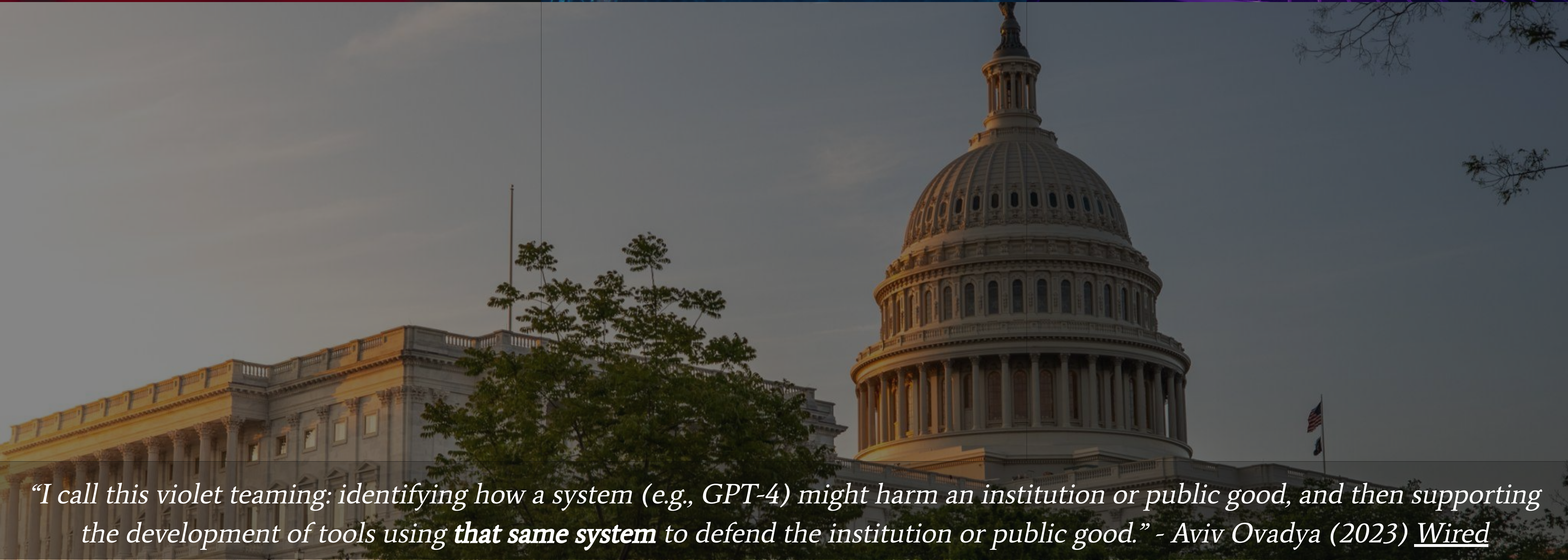
- Milestone driven
- Iterative
- Combined goals
- Business focus

