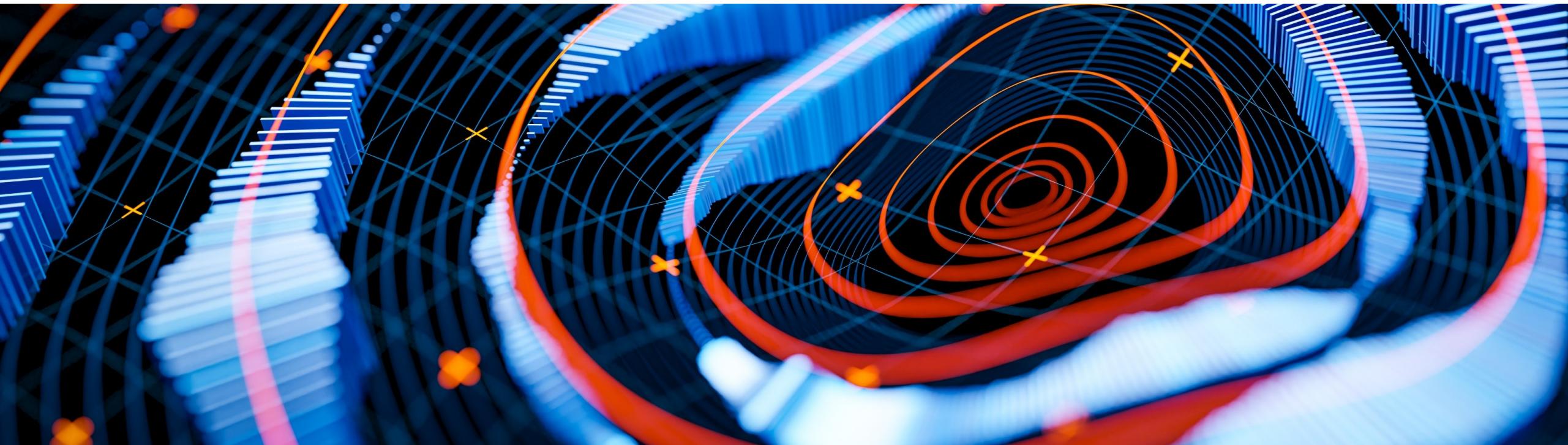


The Information Integrity Initiative:

**Delivering tools and analysis
to fight the growing threat of
misinformation for Australia**



Australian Government
Department of Home Affairs





Principal investigator:

Marian-Andrei Rizoiu (UTS)

Marian-Andrei.Rizoiu@uts.edu.au



Hany Farid (UC Berkley)

hfarid@berkeley.edu



Adam Berry (UTS)

adam.berry@uts.edu.au



Jooyoung Lee

Jooyoung.Lee@uts.edu.au



Emily Booth

Emily.Booth@uts.edu.au



Elliott Waissbluth

ewaissbluth@berkeley.edu

Project period of performance:

1 July 2022 – 31 June 2023

Organisations:

UTS Data Science Institute

University of Technology Sydney,
Ultimo, NSW 2007, Australia.



Behavioral Data Science Lab

<https://www.behavioral-ds.science>



School of Information,

University of California,
Berkeley, CA 94720, USA.



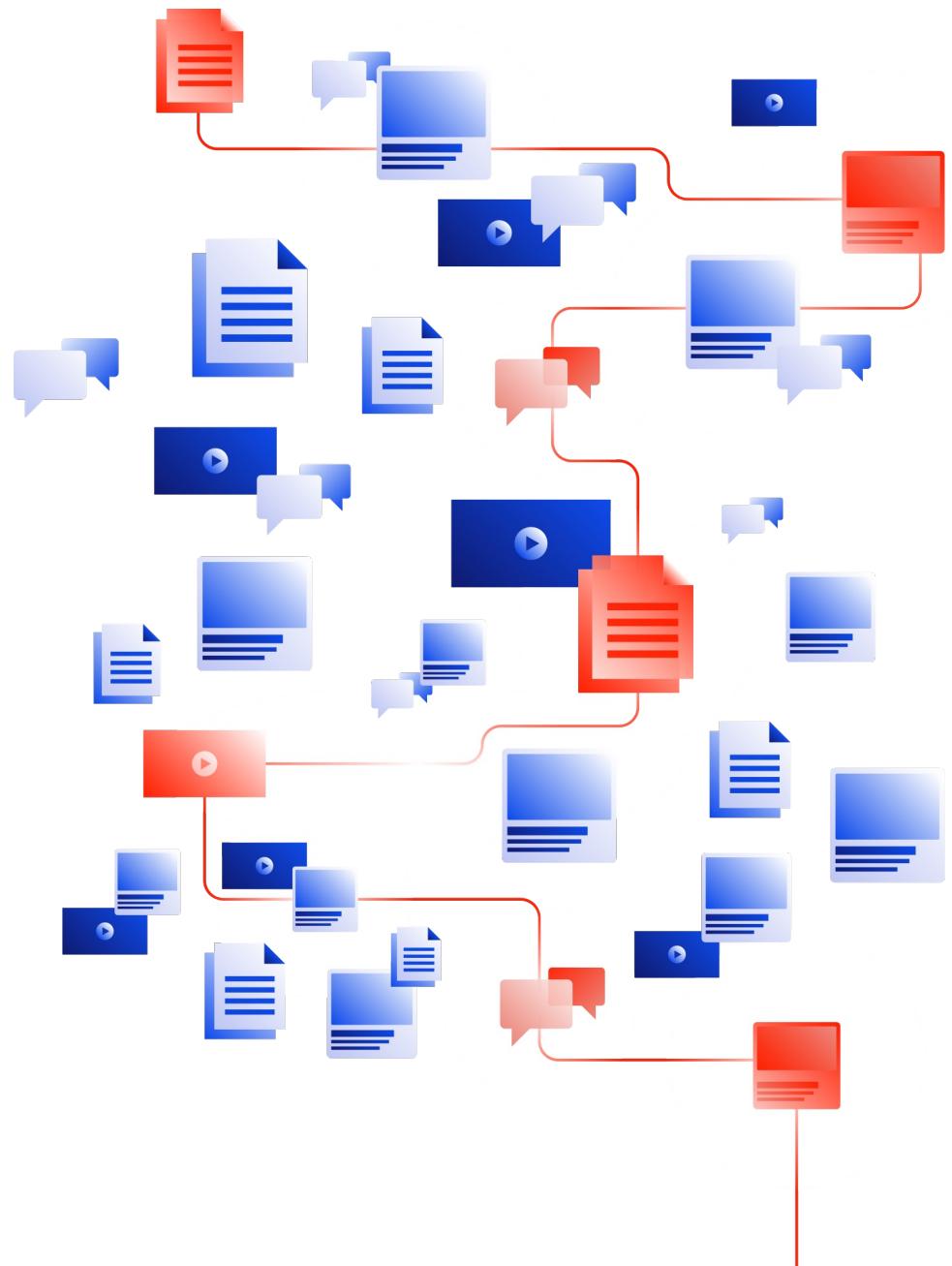
Background and Context

Objectives:

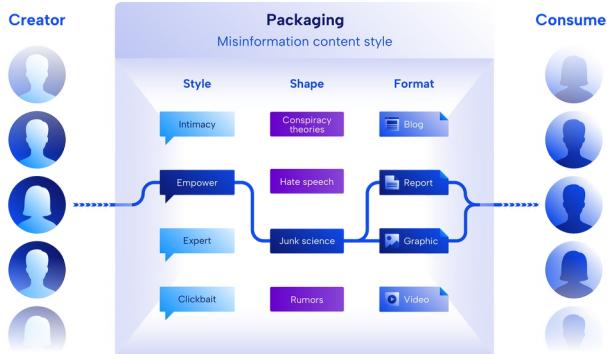
- Expand knowledge about misinformation flow and consumption in Australia.
- Identify potential responses from public and private entities.
- Develop domestic expertise in dealing with online misinformation while providing training opportunities for Australian talent.

Structure – three year project:

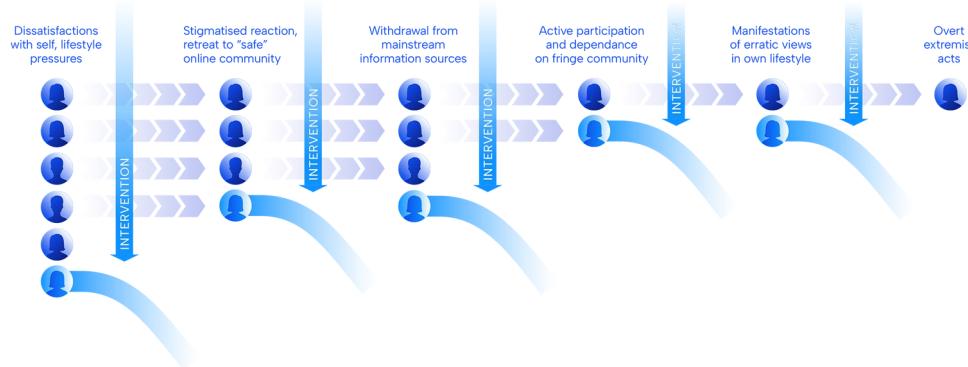
- **Performance period:** 12-month (July 2022 to June 2023) the research team works closely with the Department stakeholders.
- **Research-driven period:** 24-month (July 2023 to June 2025) PhD students continue exploring fundamental research.



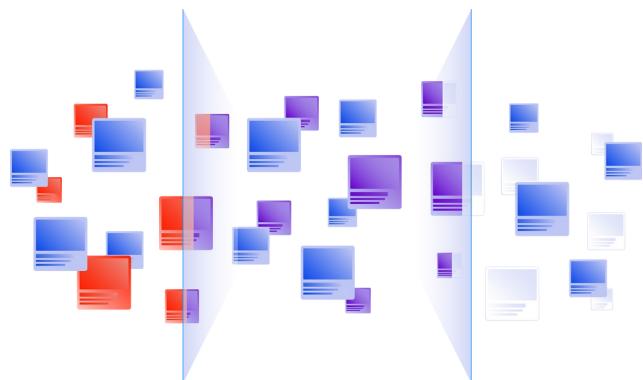
Part 1. The Information Integrity Initiative project



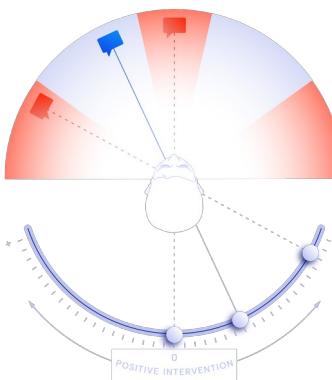
1. Misinformation Eco-System



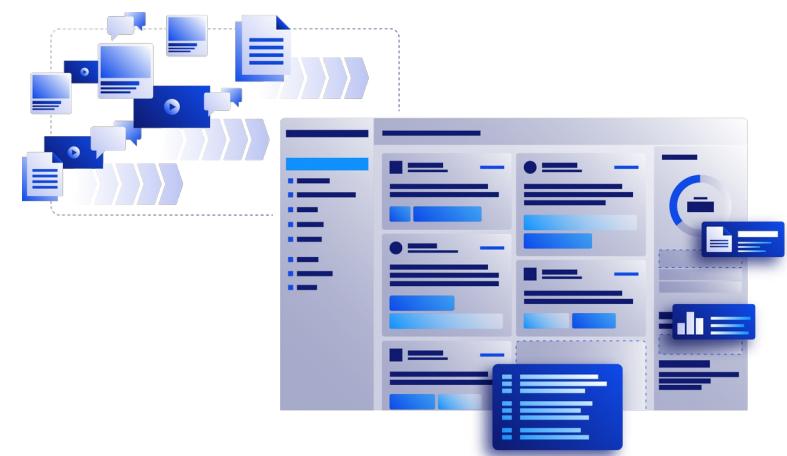
2. Misinformation Consumption as a “Radicalisation” Pathway



3. The Effectiveness of EU's Digital Services Act

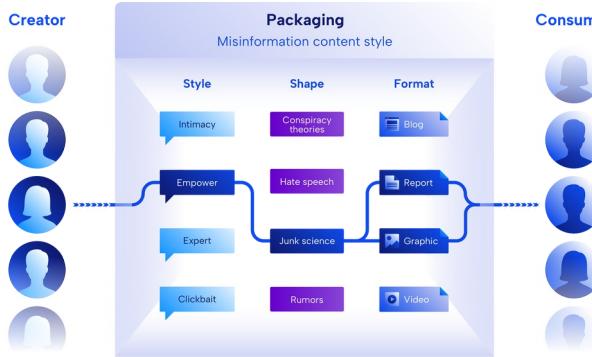


4. Synthetic Testbed for Modelling Positive Interventions



5. Prototype Software to Monitor Online Social Media Discussions

Presentation outline



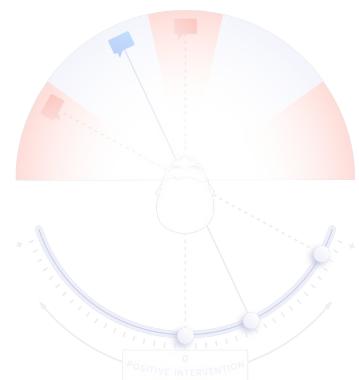
1. Misinformation Eco-System



2. Misinformation Consumption as a “Radicalisation” Pathway



3. The Effectiveness of EU's Digital Services Act



4. Synthetic Testbed for Modelling Positive Interventions



5. Prototype Software to Monitor Online Social Media Discussions

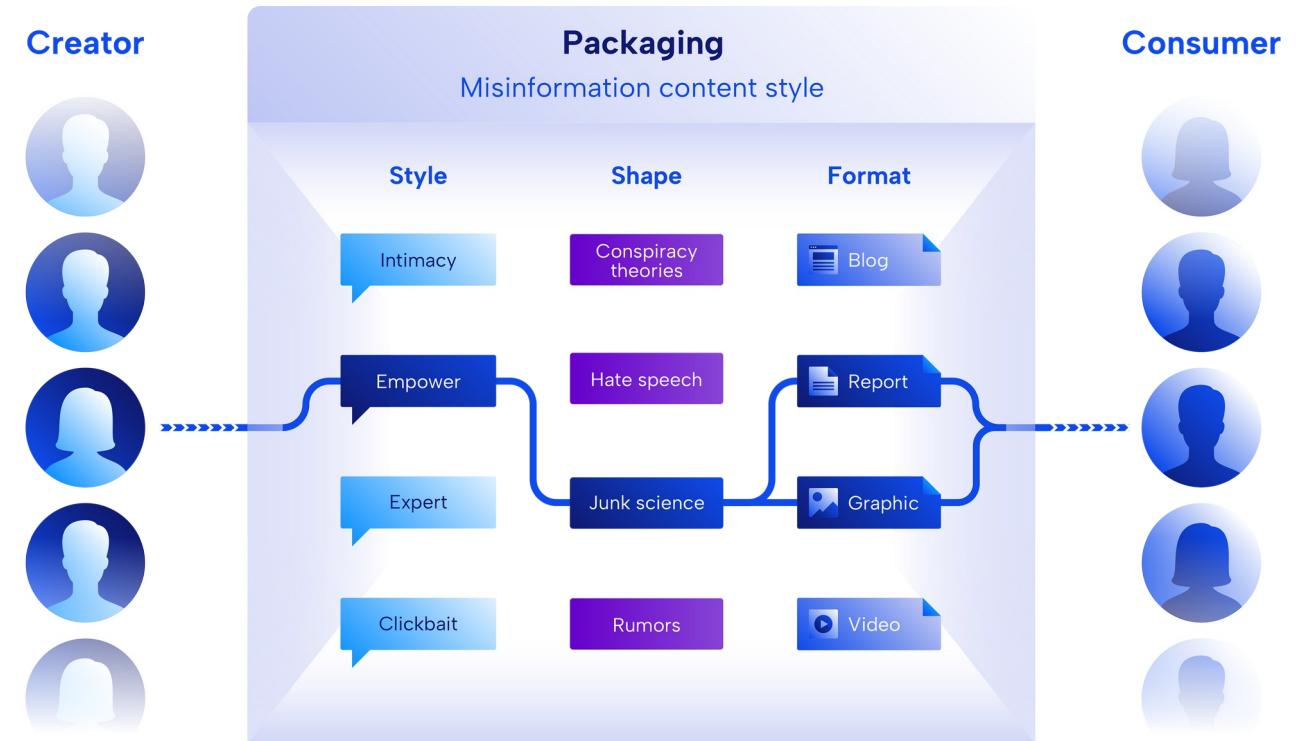
The Misinformation Eco-System

1

Aim: understand how misinformation spreads in Australia's online ecosystem.

Findings:

- Misinformation is an online product targeting specific vulnerable populations.
- People believe and share misinformation that aligns with their preexisting beliefs and communication style.
- Tailored messaging is 50-100% more effective than generic communication.
- Identify vulnerable users based on their communication style rather than content.



Datasets

Google News for Australia

Top stories >


news.com.au
Pic exposes chilling reach of Russian warlord
6 hours ago

 The New Yorker
Yevgeny Prigozhin's Rebellion: Putin's Weakness Unmasked
8 hours ago • Opinion

 Bloomberg
Mutiny in Russia Foreshadows Defeat in Ukraine
18 hours ago • Opinion


NEWS
US officials saw signs Prigozhin was planning challenge to Russian military but surprised...
5 hours ago


NEWS
Reptile believed extinct rediscovered in Aussie state
1 hour ago


NEWS
Barnaby Joyce plays down report suggesting Nationals on brink of leadership challenge
4 hours ago


The Advertiser
Second urgent warning: Another dam at high risk of failure
3 hours ago

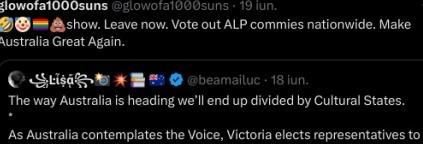
Twitter Far-right

Australian News @news_australian · 20 iun.
Do you support Donald Trump Jnr coming to Australia to Make Australia Great Again?
#auspol #DonaldTrump #donaldtrumpjnr



427 88 706 34,2 K

glowafa1000suns @glowafa1000suns · 19 iun.
Leave now. Vote out ALP commies nationwide. Make Australia Great Again.



