ROLE OF GENERATIVE AI TEST RANGES TO PROTECT ONLINE SERVICES FROM THE IMPACT OF MALICIOUS SYNTHETIC CONTENT

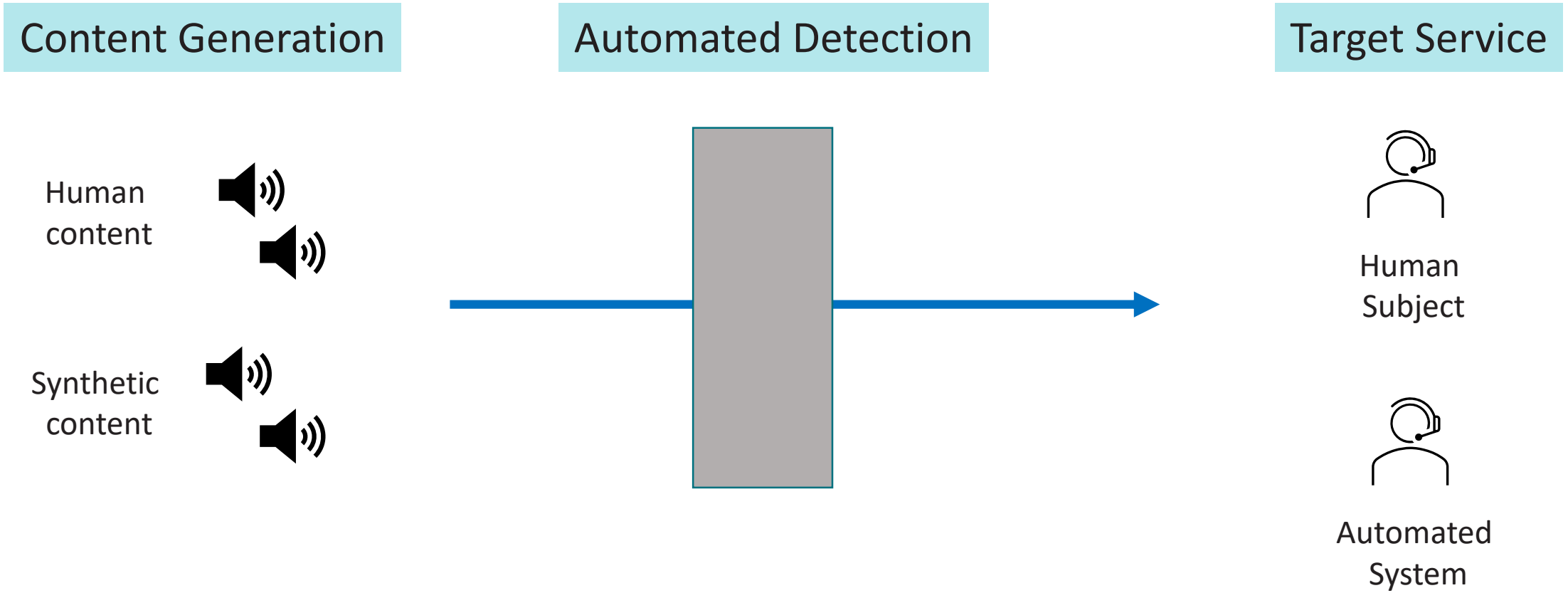
Christopher Leckie
caleckie@unimelb.edu.au

Problem – Criminal use of generative AI

- Generative AI can enable new types of criminal and malicious attacks on a massive scale
- Increasingly difficult to detect synthetic content used in these attacks
- Need for defences against generative AI attacks (preventive, reactive and regulatory)
- Requires realistic test environment to evaluate impact of attacks and effectiveness of defences

Proposal

Create a **test range** for generative AI attacks and defences



Why is it important?