AVIV OVADYA IDEAS MAR 29, 2023 7:00 AM

Red Teaming Improved GPT-4. Violet Teaming Goes Even Further

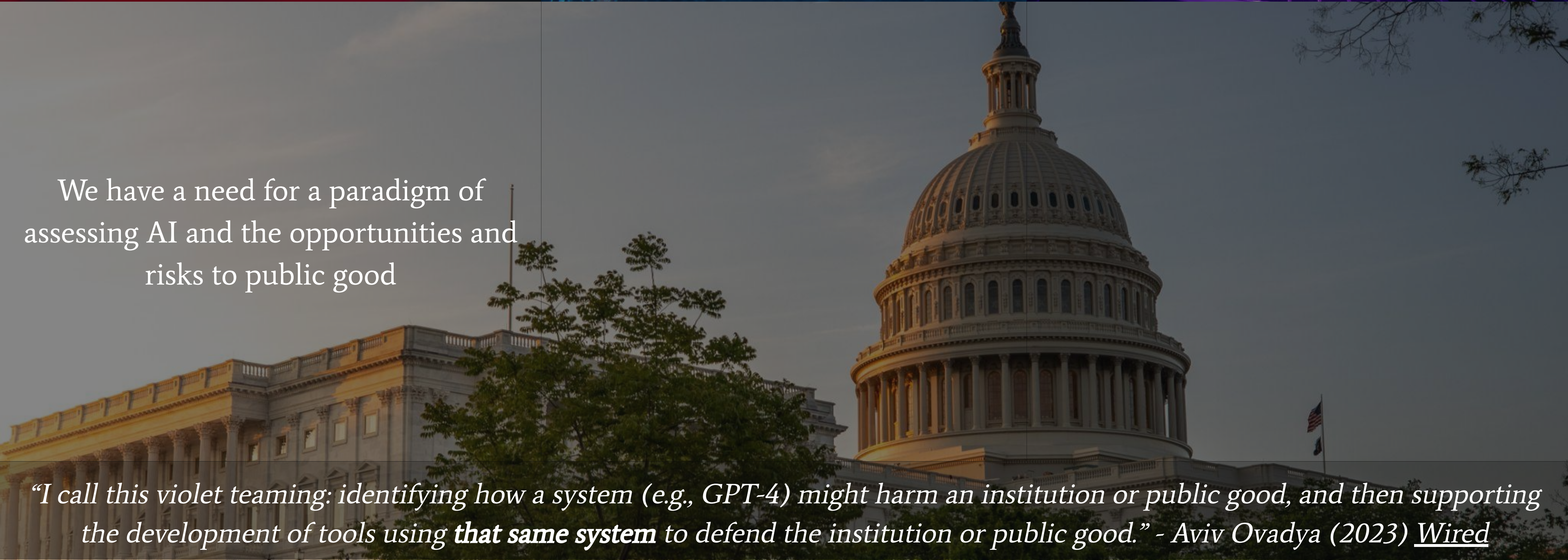
Reducing harmful outputs isn't enough. AI companies must also invest in tools that can defend our institutions against the risks of their systems.

Violet teaming brings institutions and public good into the equation



“I call this violet teaming: identifying how a system (e.g., GPT-4) might harm an institution or public good, and then supporting the development of tools using ***that same system*** *to defend the institution or public good.” - Aviv Ovadya (2023) Wired*

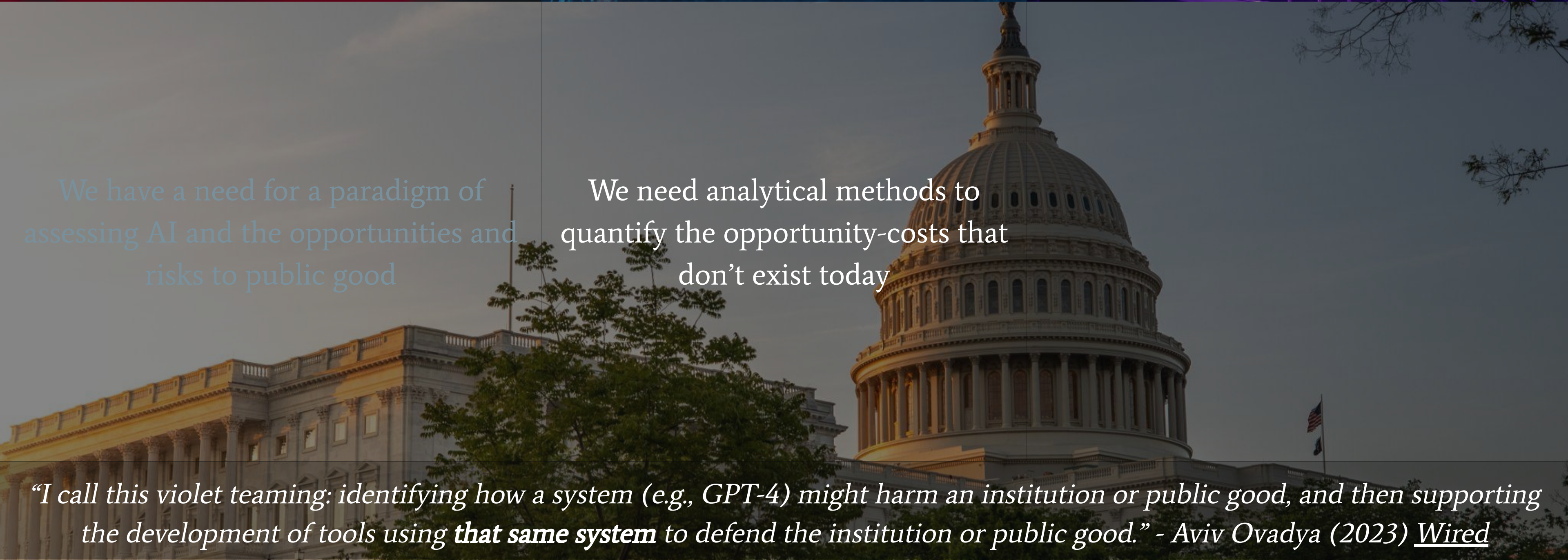
Violet teaming brings institutions and public good into the equation



