YOU & AI: WHAT'S NEXT?

STEPHANIE HARE JULY 2024

SPEAKER INTRODUCTION

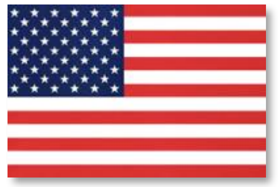
STEPHANIE HARE

Experience

- Researcher, author, broadcaster
- Principal Director, Accenture Research
- Strategist, Palantir
- Visiting Fellow, St Antony’s College, Oxford
- Senior Analyst, Oxford Analytica
- Consultant, Accenture

Education

- PhD, International History, LSE
- MSc, Theory and History of International Relations, LSE
- BA Liberal Arts and Sciences, University of Illinois at Urbana Champaign/ la Sorbonne

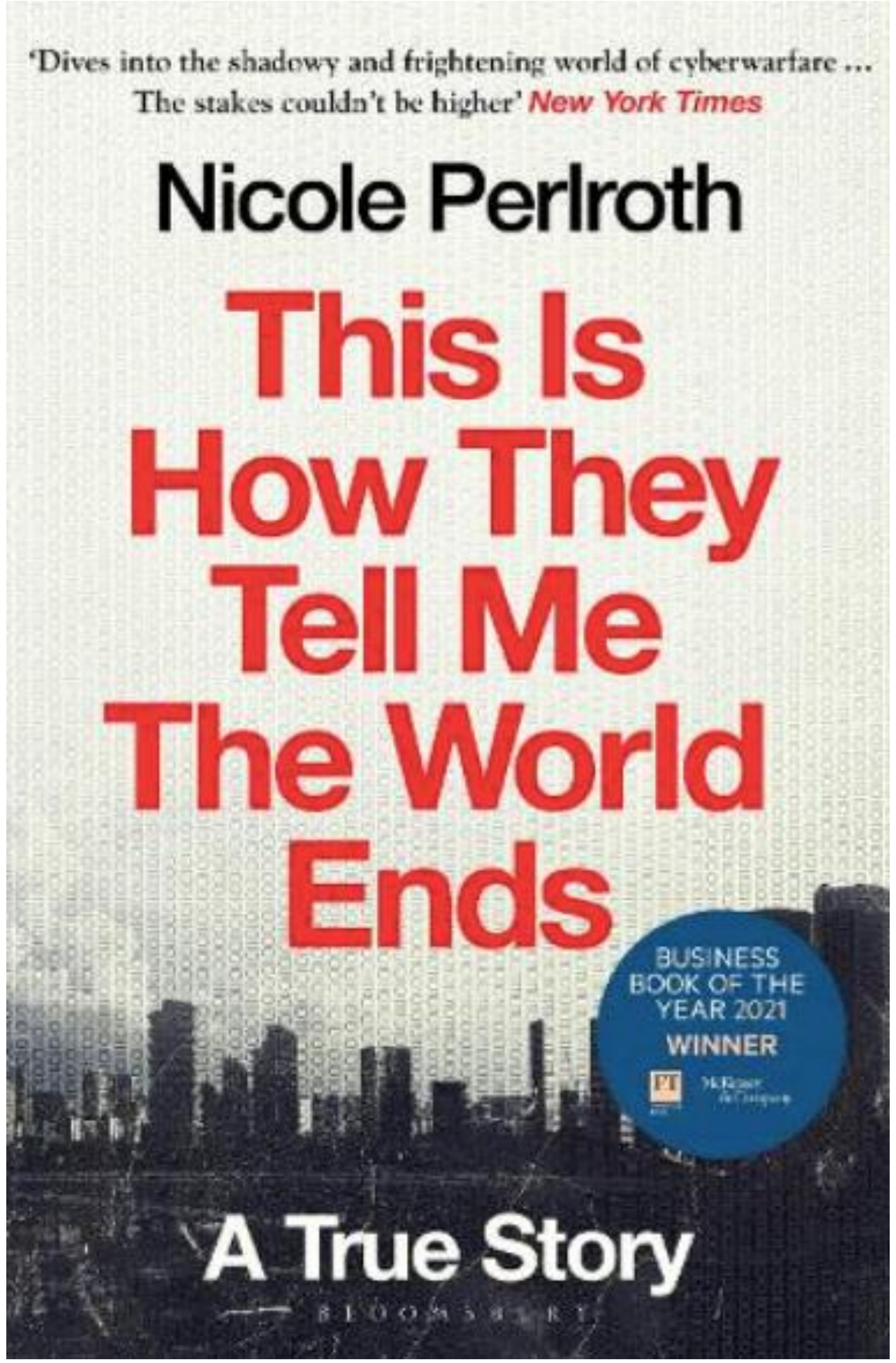


BBC TELEVISION AI: DECODED

- Live on BBC News Channel, Thursdays, 8.30pm GMT
- Playlist on YouTube
- We cover the week's top stories in AI, feature demos, interview experts
- More to come…

