UNIVERSITY OF CAMBRIDGE

67th REGULAR SESSION OF THE IAEA GENERAL CONFERENCE
SENIOR SAFETY AND SECURITY REGULATORS’ MEETING
September 28, 2023

Generative Artificial Intelligence, Disinformation and Misinformation: Addressing Current Challenges

CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE

Dr. Giulio Corsi
Leverhulme Centre for the Future of Intelligence
Centre for the Study of Existential Risk

The Problem: Disinformation and Misinformation in Online Environments

- **Disinformation and misinformation** continue to be prevalent in online environments

- There is a lack of clarity regarding how to effectively contrast disinformation and misinformation with **complex and volatile characteristics**

- Several **ethical and technical challenges** remain unresolved:
  - Objectively determining **factual accuracy** is difficult
  - Most governments are **hesitant to regulate** information flows
  - **Self-regulatory** efforts by online platforms have been largely ineffective

Framing Information Disorders: Epistemic Security

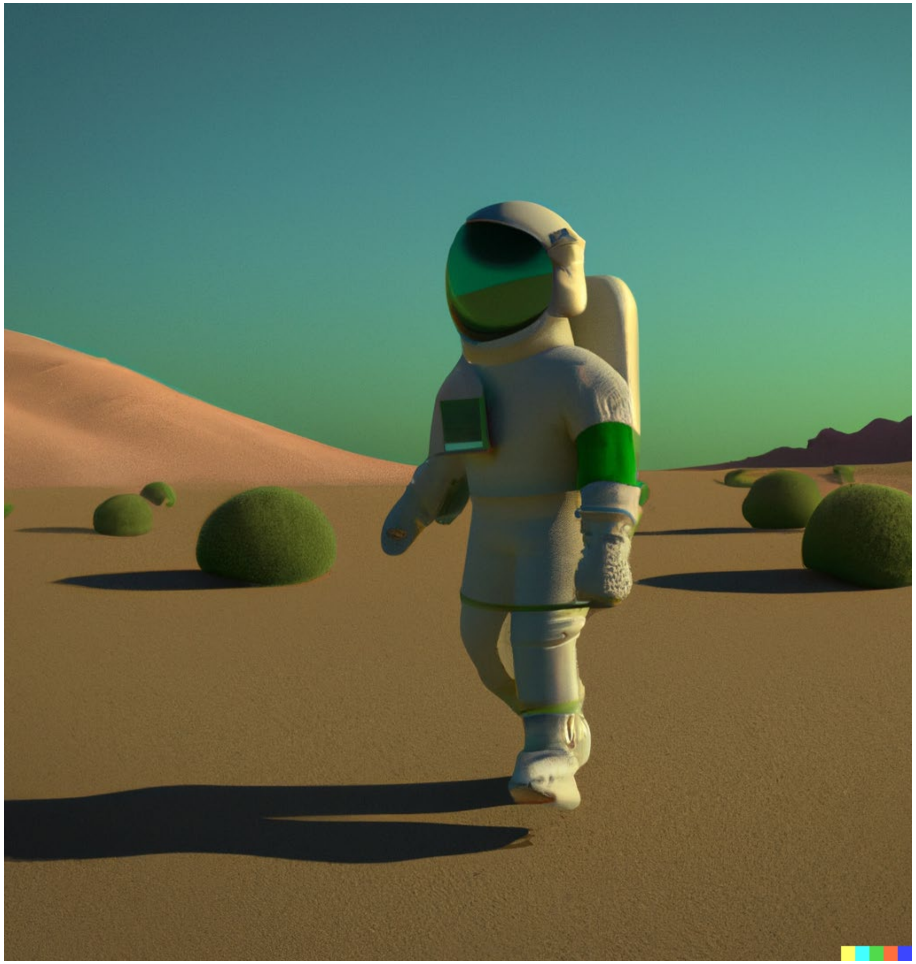
- **Epistemic security** concerns the protection and improvement of epistemic processes by which information is produced, processed and used to inform beliefs and decision-making procedures in society
- Information environments are **non-linear, complex systems**, with mechanisms such as **emergence** and **feedback loops**
- Taking a **systemic approach** to information disorders, epistemic security can be broken down into three parts:
  - Information **generation**
  - Information **circulation**
  - Information **acquisition** and **belief formation**

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



Synthetic Content Generation

- 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 and to agree on processes of information validation, potentially **endangering the integrity of information ecosystems**