We have a need for a paradigm of assessing AI and the opportunities and risks to public good

"I call this violet teaming: identifying how a system (e.g., GPT-4) might harm an institution or public good, and then supporting the development of tools using ***that same system*** *to defend the institution or public good." - Aviv Ovadya (2023) Wired*

Violet teaming brings institutions and public good into the equation

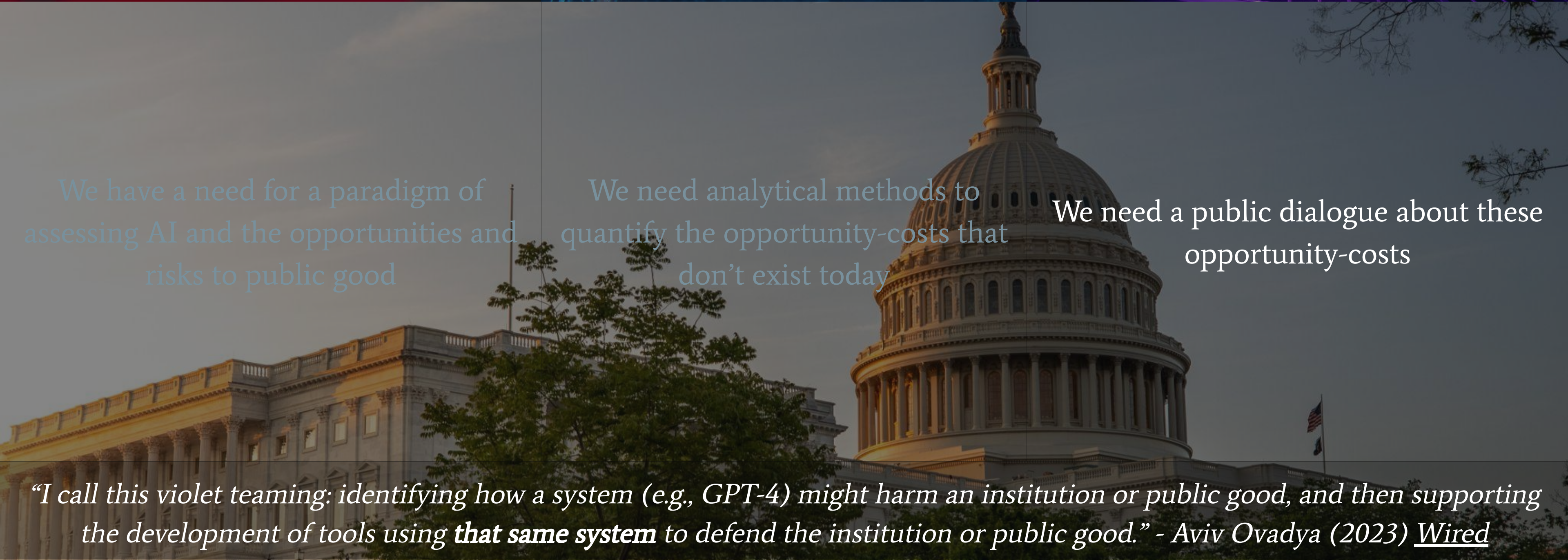


We have a need for a paradigm of assessing AI and the opportunities and risks to public good

We need analytical methods to quantify the opportunity-costs that don't exist today

"I call this violet teaming: identifying how a system (e.g., GPT-4) might harm an institution or public good, and then supporting the development of tools using ***that same system*** *to defend the institution or public good." - Aviv Ovadya (2023) Wired*