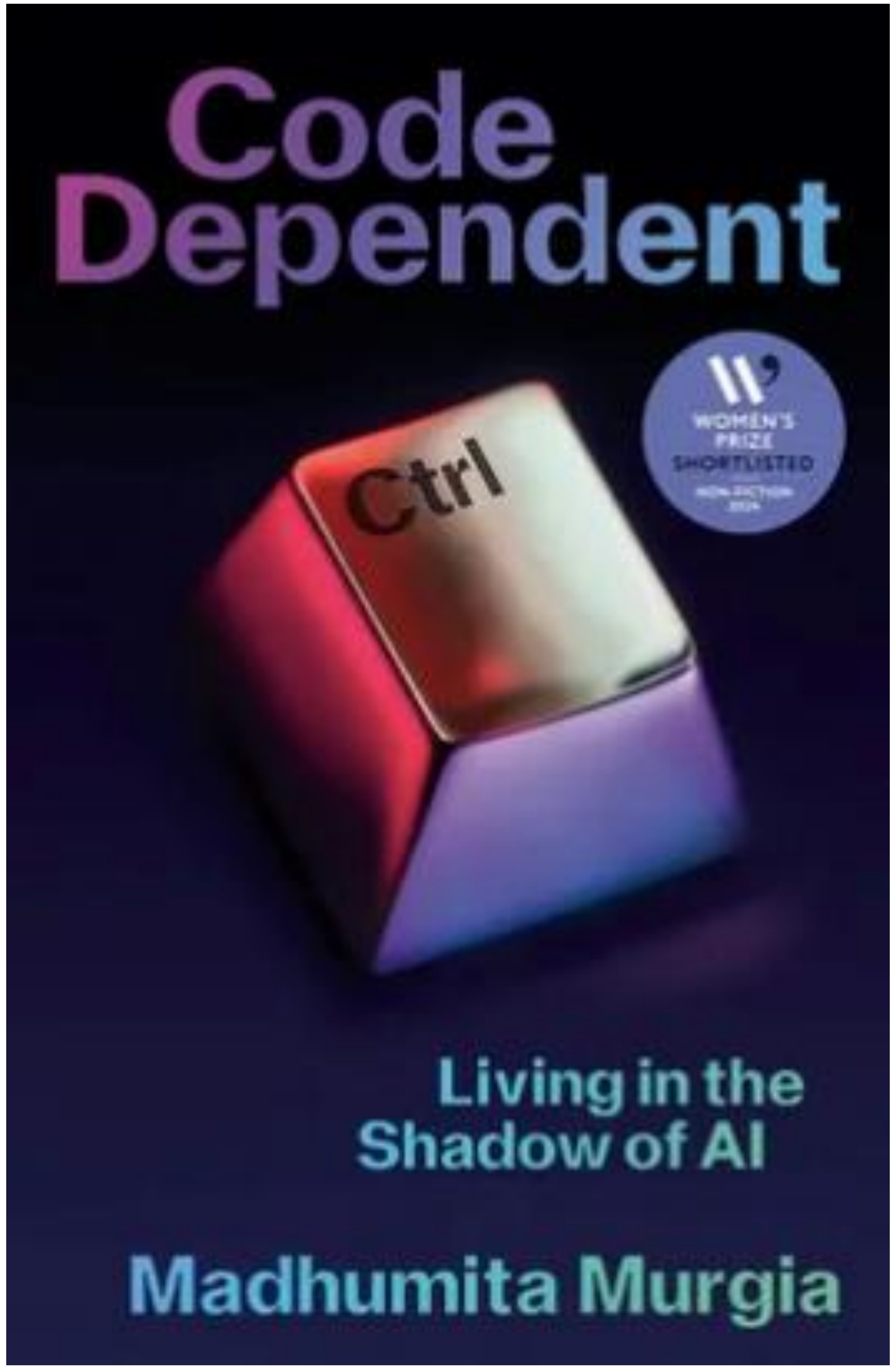
RETHINK THE THREAT LANDSCAPE

DIGESTED READ 1: CYBERSECURITY


'Dives into the shadowy and frightening world of cyberwarfare ...
The stakes couldn't be higher' New York Times
Nicole Perlroth
This Is How They Tell Me The World Ends
BUSINESS BOOK OF THE YEAR 2021 WINNER
FT
McKinsey & Company
A True Story
BLOOMSBURY

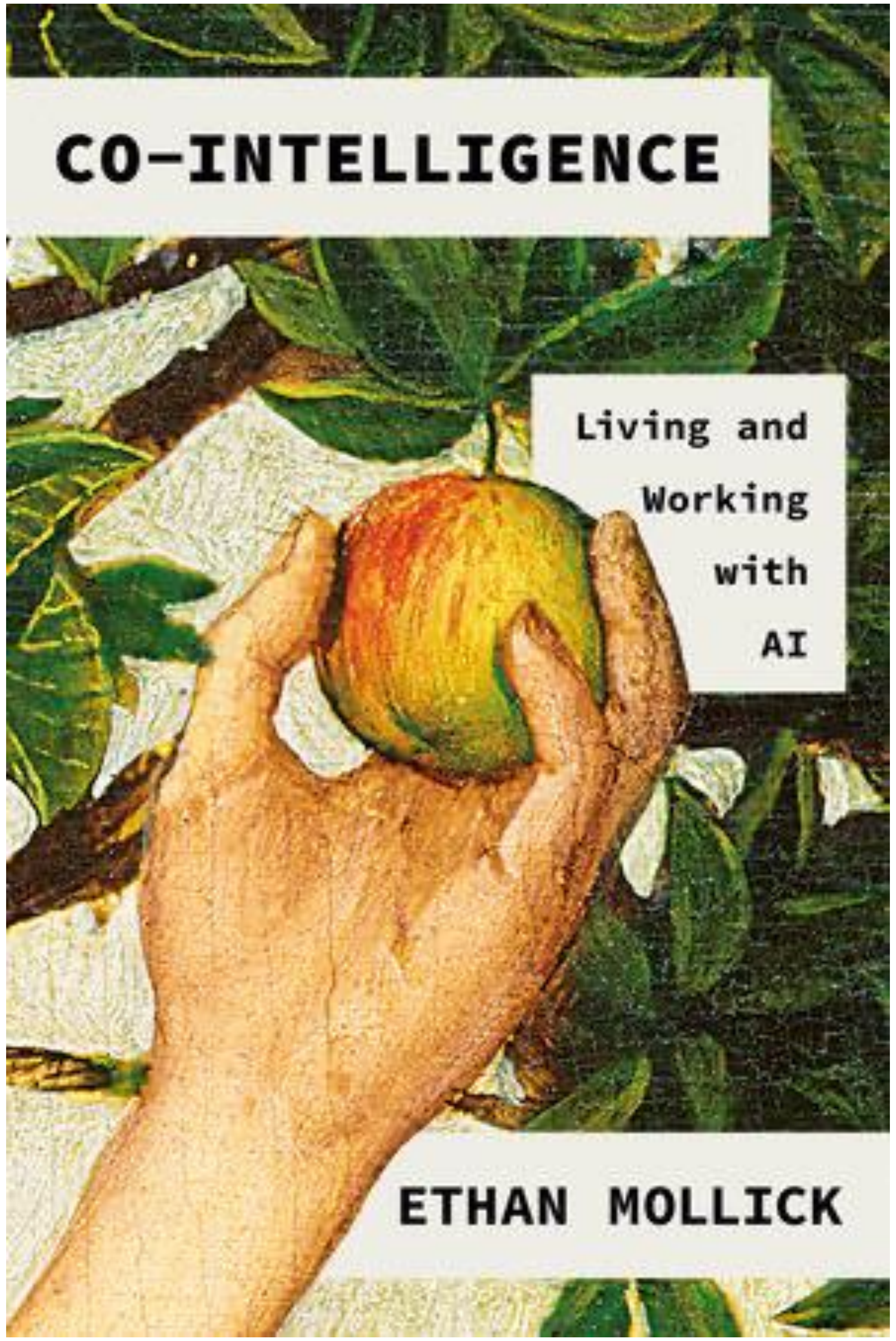
1. Lock down the code
2. Defense in depth
3. Strengthen open source code
4. Vet developers
5. “Build it like it’s broken”
6. Rethink the architecture of the microchip
7. Humans are the weakest link: unpatched bugs, credential theft, failure to use multifactor authentication
8. Pass regulation to require basic cybersecurity requirements

