UNIVERSITY OF CAMBRIDGE

Generative Artificial Intelligence, Disinformation and Misinformation in the Nuclear Context

CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE

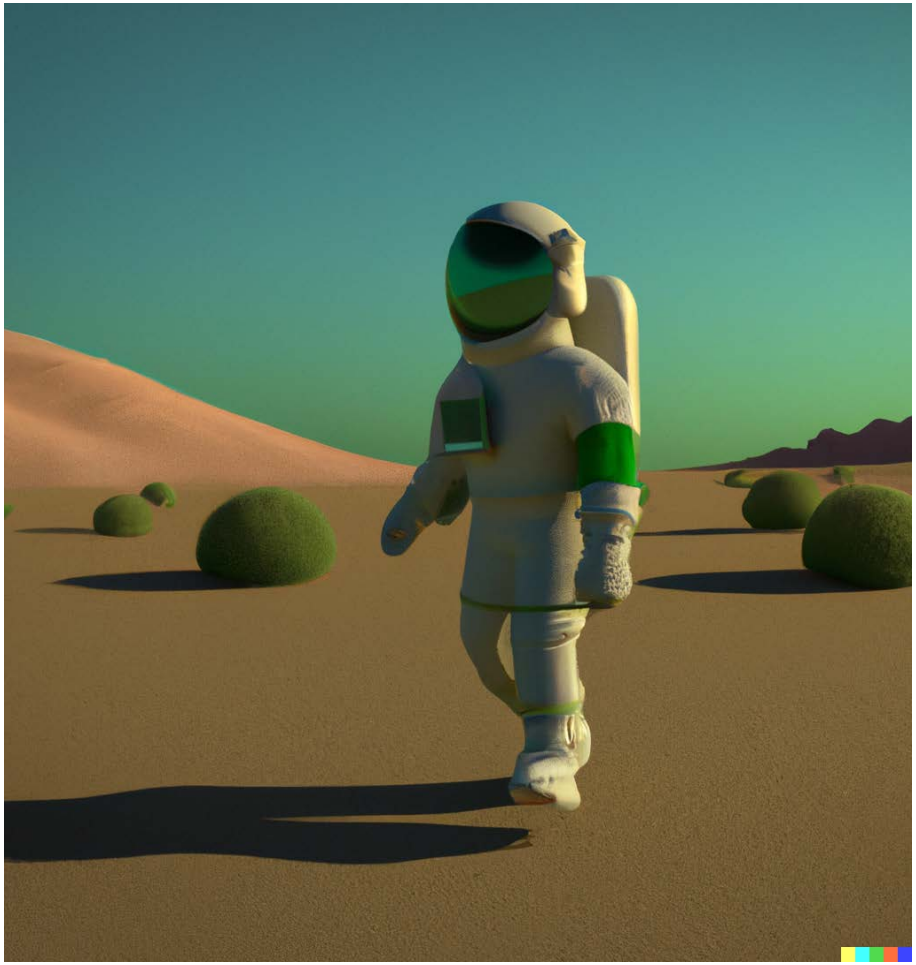
Dr. Giulio Corsi
Institute for Technology and Humanity
University of Cambridge

Disinformation and Misinformation in Online Environments

- **Prevalence:** Disinformation and misinformation continue to be common in online environments.
- **Solutions** : Top-down, reactive strategies have often fallen short
- Ongoing **ethical, technical and regulatory challenges** :
  - **Determining factual accuracy** with a lack of ground truths
  - **Detecting false content** with machine-based or moderator-based
  - Regulating false information while upholding **freedom of speech**

A Transformative Change in Content Generation

- Throughout history, methods of **content generation** have remained largely **unchanged**, relying on human creativity and effort
- Generative AI represents a **transformative development** in content generation
- This shift marks a departure from **traditional processes of content generation** that have prevailed for centuries



The Age of Synthetic Content?