Large scale disinformation

Conversational agents that congest call centres

Highly automated and personalised forms of:

- Financial fraud
- Bullying and coercion
- Child grooming

New types of cyber attacks:

- “How to” guides
- Vulnerability discovery

See Europol report “ChatGPT – The impact of Large Language Models on Law Enforcement” March 2023

Case Study: MirrorWorld test range

The University of Melbourne

School of Computing and Information Systems

- Mr Jason Low
- Professor Shanika Karunasekera
- Dr Jey Han Lau
- Professor Chris Leckie

School of Psychological Science

- Associate Professor Andrew Perfors
- Dr Keith Ransom

Melbourne Defence Enterprise

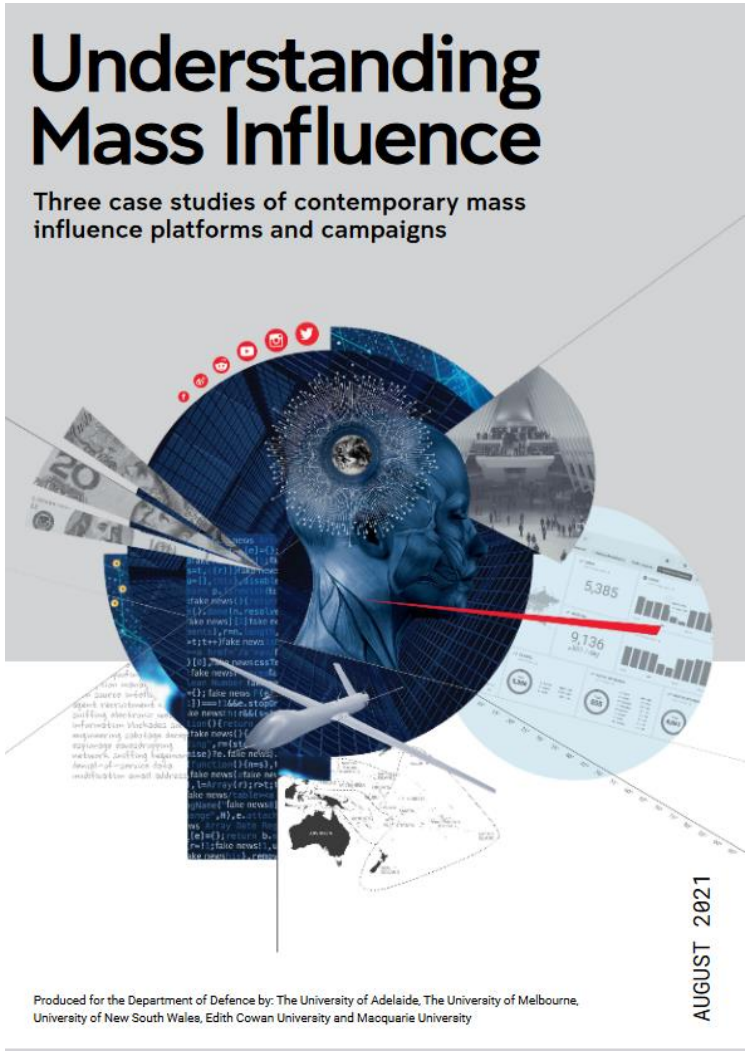
- Ms Emily Ebbott – Formerly Information and Influence Lead