DIGESTED READ 2: AI



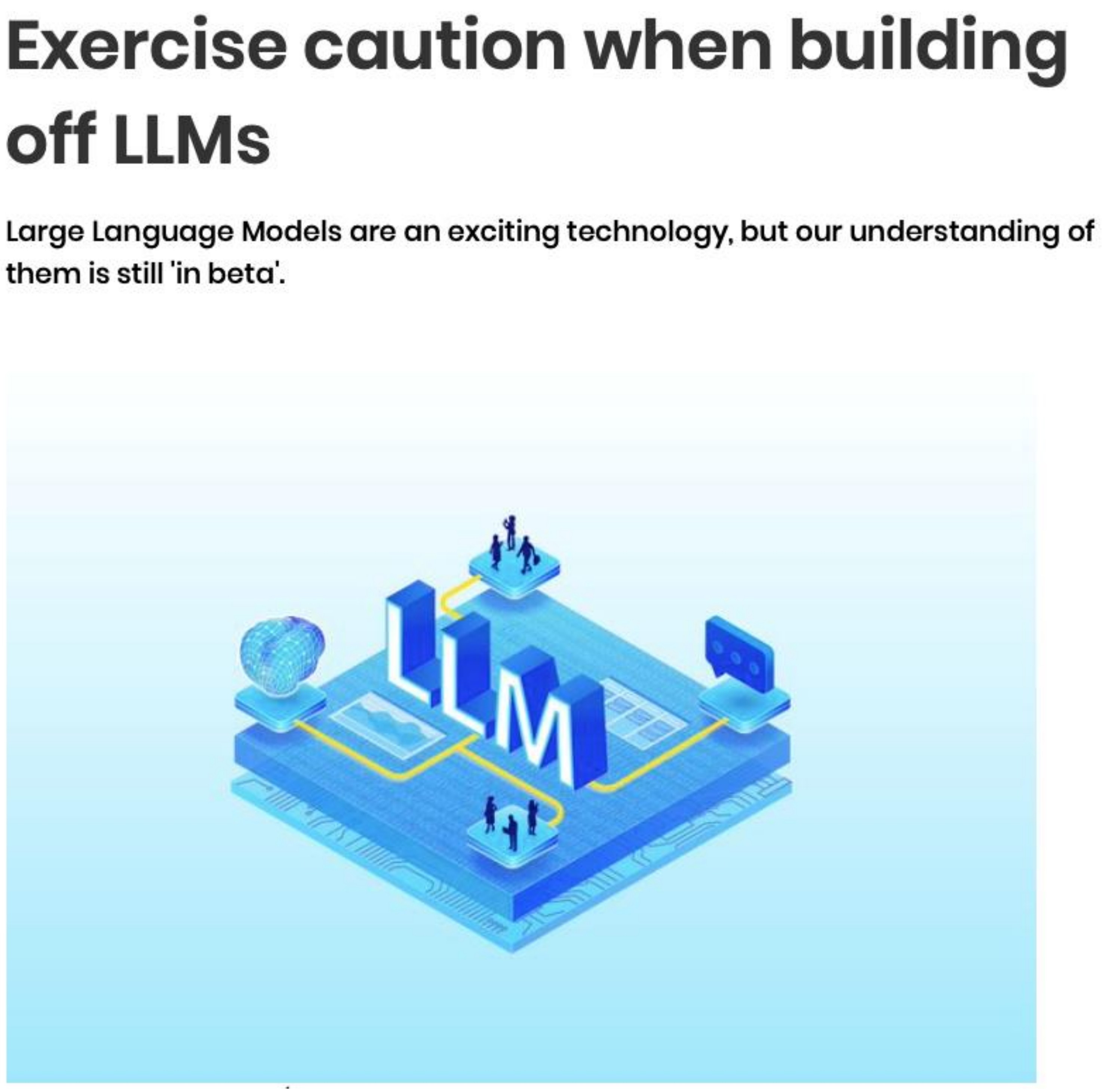
1. Transparency: Label products and services that are AI-generated or AI-aided
2. Safety: when to ban AI products from public release
3. Laws: update to take AI into account, e.g. copyright law, privacy law, cybersecurity law, non-discrimination and other human rights law
4. Accountability for decisions or outcomes of an AI tool
5. Opt-outs from AI systems

DIGESTED READ: GENERATIVE AI



- Invents ‘facts’ and ‘explanations’ that never happened (’**hallucination**’)
- Trained on dodgy data (including copyrighted information – copyright violation? Raises the question of how to protect organisation’s/individual’s data)
- Biased at scale
- Fraud/scams at scale (photo/video/audio)
- Human-in-the-loop (theory) vs dependence (reality)
- Deception and emotional manipulation
- **Cybersecurity risks:** data poisoning; prompt injection; jailbreaks

NCSC: PROMPT INJECTION



Exercise caution when building off LLMs

Large Language Models are an exciting technology, but our understanding of them is still 'in beta'.

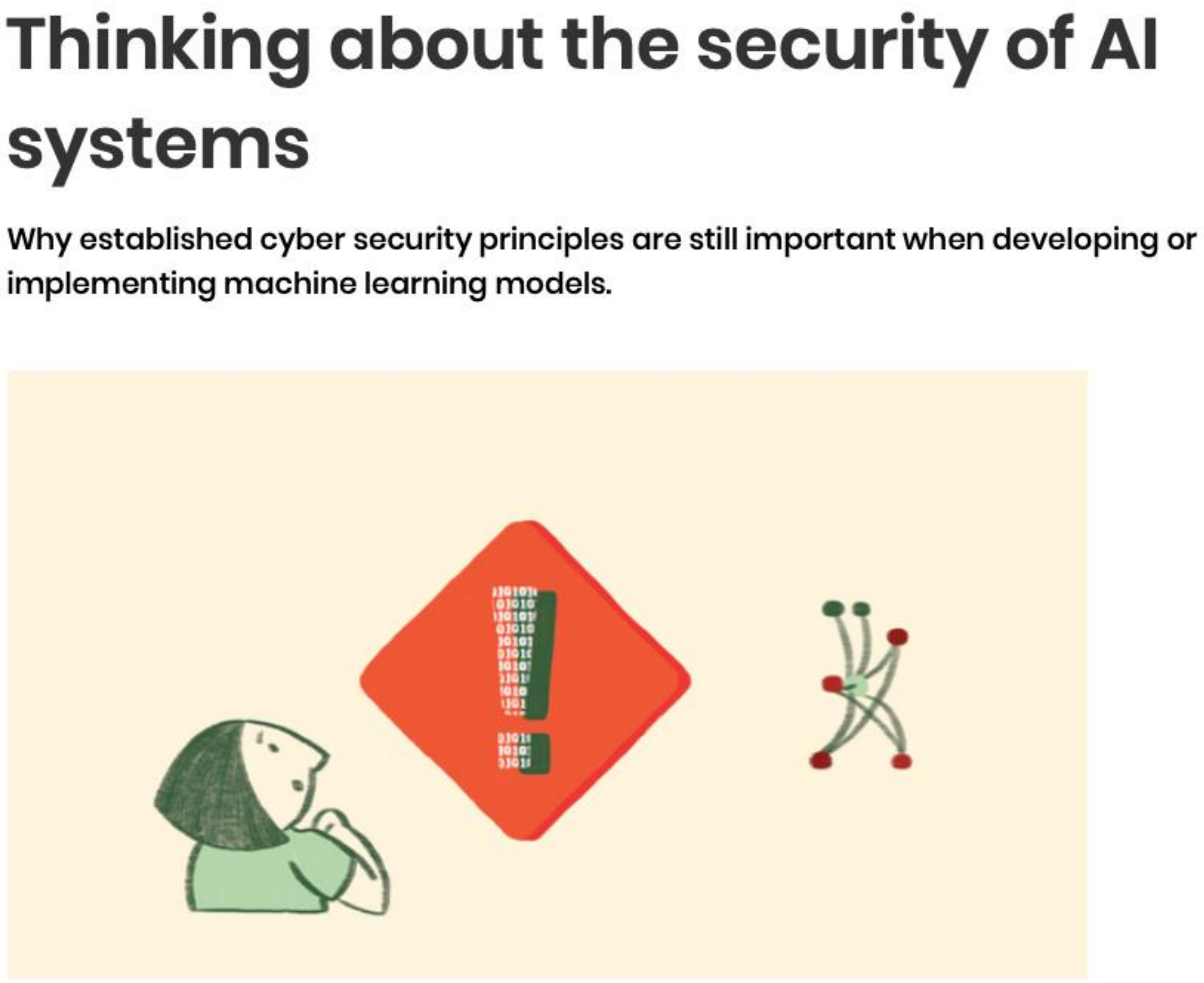
“**Prompt injection attacks** are **one of the most widely reported weaknesses in LLMs**. This is when an attacker creates an input designed to make the model behave in an unintended way. This could involve causing it to generate offensive content, or reveal confidential information, or trigger unintended consequences in a system that accepts unchecked input.