Adobe Firefly

Stable Diffusion v2.1

Disinformation in Emergency Contexts

- In times of crises, the **public relies on accurate and trustworthy information** for rapid decision-making

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

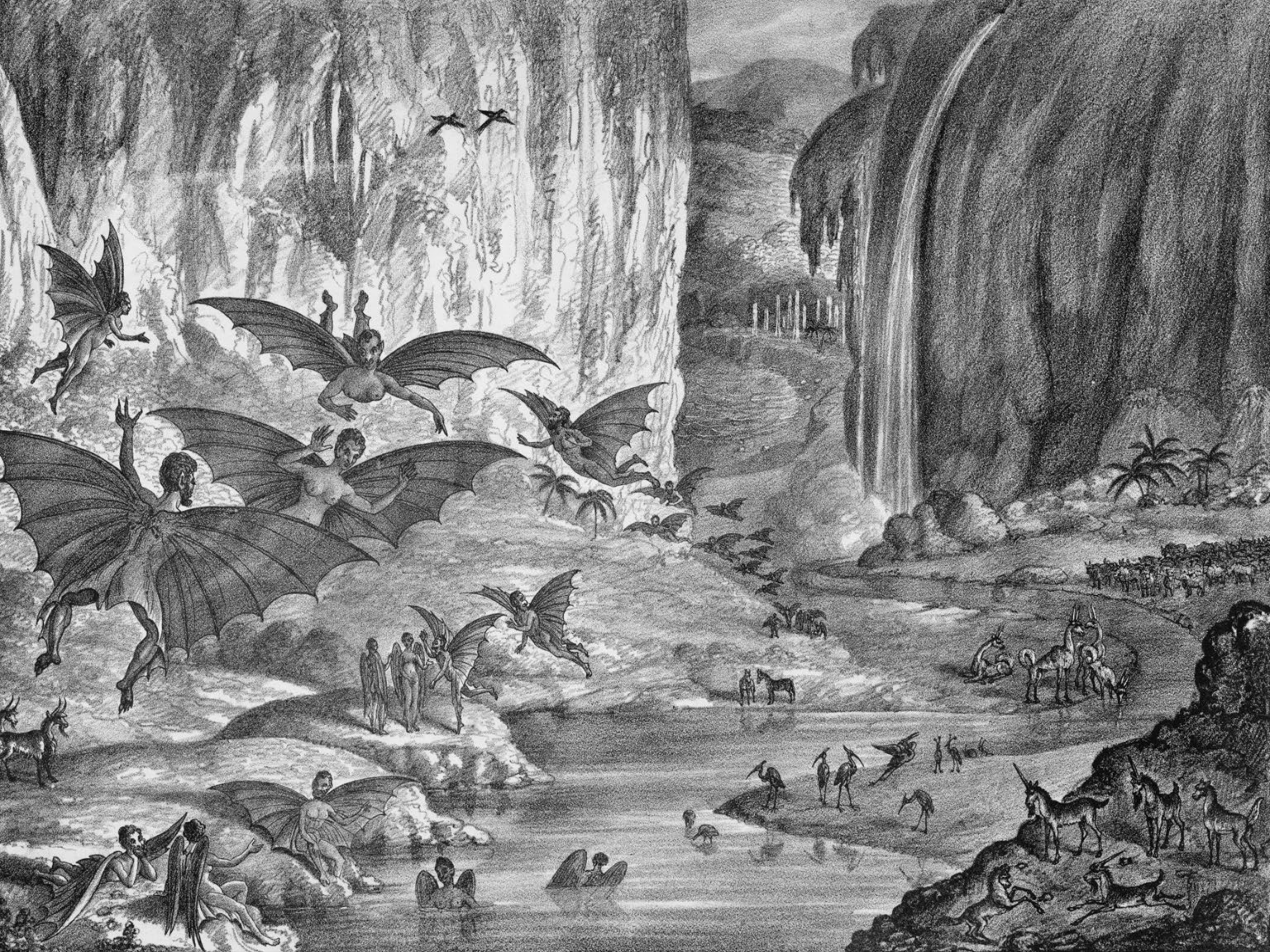
- This also applies to the **nuclear context**, where stakes are often high and misinformation could have severe consequences, such as delaying emergency responses and encouraging harmful behaviour