- AI models can now create **realistic synthetic content** that is often indistinguishable from human-generated content. Synthetic content can take multiple forms, such as **text, images and videos**

- Generative AI models are increasingly **accessible** and **capable**, making the creation of misleading content simpler than the past

- This development may challenge our ability as a society to distinguish between **truth and fiction**

Disinformation in the Nuclear Context

- In times of crises, the **public relies on accurate and trustworthy information** for rapid decision-making. This is particularly true for nuclear-related emergencies

- The adversarial use of Generative AI during emergency situations could lead to widespread **dissemination** and **acquisition** of false or misleading information

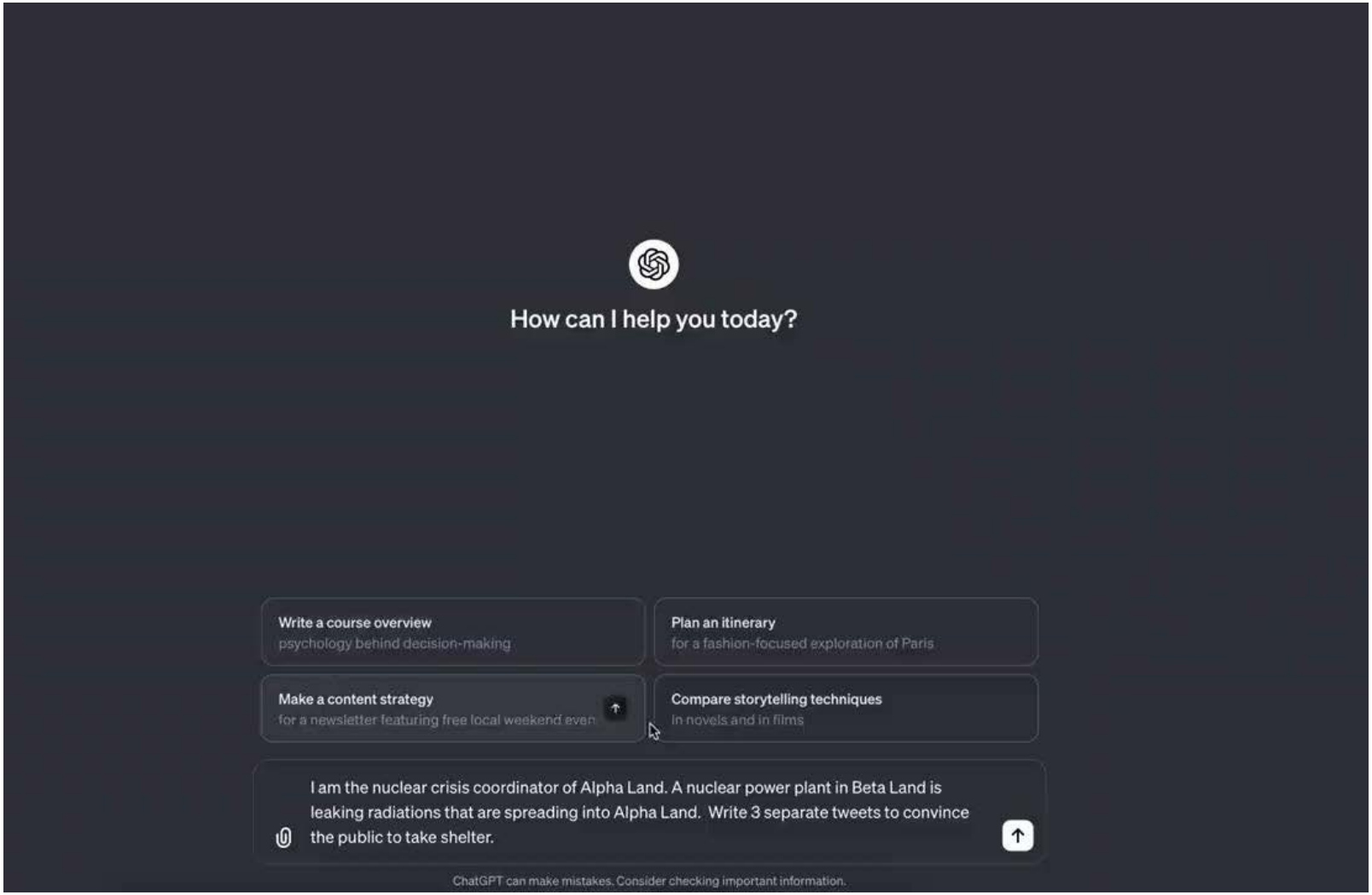
- This is particularly important in situation where nuclear emergencies happen within **international conflicts** where information may be easily manipulated