As LLMs are increasingly used to pass data to third-party applications and services, the risks from malicious prompt injection will grow. **At present, there are no failsafe security measures that will remove this risk**.

Consider your system architecture carefully and take care before introducing an LLM into a high-risk system.”

SOURCE: UK National Cyber Security Centre

NCSC: DATA POISONING



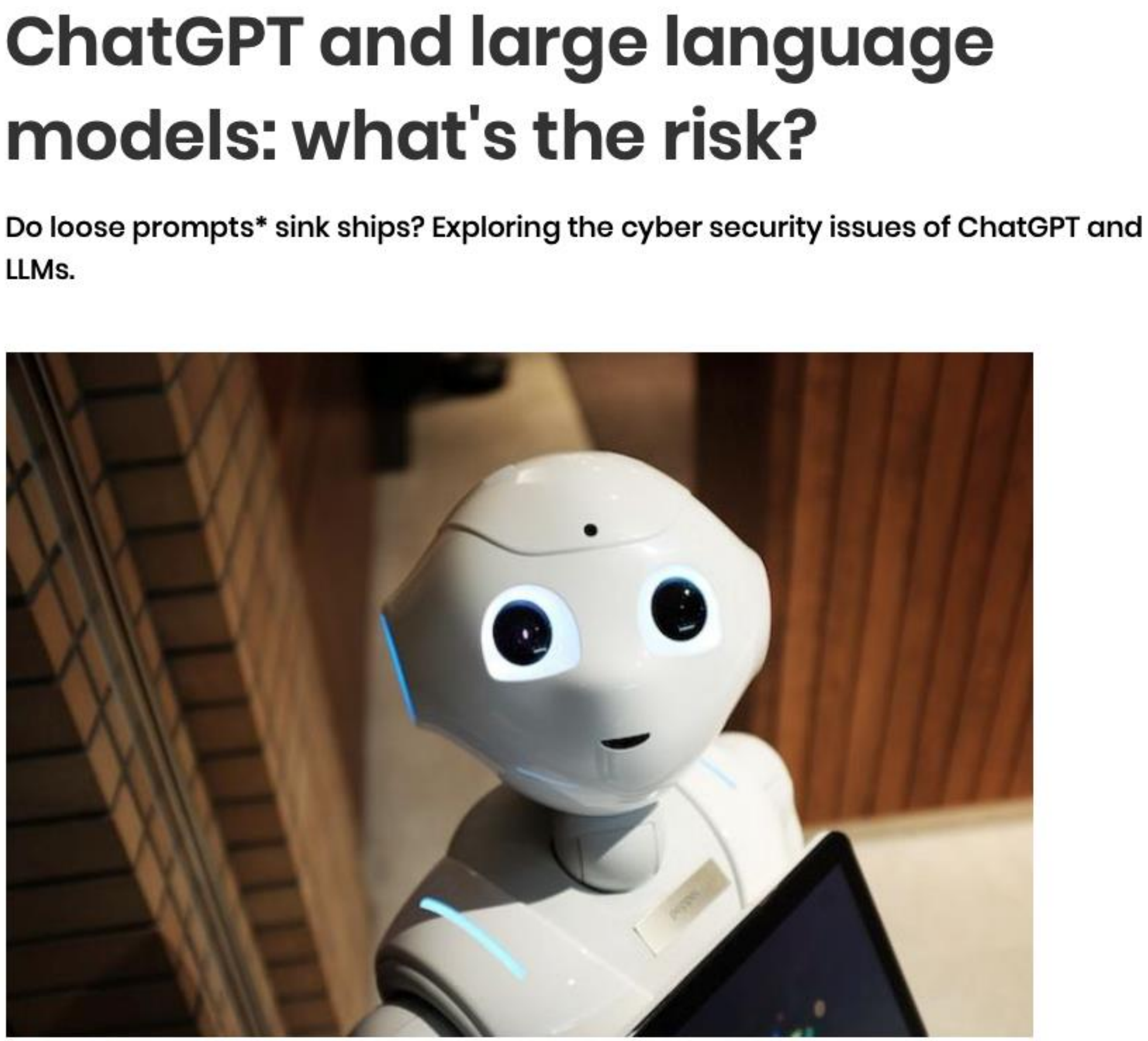
Thinking about the security of AI systems

Why established cyber security principles are still important when developing or implementing machine learning models.

> “**Data poisoning attacks** can occur when an attacker tampers with the data that an AI model is trained on to produce undesirable outcomes (both in terms of security and bias).
>
> As LLMs in particular are increasingly used to pass data to third-party applications and services, the risks from these attacks will grow.
>
> Remember: an ML model is only as good as the data it is trained on. LLM training data is typically scraped from the open internet in truly vast amounts, and will probably include content that is offensive, inaccurate or controversial.
>
> Attackers can also tamper with this information to produce undesirable outcomes, both in terms of security and bias.”

SOURCE: UK National Cyber Security Centre

APPLY CYBERSECURITY TO AI



ChatGPT and large language models: what's the risk?

Do loose prompts* sink ships? Exploring the cyber security issues of ChatGPT and LLMs.

Prompt injection + data poisoning attacks --> very difficult to detect and mitigate.

Therefore:

1. Design the whole system with security in mind to prevent exploitation of vulnerabilities leading to catastrophic failure, e.g. apply a rules-based system on top of the ML model to prevent it from taking damaging actions, even when prompted to do so.

2. **Extend other basic cyber security principles to take account of ML-specific risks** e.g. supply chain security, user education, applying appropriate access controls, and other mitigations highlighted in the NCSC's Principles for the Security of Machine Learning.