AI as an Epistemic Threat-Multiplier

- **Disinformation** and **misinformation** have been a constant feature of humankind through **history**

- AI enhances the capacity to create false information, acting as a **threat-multiplier for existing epistemic 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 will largely be 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)

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

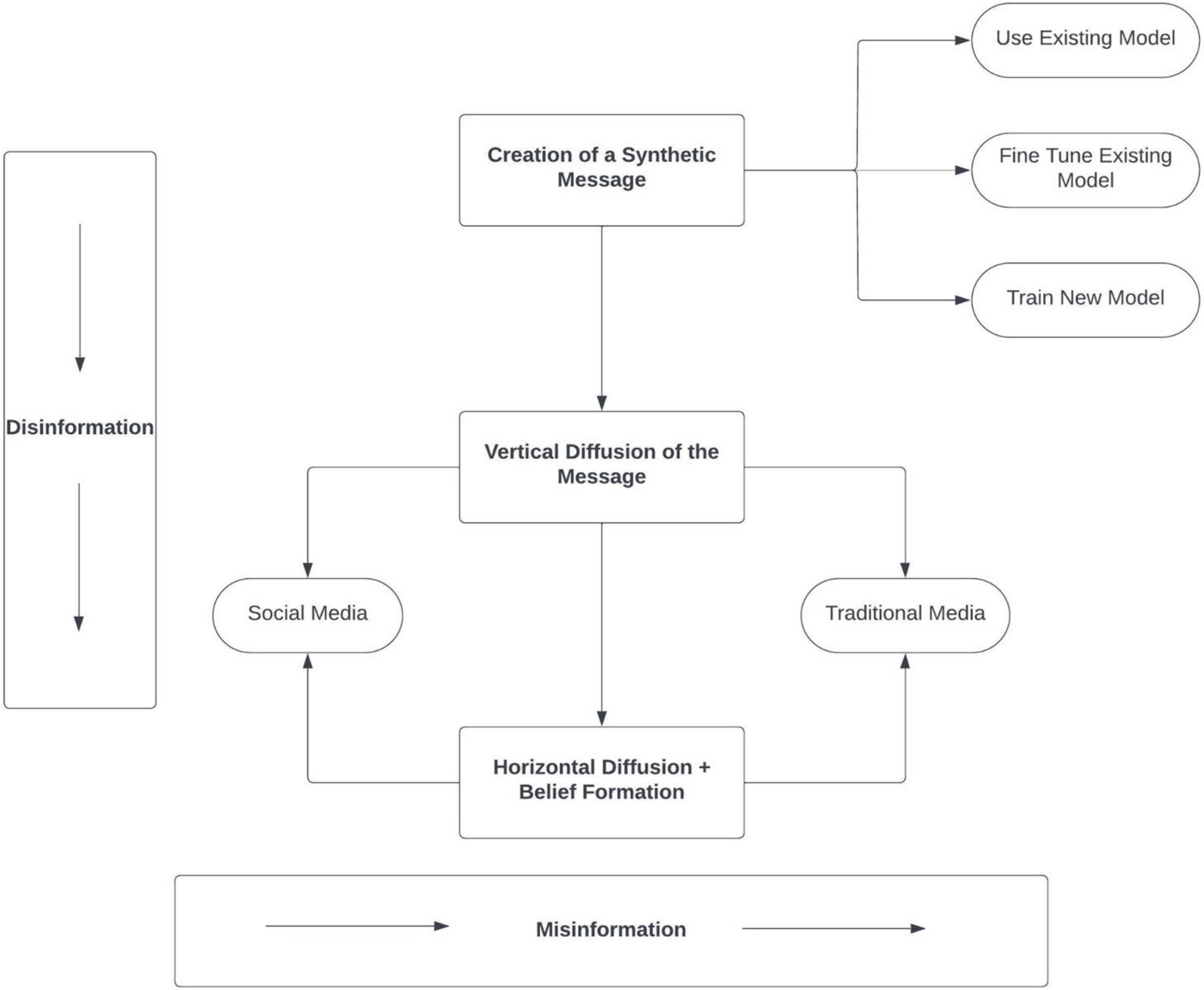
Why Does AI-Generated Disinformation Matter?

