Cognitive hacking and Big Tech platforms

Steven Coomber

IAA - Public Sector Assurance Forum

4 August 2022

Synergy

DEFINITIONS

Disinformation:

False information spread deliberately to deceive

Misinformation:

False information but not deliberate

Malinformation:

Information stemming from the truth but exaggerated in a way that misleads and causes harm

Information warfare:

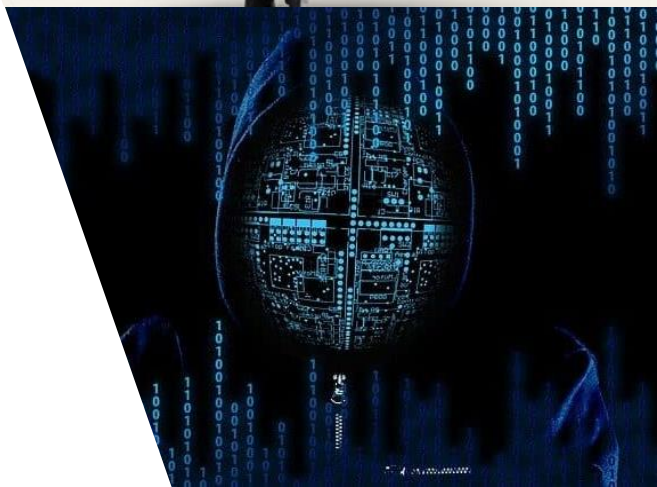
Contest for provision and assurance of information supporting friendly decision-making, while denying and degrading adversaries

Fake news:

False or misleading information presented as news

Propaganda:

Misleading or biased information used to promote a political cause or point of view



COGNITIVE HACKING IS CYBER

Cognitive hacking

- Cyberattack using disinformation manipulating perception and exploiting psychological vulnerabilities to change behaviour
- Against data, systems, communications platforms, networks, infrastructure, network security
- Cybersecurity threat
- Solutions part of cyber domain

Synergy

TECHNIQUES

False reports	Out of context	Memes	Deepfakes
AI bots	Trolling	Flooding	Astroturfing
‘Testimonials’	Distorted images	Favourable influencers	Cross-platform activity
Repurposed spam	Coordinated posting	Narrative laundering	Edited audio

Synergy

WHAT DOES IT DO?

To what degree is disinformation a threat and be intentionally weaponised against a population without it's knowledge?

- Creates disharmony
- Shapes views
- Amplifies grievances
- Exacerbates tensions
- Attracts attention
- Encourages protests
- Promotes bad health advice
- Sows discord
- Unsettles communities
- Influences corporate decisions
- Weakens brands
- Tactical military advantage
- Damages reputations
- Financial crime
- Fraud
- Theft
- Inaccurate records
- Insider threats
- Insecure information
- Impacts stock market
- Government mistrust
- Destabilises governments
- Influences elections

Synergy

HOW IT MIGHT WORK

- **Multiple sources** more persuasive
- **Endorsement** by large numbers
- **Familiar themes** appealing even if false
- From **groups recipient belongs**
- **Appearance** of expertise/trustworthiness
- Confirmation bias **reaffirm beliefs**
- Backed by **apparent evidence**
- ***Sleeper-effect*** **low credibility source** remembered true when source forgotten
- **Initially assumed valid** but proven false
- **Rapid, continuous** and **repetitive**

Synergy

WHAT IS THE IMPACT?

2022 Philippines Election – Ferdinand 'Bongbong' Marcos Jr

- Officially launched campaign Facebook video
- False social media undermining reputation, whitewashing scandals and enhancing family
- Most viral news driven by social media shares

2018 Brazilian Election – Jair Bolsonaro

- Relied heavily on social media
- Brazil 120m Whatsapp users – 44% electorate
- Followers created Whatsapp groups
- Disseminated misinforming content against political rivals and minority groups

Synergy

WHAT IS THE IMPACT?