@beamailuc · 18 iun.
The way Australia is heading we'll end up divided by Cultural States.
As Australia contemplates the Voice, Victoria elects representatives to negotiate state treaty - ABC News apple.news/AzcXw9tfcRLOBE...

Make Australia Great Again @MakeAUGr8Again · 20 iun.
VOTE NO, BOOKIES FAVOURING THE NO VOTE



@AlboMP · 19 iun.
BREAKING: Parliament has just said yes to holding a referendum.
Now the Australian people will have a chance to say yes.
Together, we can make history by enshrining recognition of Aboriginal and Torres Strait Islander people in our constitution.

Facebook Antivax

Know Your Rights Group
8 months ago

URGENT NOTICE!!! Please stop cancelling your podcast memberships! We have had a huge number of people cancel their podcast memberships over the past few days because they received notification that PayPal was going to start "fining people \$2,500 for spreading mis-information" – that is NOT TRUE and is, in fact, misinformation itself! There are numerous... See More

"Insanity": PayPal renege on US\$2500 fine for users spreading misinformation
PayPal has apologised for a new policy that would've seen users penalised US\$2500 (\$3400) from their accounts for...
SMARTCOMPANY.COM.AU

22.5x 739 +688 529 +525 57 +53

Informed Medical Options Party - IMOP posted a video.
7 months ago

Russell Broadbent MP speaking out in parliament about the possible cause of Australia's increasing excess deaths 🙏🙏



46.2K +43.3K Post Views 74.8K Total Views

Facebook Far-Right

Freedom First Australia
5 months ago

Freedom First Australia


1970


2020

9 Nov 2022 – 11 Jan 2023

9929 news publishers

1496 users
2969 posts
1,827,162 external links

7 pages
2969 posts
558 external links

16 pages
6017 posts
846 external links

Misinformation funding



TRENDING: RUSSIA REVOLT BIDEN CRIME FAMILY TITAN SUB 2024 RACE TRUMP INDICTMENT PRIDE PROPAGANDA BORDER CRISIS

TRUMP: 'BIDEN CRIME FAMILY' CORRUPTION '100 TIMES BIGGER THAN WATERGATE'



Former President Donald Trump believes President Joe Biden's family corruption is "100 times bigger than Watergate," according to his prepared remarks for an address at the Faith and Freedom Coalition. Just days after an IRS whistleblower made shocking revelations about the Biden family.

by JORDAN DIXON-HAMILTON

898



advertisement

...WE WILL DEFEAT RADICAL DEMOCRAT POLICY OF LATE TERM ABORTION
by NICK GILBERTSON

38

'OBSCENE INEQUALITY' BARACK OBAMA BLASTS MEDIA'S WALL-TO-WALL COVERAGE OF MISSING SUB
by ELAINE MALLON

455

GROOMING FAIL: DISNEY LOST \$890 MILLION ON LAST EIGHT THEATRICAL RELEASES
by DAVID NG

348

NOLTE: MR. THEY/ THEM'S 'FLASH' IS A CERTIFIED BOX OFFICE CATASTROPHE
by JOHN NOLTE

262

EXCLUSIVE — SCHWEIZER: THE DAM IS STARTING TO BREAK ON BIDEN FAMILY CORRUPTION
by HANNAH BLEAU

998

FORTUNE SOLAR
#1 ONE OF THE AUSTRALIA'S LARGEST SOLAR COMPANY
6.6 PREMIUM kW system
STARTING FROM \$2921*
GET A FREE QUOTE

advertisement

Misinformation funding



TRENDING: RUSSIA REVOLT BIDEN CRIME FAMILY TITAN SUB 2024 RACE TRUMP INDICTMENT PRIDE PROPAGANDA BORDER CRISIS

TRUMP: 'BIDEN CRIME FAMILY' CORRUPTION '100 TIMES BIGGER THAN WATERGATE'



Former President Donald Trump believes President Joe Biden's family corruption is "100 times bigger than Watergate," according to his prepared remarks for an address at the Faith and Freedom Coalition, just days after an IRS whistleblower made shocking revelations about the Biden family.

by JORDAN DIXON-HAMILTON 898



FORTUNE SOLAR
#1 ONE OF THE AUSTRALIA'S LARGEST SOLAR COMPANY
6.6 PREMIUM kW system
STARTING FROM \$2921*

GET A FREE QUOTE

advertisment

DON'T MISS OUT BRILLIANT EOFY SOLAR DEALS
6.6KW PREMIUM SOLAR PACKAGE
16 PANELS HIGH EFFICIENCY WAS \$3991 \$2921*
5KW WIFI ENABLED INVERTER \$1070 OFF
25 YEARS WARRANTY ON PANELS 10 YEARS WARRANTY ON INVERTER
LOWEST EVER PRICE
CHECK BIG PACKAGES
LIMITED STOCK AT THIS PRICE! GET A QUOTE

Lightinthebox

Toate Categoriele New In Best Sellers Wedding Season Plus & Curve Custom Clothing

HOME > Haine de Bărbați > Maiu & Tricou Bărbați > Tricou grafic pentru bărbați (911 Produse)

animal & desen animat credință cuvinte amuzante Exterior familie Hobby Modă oameni batrani

Tricou grafic pentru bărbați animal & desen animat credință cuvinte amuzante Exterior familie Hobby Modă oameni batrani

Still Top-uri

Model

5 THINGS YOU SHOULD KNOW ABOUT MY WIFE

I WOULD RATHER STAND WITH GOD AND BE JUDGED BY THE WORLD

I AM PROUD OF MANY THINGS IN LIFE

BEING A DAD IS AN HONOR

BEING GRANDPA IS AN HONOR

YOU CAN'T SCARE ME I HAVE TWO DAUGHTERS

advertisment

Vulnerable target personae – Jennifer

Jennifer Matthews

age: 25-40

residence: Australia

education: BA

occupation: Formerly a communications specialist for a non-profit, but retired after the birth of her first child

marital status: Married with two children (a newborn and a 2 year old)



"A mother's work is never done"

Jennifer wakes up at 6am each day, often after a broken night's sleep due to her newborn still waking up. She prepares breakfast for her husband and does some yoga, and misses the morning class she used to attend. Jennifer's peace ends when the children wake up, and she spends most of the day juggling their needs. If she can successfully get them to have a nap, she checks Facebook and Instagram to try and keep up with her friend's lives and stay in touch. She might also browse Pinterest, saving pictures to boards about yoga, healthy and quick meals, motherhood advice, and painting tips. Twice a week, she has an afternoon walk with the kids to the nearby supermarket for groceries. In the evening, she makes dinner for her husband. He enthusiastically talks about work, but she feels like she has very little to say about her life and it's not very interesting to him.

Interests:

Health, children, creativity, simple pleasures

Fears

Her children will get sick
She is not the best mother she could be
Her friends at work forget her
She become unattractive to her husband after having two children

Preferred style:

Intimacy + Junk Science

Barrier to generic communication:

formal, masculine style

Vulnerable target personae – Patrick

Patrick Bear

age: 67

residence: Suburban NSW

education: Bachelor of Education

occupation: Retired primary school teacher

marital status: Married for 45 years, 3 children, 4 grandchildren



"Kids these days"

Patrick wakes early and takes a walk around the neighbourhood, before helping his wife up and organising breakfast for her. She's become quite frail, and he has to help her get dressed and walk. After, they often sit outside in the garden and he reads to her. He organises lunch and they both have a short rest, before the grandkids come over after school until their parents finish work. These frantic few hours are their favourite part of the day, even if it gives them a bit of a headache. He helps the kids with homework, and then takes care of chores as his wife supervises them watching TV or spending time in the garden. When the grandkids have been retrieved by their parents he organises dinner, and as his wife watches late night TV, he logs on to the Facebook his daughter signed him up for. He tries to keep up with the posts there, and also finds interesting news items about issues he never hears about on TV. His wife thinks the internet is silly, but he's excited about the new things he can learn there. As he gets ready for bed, he often finds himself still thinking about the things he's read.

Interests:

Humour, Hard work, Family

Fears

New changes in the world

His family forgetting him

Financial security

Being alone

Illness and death for him and his wife

Preferred style:

Expert + Conspiracy

Barrier to generic communication:

belief in own expertise, difficulty in understanding long complex text

Vulnerable target personae – Aaron

Aaron Turner

age: 21

residence: Australia

education: Law student (final year)

occupation: Summer law clerk

marital status: Single



"I object!"

Aaron wakes at around 8am and makes a mad rush for either university or his summer job, skipping breakfast. He is enthusiastic about both and works hard at each. By lunch time, he is feeling exhausted and deserving of a "treat", often buying a burger or pizza meal--and after all, he skipped breakfast, which makes up for the extra calories. He works intently through the afternoon, and leaves work feeling exhausted again. He often tells himself that he's too tired to hit the gym now, but he'll go tomorrow. On Fridays, he spends the night with his Church youth group, but it's increasingly bittersweet, as he will have to leave it after his upcoming birthday. On nights when he doesn't have Church, he often winds up having a few beers and scrolling through Reddit, where he can vent his frustration about still being single and find other interesting blogs to explore.

Interests:

Strength and confidence in men

Beauty in women

Christianity

Recognition for his hard work

Fears

Professional failure

Personal failure – He will not get a girlfriend

He will lose his connection to his Church

Preferred style:

Empower + Hate Speech

Barrier to generic communication:

authoritative tone

Effectiveness of communication styling

Testing protocol:

- Facebook ads with A/B testing facility
- Stock images
- Ads link to Government information sites
- A/B test: targeted (A) vs generic styling (B). All other indicators stay the same: target demographic, image, message

Tested variables:

- Three target demographics – consumer personae
- Four message styling – generic, Jennifer-, Patrick- and Aaron-styled
- 5 topics of interest: children vaccination, children sexual education, 5G, climate change, nutrition supplements for physical fitness

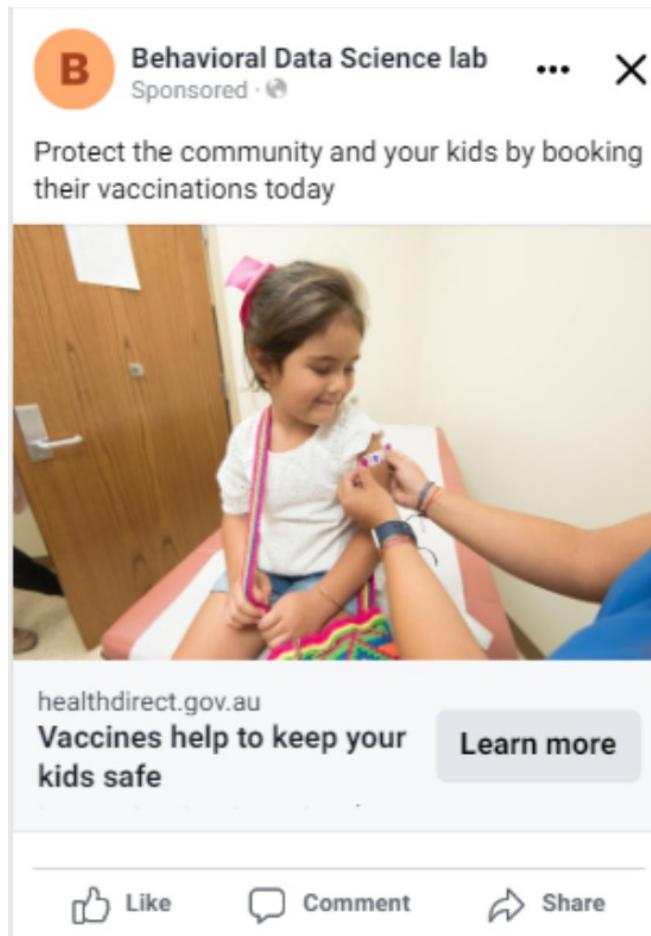
Success measurement

- Click-Through-Rate (CTR): number of clicks vs number of impressions
- **Higher (CTR) – higher message effectiveness**

Effectiveness of communication styling

Children vaccination (Jennifer)

A: Generic



B: Styled



Results:

A: Generic

Reach: 10,346 users

Impressions: 15,567

Clicks: 101

CTR: 0.65%

B: Styled

Reach: 10,020 users

Impressions: 13,107

Clicks: 129

CTR: 0.98%

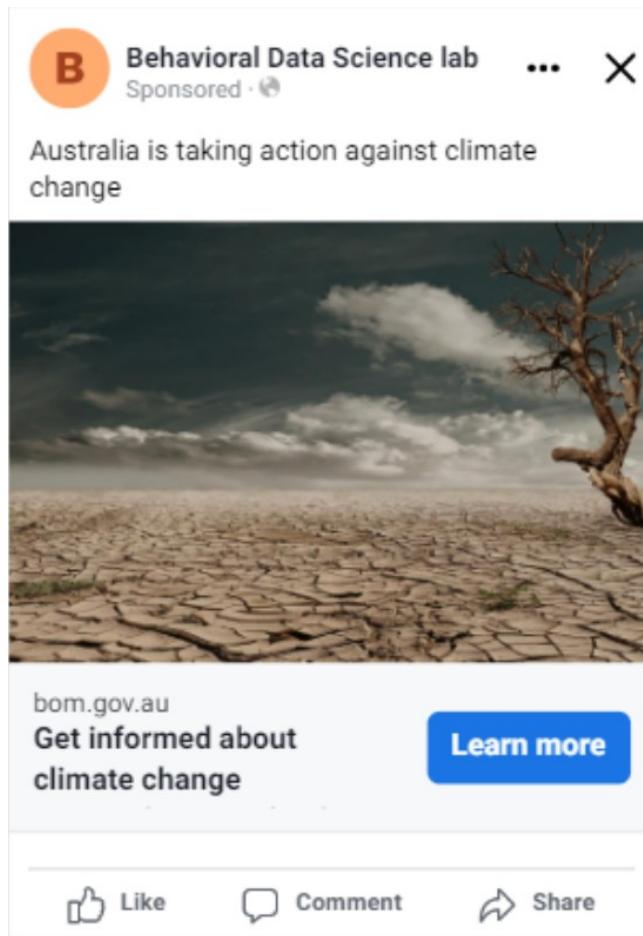
Improvement:

51%

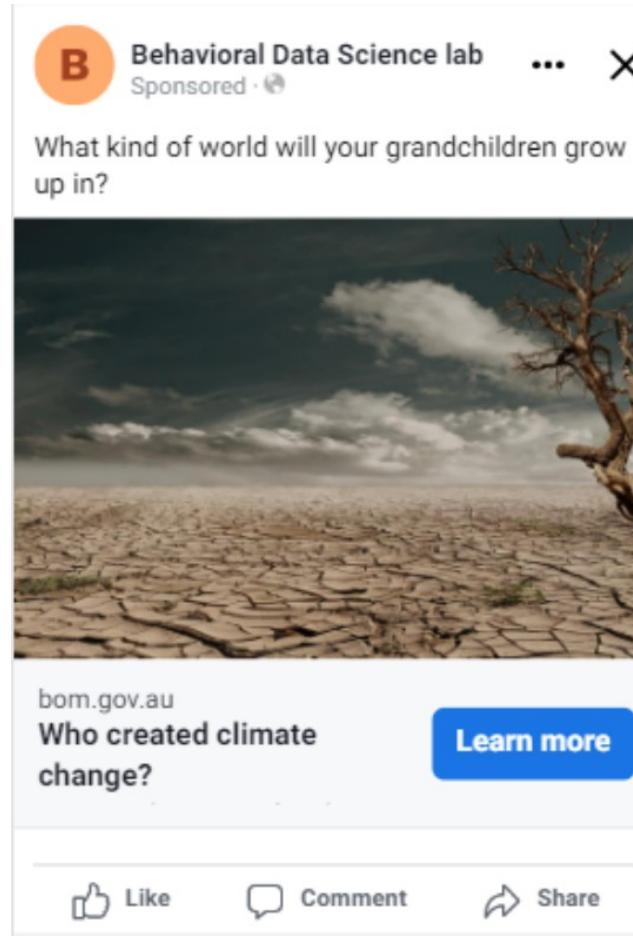
Effectiveness of communication styling

Climate change (Patrick)

A. Generic



B. Styled



Results:

A: Generic

Reach: 5,362 users
Impressions: 10,095
Clicks: 170
CTR: 1.68%

B: Styled

Reach: 5,952 users
Impressions: 10,082
Clicks: 334
CTR: 3.31%

Improvement:
97%

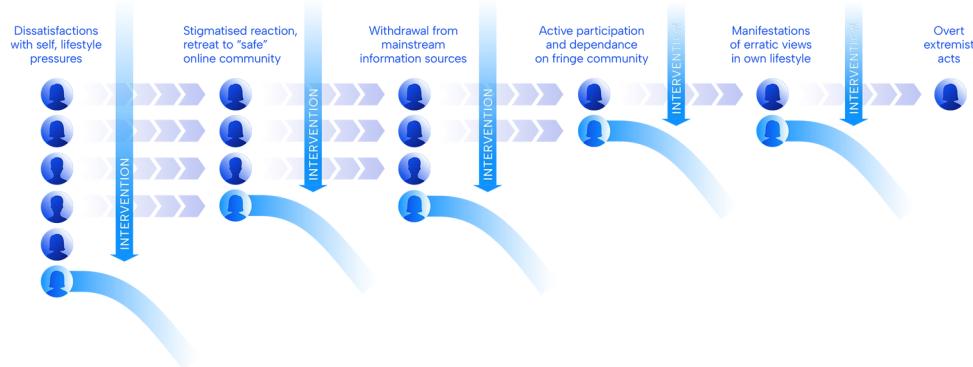
Policy implications for countering online misinformation:

1. Create more effective communication campaigns by customising messages for the intended audience.
2. Disrupt the spread of misinformation by defunding advertisements that support it.
3. Detect misinformation producers and consumers by analysing their adopted styles and topics.