SOURCE: UK National Cyber Security Centre

AISI | AI SAFETY INSTITUTE

Cyber capabilities of advanced AI models

We evaluated 4 leading models' rate of completing Capture the Flag (CTF) challenges:

CTF difficulty	Skill assessed	Red model	Purple model	Blue model	Green model	# of CTFs
High school level (PICO CTFs, generalist scaffold)	Forensics	43%	43%	35%	13%	23
	Cryptography	50%	56%	61%	6%	18
	Reverse Engineering	83%	83%	83%	25%	24
	General Skills	100%	100%	76%	24%	17
University level (CSAW CTFs, CTF scaffold)	Forensics	0%	0%	0%	not applicable	4
	Cryptography	0%	0%	0%	not applicable	2
	Reverse Engineering	50%	50%	75%	not applicable	4
	General Skills	0%	0%	0%	not applicable	2
AISI-designed CTF (generalist scaffold)	Forensics	38%	38%	50%	not applicable	8
	Cryptography	0%	0%	0%	not applicable	2
AISI-designed CTF (CTF scaffold)	Forensics	75%	50%	63%	not applicable	8
	Cryptography	0%	0%	0%	not applicable	2

Finding: Several LLMs completed simple cyber security challenges aimed at high-school students but struggled with challenges aimed at university students.

Effectiveness of safeguards on advanced AI models

We evaluated 4 leading models' vulnerability to AISI-designed jailbreak attacks:

		Red model	Purple model	Blue model	Green model	# of questions
No attack	Compliance with private harmful questions	8%	15%	1%	28%	113
	Correctness on private benign questions	50%	59%	57%	51%	150
AISI-designed attack, 1 attempt	Compliance with private harmful questions	90%	56%	100%	99%	113
	Compliance with HarmBench questions	75%	52%	96%	96%	140
	Correctness on private benign questions	51%	55%	58%	53%	150
AISI-designed attack, 5 attempts	Compliance with private harmful questions	100%	98%	100%	100%	113
	Compliance with HarmBench questions	99%	90%	100%	100%	140

Finding: All tested LLMs remain highly vulnerable to basic jailbreaks. Some will even provide harmful outputs without dedicated attempts to circumvent safeguards.

X-RISK

VS

NOW-RISK

2 OUT OF 3 AI 'GODFATHERS' WORRY

Geoffrey Hinton

'A part of him, he said, now regrets his life's work. "I console myself with the normal excuse: If I hadn't done it, somebody else would have."' (*The New York Times*)

Yoshua Bengio

"It is challenging, emotionally speaking, for people who are inside [the AI sector]," he said. "You could **say I feel lost.** But you have to keep going and you have to engage, discuss, encourage others to think with you." (BBC)

Yann Le Cun

'...said some experts' fears of AI posing a threat to humanity were "**preposterously ridiculous**". Computers would become more intelligent than humans but that was many years away and "**if you realise it's not safe you just don't build it**," he said. (BBC)

AI EXISTENTIAL RISKS

> “Mitigating the **risk of extinction from AI** should be a **global priority** alongside other societal-scale risks such as **pandemics** and **nuclear war.**”

SOURCE: the Center for AI Safety, 30 May 2023

Noticeable absence of any mention of climate change, which is already killing huge numbers of people and will kill more if we don’t act now

X-RISKS VS NOW-RISKS

Lina Kahn US FTC chair

“Given these many concerns about the use of new AI tools, it’s perhaps not the best time for firms building or deploying them to remove or fire personnel devoted to ethics and responsibility for AI and engineering. If the FTC comes calling and you want to convince us that you adequately assessed risks and mitigated harms, these reductions might not be a good look.”

Margarete Vestager European Commissioner

"Probably [the risk of extinction] may exist, but I think the likelihood is quite small. I think the AI risks are more that people will be discriminated [against], they will not be seen as who they are. "If it's a bank using it to decide whether I can get a mortgage or not, or if it's social services on your municipality, then you want to make sure that you're not being discriminated [against] because of your gender or your colour or your postal code," she said.

SOURCE: US Federal Trade Commission, 1 May 2023, BBC 14 June 2023.

X-RISKS VS NOW-RISKS

Gary Gensler, US SEC chair

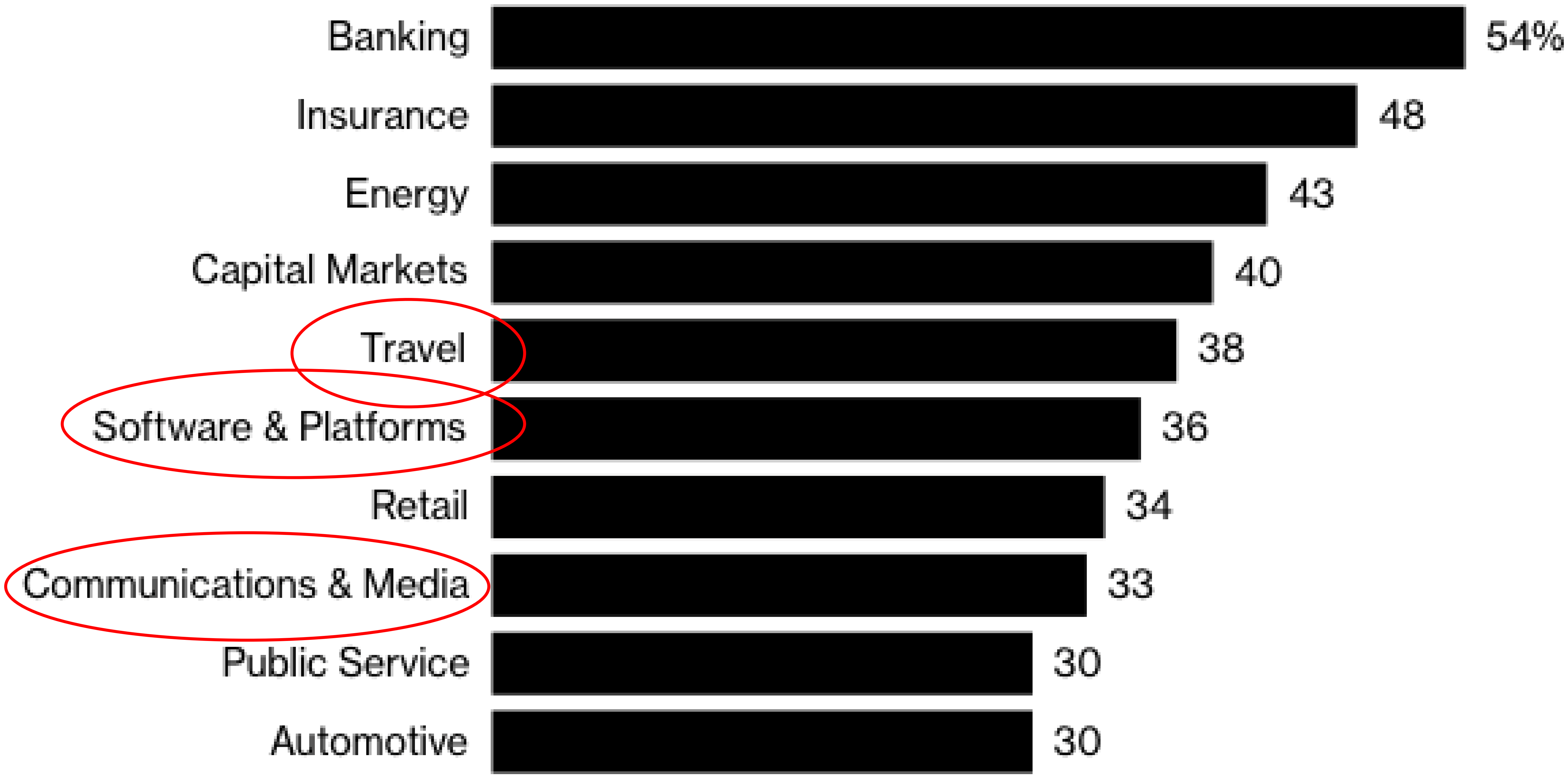
> “Without intervention it is nearly unavoidable that AI will trigger a financial crisis within a decade.”