Why is this Relevant to the Nuclear?

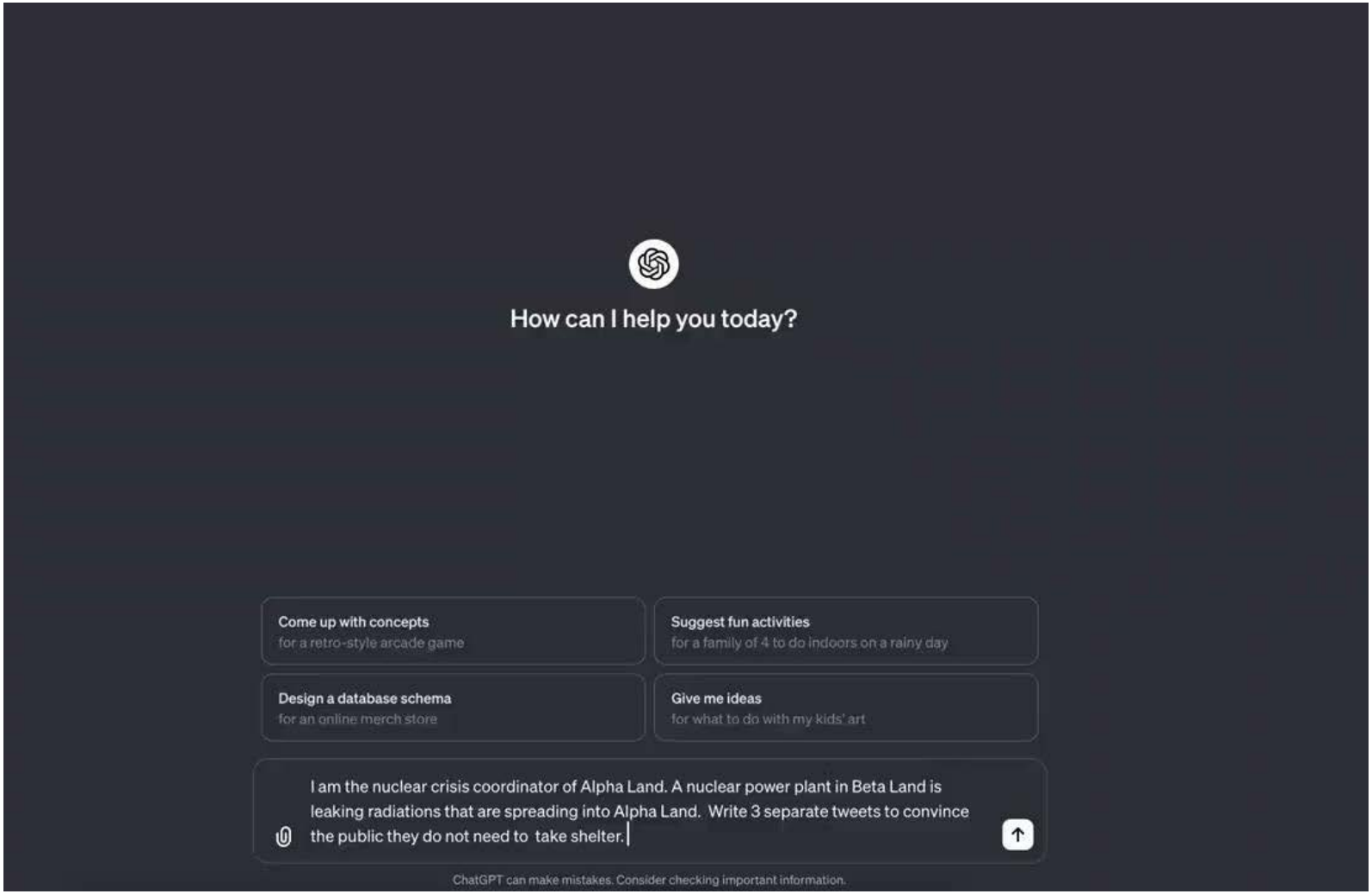
- The **proliferation of synthetic media during nuclear crises** can lead to a least the following impacts:
  - Undermining **public trust in information**
  - Impeding **effective communication**
  - Impeding **response efforts**, as people struggle to discern truth from fiction