Presentation outline



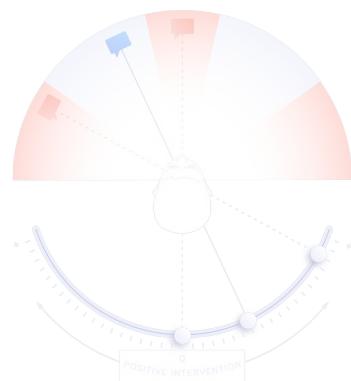
1. Misinformation Eco-System



2. Misinformation Consumption as a “Radicalisation” Pathway



3. The Effectiveness of EU's Digital Services Act



4. Synthetic Testbed for Modelling Positive Interventions



5. Prototype Software to Monitor Online Social Media Discussions

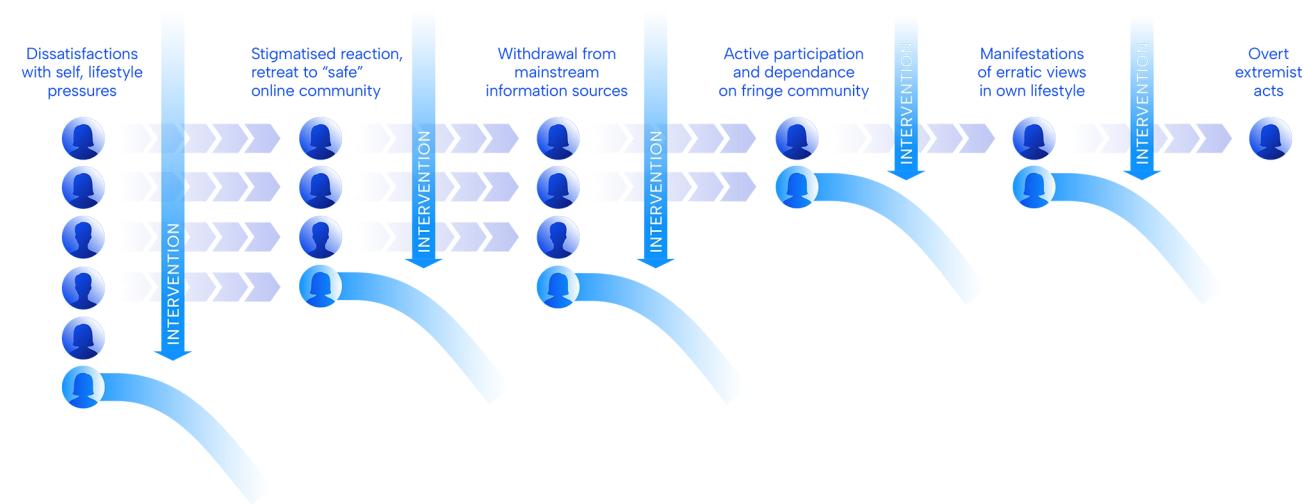
Misinformation Consumption as a “Radicalisation” Pathway

2

Aim: theoretical understanding of the process by which online users encounter and adopt misinformation narratives.

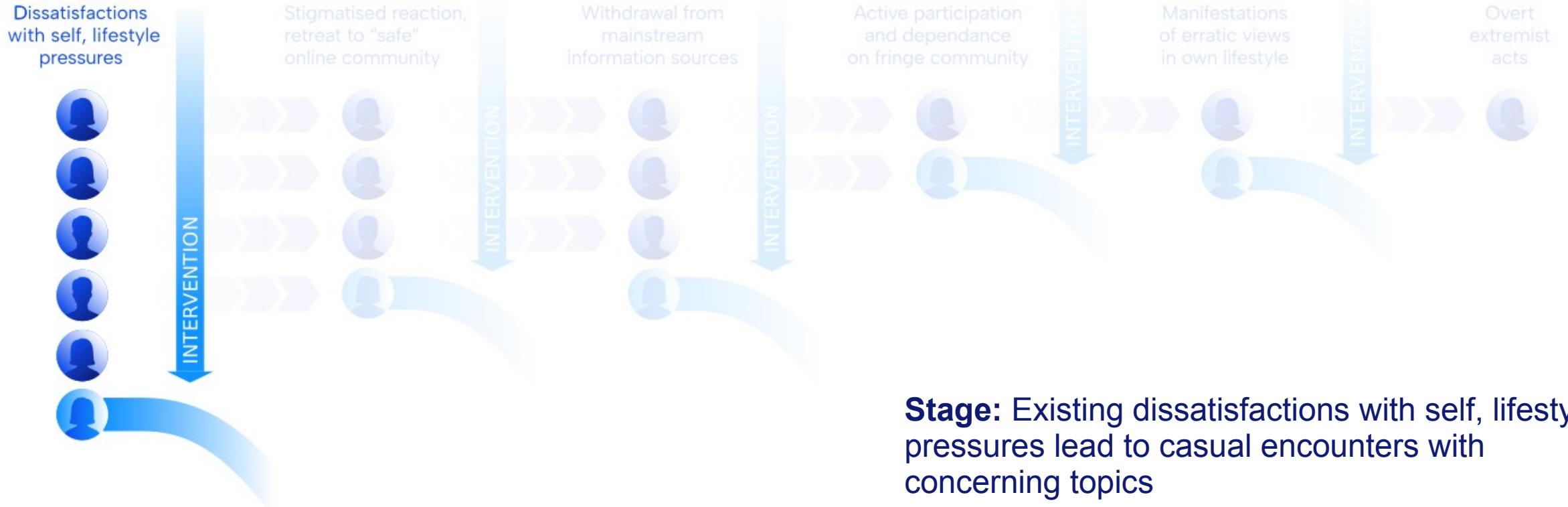
Findings:

- Six-stage pathway of misinformation “radicalisation”
- Ranges from initial confusion and grudges (stage 1) to extreme violence (stage 6).
- Each stage – specific symptoms
- Intervention strategies and off-ramps to help individuals disengage from radicalisation.
- Evaluated against existing literature and real-life examples of former QAnon sympathisers’.



Six stages of “Radicalisation” Pathway

1



Stage: Existing dissatisfactions with self, lifestyle pressures lead to casual encounters with concerning topics

Intervention: Plain-language, styled resources and shared from trusted source via online platforms and websites

Off-ramps: Receives reassuring advice or information from a trusted source

Six stages of “Radicalisation” Pathway

2



Stage: Increased curiosity in encountered topics, tentative disclosure to peers and stigmatised reaction (disbelief/mockery/shaming), retreat to “safe” online community for assurance

Intervention: Resources to support peers and mentors (e.g., GP) in talking about misinformation and fringe views with friends

Off-ramps: Receives an empathetic, non-stigmatising response from peers, no longer requires support from online community

Six stages of “Radicalisation” Pathway

3



Stage: Self-censorship with existing social networks, withdrawal from mainstream information sources, reliance on fringe community for social and informational needs

Intervention: Recommend organic deradicalisation spaces. Public messaging campaigns about reducing time online.

Off-ramps: Engages with organic deradicalisation spaces; 'talked out of' views by former believers; limit or leave social media

Six stages of “Radicalisation” Pathway

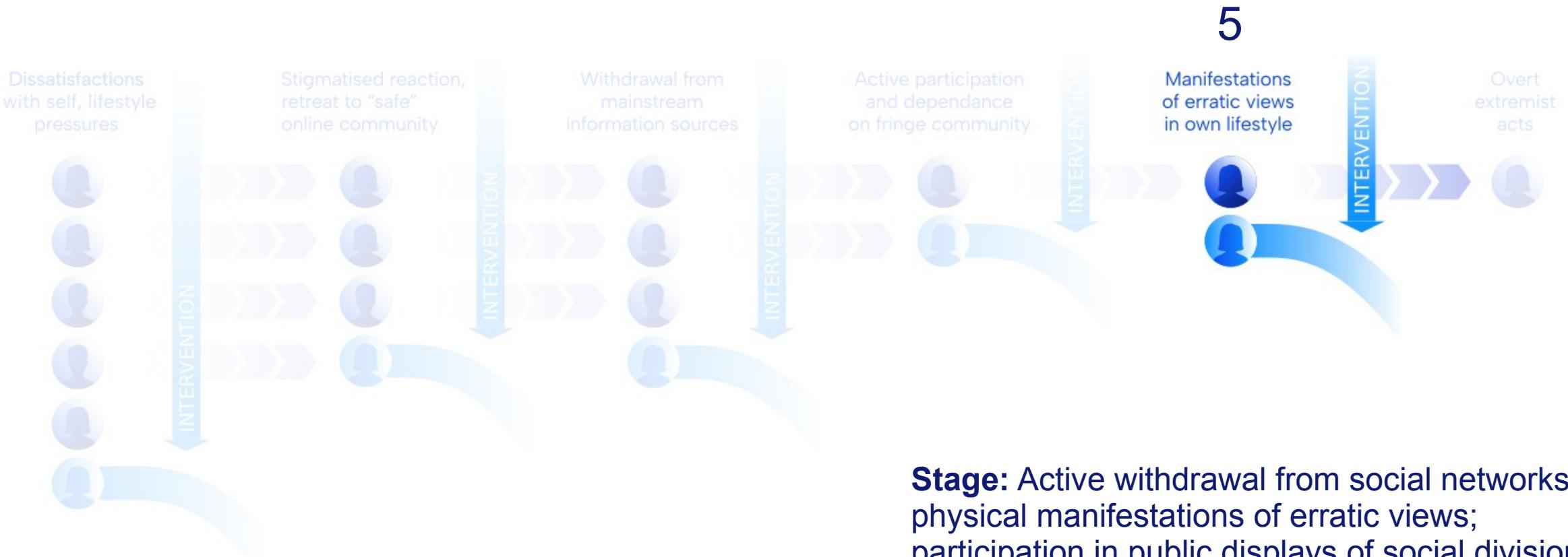


Stage: Active participation and reliance on fringe community; contributing own interpretations and sharing own sources, increased stigma from existing networks

Intervention: Promote user-generated content that objectively highlights illogic of misinformation; reflection on fringe beliefs

Off-ramps: Recommended third-party content that objectively highlights illogic startles person out; Shock highlights the cost of misinformation

Six stages of “Radicalisation” Pathway



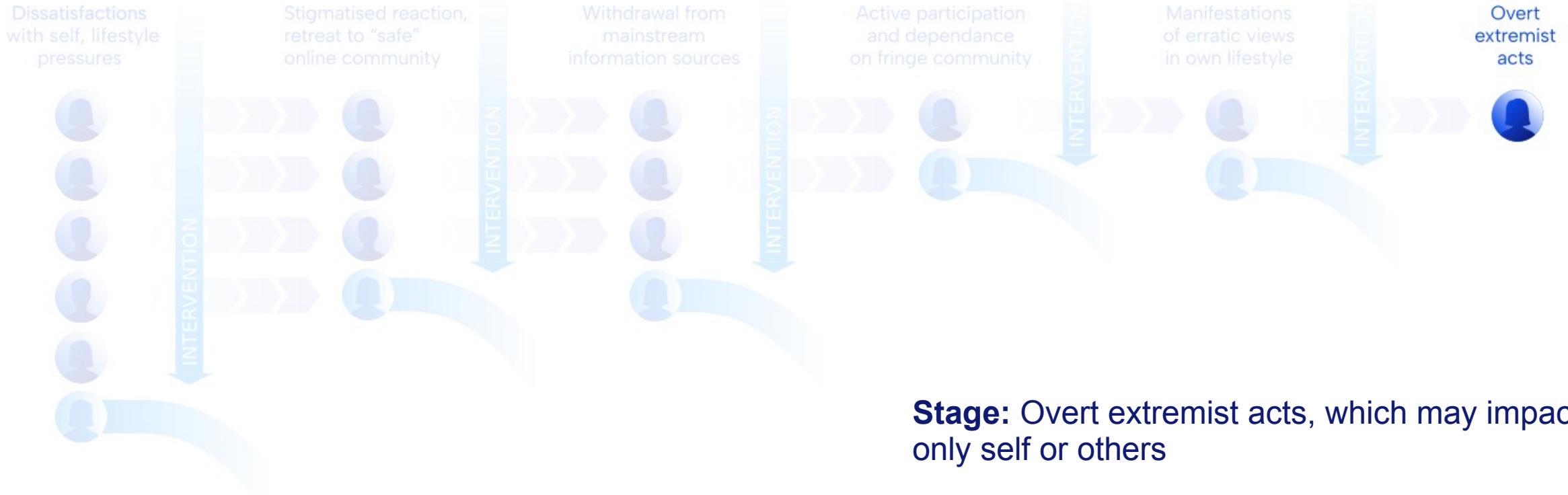
Stage: Active withdrawal from social networks; physical manifestations of erratic views; participation in public displays of social division.

Intervention: Online resources and support groups for family members to connect with affected individual

Off-ramps: Institutionalisation by family members and peers for therapy and recovery; medical intervention addressing underlying issues

Six stages of “Radicalisation” Pathway

6

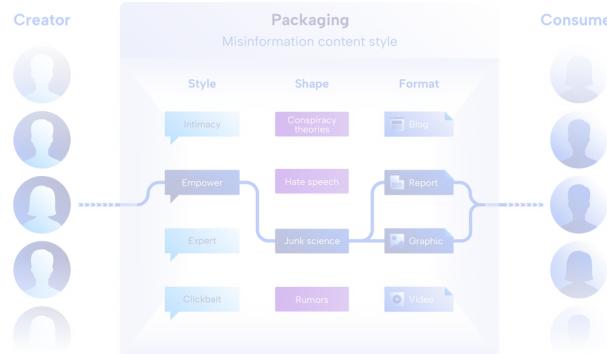


Stage: Overt extremist acts, which may impact only self or others

Intervention: Government-funded de-radicalisation program

Off-ramps: Admittance to correctional and mental health institutions

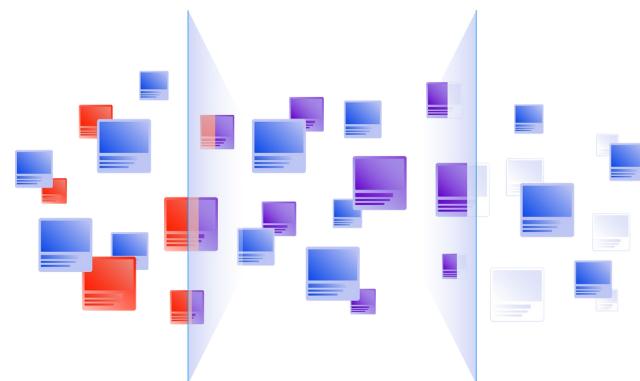
Presentation outline



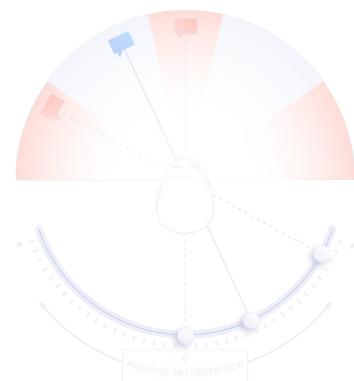
1. Misinformation Eco-System



2. Misinformation Consumption as a “Radicalisation” Pathway



3. The Effectiveness of EU's Digital Services Act



4. Synthetic Testbed for Modelling Positive Interventions



5. Prototype Software to Monitor Online Social Media Discussions

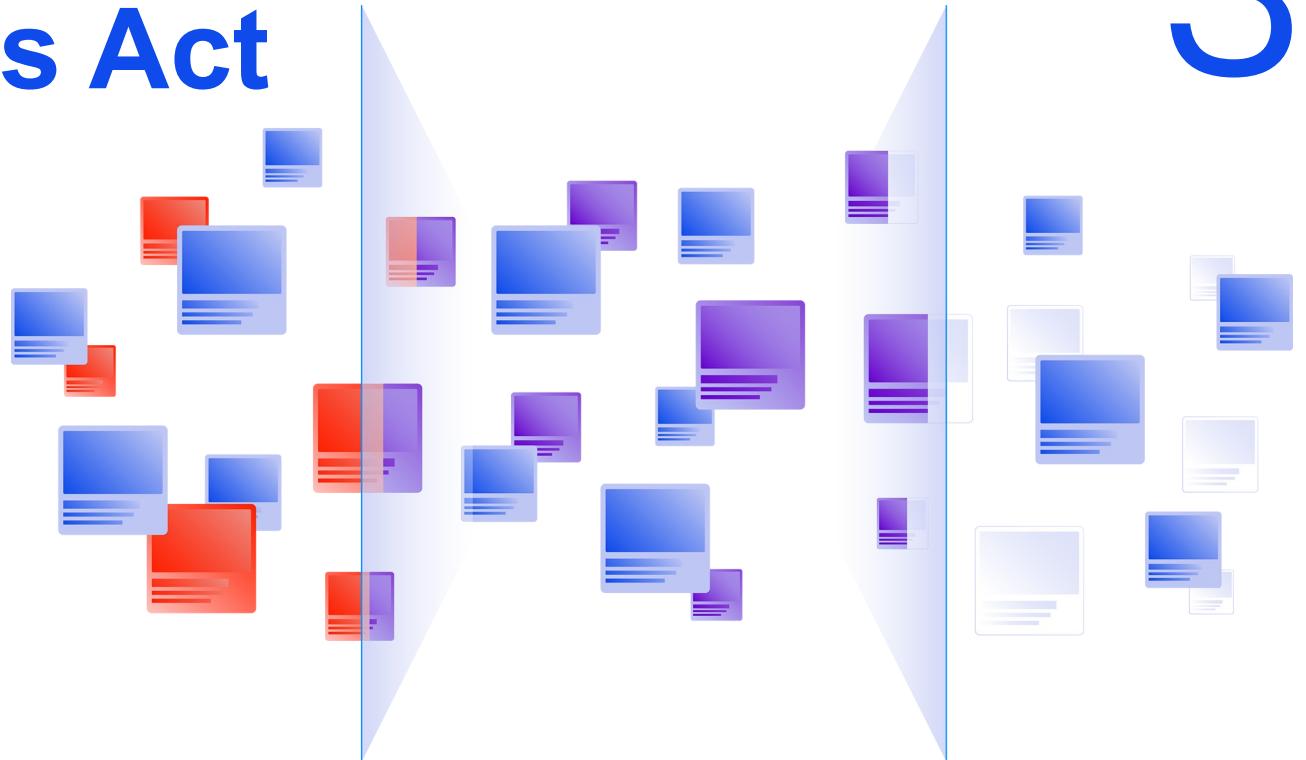
The Effectiveness of EU's Digital Services Act

3

Aim: examines the effectiveness of the Digital Services Act (DSA).