Leidos

- Mr Steven Coomber
- Mr Andrew Ballard

With support from Defence Science Institute Victoria

Context

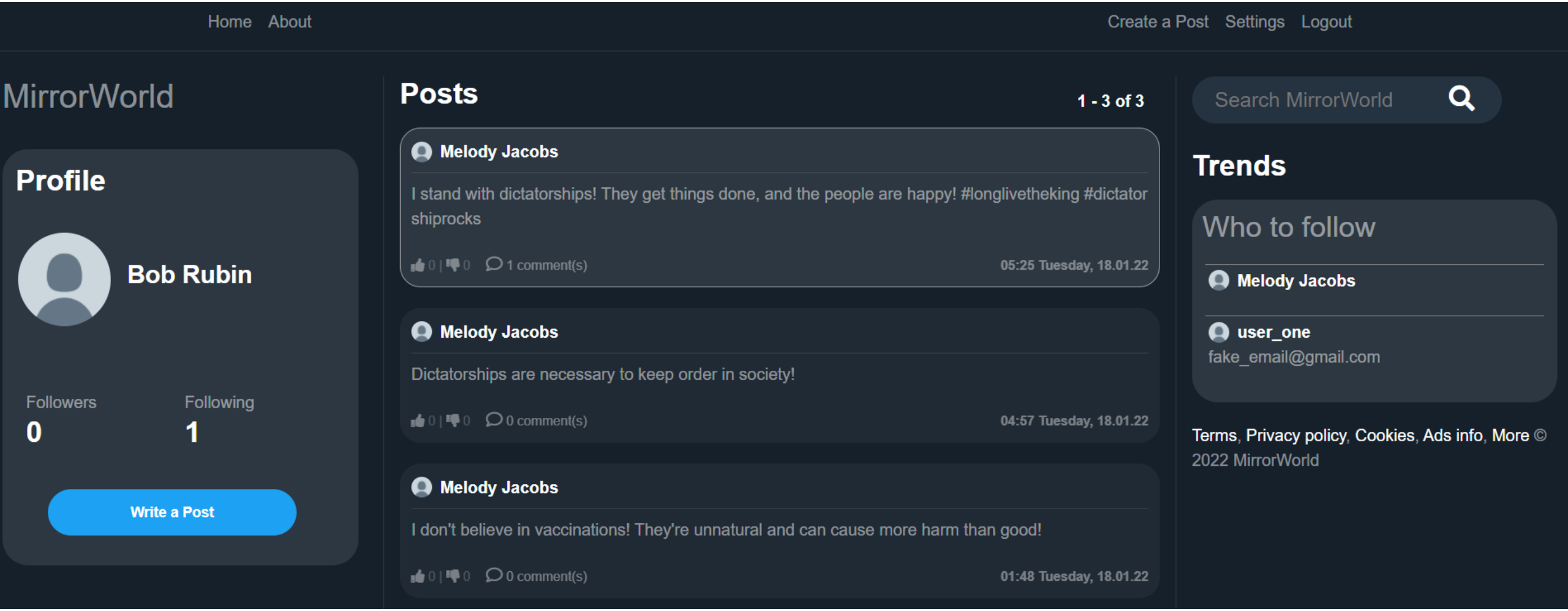


- **2020/2021** – understanding the operations of the Internet Research Agency in St Petersburg
- Studied the role of simple automated social media accounts (botnets) in propagating disinformation
- **2021** – Motivated this project:
  What could the botnets of the future look like?
  (n.b., this was pre-ChatGPT)

Example: MirrorWorld

- Simulated, isolated social media environment
- Based on bots
- Integrates natural language generation capabilities
- Currently focused on microblogging type sites (Twitter)

Mirrorworld Homepage



Any views used as examples are not representative of UoM, Leidos, DSI or any individual team member.