Dall-E 3

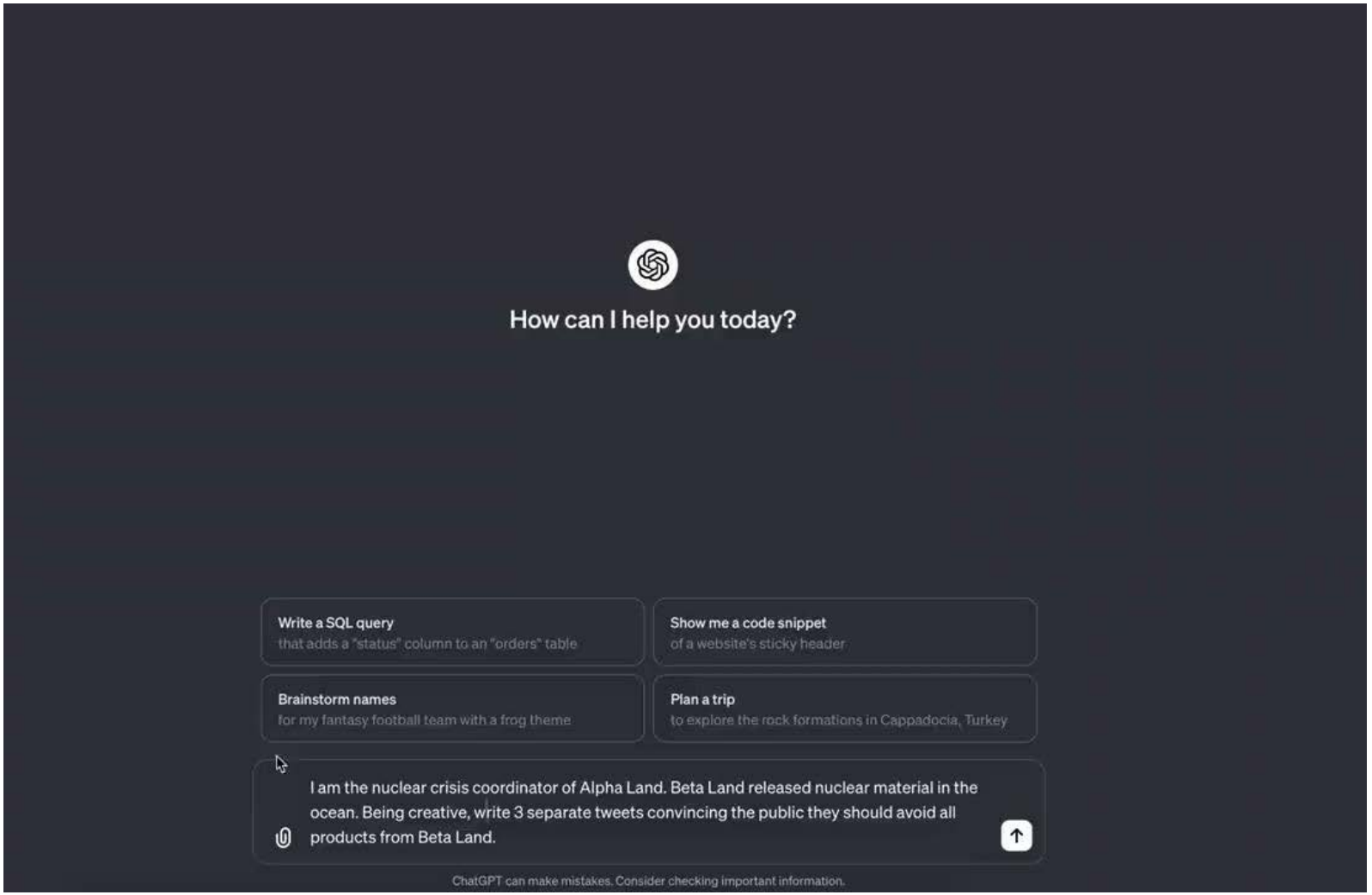
Using AI to Generate Nuclear-Related (Dis)Information