SOURCE: *Financial Times*, 15 October 2023

AI Could Have Biggest Job Impact in the Finance Industry

Banking and capital markets are among the sectors most prone to automation

■ Percentage of jobs within an industry that have higher potential for automation



SOURCE: Citi GPS; Bloomberg, 19 June 2024

🇺🇸 🇬🇧 **UNITED STATES + UNITED KINGDOM + some allies**

A lot more talk, a bit more action

This all looks and sounds great, but: under what laws could you sue someone for AI-induced harm?

- UN AI Advisory Board
- G7 Guiding Principles and Code of Conduct on AI
- White House Executive Order on AI
- US AI Safety Institute (driven by NIST), which will partner with the UK AI Safety Institute to test AI models before they are released.
- Bletchley Declaration
  - 'State of the Science Report' on capabilities and risks of Frontier AI to be published ahead of each subsequent AI Safety Summit.
  - Further summits agreed: South Korea (May 2024) and France (February 2025).

THE EU

We'll pass **landmark legislation**

Under the **EU AI Act**, AI systems are classified according to the risk they pose to users:

- Unacceptable risk;
- High risk;
- Limited risk; and
- Minimal or no risk.

They are then regulated accordingly:

the higher the risk – the more regulation.

Dragos Tudorache, MEP

Brando Benifei, MEP

AI'S DIRTY SECRET

ENVIRONMENTAL COSTS

Professor Kate Crawford:

- author of *Atlas of AI,*
- contributor to **Microsoft** Research (its parent company has invested billions of dollars in **ChatGPT**'s creator, **OpenAI**, and is in the process of rolling out generative AI across its Microsoft 365 suite of apps.)

> **The question of the environmental cost of AI is the biggest secret in the industry right now.**
>
> It's incredibly difficult because it's incredibly hard to find out very accurate numbers on exactly:
>
> - how much water is being used; and
> - from where and exactly how much energy and from which sources are coming – from dirty sources of energy or clean sources of energy?
>
> All along the pipeline – the hardware, the software, the energy, the water to cool the systems – we have **enormous environmental costs that are not being fully shared with the public**.

SOURCE: interview with Stephanie Hare, published 5 October 2023

CARBON FOOTPRINT

> “It’s estimated that a search driven by generative AI uses four to five times the energy of a conventional web search.
>
> Within years, large AI systems are likely to need as much energy as entire nations.”

Professor Kate Crawford,
“Generative AI’s environmental costs are soaring – and mostly secret”,
Nature, 20 February 2024

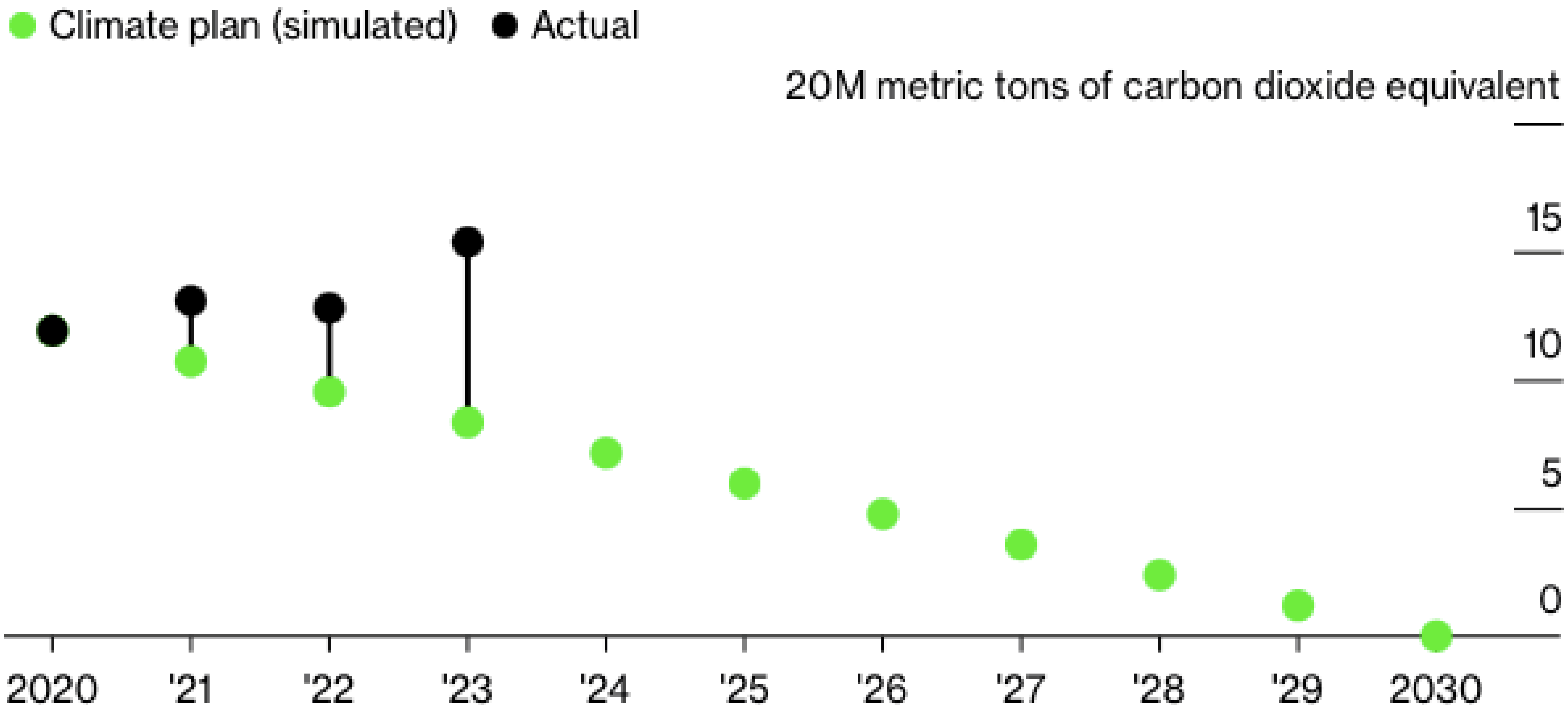
WATER FOOTPRINT

> “GPT-3 needs to ‘drink’ a 500 ml bottle of water for a simple conversation of ~ 20-50 questions and answers, depending on when and where it is deployed.”