Violet teaming brings institutions and public good into the equation



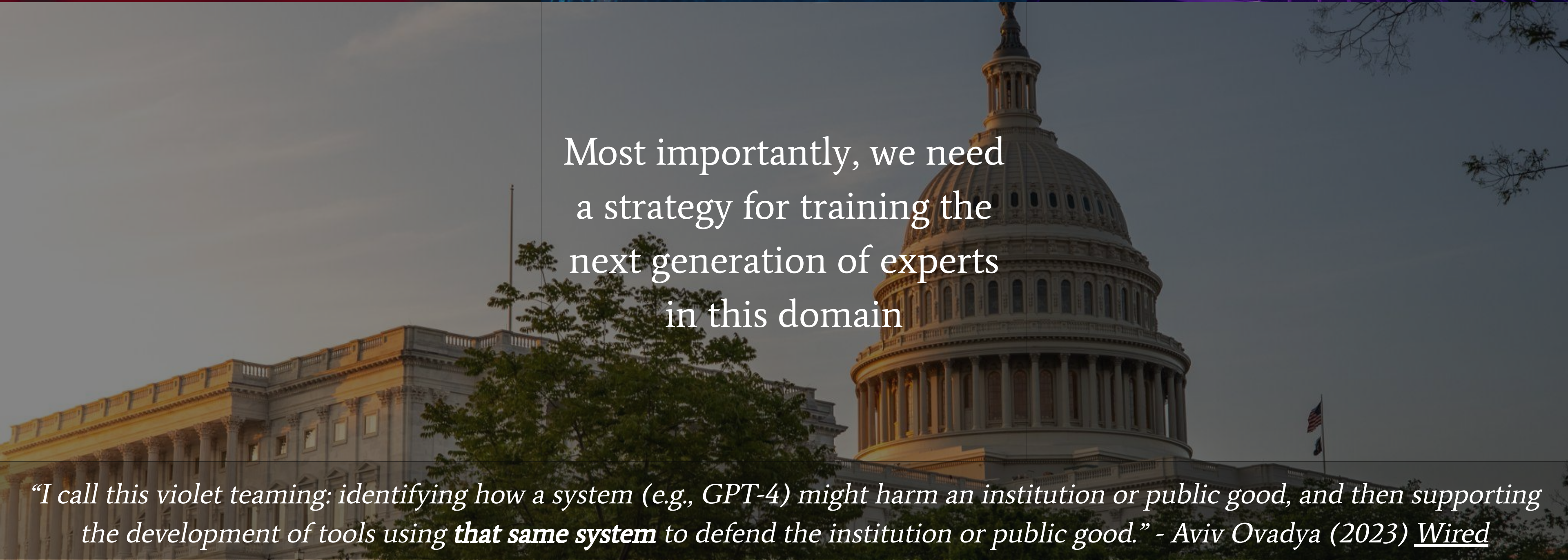
We have a need for a paradigm of assessing AI and the opportunities and risks to public good

We need analytical methods to quantify the opportunity-costs that don’t exist today

We need a public dialogue about these opportunity-costs

“I call this violet teaming: identifying how a system (e.g., GPT-4) might harm an institution or public good, and then supporting the development of tools using ***that same system*** *to defend the institution or public good.” - Aviv Ovadya (2023) Wired*

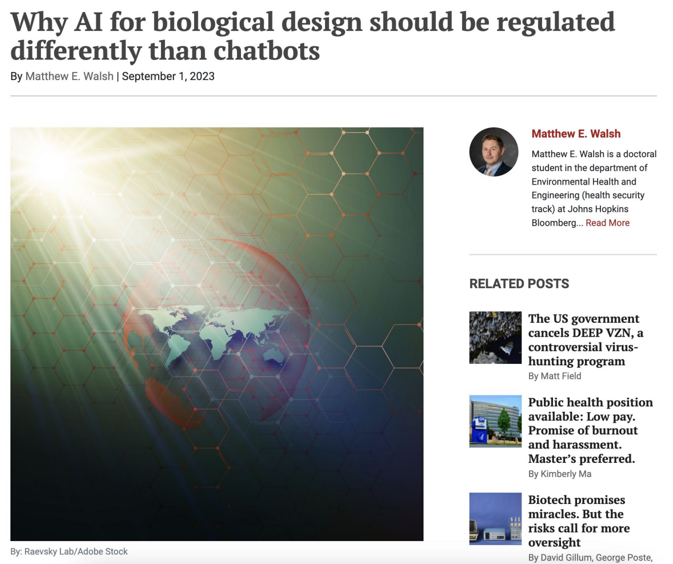
Violet teaming brings institutions and public good into the equation