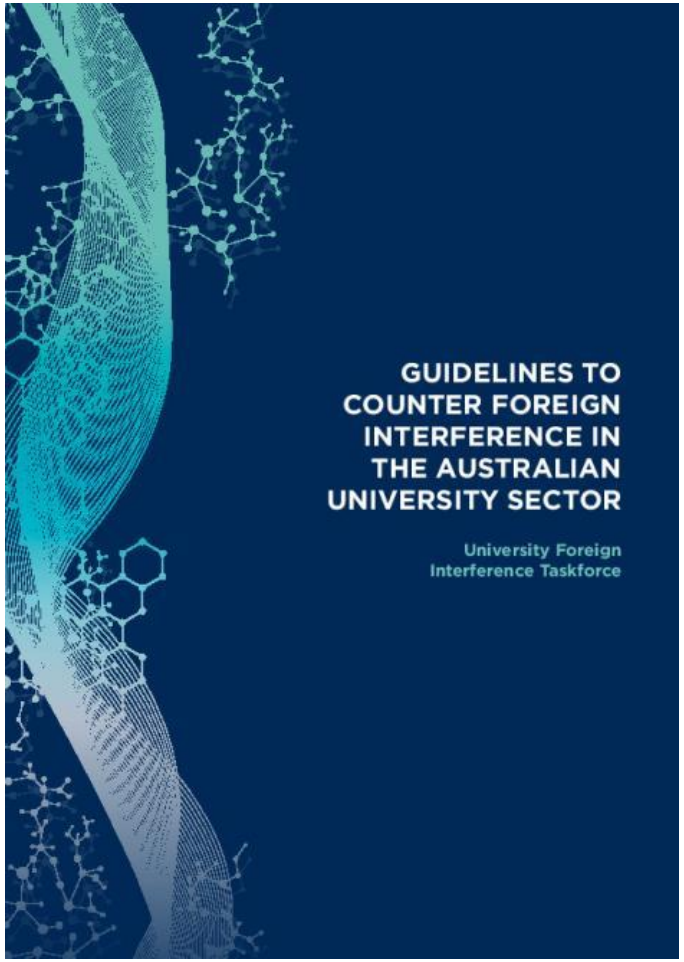
2016 US Presidential Election – Donald Trump

- Exploited online media to agitate and motivate vast, unseen audience
- Mastered Twitter
- Redefining its power as tool of political promotion, distraction, score-settling and attack

Synergy

GOVERNMENT

- DFAT disinformation taskforce
- DHA counter foreign interference taskforce
- Guidelines to counter foreign interference in Australian university sector
- **Defence Information Warfare Division** – Joint Influence Activities
- **Foreign influence transparency scheme**
- **Legislation**
  - ACMA regulatory powers holding big tech accountable for harmful content
  - Platforms developing voluntary Australian code of practice
  - Gamified e-smart digital literacy course for primary and high school



Synergy

Meta

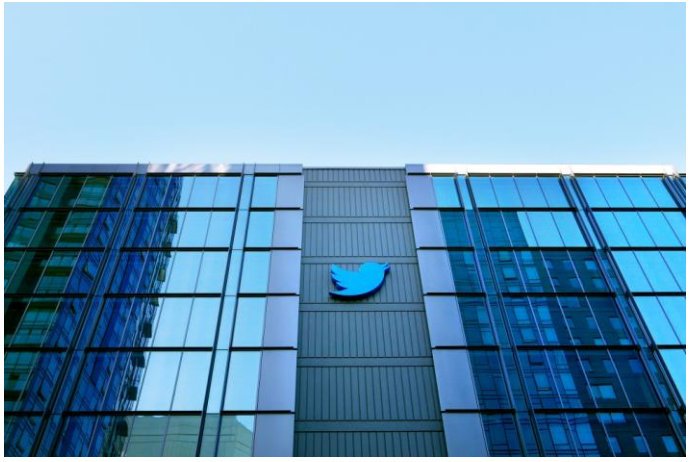
- Parent company of Facebook, Instagram, and WhatsApp