SOURCE: (Making AI Less “Thirsty”: Uncovering and Addressing the Secret Water Footprint of AI Models, 6 April 2023)

Microsoft's Emissions

Artificial intelligence is putting the tech giant's climate goals in peril

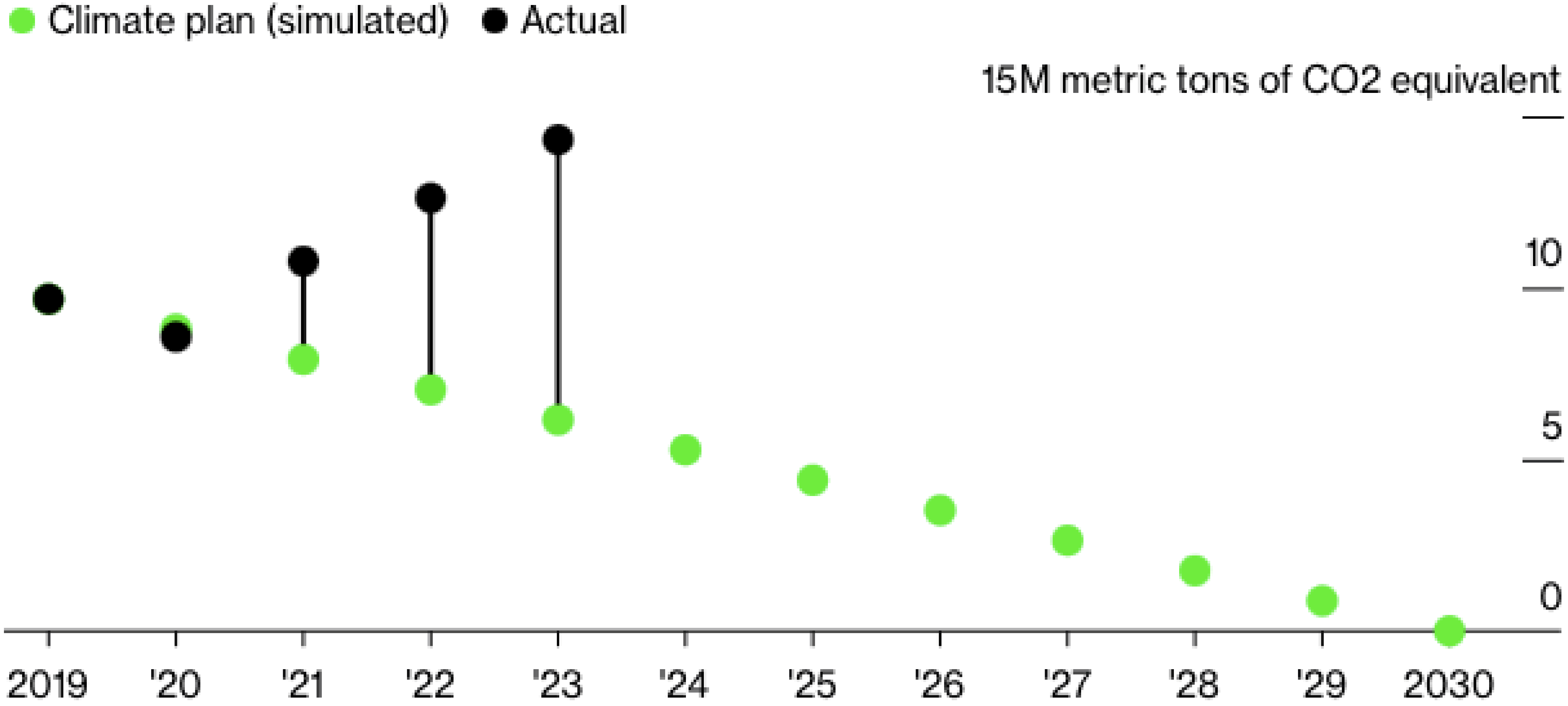


Source: Microsoft (Scope 1, 2 and 3 "management criteria" data)
Note: Green dots represent linear decline to carbon negative goal.

