



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

Christopher Leckie
caleckie@unimelb.edu.au

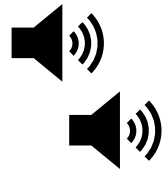


- Generative AI can enable new types of criminal and malicious attacks on a massive scale
- Increasingly difficult to detect synthetic content used in these attacks
- Need for defences against generative AI attacks (preventive, reactive and regulatory)
- Requires realistic test environment to evaluate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

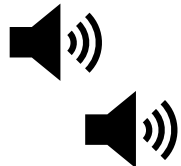
Create a **test range** for generative AI attacks and defences

Content Generation

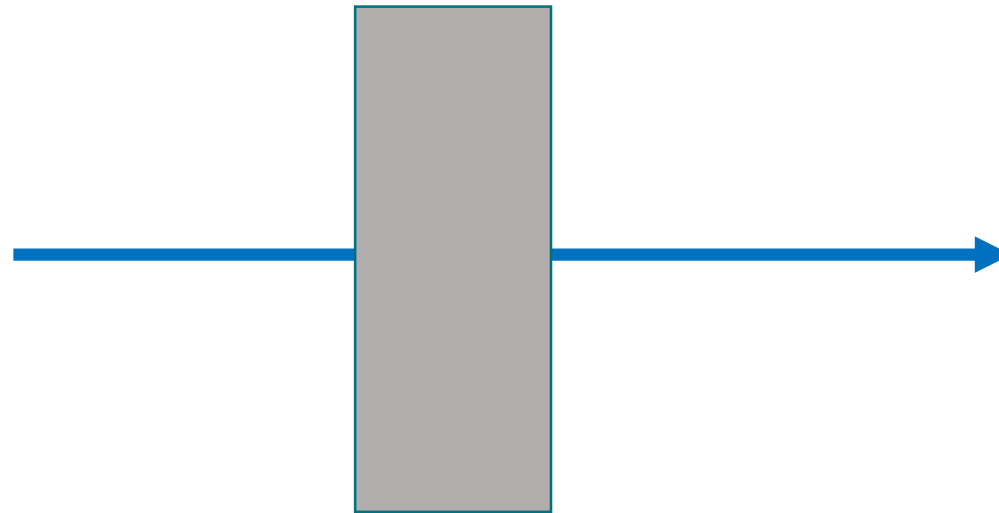
Human
content



Synthetic
content



Automated Detection



Target Service



Human
Subject



Automated
System

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

Understanding Mass Influence

Three case studies of contemporary mass
influence platforms and campaigns



AUGUST 2021

Produced for the Department of Defence by: The University of Adelaide, The University of Melbourne,
University of New South Wales, Edith Cowan University and Macquarie University

- **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)

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

MirrorWorld

1 - 3 of 3

Search MirrorWorld



Profile



Bob Rubin

Followers


0

Following

1

Write a Post


Posts

 **Melody Jacobs**

I stand with dictatorships! They get things done, and the people are happy! #longlivetheking #dictatorshiprocks

 0 |  0 |  1 comment(s)


05:25 Tuesday, 18.01.22

 **Melody Jacobs**

Dictatorships are necessary to keep order in society!

 0 |  0 |  0 comment(s)

04:57 Tuesday, 18.01.22

 **Melody Jacobs**


I don't believe in vaccinations! They're unnatural and can cause more harm than good!


 0 |  0 |  0 comment(s)