Facebook

- Online social media and social networking service
- Psychological persuasion techniques facilitate user engagement
- Very efficient propagating emotionally resonating content appealing to user preferences/beliefs
- **Algorithms prioritise content aligning user interests over accuracy more likely to be shared regardless of veracity**
- Uses algorithms for targeted advertising based on user data
- **More likes, shares and comments more engagement**
- Not financially incentivised to weed out dis/misinformation
- **Users don't need to create content, longer users spend longer time for ads**

TWITTER

- Microblogging social networking communications service
- **Extremely important global public discourse platform**
- **Evolved into very powerful influential private entity for political campaigns**
- Algorithms and users reporting to detect objectionable content
- Own censorship rules deciding who gets banned
- **Content removed on ideology** creating perceived bias
- After claims tweets sparked violence and mob stormed U.S. Capitol 6 January 2021 Trump permanently suspended
  - **Political censorship vs free-speech**



Synergy

Bespoke algorithms

- Short-form video hosting service owned by Chinese company ByteDance
- **Powerful AI behavioural profile algorithms**
- Predicting friends
- Positive-negative AI stimulating emotions to redirect
- **Repeated exposure to positive emotions subconsciously links to propaganda/politics**
- **User-specific profile learning trigger stimuli**
- Addictive using compelling stimuli to spend hours on
- Ukraine invasion world's first 'TikTok war'
- Zelenskiy asked 'TikTokers' to help war and Joe Biden briefed top users
- Features and opaque algorithms prime to remix disinformation media

Synergy

MEDIA REACTIONS

Left vs Right

- January 2018 then POTUS reportedly referred to African nations and Haiti as ***'sh*thole countries'*** when discussing immigration deal

Reactions

- **CNN's Don Lemon** called President Trump a *'racist'* saying he wasn't shocked

- **CNN's Anderson Cooper** emotional saying White House could learn from *'dignity'* of Haitian people

- **Fox's Jesse Watters** defended President Trump saying many American people *'speak similarly'*

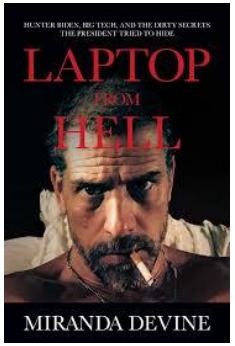
- **Fox's Tucker Carlson** defended President Trump saying *' awful lot of immigrants come from dangerous and dirty countries'*

Purr and Snarl: *terrorist vs freedom fighter, seal harvest vs sea pup slaughter, vandalism vs street art, fetus vs unborn child, management offers vs union demands, protester vs rioter*

Synergy

HUNTER BIDEN

- Laptop left at Delaware repair shop 2019
- October 2020 New York Post published story
- Twitter blocked newspaper's account and users sharing
- Hunter Biden introduced then VP Joe Biden to Ukrainian energy company Burisma
- Biden pressured Ukraine Government into firing investigating prosecutor
- Twitter said decision to block was mistake



'10 held by H for the big guy?'

Synergy

CANBERRA CONVOY

- Anti-vaccine mandate protests February 2022
- Disparate groups protesting different issues part protest/reality TV
- Influencers para-social intimate relationship generating loyalty, affection and support
- More social media content than protest
- Selling merchandise, livestreaming and latest interpersonal dramas
- Hothouse minute-by-minute content creating visible divisions

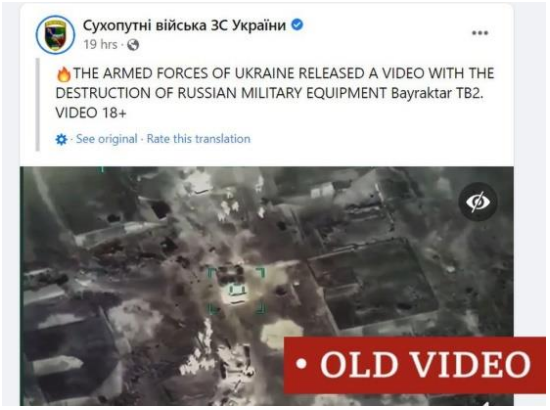
Anti-vaccination, anti-vaccine mandate activists, Sovereign citizens, ultra religious groups, United Australia Party, indigenous rights activists