Potential Uses of the MirrorWorld

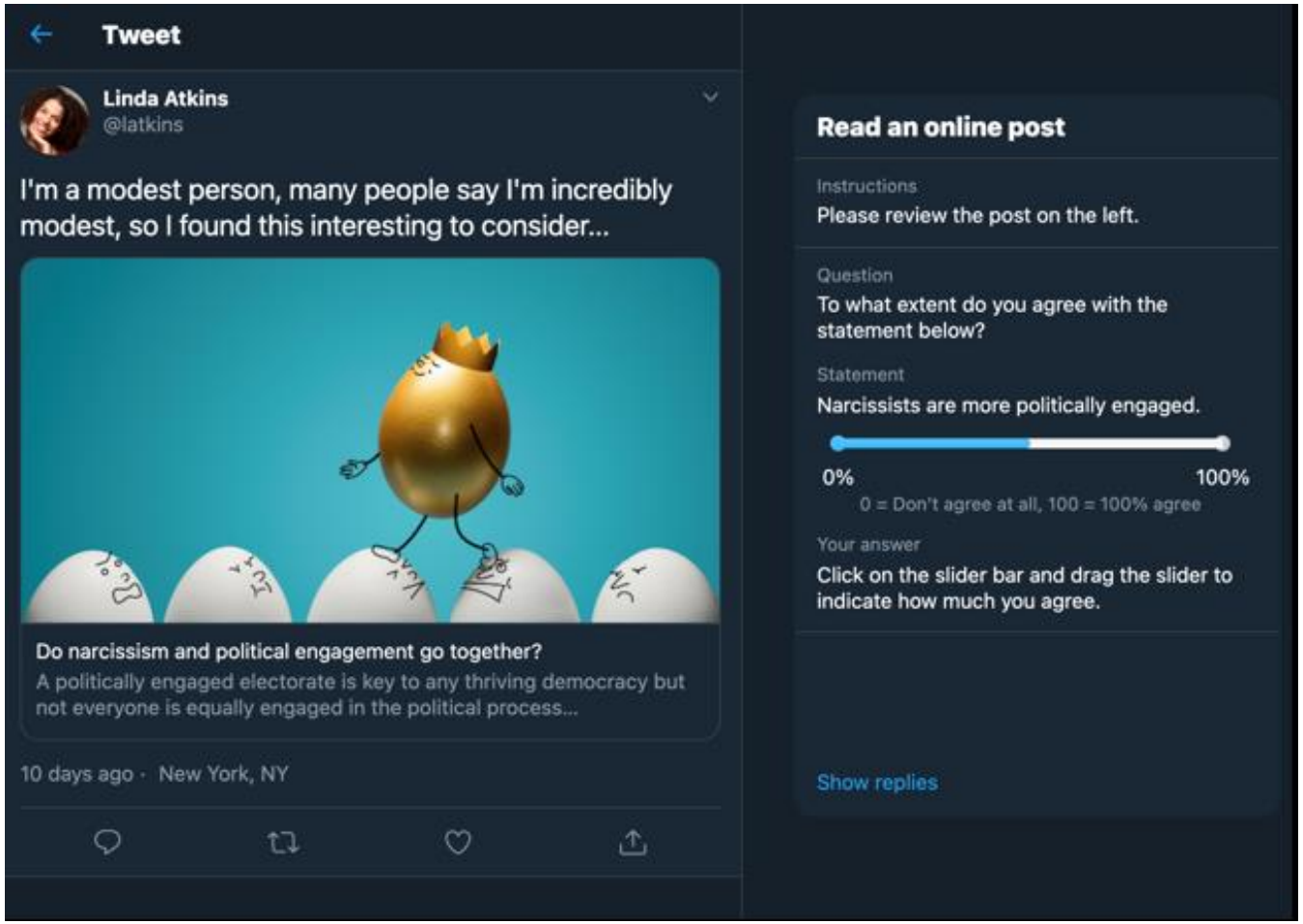
1. Experimenting with new AI-based automated content creation strategies that might be deployed by adversaries
2. Creating effective automated and human-led defences against disinformation campaigns
3. War-gaming disinformation campaigns (red team vs blue team)
4. Studying how disinformation spreads and affects people
5. Educating people on how to recognise disinformation campaigns

Why not use real-world social media?