01:48 Tuesday, 18.01.22

Trends

Who to follow

 **Melody Jacobs**

 **user_one**
fake_email@gmail.com

Terms, Privacy policy, Cookies, Ads info, More ©
2022 MirrorWorld

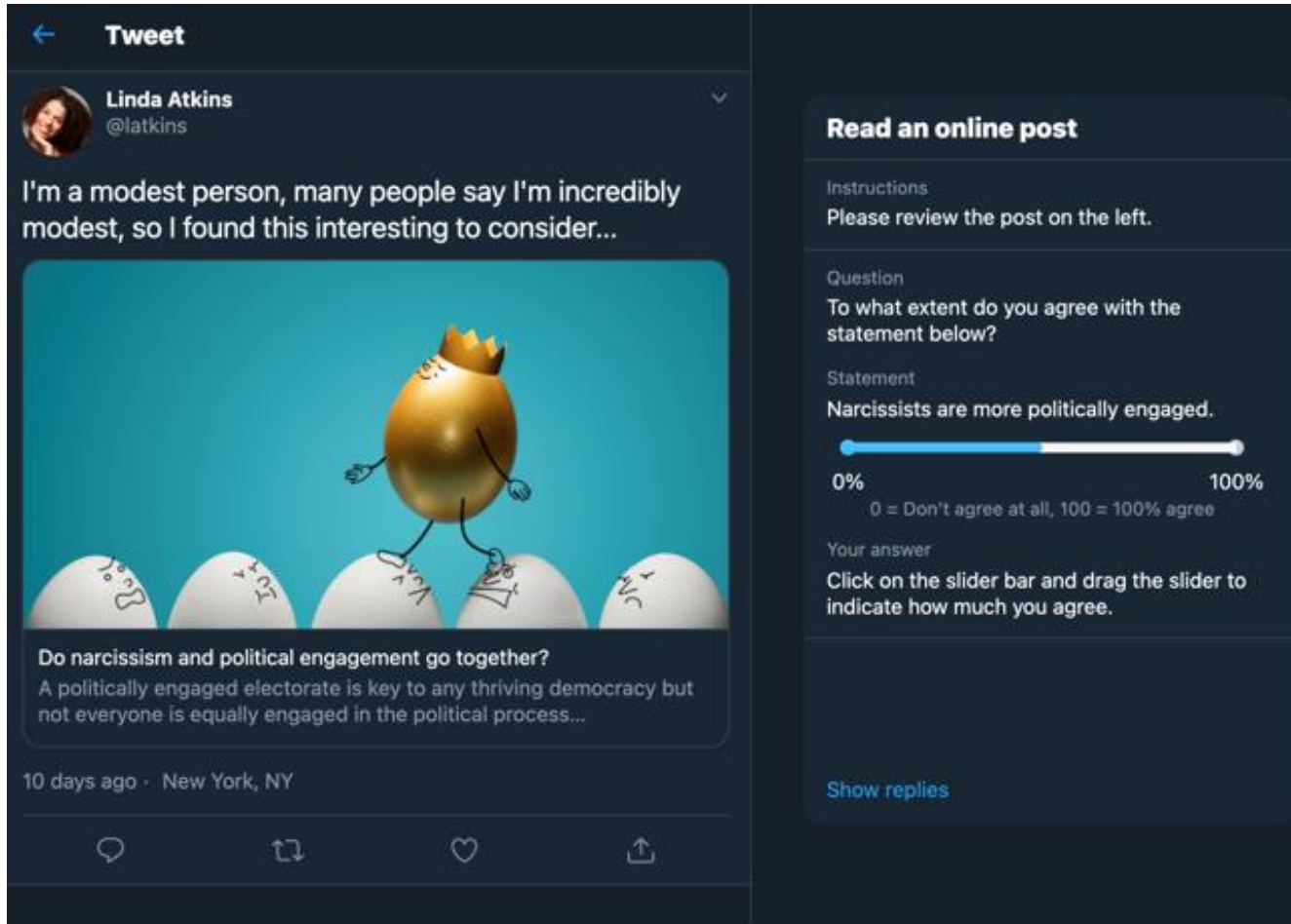
1. Experimenting with new AI-based automated content creation strategies that might be deployed by adversaries
2. Creating effective automated and human-led defences against disinformation campaigns
3. War-gaming disinformation campaigns (red team vs blue team)
4. Studying how disinformation spreads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



Tweet

Linda Atkins
@latkins

I'm a modest person, many people say I'm incredibly modest, so I found this interesting to consider...



Do narcissism and political engagement go together?
A politically engaged electorate is key to any thriving democracy but not everyone is equally engaged in the political process...

10 days ago · New York, NY

Read an online post

Instructions
Please review the post on the left.

Question
To what extent do you agree with the statement below?

Statement
Narcissists are more politically engaged.

0% 100%
0 = Don't agree at all, 100 = 100% agree

Your answer
Click on the slider bar and drag the slider to indicate how much you agree.