DSA provisions:

- Implements notice and action mechanisms (cf. Art. 16) to report harmful online content
- Appoints trusted flaggers to detect harmful content (cf. Art. 22).
- When content is flagged, platforms must promptly remove the content (max 24 hours).



Question:

Is such a mechanism effective in stemming the spread of harmful content?

The mathematical tools

Important quantities:

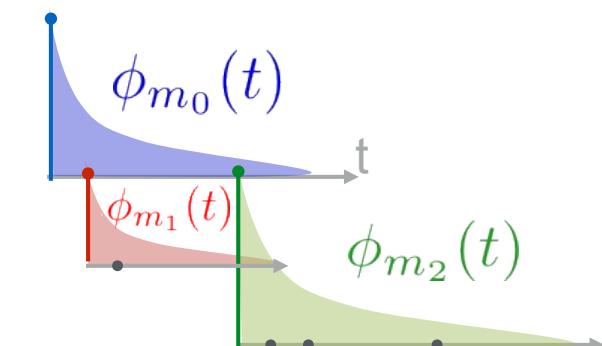
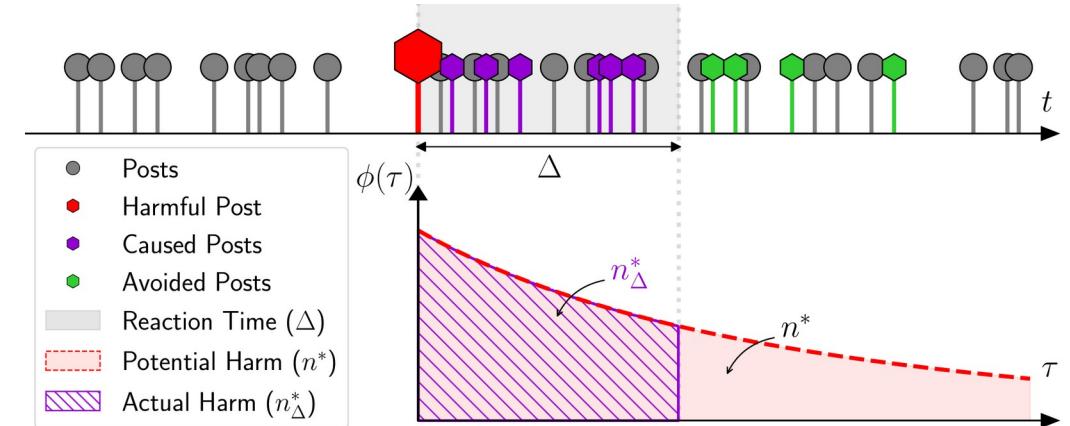
n^* Potential harm – The number of additional harmful posts a content generates directly.

$\tau_{1/2}$ Content half-life – Time until a content generated half the direct reactions.

Twitter (24 mins), Facebook (105 mins), Instagram (20 hours), LinkedIn (24 hours), YouTube (8.8 days), Pinterest (3.75 months).

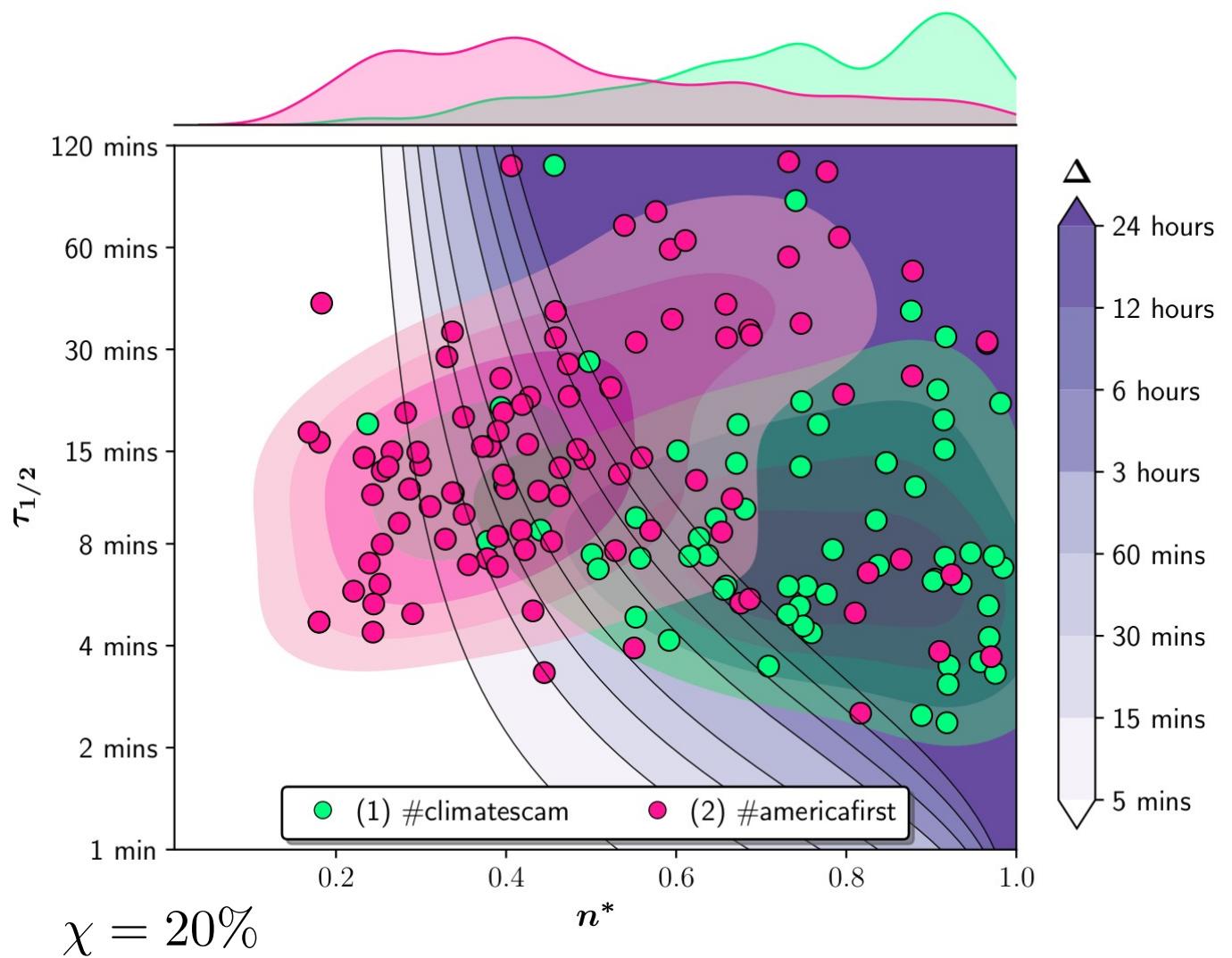
Δ Reaction time – Mandated time to remove flagged harmful content on social media platforms.

χ Harm reduction – Percentage of direct and indirect harmful offspring avoided by content moderation.



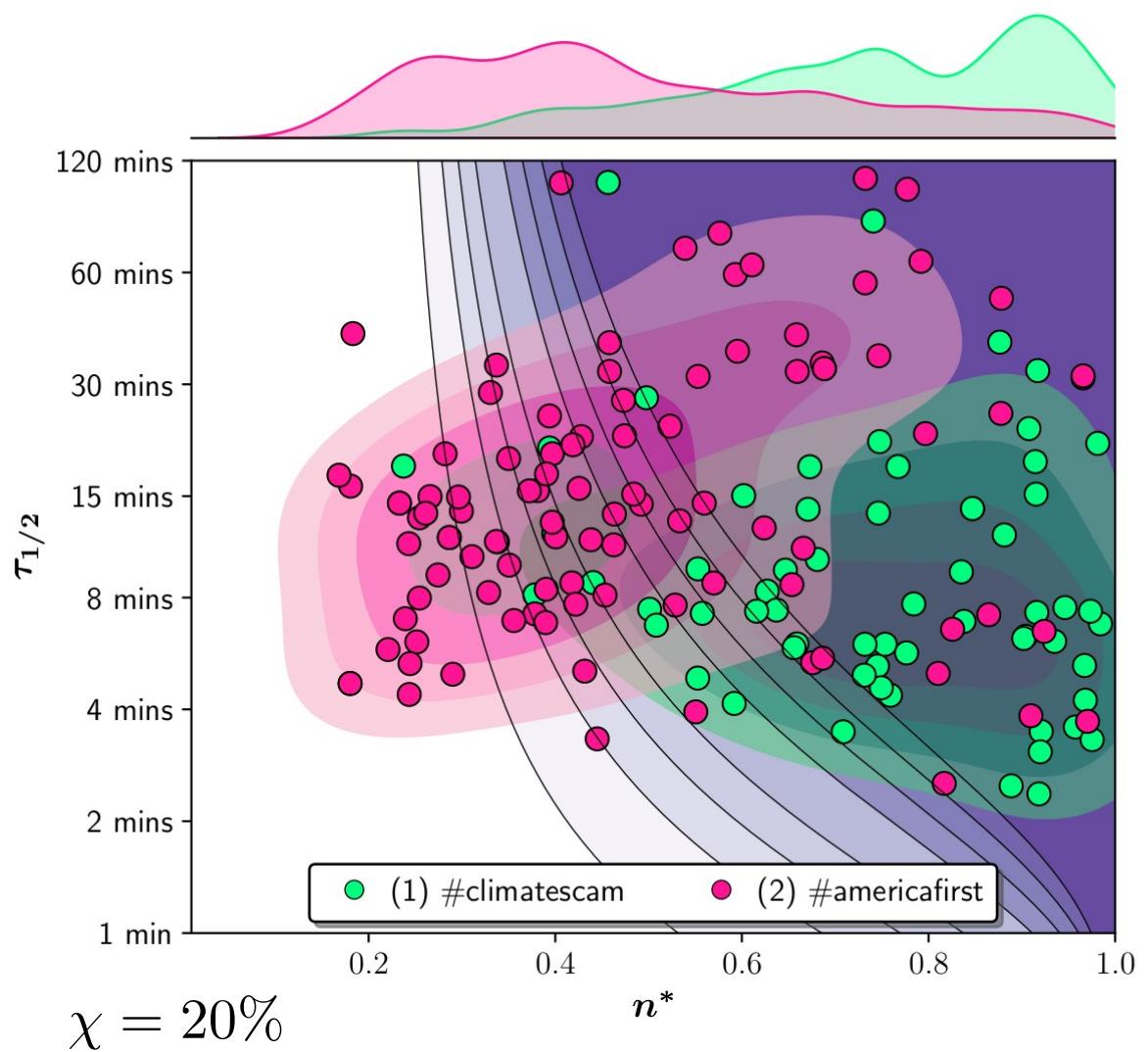
Modelling word-of-mouth spreading process using self-exciting point processes

Moderation is effective to control the spread of misinformation



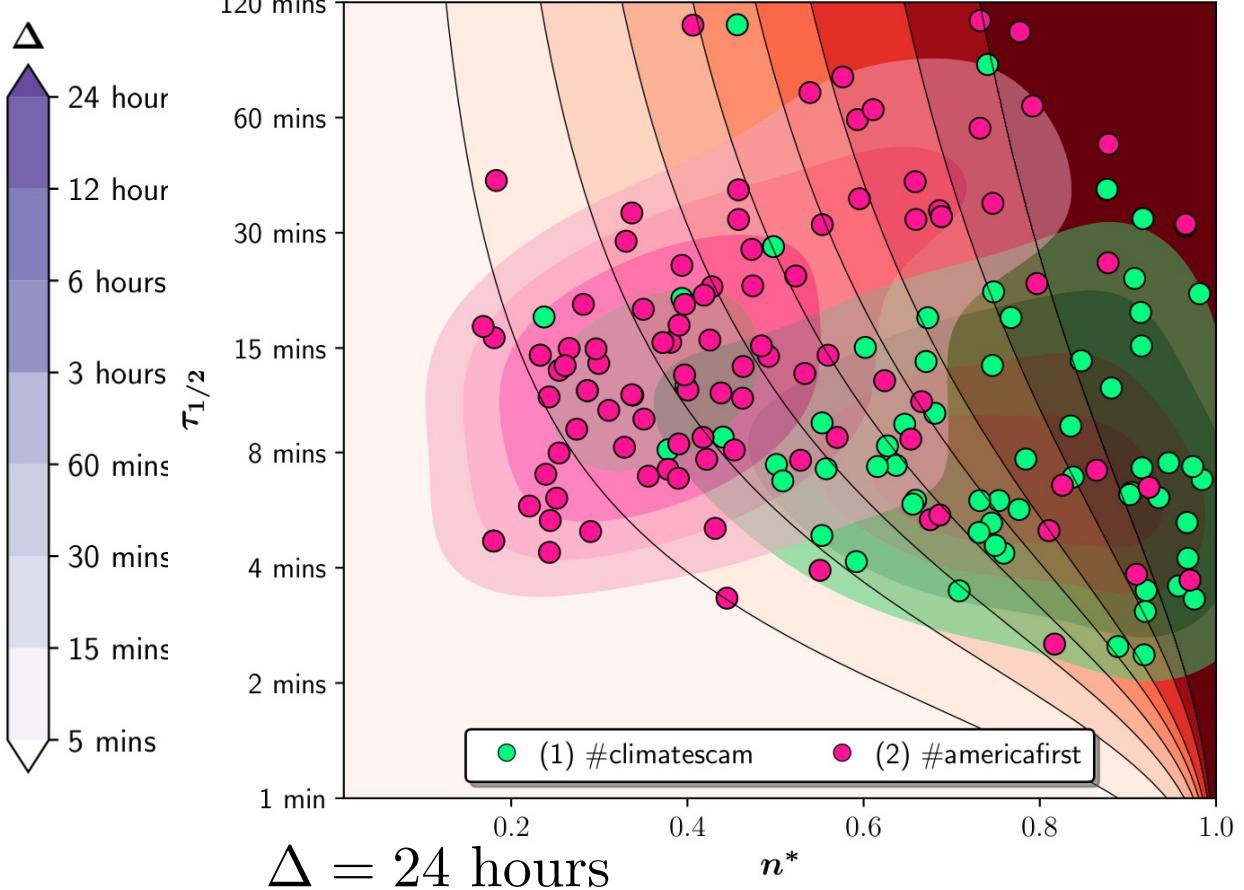
- Two Twitter case studies: #AmericaFirst and #ClimateHoax
- Maximum reaction time increases with half-life and potential harm

Moderation is effective to control the spread of misinformation



$\chi = 20\%$

- Two Twitter case studies: #AmericaFirst and #ClimateHoax
- Maximum reaction time increases with half-life and potential harm



- Harm reduction can be achieved for highly viral and harmful content.
- Rule of thumb for selecting content for managing flaggers' workload.