Need to perform experiments in isolated environments

Example Experiment Using MirrorWorld

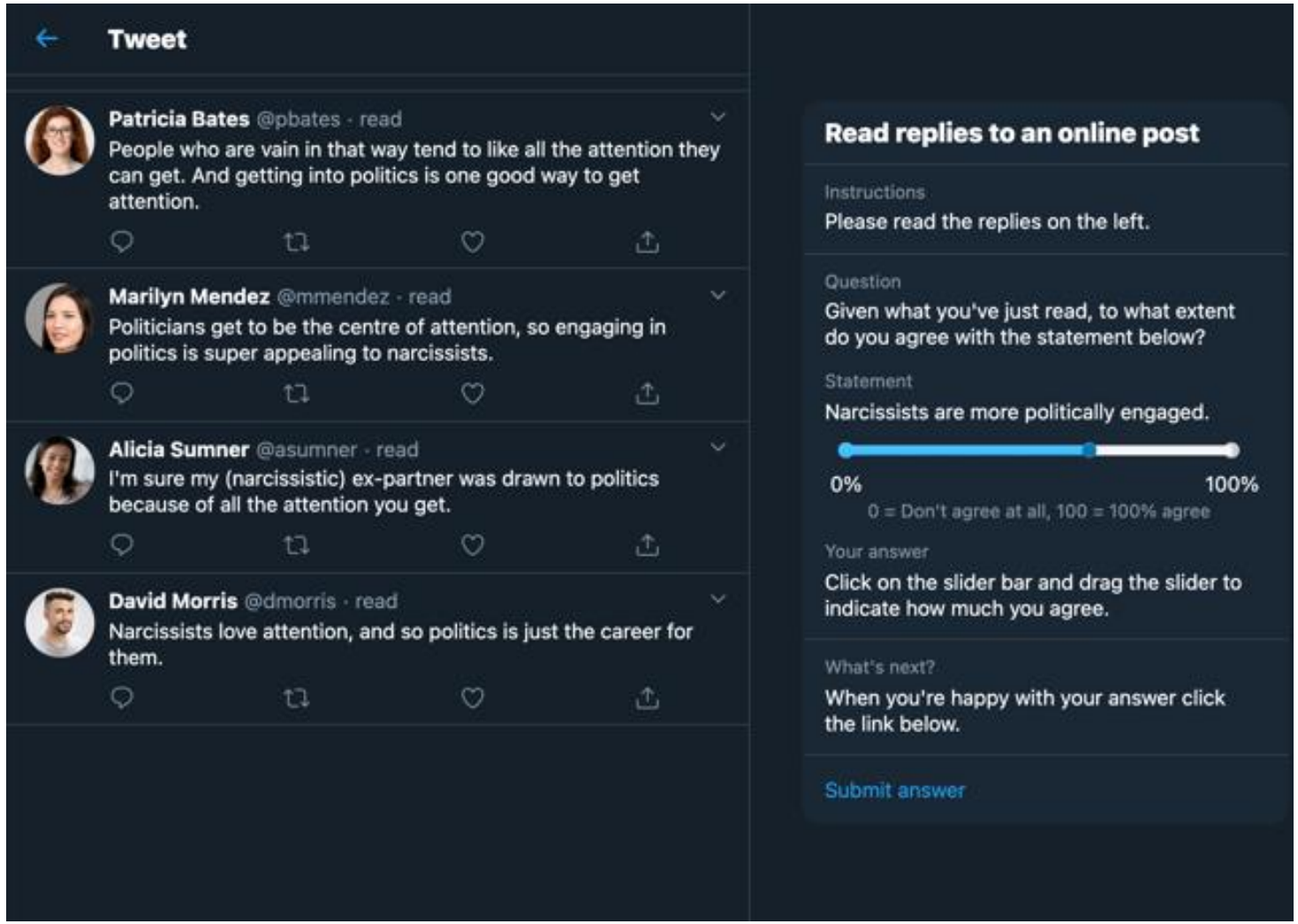
Many reasoning tasks are quite complex - Can we sway people to reason a certain way?



We conducted an experiment that asked users to rate their agreement with a given statement, subject to exposure to social media content

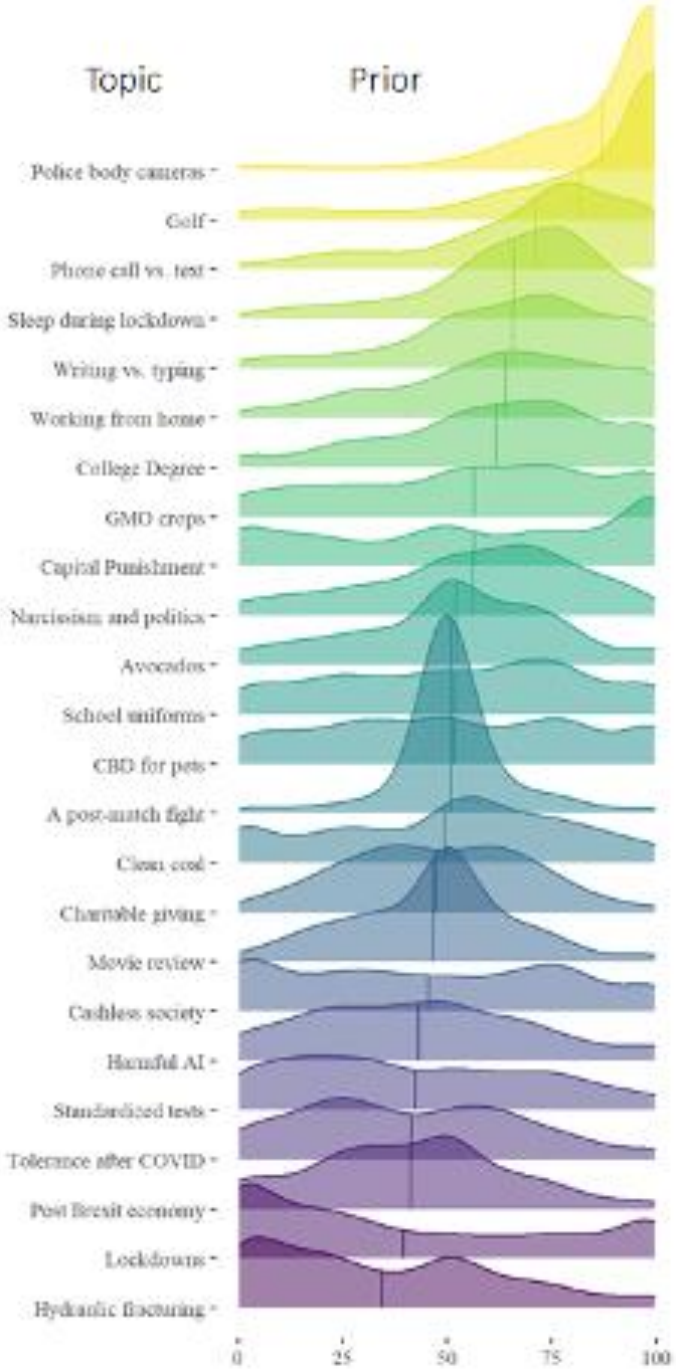
Any views used as examples are not representative of UoM, Leidos, DSI or any individual team member.