Most importantly, we need a strategy for training the next generation of experts in this domain

"I call this violet teaming: identifying how a system (e.g., GPT-4) might harm an institution or public good, and then supporting the development of tools using ***that same system*** *to defend the institution or public good." - Aviv Ovadya (2023) Wired*

Violet teaming brings institutions and public good into the equation



Why AI for biological design should be regulated differently than chatbots

By Matthew E. Walsh | September 1, 2023

By: Raevsky Lab/Adobe Stock

Matthew E. Walsh

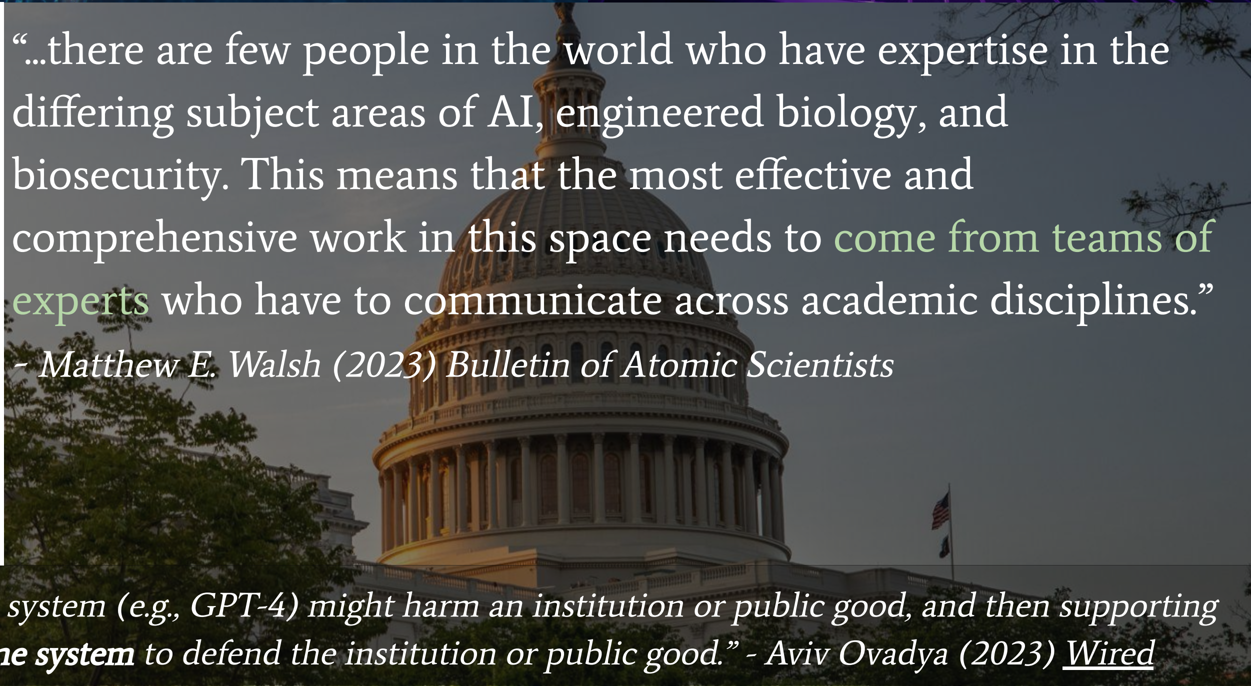
Matthew E. Walsh is a doctoral student in the department of Environmental Health and Engineering (health security track) at Johns Hopkins Bloomberg... Read More

RELATED POSTS

The US government cancels DEEP VZN, a controversial virus-hunting program
By Matt Field

Public health position available: Low pay. Promise of burnout and harassment. Master's preferred.
By Kimberly Ma

Biotech promises miracles. But the risks call for more oversight
By David Gillum, George Poste,