Presentation outline



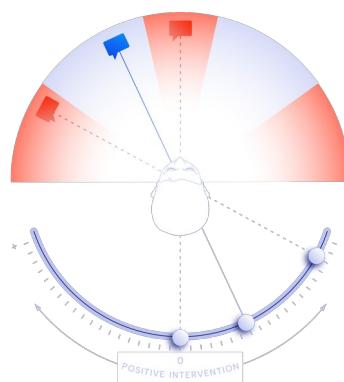
1. Misinformation Eco-System



2. Misinformation Consumption as a “Radicalisation” Pathway



3. The Effectiveness of EU's Digital Services Act



4. Synthetic Testbed for Modelling Positive Interventions



5. Prototype Software to Monitor Online Social Media Discussions

Synthetic Testbed for Modelling Positive Interventions

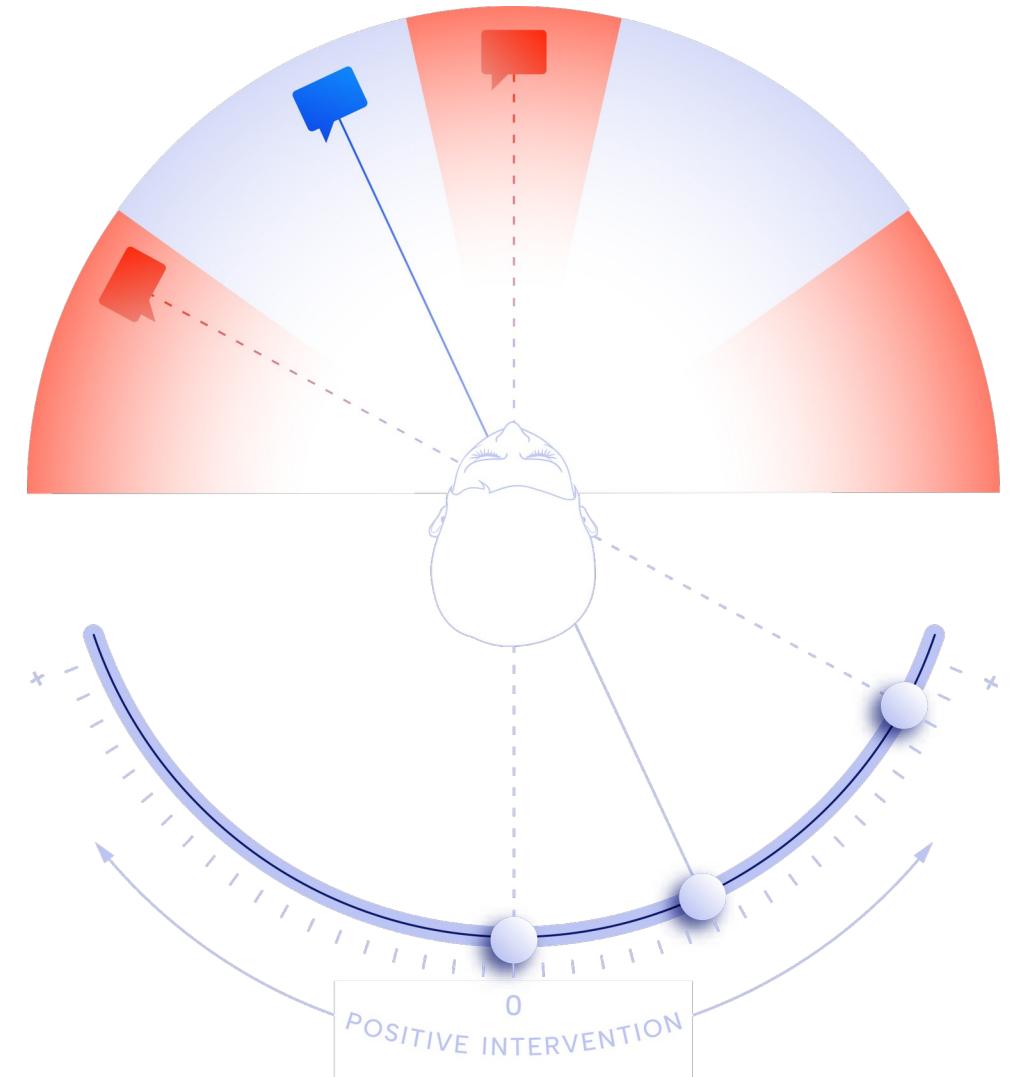
4

Aim: design interventions against misinformation using positive interventions.

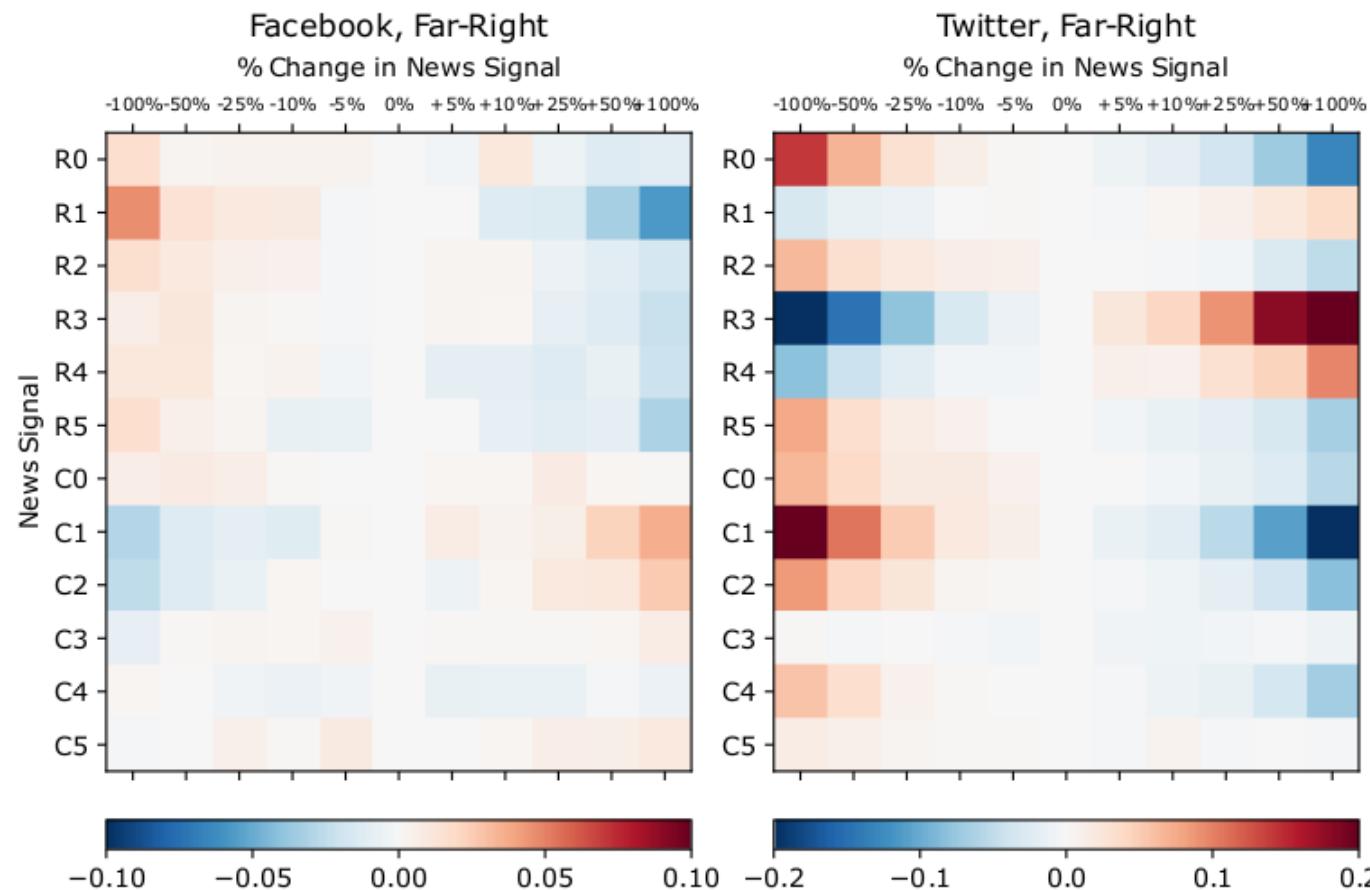
Contributions and findings:

- Negative interventions are moderation; positive interventions increase attention to debunking opinions.
- Opinion market model – misinformation and debunking opinions compete or cooperate for attention.
- Can uncover interactions between far-right and moderate opinions.

debunking far-right ideas on Twitter unexpectedly reinforces instead of countering them.

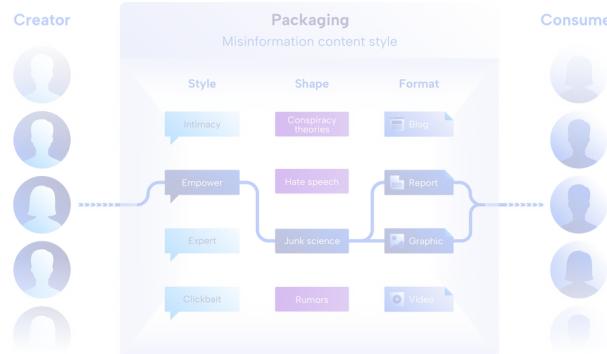


Interventions sometimes backfire

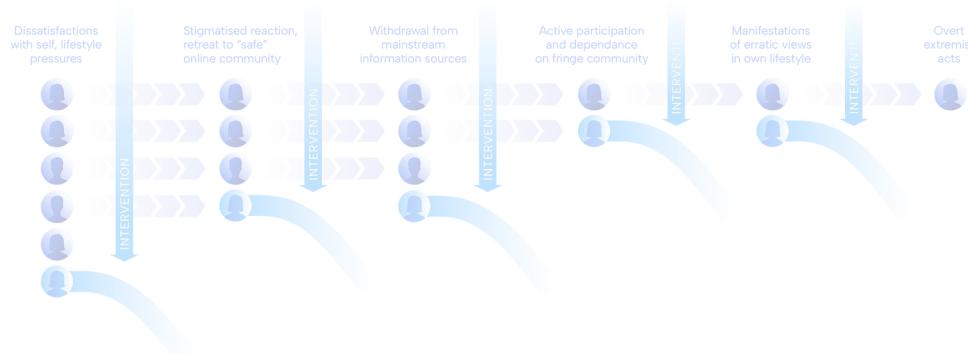


Some reputable media coverage increases far-right opinion prevalence

Presentation outline



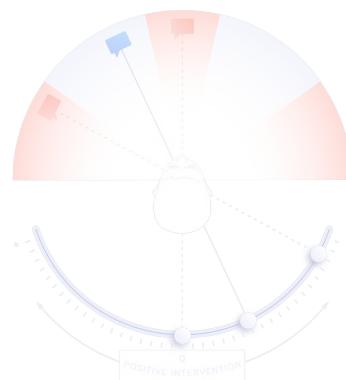
1. Misinformation Eco-System



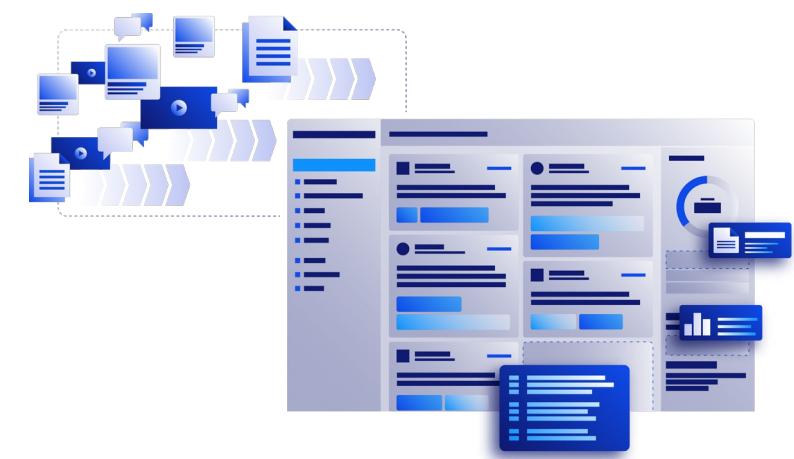
2. Misinformation Consumption as a “Radicalisation” Pathway



3. The Effectiveness of EU's Digital Services Act



4. Synthetic Testbed for Modelling Positive Interventions



5. Prototype Software to Monitor Online Social Media Discussions

Prototype Software to Monitor Online Social Media Discussions and Ads Collection

5

Aim: design software prototypes – practical approaches to online misinformation.

Contributions and findings:

- Two software prototypes
 - monitoring online discussions about specific topics, on multiple social media platforms
 - extracts ads from the web pages shared by misinformation and extreme-leaning users



Monitor online discussions

Information Dashboard

Topics Sort by attention ▾

Add new topic

ChatGPT

Labels

Facebook Pages

Settings

Logout

ChatGPT

Throwback Store https://www.facebook.com/111566542282132 25 Jun / 09:58 A.I. coming in strong with the Answer. #chatgpt

Dagny Taggart https://www.instagram.com/dagny.taggart.vp/ 25 Jun / 09:53 Very few memes have me roaring like this one did. I'm going to start asking ChatGPT to generate my bingo cards bc lately I'm not even in the same ballpark. 🦁

Bard Ai For Business https://www.facebook.com/703694553093530 25 Jun / 09:50 Bard Google will replace ChatGPT in the present and in the future because Bard Google is a super AI with human thinking, 80 times smarter than ChatGPT . Trial version before release here =>>...

Music Managers Forum - US https://www.instagram.com/mmf_us/ 25 Jun / 09:48

Rebecca Hains, Author https://www.facebook.com/316579058382863 25 Jun / 09:54 I asked ChatGPT to write us a story about the Muskerberg cage fight. The results seem legit! It wrote: In the heart of Las Vegas, a crowd gathered for an unprecedented event—a cage fight between...

Onefootball https://www.facebook.com/247792341912718 25 Jun / 09:53 The best all-time Argentina XI according to ChatGPT 🇦🇷

ALEXANDRIA,DMV, DC, VA LASHES https://www.instagram.com/winklashart/ 25 Jun / 09:50 Customized Lashes at Wink Write caption for instagram about how at Wink lash art we costumize every lash extension set . Keep it to max 3 paragrafts ChatGPT "🌟 Customized Beauty at Wink Lash...

ChatGPT Community https://www.facebook.com/groups/495687532539367 25 Jun / 09:45

☆ Saved views Filter Sort by Posting Date DESC

Summary Hour Day Week All

Total 250

Source	Posts	Value
Instagram	133	53.2%
Facebook	117	46.8%
Twitter	0	0.0%
News	0	0.0%

Keywords chatgpt

Excluded Keywords

Try the dashboard here: <https://dashboard.behavioral-ds.science/>
Video demonstration: https://youtu.be/DcW_qNJ1UFg

Extract ads funding misinformation

Challenges:

- Ads are automatically placed by an engine (Google, Facebook, Twitter)
- Ads buyers have limited control where the ads are shown
- Social media platforms are reluctant to disclose exact placements of ads.

Our software:

- Visits given web pages and record ads placed
- Link the host pages to the ads shown circumventing the platform
- Can be configured with specific identities (e.g., one of the three personas)

Merits:

- Identifying companies that **unintentionally** fund harmful content
- Intervention against online misinformation by disrupting its funding flow
- Protect brand reputation.

The screenshot shows a web browser interface with a table titled "All URLs - Table View". The table lists various URLs with columns for Rank, URL, Title, Start, End, Date, Previous, This Year, Average, Max, Min, Avg, Std Dev, Sess, Err, and Size. A red arrow points from the top right of the table area down towards a solar panel advertisement.

BREITBART

TRENDING: RUSSIA REVOLT BIDEN CRIME FAMILY TITAN SUB 2024 RACE TRUMP INDICTMENT PR

TRUMP: 'BIDEN CRIME FAMILY' CORRUPTION '100 TIMES BIGGER THAN WATERGATE'

Former President Donald Trump believes President Joe Biden's corruption is "100 times bigger than Watergate," according to his prepared remarks for an address at the Faith and Freedom Coalition, just days after an IRS whistleblower made shocking revelations about the Biden family.

FORTUNE SOLAR

#1 ONE OF THE AUSTRALIA'S LARGEST SOLAR COMPANY

6.6 kW system STARTING FROM \$2921*

GET A FREE QUOTE

ADVERTISEMENT

Part 2. The bigger picture

Generative AI and misinformation

Generative AI fueled disinformation

In under two minutes,
this fake image led to a
\$500 billion (USD) drop
in the U.S. stock market



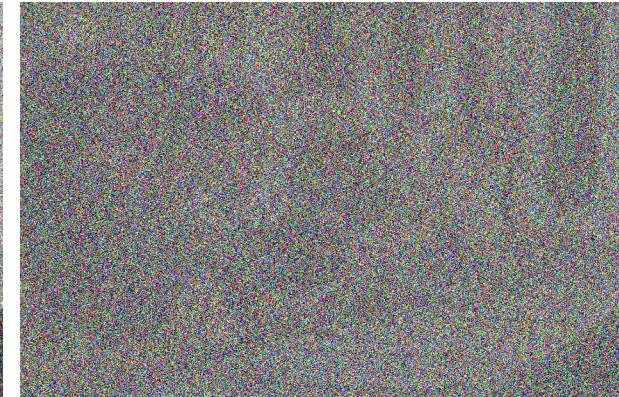
Generative AI fueled disinformation

This fake image was used by the Ron DeSantis campaign in an anti-Trump ad.



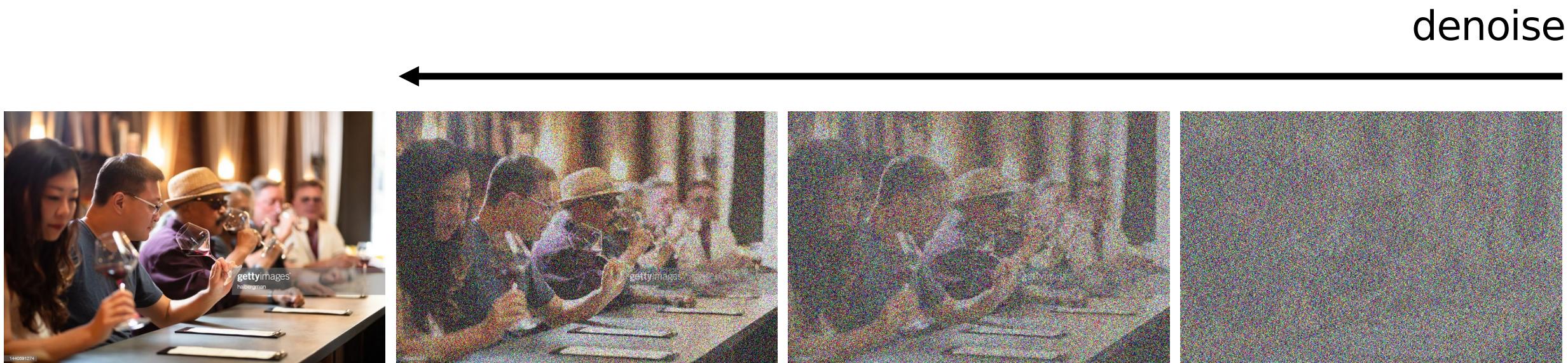
Diffusion-Based Image Synthesis

additive noise



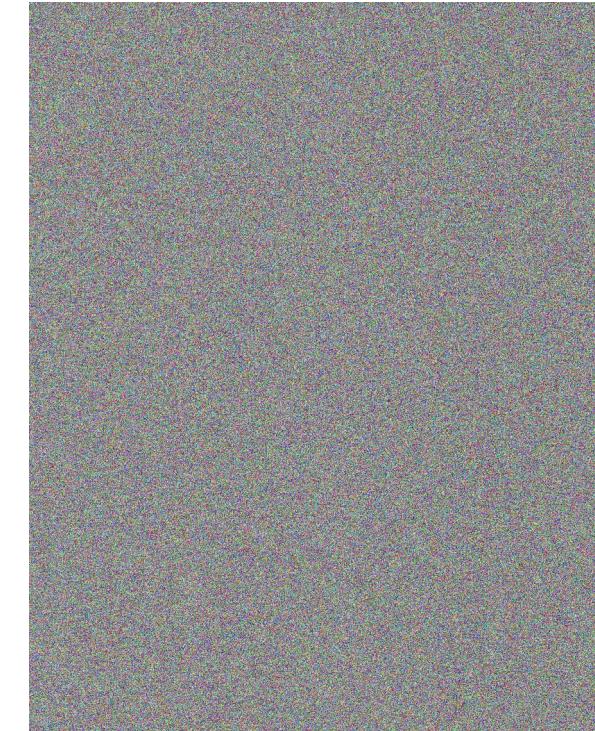
A diverse group of people sitting at the bar in an elegant wine tasting lounge in Northern California, drinking and chatting.