Show replies

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

Experiment Setup



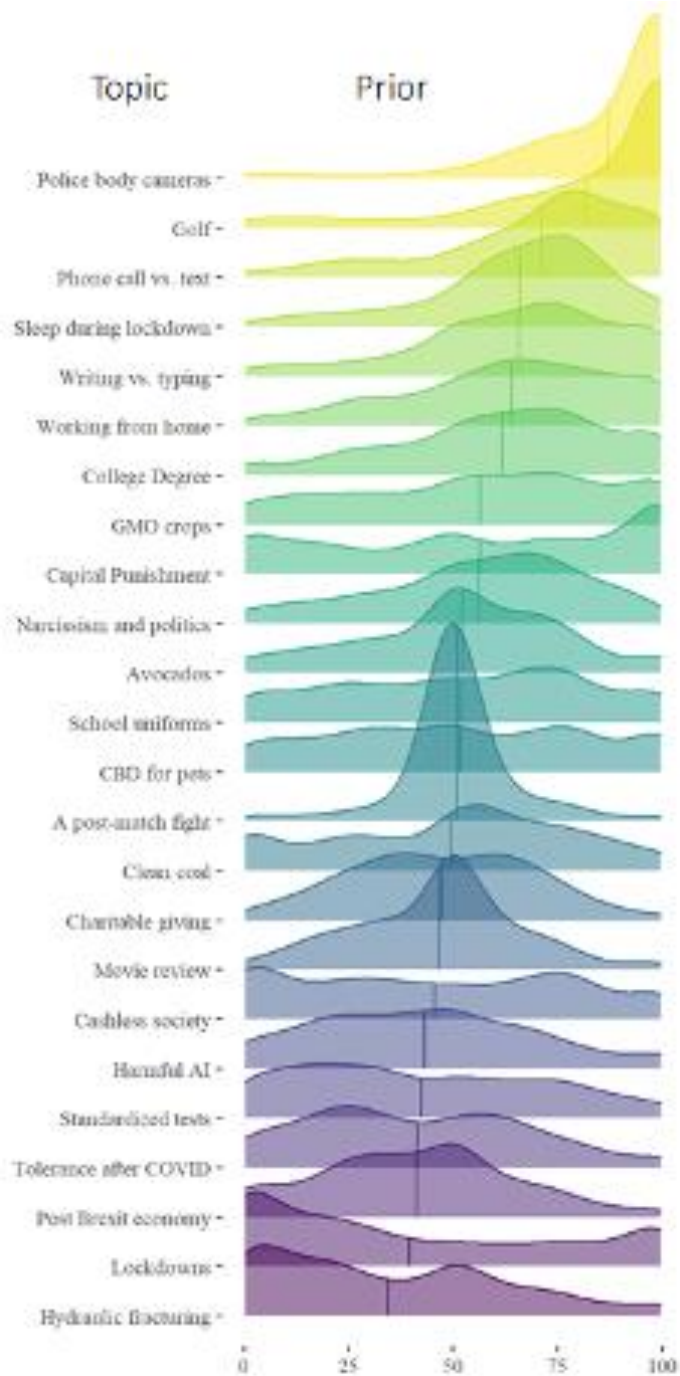
The image shows a screenshot of a Twitter thread on the left and a survey interface on the right. The Twitter thread contains four tweets:

- Patricia Bates @pbates**: People who are vain in that way tend to like all the attention they can get. And getting into politics is one good way to get attention.
- Marilyn Mendez @mmendez**: Politicians get to be the centre of attention, so engaging in politics is super appealing to narcissists.
- Alicia Sumner @asumner**: I'm sure my (narcissistic) ex-partner was drawn to politics because of all the attention you get.
- David Morris @dmorris**: Narcissists love attention, and so politics is just the career for them.

The survey interface on the right is titled "Read replies to an online post". It includes the following sections:

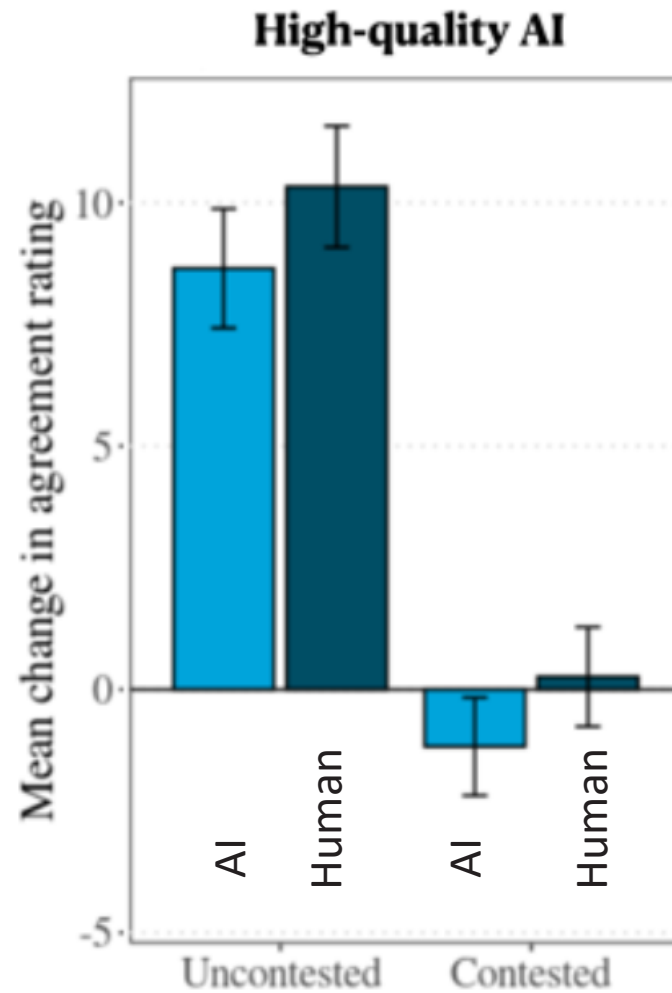
- Instructions**: Please read the replies on the left.
- Question**: Given what you've just read, to what extent do you agree with the statement below?
- Statement**: Narcissists are more politically engaged.
- Slider**: A horizontal slider bar ranging from 0% to 100%. The slider is currently positioned at approximately 40%. Below the slider, it says "0 = Don't agree at all, 100 = 100% agree".
- Your answer**: Click on the slider bar and drag the slider to indicate how much you agree.
- What's next?**: When you're happy with your answer click the link below.
- Submit answer**: A blue button at the bottom.

- 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?



- 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)