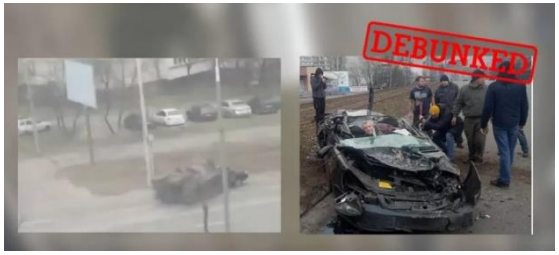
FALL OF MOSUL

Mosul

- Iraqi security **fell to ISIS (<1500) in three days** 2014
- **Choreographed marketing campaign** – Twitter and Instagram selfies
- **Exciting edited pictures** truck convoys black flags
- **Bots amplifying and exploiting social media algorithms**
- **Mobile apps watch at home** feeling part of battle
- **ISIS arrived defenders panicked** spreading fear
- Saw convoys and suicide bombings **deserting flood**
- **Better equipped soldiers abandoned US weapons**
- **Transformed narrative through media exaggerating capabilities** and territorial gains polarising population **driving rivals away**
- Online atrocities in **psychological operations**
- **No sophisticated military of cyber** capability
- **Manipulated information mitigating** shortcomings
- Impression unstoppable mission **making minor firefights appear heroic victories**

GHOSTS OF KYIV





CITIZEN WARFARE

Elon Musk

- Starlink satellite internet services

Anonymous

- Declared cyber war on Russia
- Denial of services hacking Defence databases and hijacking state TV

Civilian drones

- Consumer drones used video footage shared social media

Bellingcat

- Volunteers fact-checking and citizen investigations in war zones, human rights abuses and crime

Baltic Elves

- Volunteers highlighting false news using scientific and academic sources

Synergy

FRAUDSTERS AND CON ARTISTS

Bernie Madoff

- Former NASDAQ chairman ran $64b Ponzi scheme

Belle Gibson

- Scammer and pseudoscience advocate

Elizabeth Holmes

- Biotech entrepreneur and Theranos CEO

Melissa Caddick

- Fraudulent financial advisor

Charismatic, visionaries, grandiose, fanciful claims, unrealistic results, unpaid bills, vague assertions, unkept promises, determined denials, small unusual errors, unflustered, affinity fraud

Synergy

PAST

Aircraft

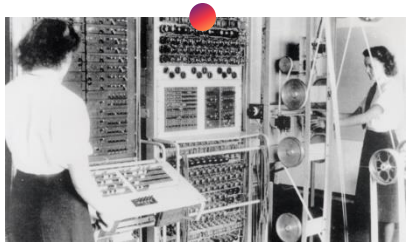
1900s

1940s

Now

Computer

1950s

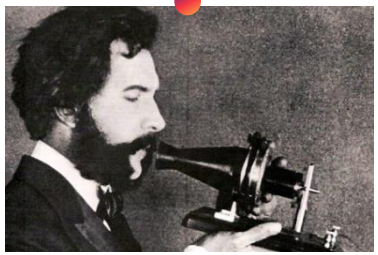


1980s

Now

Telephone

1870s



1970s

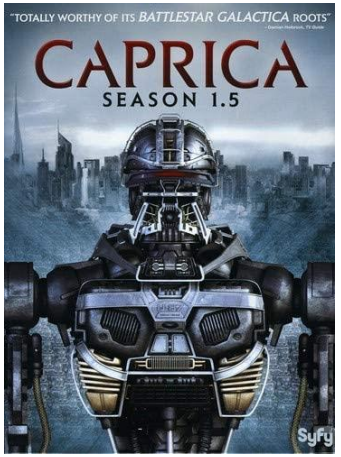
Now

Future?