Diffusion-Based Image Synthesis



A diverse group of people sitting at the bar in an elegant wine tasting lounge in Northern California, drinking and chatting.

Diffusion-Based Image Synthesis



denoise

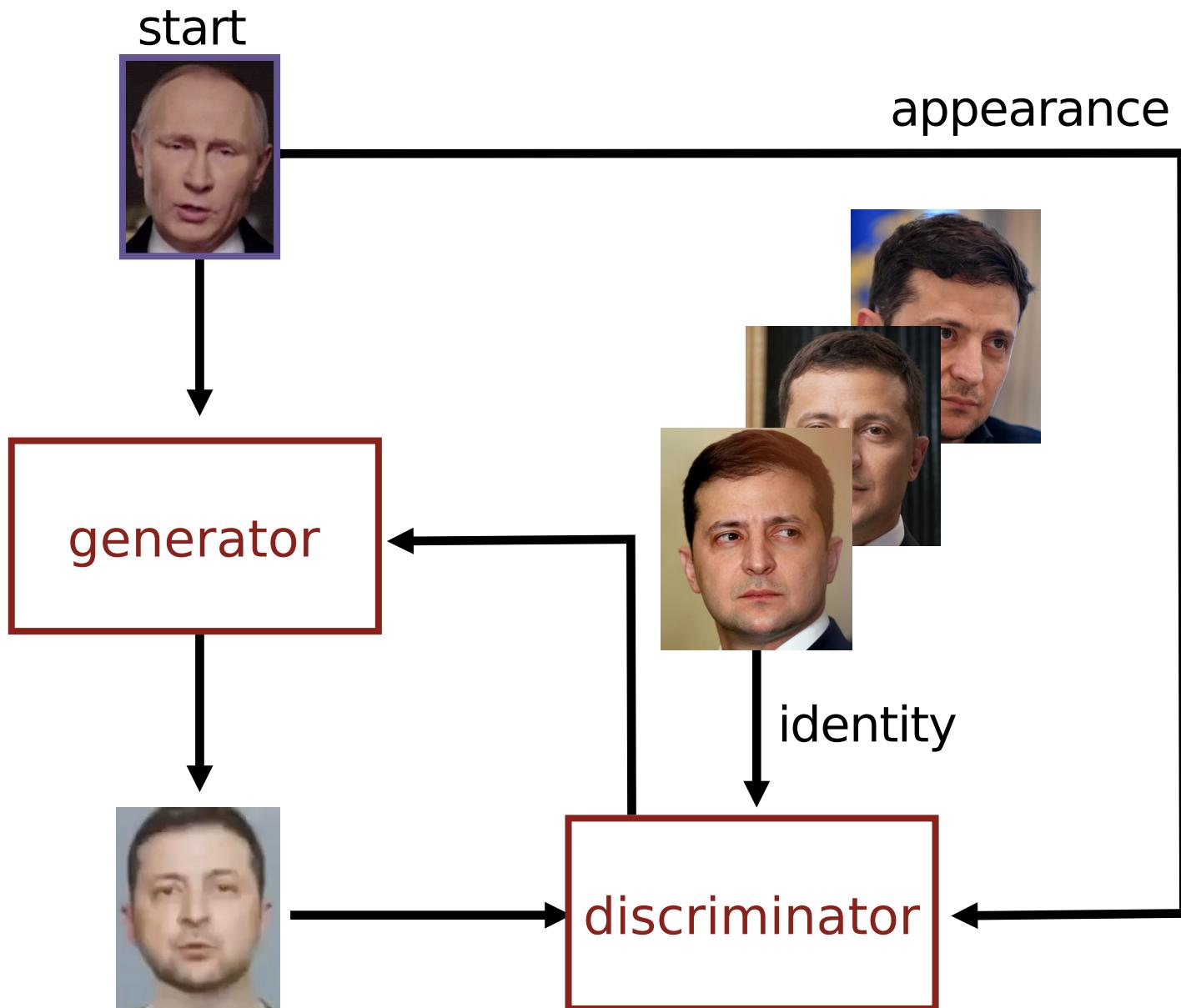
Pope Francis in a white puffer coat

Generative AI fueled disinformation

A fake video was broadcast on Ukrainian television in the early days of the Russian invasion.



GAN-based deepfake videos



GAN-based deepfake videos

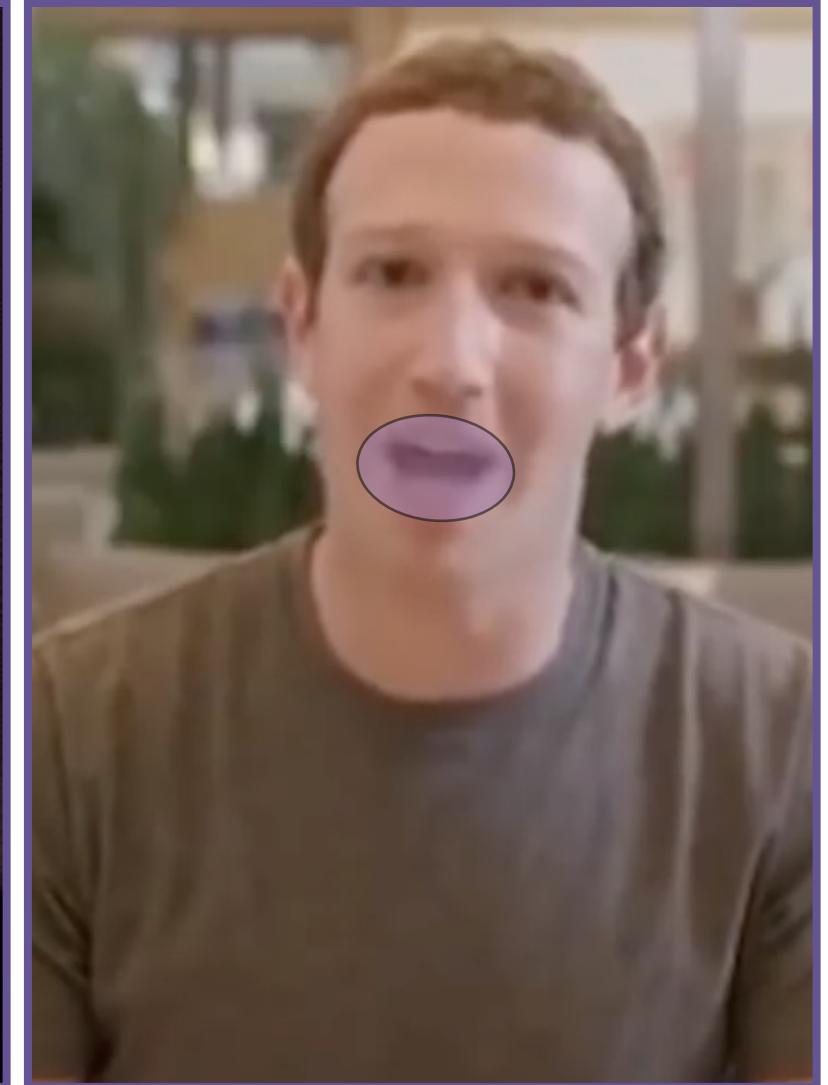
face swap



puppet master



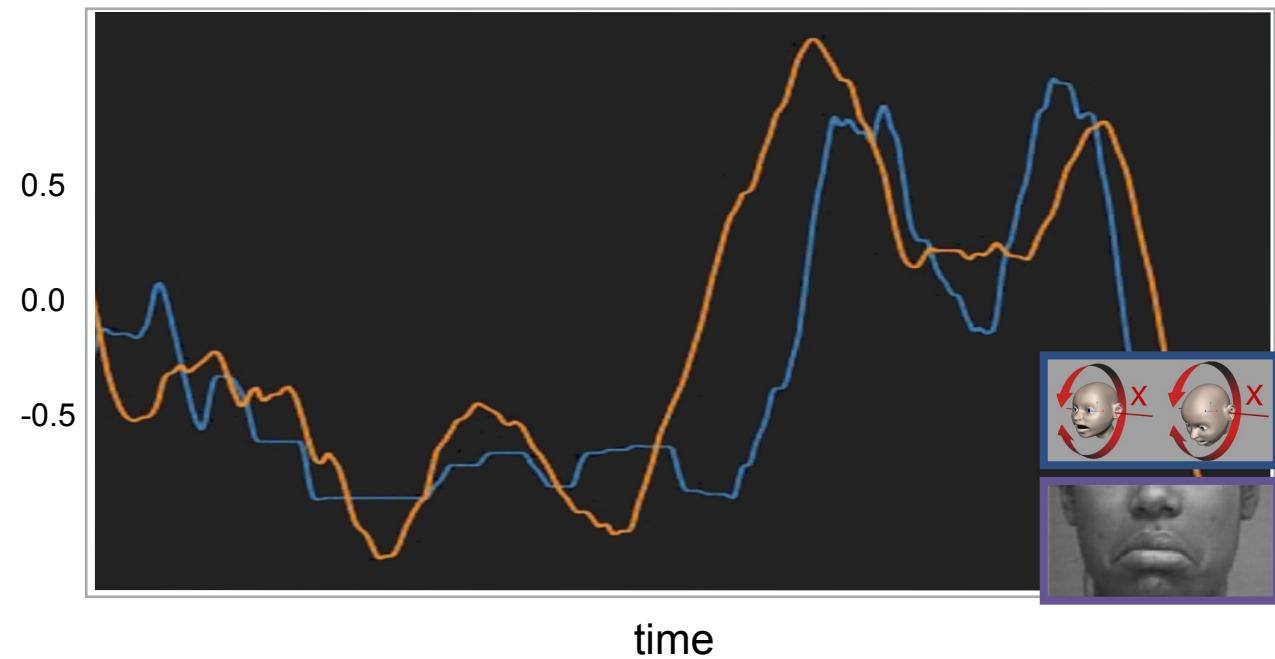
lip sync



Detecting deepfake videos



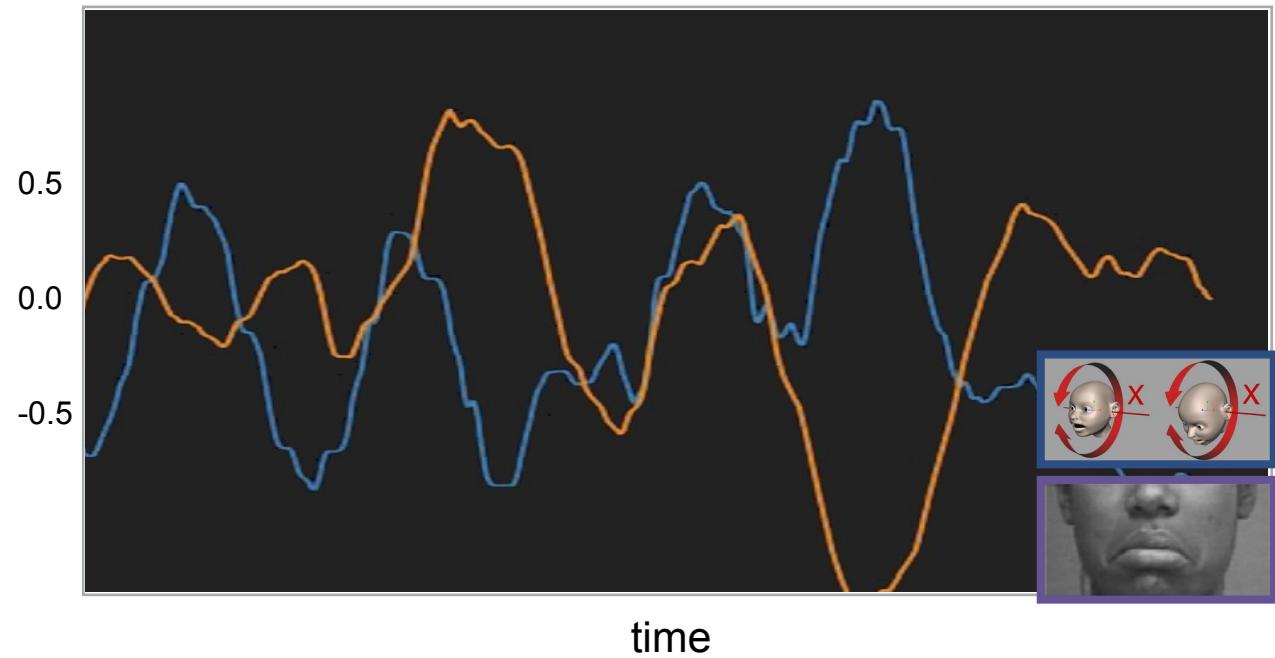
real



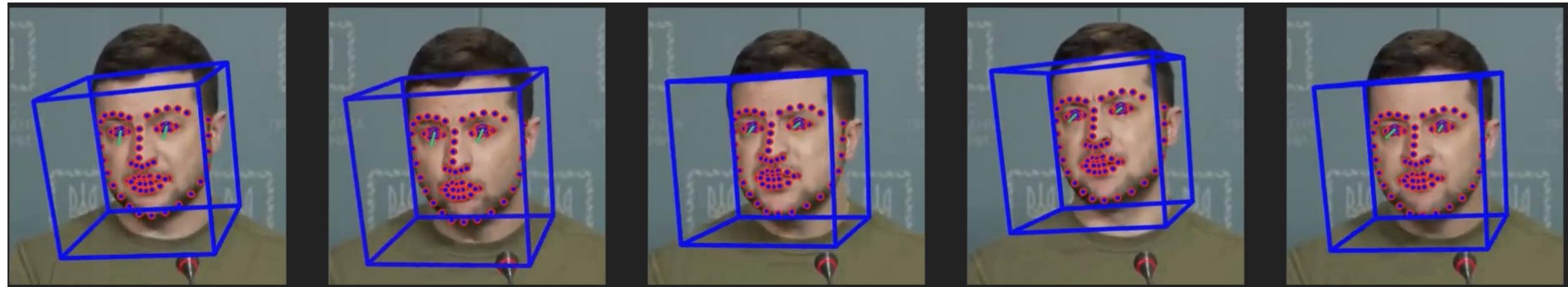
Detecting deepfake videos



fake

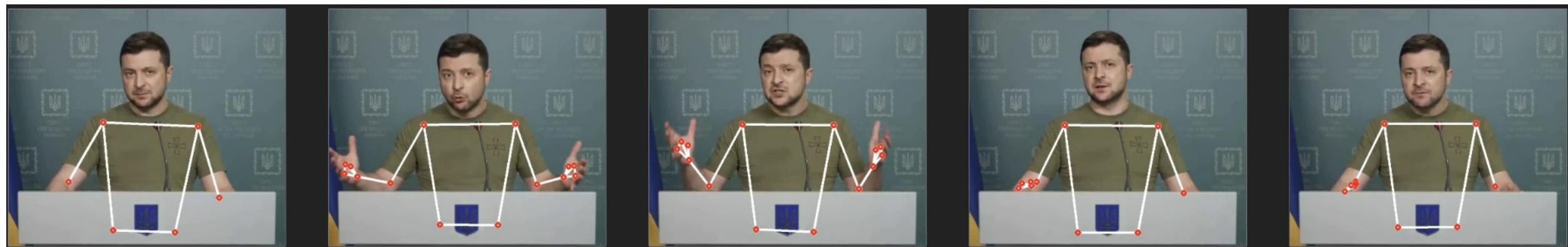


Detecting deepfake videos



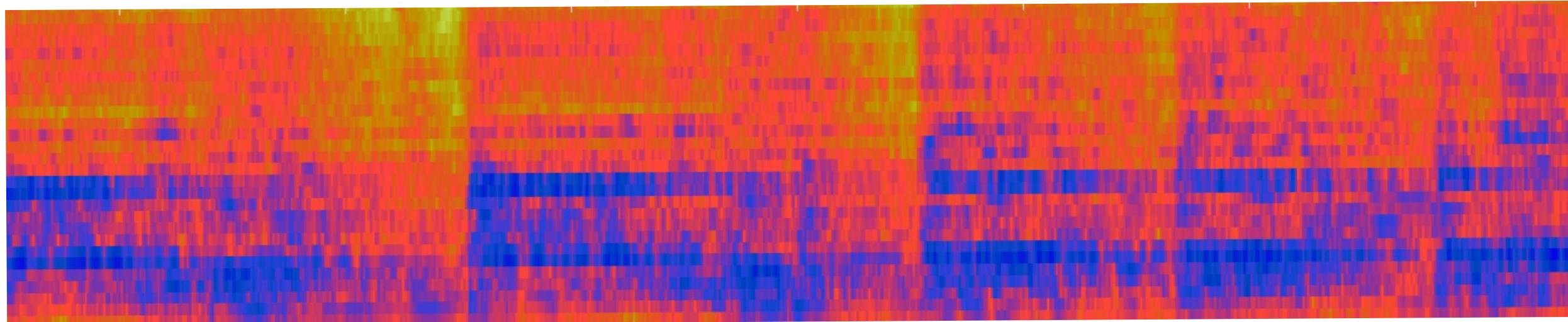
Facial expressions and head rotation

Detecting deepfake videos



Hand gestures

Detecting deepfake videos



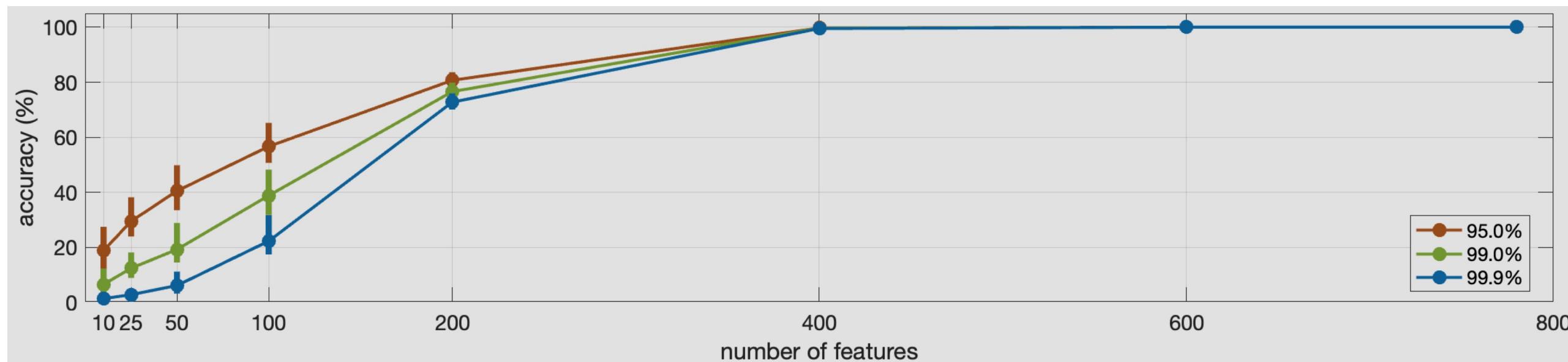
Voice patterns

Detecting deepfake videos



Data collection of real Zelenskyy

Detecting deepfake videos



Generative AI fueled disinformation, but... social media is the delivery mechanism