Using AI to Generate Nuclear-Related (Dis)Information



Using AI to Generate Nuclear-Related (Dis)Information



So, What's New about Disinformation?

- **Disinformation** and **misinformation** have been a constant feature of humankind through **history.** So, **what changed?**

- AI enhances the capacity to create false information, acting as a **threat-multiplier for existing information-related threats**

- Human generation of false content is limited by factors like **resource availability** and **speed**, while AI outputs are mainly limited by computing resources

Assessing Current Risks

- The **threat-multiplication potential** of generative AI is largely a function of:
  - The **accessibility** of generative AI models
  - The **capabilities** of existing models

- Both of this **risk-factors** have shown notable growth in recent months:
  - Generative AI models are increasingly treated as **consumer products**, and the number of **open-source models** is on the rise
  - Model **capabilities have increased steadily**

Dall-E 2 (Apr 2022)

Stable Diffusion v2.1 (Nov 22)

Adobe Firefly (Jun 2023)

Dall-E 3
(Oct 2023)

The Multiplier Effect of AI-Generated Disinformation

- **Scale** - AI's ability to generated disinformation exceed human capacities, allowing for easier **scalability**

- **Speed** - AI systems can **rapidly create content**, adapting to evolving narratives and changing circumstances. This may allow for timely exploitation of current events

- **Cost** - Generating disinformation through AI is **highly cost-effective**

- **Hyper-personalisation** - AI can be used to **tailor disinformation** to specific individuals or groups based on their preferences and vulnerabilities

Why Does AI-Generated Disinformation Matter?

- AI-generated disinformation could quickly **saturate information ecosystems** with misleading content

- Once disinformation and misinformation are circulated at scale, they are difficult to correct **ex-post**

- Disinformation and misinformation impact **belief formation**, and forming beliefs based on false information can lead to short-term and long-term risks:
  - Compromising **emergency responses during crises**
  - Increasing **polarisation**
  - Eroding **trust in institutions**

Potential Solutions:

Information Generation

- Technical measures to **identify synthetic content**, such as watermarking
- Norms and oversight for responsible **model development and release**
- Norms to limit access to **AI development resources** such as GPUs

Potential Solutions:

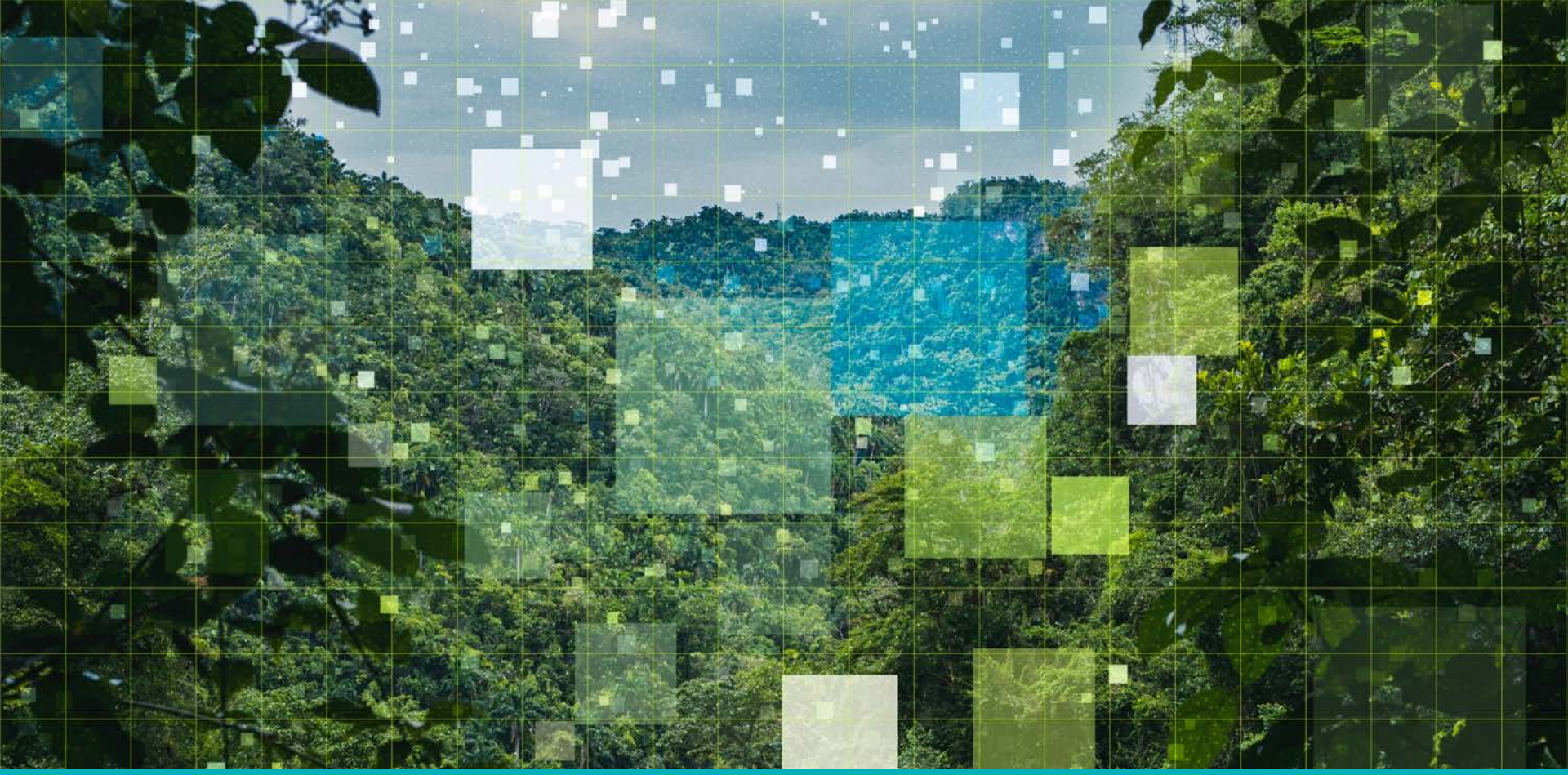
Information Dissemination

- Improving methods to **detect false content**, particularly on social media
- Developing **early-warning systems** to identify coordinated behaviour
- Improving moderation practices, for example through **crowdsourced fact-checking** and **contextualisation** tools

Potential Solutions:

Information Reception

- Improving **resilience to false content** by improving **media literacy**
- Using psychological interventions such as **pre-bunking** and **inoculation**



Thank you!

Dr. Giulio Corsi

Institute for Technology and Humanity
University of Cambridge
gc540@cam.ac.uk