Virtual lives

Implanted memories

Digital immortality

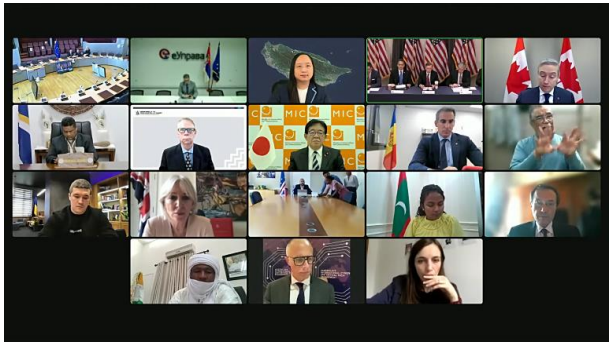


Cyborgs

Holographic partners

Precognition hiveminds

EVOLVING



- Specific **'naming and shaming' exposes but won't always lead to behaviour change** or embarrassment
- When defence improves, disinformation innovates in cat-and-mouse
  - AI fake profiles, handle switching and new perpetrators
- Censorship ramped up alternative methods embraced
- **Political outsourcing**
  - Marketing and PR companies
  - Indonesia deepfake 'Jasmine'
- Fake news outlets remain

Synergy

WHY DO SOMETHING ABOUT IT?

- **Different viewpoints**
- **Unelected company employees running systems how they see fit with near monopoly on public discourse**
- People disagreeing with certain narratives excised or shadow banned
- Algorithmically curated information affects exposure to ideologically cross-cutting news - embolden abuse or encourage other views?
- **Companies controlled by private citizens banning people**
  - Trump not Putin, Iran's Supreme Leader or Taliban
- Mainstream media
  - Want balanced reporting
- Mistakes made
  - COVID origin, treatment, isolation rules, infection after vaccination and mask effectiveness
  - Platforms updating standards and policies

REFORM

Labelling

- **Warnings labels** at initial exposure with extra click to see
- Repetition, retraction or refutation
- Corrections providing alternative story

Suspensions

- **People don't back down**, they want to talk, cancelling may move but won't silence
- **Banning creates one side echo chambers**
- Temporary suspensions not permanent
- Remove or hide offensive posts

Regulations

- **Independent government regulatory body**
- Legislation amendments

Algorithms

- Instead of liable for user content, **accountable for content amplification/targeting**
- **Require by law to meet baseline algorithmic open standards**

Transparency

- **Terms of service rules** for permitted user-generated content and behaviour
- **Disclose algorithms and processes** identifying recommendations and violations
- Regular transparency reports with volume and actions taken
- **Explain algorithm optimisation and variables** allowing users to shape experience
- **Platforms or publishers**

Synergy

JOE ROGAN

Joe Rogan Experience

- Comedian and MMA commentator
- **Promotes show as conversation** not source of accurate information
- Diverse accomplished scientists, politicians, adventurers, comedians and sports people to conspiracy theorists
- Openly engages in **array of boundary free ideological subjects**
- Current affairs, politics, comedy, MMA, philosophy and psychedelics

Recent trouble 2021

- Guests discussing COVID and not getting vaccinated
- Apologised for racially insensitive language and be more balanced
- 113 episodes removed

Synergy

JOE ROGAN EXPERIENCE

Critics

- **Financially sponsoring misinformation**
- Guests discuss nonsense with misinformation treated as fact
- Popularity and money with **no real pressure for mistakes**
- Portraying open-mindedness giving conspiracies/dangerous ideas platform

What to do

- For disinformation to cause harm needs **transmitter and receiver**
- Receivers don't have research/literacy to evaluate lacking critical thinking
- Many minds made up beforehand
- Greater content and advisory warnings
- **If de-platformed will move**
- Others will fill void
- Let interview who he wants
- **Public decides if you like listen if not don't**

Synergy