In under two minutes,
this fake image led to a
\$500 billion (USD) drop
in the U.S. stock market



Authentication at the source

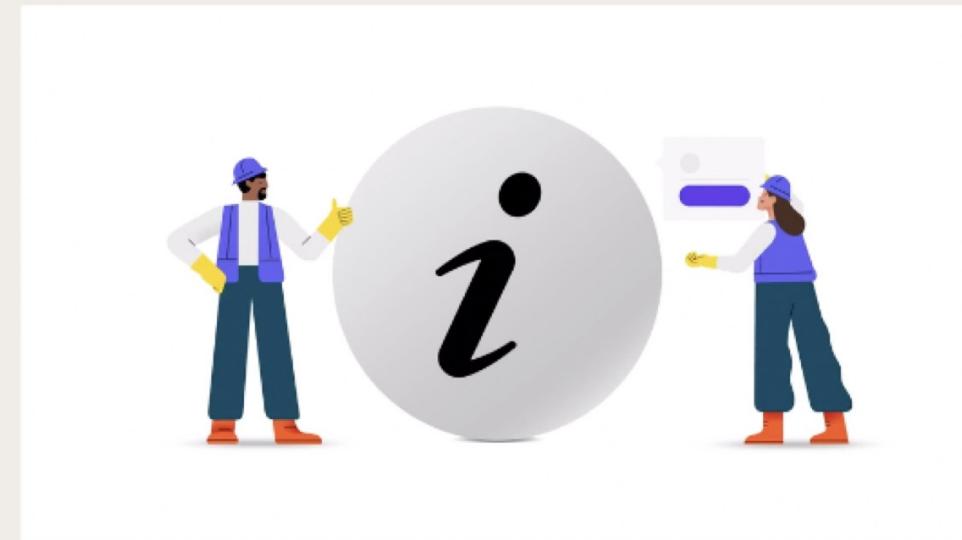


Content
Authenticity
Initiative

[How it works](#) [Membership ▾](#) [Resources ▾](#) [Blog](#) [Open-source tools](#) [Go to Verify](#)

**Authentic
storytelling
through digital
content
provenance.**

We are a community of media and tech companies, NGOs, academics, and others working to promote adoption of an open industry standard for content authenticity and provenance.





Supplementary Material for Q&A

Who are our online opinion leaders?



Jenna Abrams

@Jenn_Abrams

Politics is a circus of hypocrisy. I DO care. Any offers/ideas/questions? DM or email me jennnabrams@gmail.com (Yes, there are 3 Ns, this is important)

📍 USA

🔗 jennabrams.com

📅 Joined October 2014

⌚ Born on October 02



Tennessee GOP

@TEN_GOP

I love God, I Love my Country

📍 Tennessee, USA

📅 Joined November 2015

Common traits:

- Highly influential;
- Highly followed and retweeted;
- Opinion leaders;
- ...

General Flynn @GenFlynn
Believe in #AmericanException it's for real. NYT bestselling au @fieldoffight (Read it) I stand for American principles & values.

GLOBAL
mikeflynndefensefund.org
S-a alăturat în Ianuarie 2014

Sebastian Gorka DrG @SebGorka
@FOXNews Nat. Sy. Strategist NYT Bestseller: DEFEATING JIHAD amzn.to /2zTuXyl Fmr Strategist to Pres. Trump Order my NEW book: amzn.to /2JkuSGJ

Washington, DC
S-a alăturat în martie 2014

Who are our online opinion leaders?



Jenna Abrams

@Jenn_Abrams

Politics is a circus of hypocrisy. I DO care. Any offers/ideas/questions? DM or email me jennabrams@gmail.com (Yes, there are 3 Ns, this is important)

📍 USA

🔗 jennabrams.com

📅 Joined October 2014

📍 Born on October 02



Tennessee GOP

@TEN_GOP

I love God, I Love my Country

📍 Tennessee, USA

📅 Joined November 2015

Common traits:

- Highly influential;
- Highly followed and retweeted;
- Opinion leaders;
- ...

**Russian-controlled trolls
operated by the Internet Research Agency in St. Petersburg**

More challenges: Red Queen effect and generative models

Content- and user-based detection tools:

- language nuances, language drift, foreign language-dependent
- sensitive to adversarial training attacks – use the detector to train the attacker.

ChatGPT-like introduced problems:

- **Increase scalability of troll presence** – # accounts manned by a single operator increase multifold
- Increased awareness of **social context knowledge**, adapted to local language patterns
- **Adaptability for detection** – generative models can spit out infinite variations of the same message.



"Now, here, you see, it takes all the running you can do just to keep in the same place. If you want to get somewhere else, you must run at least twice as fast!"

- Red Queen from *Through the Looking Glass*

Solution: not what they say, but what they stir

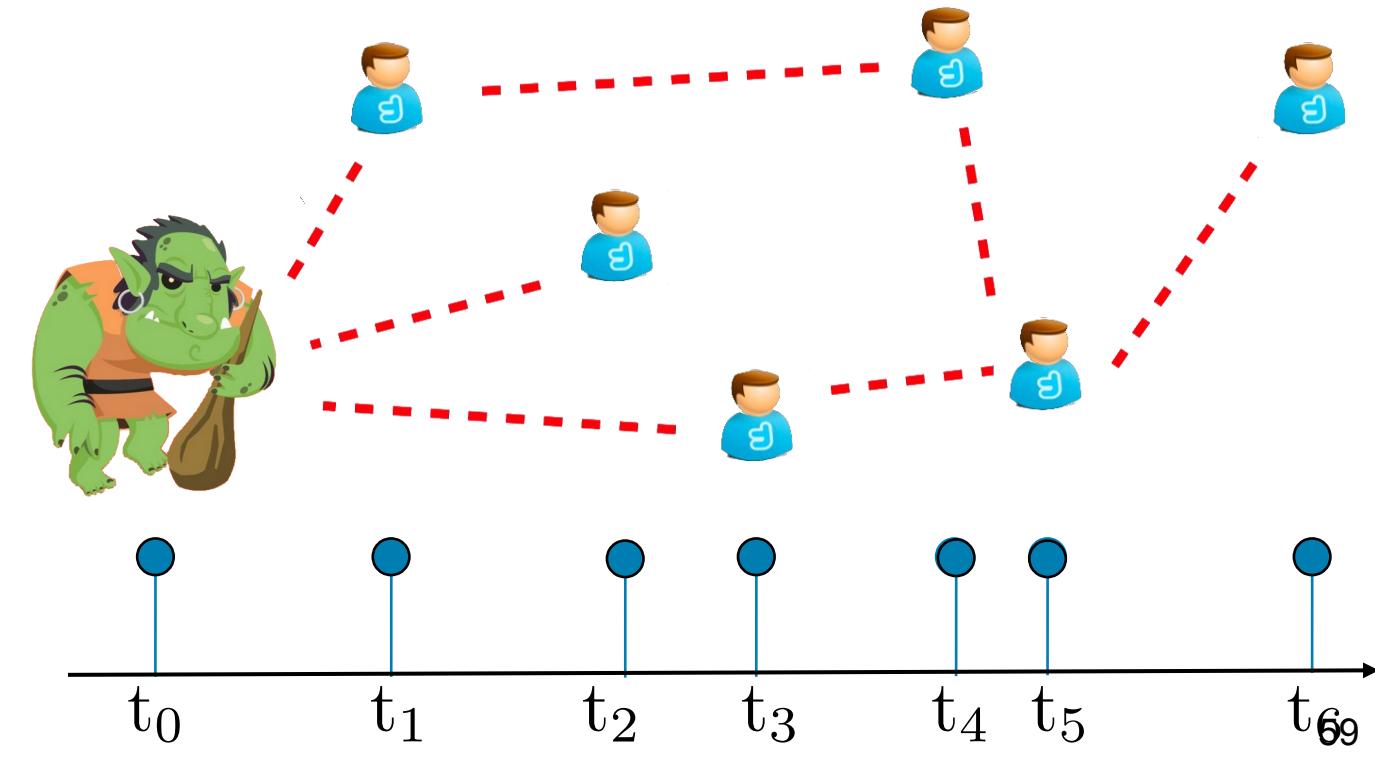
IO are designed to elicit particular reactions from the target audience

- use these reactions to detect
- cannot be faked or hidden, as they are the purpose of the campaign

Distinguish users and content types
based on the reaction of online
social systems (**no content**)

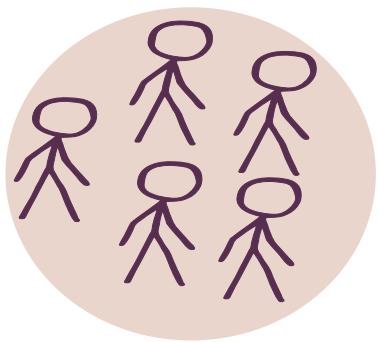
**Build early detection systems
based on information spread
patterns within the user population**

Language invariant – the model
doesn't need to “speak” the
language

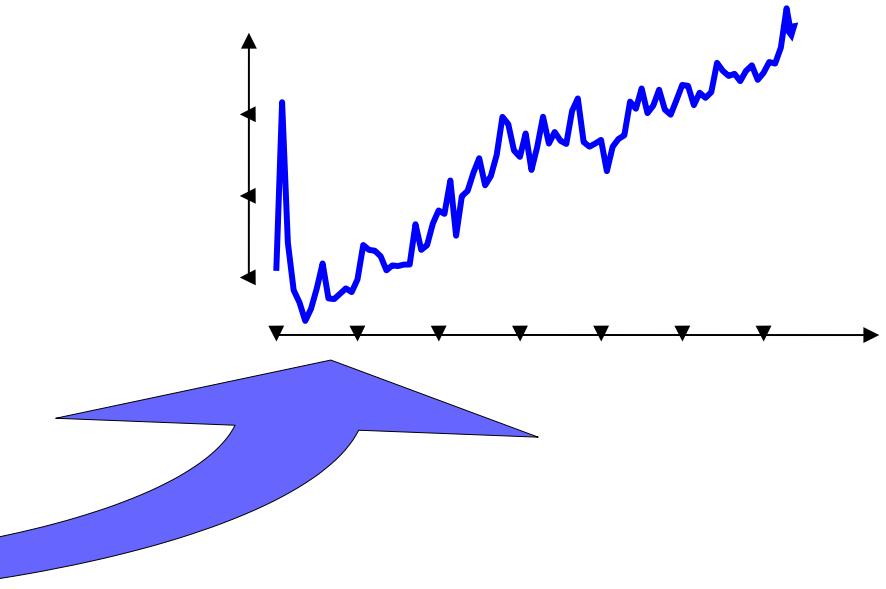


The Behavioral Data Science

1.

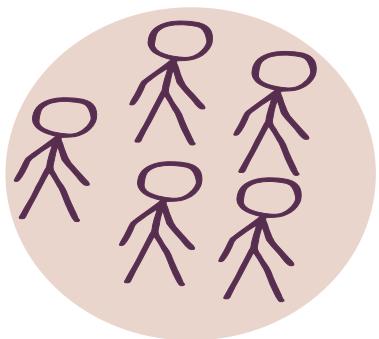


information diffusion
epidemics spreading
behavioral modeling

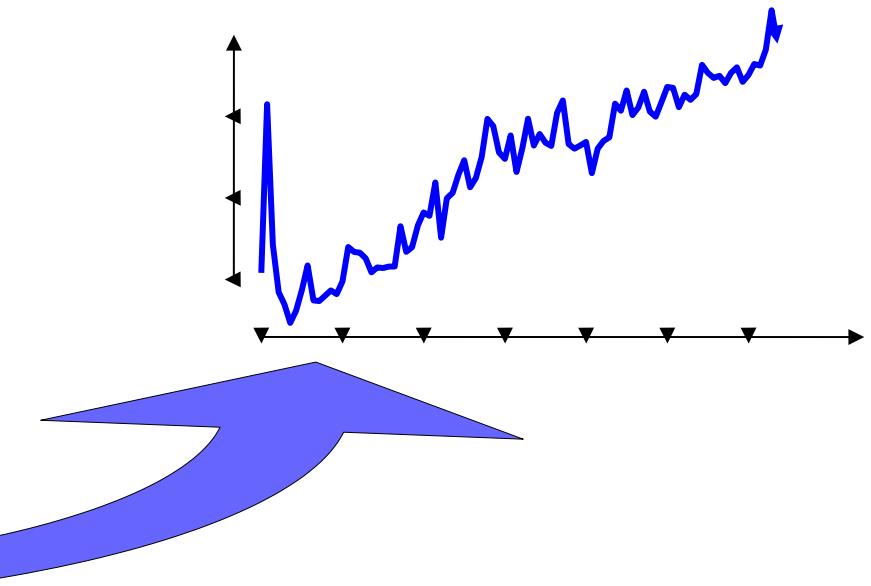


The Behavioral Data Science

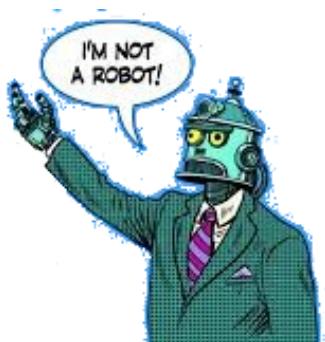
1.



information diffusion
epidemics spreading
behavioral modeling



2.

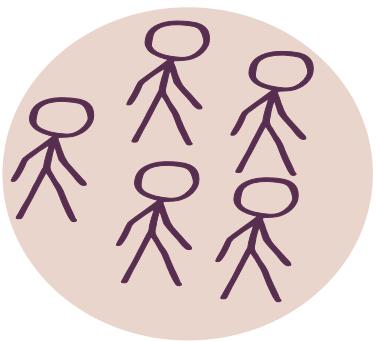


[Rizoiu et al ICWSM'18]

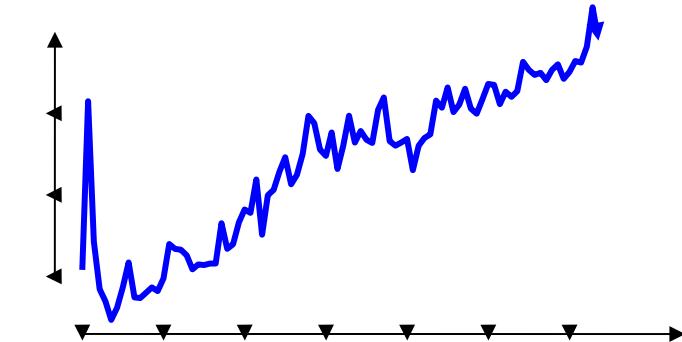
[Kim et al Journ.Comp.SocSci'19]

The Behavioral Data Science

1.



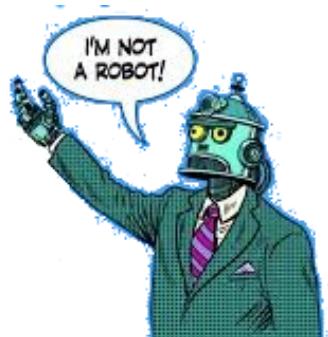
information diffusion
epidemics spreading
behavioral modeling



3.



2.



[Rizoiu et al ICWSM'18]

[Kim et al Journ.Comp.SocSci'19]

FAKE FACT

Behavioral DS action plan in Influence Operations space



Monitor

Response level



Objective



Approach

How can we develop and deploy dashboards to monitor discussion on both the social media and traditional media outlets, in which the adversaries are most likely to deploy the influence operations?

Monitor discussions on social and traditional media

Characterising the dynamic interaction between traditional and social media ecosystems in the flow and spread of disinformation and problematic content.