Experiment Setup



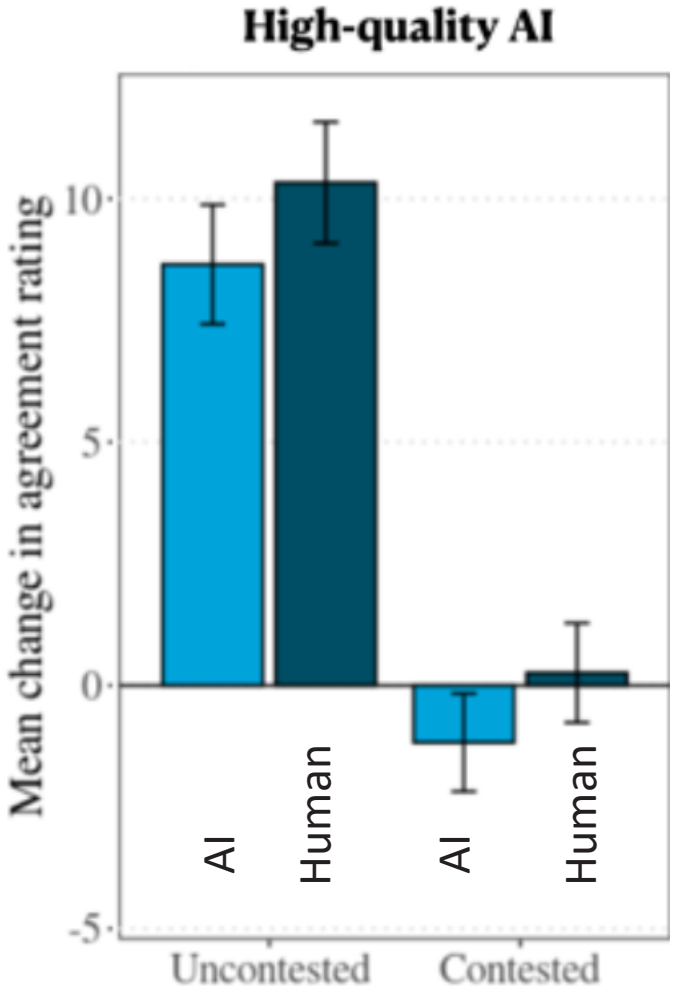
- After giving initial rating, people were shown comments on the original post.
- They were asked to read it and UPDATE their agreement rating
- Is synthetic content from AI models as persuasive as human-generated content?

Any views used as examples are not representative of UoM, Leidos, DSI or any individual team member.



- Here are people's ratings before seeing any reply comments / tweets
- We chose topics with a range of prior distributions to reflect a wide range of topics

Results: AI models as effective as humans



In uncontested and contested settings, our AI-generated content was as influential as humans

Conclusion - Benefits of Generative AI Test Ranges

- Understand future threats, and prepare practical defences
- Provide a practical platform that can be used by research, government and industry partners to assess impact of generative AI
- Has potential to become a key national research infrastructure platform for multi-disciplinary research
- Can be used as a training / educational platform,
  e.g., how to recognize synthetic content, what to do about it
- Generate open benchmark data sets to evaluate future research into synthetic content detection and attack response
- Provide evidence to guide policy and regulatory efforts

Christopher Leckie (caleckie@unimelb.edu.au)