- 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**
    - Increasing **polarisation**
    - Eroding **trust in institutions**

Credits: Encyclopedia Britannica


Use Existing Model
Creation of a Synthetic Message
Fine Tune Existing Model
Disinformation
Train New Model
Vertical Diffusion of the Message
Social Media
Traditional Media
Horizontal Diffusion + Belief Formation
Misinformation

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

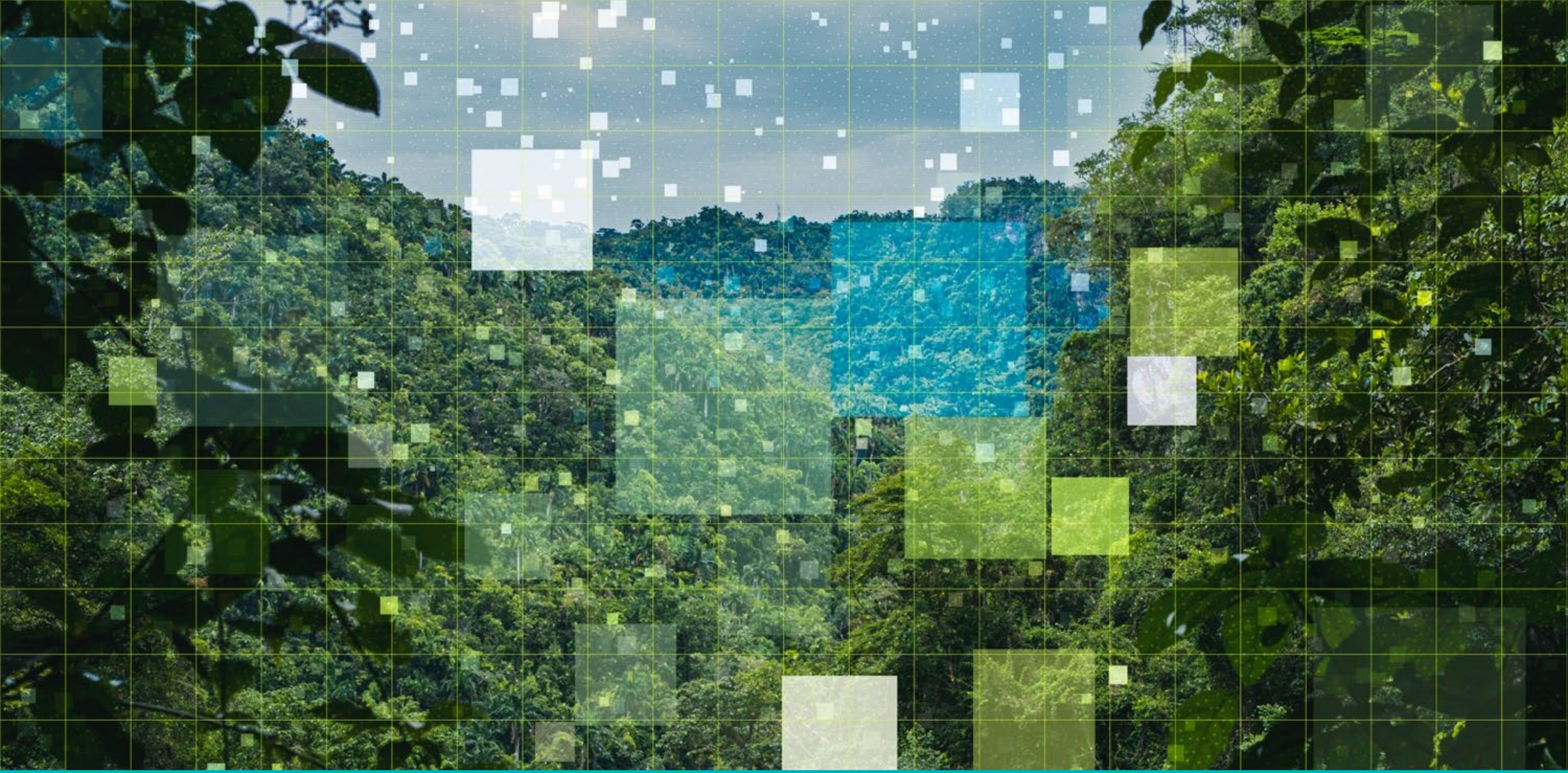
- Improving moderation practices, for example through **crowdsourced fact-checking** and **contextualisation** tools

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

Leverhulme Centre for the Future of intelligence
Centre for the Study of Existential Risk
gc540@cam.ac.uk