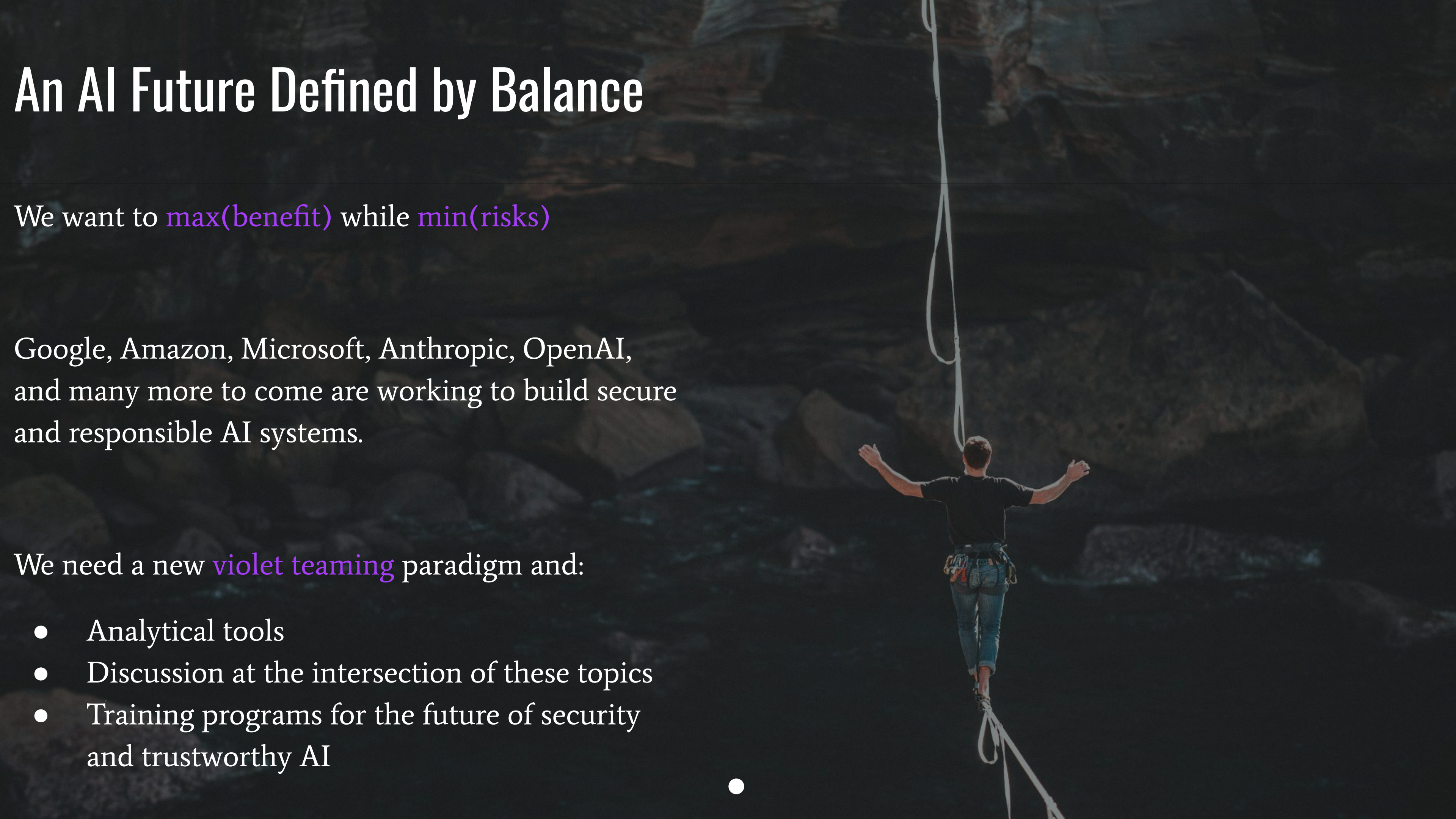


"...there are few people in the world who have expertise in the differing subject areas of AI, engineered biology, and biosecurity. This means that the most effective and comprehensive work in this space needs to come from teams of experts who have to communicate across academic disciplines."

- Matthew E. Walsh (2023) Bulletin of Atomic Scientists

*"I call this violet teaming: identifying how a system (e.g., GPT-4) might harm an institution or public good, and then supporting the development of tools using **that same system** to defend the institution or public good." - Aviv Ovadya (2023) Wired*

An AI Future Defined by Balance



We want to max(benefit) while min(risks)

Google, Amazon, Microsoft, Anthropic, OpenAI, and many more to come are working to build secure and responsible AI systems.

We need a new violet teaming paradigm and:

- Analytical tools
- Discussion at the intersection of these topics
- Training programs for the future of security and trustworthy AI

Thank you!

●●●

Alexander Titus, PhD

GUIRR Meeting @ the National Academies

11 October 2023