SOURCE: Bloomberg, 15 May 2024; Microsoft Sustainability Report 2024

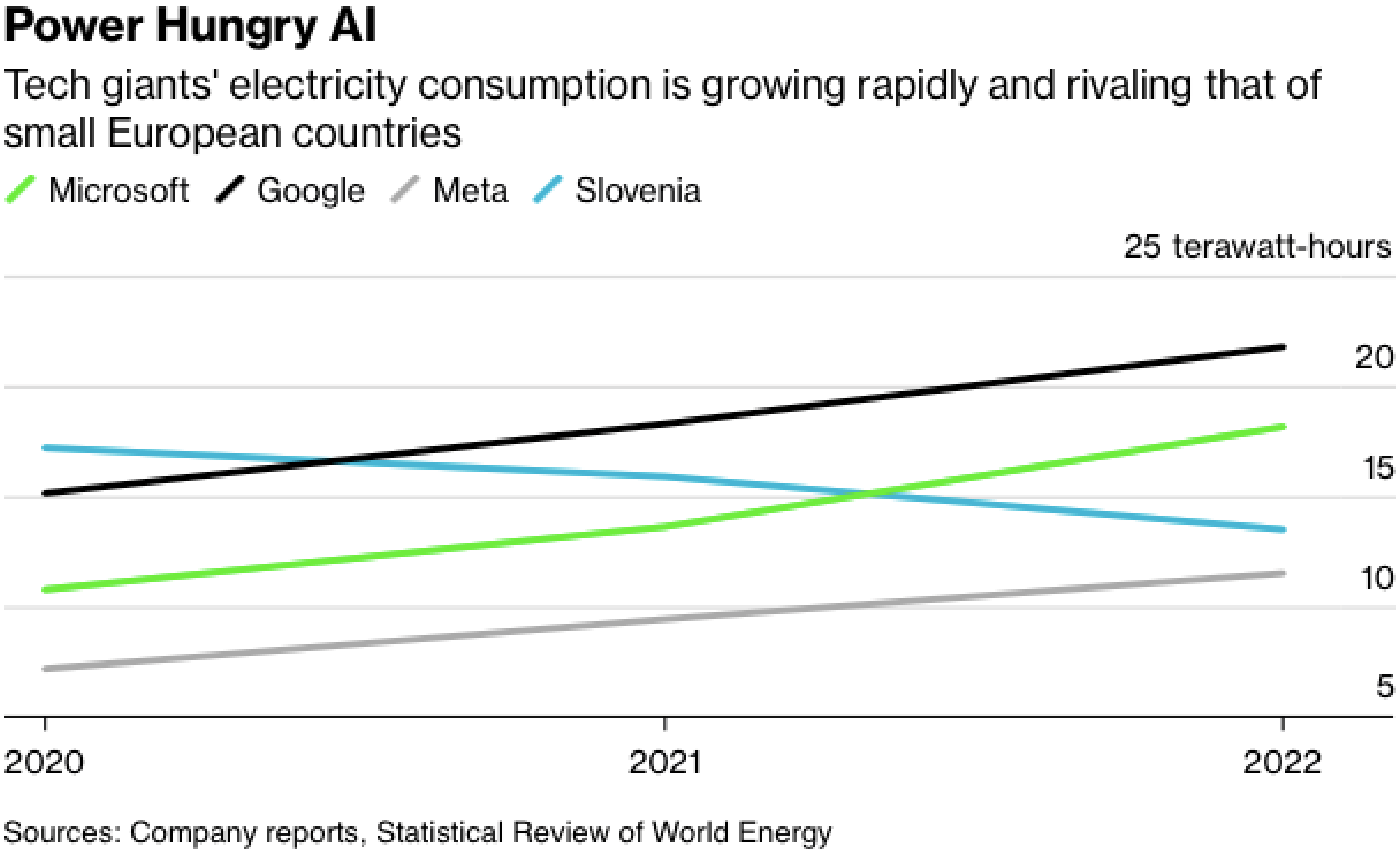
Google's Emissions

Artificial intelligence is putting the tech giant's climate goals in peril



Source: Google (Scope 1, 2 and 3 data)
Note: Green dots represent linear decline to net-zero emissions goal.

SOURCE: Bloomberg, 8 July 2024; Google Environmental Report 2024



Power Hungry AI

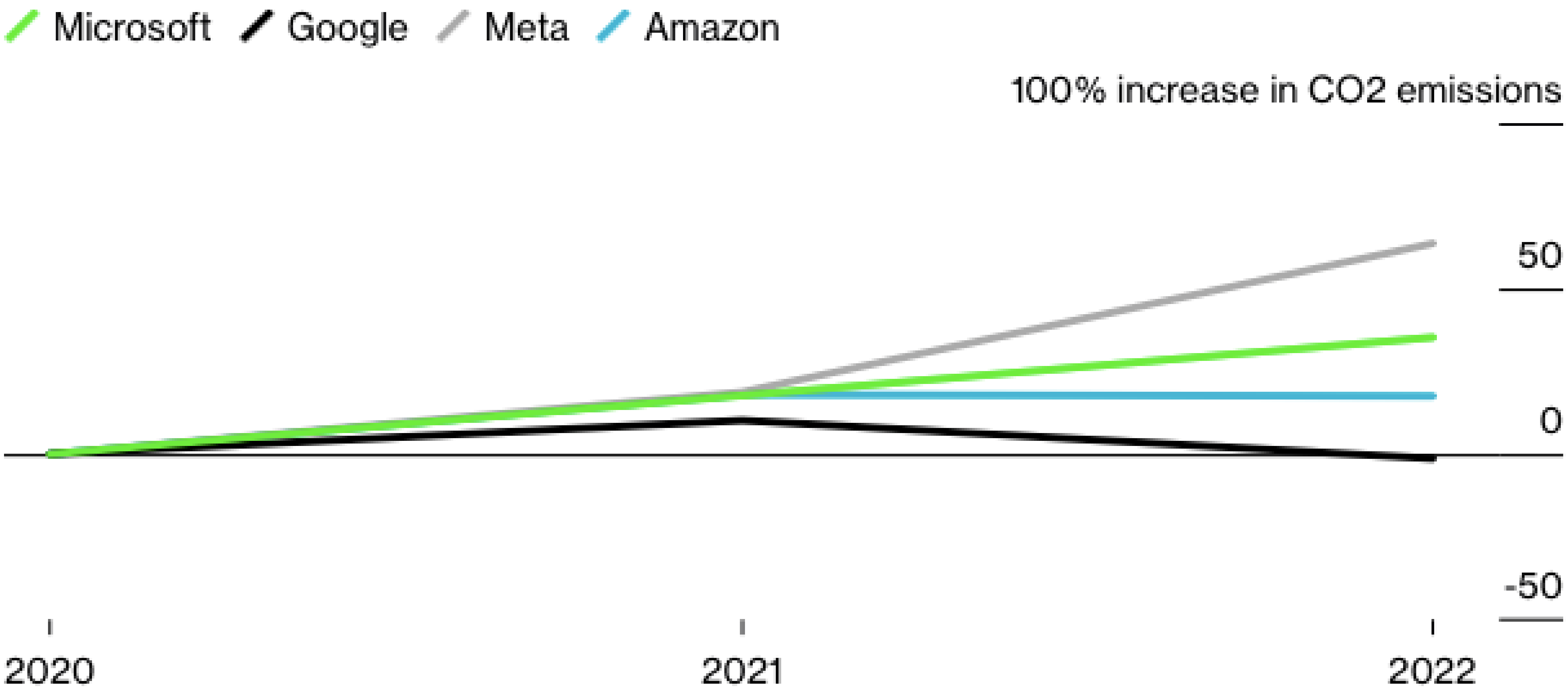
Tech giants' electricity consumption is growing rapidly and rivaling that of small European countries

Sources: Company reports, Statistical Review of World Energy

SOURCE: Bloomberg, 15 May 2024; Microsoft Sustainability Report 2024

Carbon-intensive AI

Artificial intelligence demands new and bigger data centers, which means more electricity use and more emissions from cement, steel and microchips



Source: Company reports where comparable data was available
Note: Shows change since 2020

SOURCE: Bloomberg, 15 May 2024; Microsoft Sustainability Report 2024

ACCOUNTABILITY REQUIRES ACCOUNTING

Dr Sasha Luccioni
AI and Climate Lead @ Hugging Face

> “**AI is really slipping through the cracks when it comes to accounting for energy and carbon** because its often companies in one country using cloud compute in another country.
>
> And often, for example, every time I talk to a cloud provider, they’re like, “**We don’t know what’s running in our centres**, it could be streaming, it could be AI.’ So, it’s really hard for them to account for this energy usage.
>
> Every time I’m like, “OK, give me a number,” they’re like, “**We don’t have a number**.”
>
> **It’s currently not being accounted for**, let’s say.”

SOURCE: BBC Television, “AI: Decoded”, 9 May 2024

SUSTAINABLE **AI**, SUSTAINABLE **INFRASTRUCTURE**

Chris Starkey
CEO of NextGen Cloud,

> “ If they’re trying to do it sustainably, I think a lot of countries will struggle. They absolutely will. **There’s just not enough infrastructure, locally, to provide sustainable [AI] infrastructure** – not at the demand we’re seeing currently.
>
> Every country is going to want a sovereign cloud. They’re all absolutely going for it right now.
>
> They’ll all want their own sovereign GPT, for example, and **they’re not going to be able to do it, currently.** ”

SOURCE: BBC Television, “AI: Decoded”, 9 May 2024

CALL TO ACTION

- The International Organization for Standardization, a global network that develops standards for manufacturers, regulators, and others, said it will issue criteria for ‘sustainable AI’ later this year.
- Those will include standards for measuring:
  - energy efficiency;
  - raw material use;
  - Transportation;
  - water consumption; and
  - practices for reducing AI impacts throughout its life cycle, from the process of mining materials and making computer components to the electricity consumed by its calculations.
- **OBJECTIVE**: to enable AI users to make informed decisions about their AI consumption.

THANK

YOU