Develop and deploy a "mission control" dashboard to retrieve content from a constantly updating list of traditional media and Internet sources.

Detect

How do we most effectively identify and triage information campaigns based on the characteristics of the message, how it spreads, who is communicating it, and where it is being communicated?

Detect adversarial information campaigns

Utilise information diffusion techniques to identify problematic content based on the way it moves through and across online channels

Deploy natural language processing techniques to automate the detection of problematic online messages based on the structure and content of the message

Predict

What factors accelerate and intensify the communication and reach of weaponized messages within and across online environments, and which factors lead to the most significant real-world harms?

Estimate the effectiveness of influence operations

Model the impact of networks and influencers on the virality and reach of problematic messages

Track the spread of problematic messages across and between online platforms and into the real-world

Mitigate

What are practical approaches that allow us to both pro-actively and re-actively limit the harms of problematic messaging, including identifying where, when and how counter-messaging should be deployed?

Design and apply countermeasures

Use natural language processing to automatically generate counter-messaging that is tuned for the platform and target group of interest

Identify key message inoculation points in social networks based on how information flows and gains velocity

Our founders in the mis-, dis-, IO and IW spaces



Australian Government
Department of Defence
Defence Science and
Technology Group

Real-time detection of
disinformation campaigns



Information integrity initiative:
fighting misinformation in Australia



Effectiveness of Information
Operations in the Pacific



Australian Government
Department of Defence
Defence Science and
Technology Group

Information Warfare
STaR Shot “Developing
Situational Awareness”



Hate Speech propagation
on Social Media

Action plan



**Response
level**



Objective



Approach

Monitor

Detect

Predict

Mitigate

How can we develop and deploy dashboards to monitor discussion on both the social media and traditional media outlets, in which the adversaries are most likely to deploy the influence operations?

Monitor discussions on social and traditional media

Characterising the dynamic interaction between traditional and social media ecosystems in the flow and spread of disinformation and problematic content.

Develop and deploy a "mission control" dashboard to retrieve content from a constantly updating list of traditional media and Internet sources.

How do we most effectively identify and triage information campaigns based on the characteristics of the message, how it spreads, who is communicating it, and where it is being communicated?

Detect adversarial information campaigns

Utilise information diffusion techniques to identify problematic content based on the way it moves through and across online channels

Deploy natural language processing techniques to automate the detection of problematic online messages based on the structure and content of the message

What factors accelerate and intensify the communication and reach of weaponized messages within and across online environments, and which factors lead to the most significant real-world harms?

Estimate the effectiveness of influence operations

Model the impact of networks and influencers on the virality and reach of problematic messages

Track the spread of problematic messages across and between online platforms and into the real-world

What are practical approaches that allow us to both pro-actively and re-actively limit the harms of problematic messaging, including identifying where, when and how counter-messaging should be deployed?

Design and apply countermeasures

Use natural language processing to automatically generate counter-messaging that is tuned for the platform and target group of interest

Identify key message inoculation points in social networks based on how information flows and gains velocity

Monitor: Monitoring discussion spaces

Information Dashboard

Topics Sort by attention ▾

Add new topic

- # climate change
- # 2019-20 Australian bushfire ...
- # LGBT
- # Off-topic
- # vaccination

Labels

Facebook pages

Settings

2019-20 Australian bushfire season

Saved views Filter Sort by Posting date

Summary

Hour Day Week

Total 2 456

Source	Posts	Value
Twitter	818	33.3%
Facebook	818	33.3%
News	818	33.3%

Keywords

garden, seed, landscape, lawn
mower, lawn mower, mulch, prunes,
aerator, planters

Excluded Keywords

sweet, peanut butter, pizza rolls, ice
cream sandwich

Washington Post Jul 21 / 18:01

Australian fires had bigger impact on climate than covid-19 lockdowns in 2020

Theme Sexism Opinion Do not agree China is responsible for Covid-19

Elaine Johnson Jul 21 / 18:01

"We'd just finished building our house before Christmas and by New Years' Eve, it was gone. There was no time to stop and take it all in. We were needed somewhere else."

— Michael Pratt, Fire and Rescue Tumbarumba

Theme Vaccine government tracking/controlling conspiracy theory Opinion Agree China is responsible for Covid-19

ABC Adelaide Jul 21 / 18:01

How cool is this? 🤯

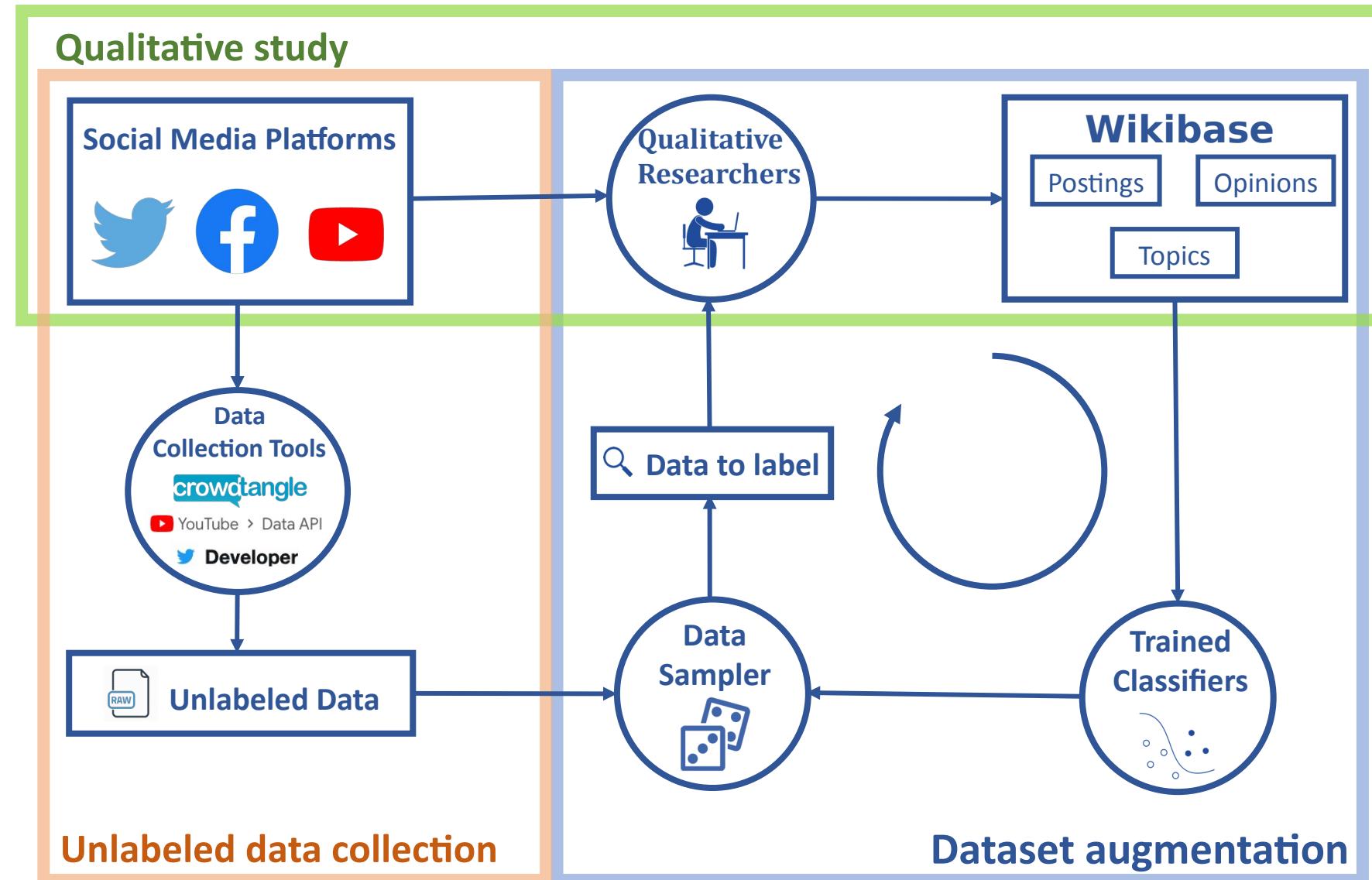
The first Australian owned and operated Black Hawk helicopter has been fitted with firefighting capabilities ahead of next bushfire season. Adelaide based aviation company...

Theme White nationalism Opinion Agree 5G/smart tech is unsafe/a scam/a way of controlling people

Washington Post Jul 21 / 18:01

Graphical interface of the Information Dashboard

Monitor: Semi-automatic data labelling



Action plan



**Response
level**



Objective



Approach

Monitor

Detect

Predict

Mitigate

How can we develop and deploy dashboards to monitor discussion on both the social media and traditional media outlets, in which the adversaries are most likely to deploy the influence operations?

Monitor discussions on social and traditional media

Characterising the dynamic interaction between traditional and social media ecosystems in the flow and spread of disinformation and problematic content.

Develop and deploy a "mission control" dashboard to retrieve content from a constantly updating list of traditional media and Internet sources.

How do we most effectively identify and triage information campaigns based on the characteristics of the message, how it spreads, who is communicating it, and where it is being communicated?

Detect adversarial information campaigns

Utilise information diffusion techniques to identify problematic content based on the way it moves through and across online channels

Deploy natural language processing techniques to automate the detection of problematic online messages based on the structure and content of the message

What factors accelerate and intensify the communication and reach of weaponized messages within and across online environments, and which factors lead to the most significant real-world harms?

Estimate the effectiveness of influence operations

Model the impact of networks and influencers on the virality and reach of problematic messages

Track the spread of problematic messages across and between online platforms and into the real-world

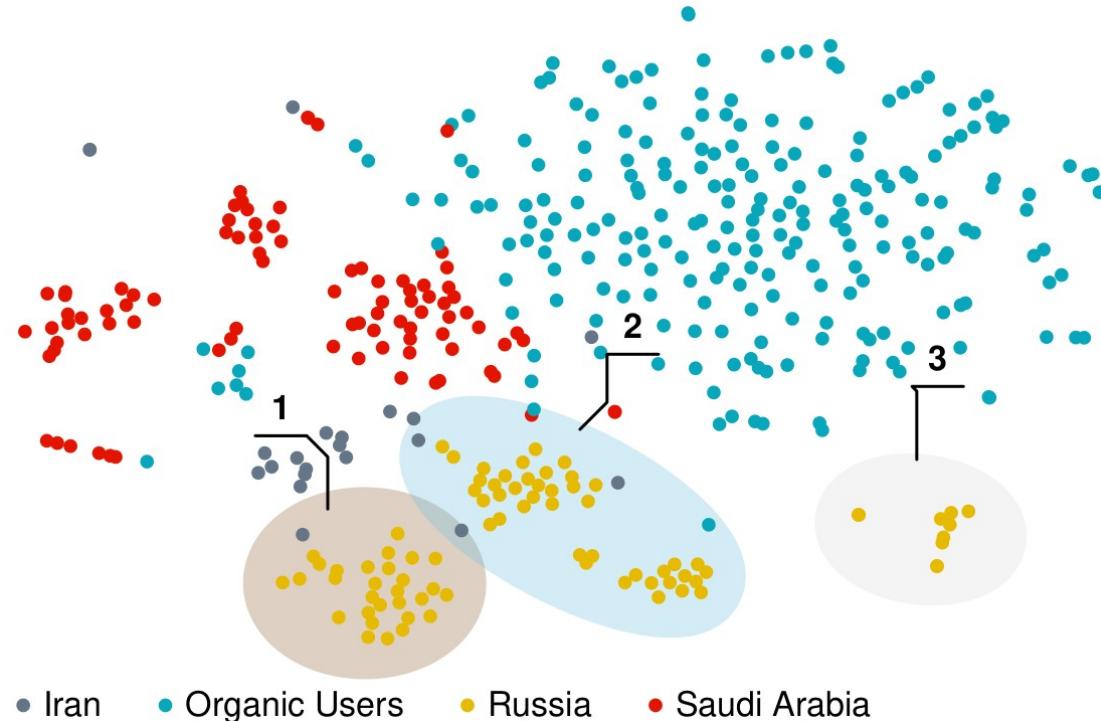
What are practical approaches that allow us to both pro-actively and re-actively limit the harms of problematic messaging, including identifying where, when and how counter-messaging should be deployed?

Design and apply countermeasures

Use natural language processing to automatically generate counter-messaging that is tuned for the platform and target group of interest

Identify key message inoculation points in social networks based on how information flows and gains velocity

Detect: identify agent types and coordinated behavior



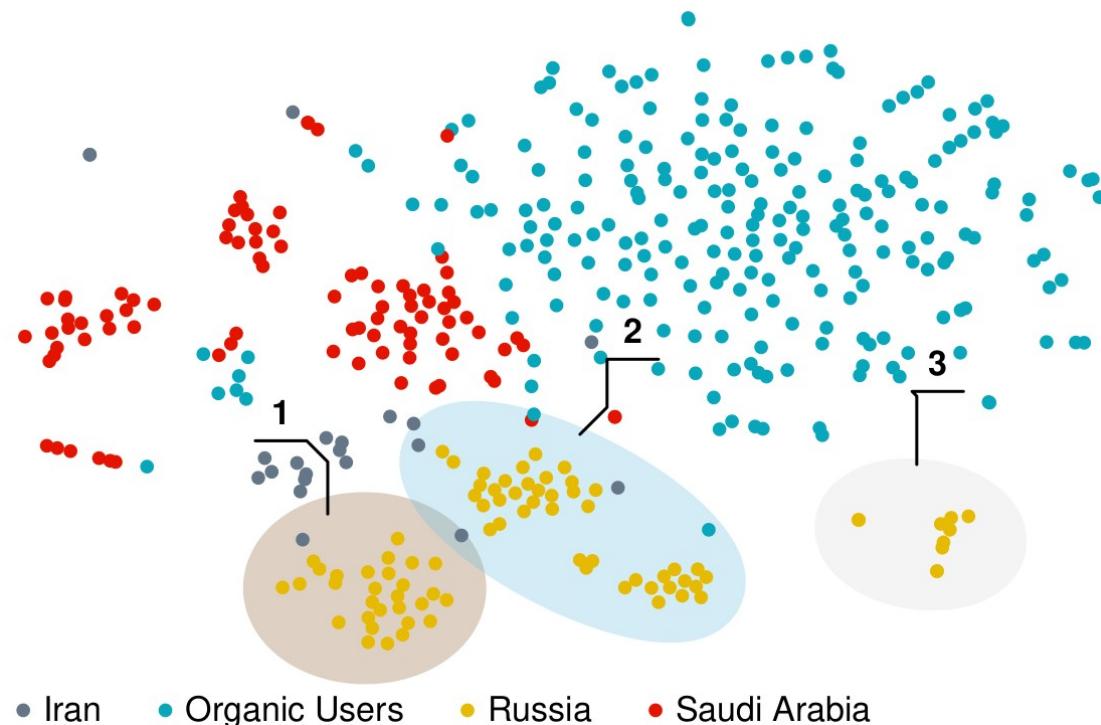
IC-TH clusters IO agents from specific countries based solely on the timing of the cascades in which they participate; it identifies even individual “troll farms”.



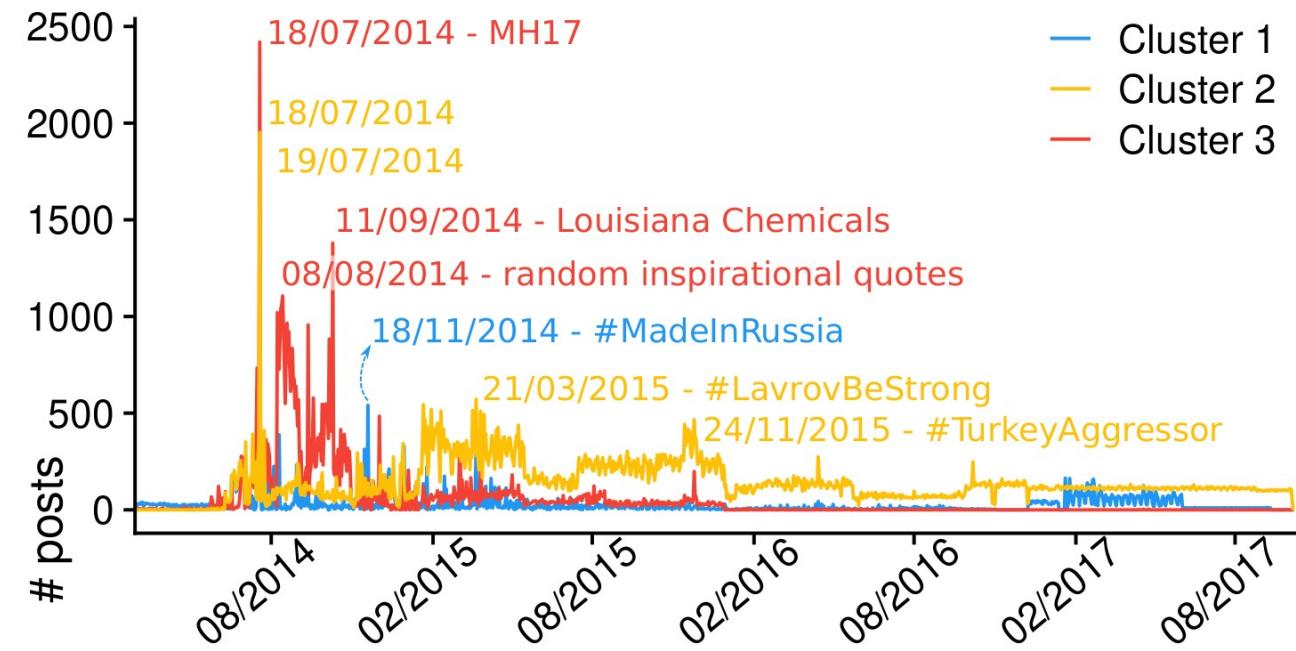
The technical detail:

Interval-censored Transformer Hawkes; Twitter Moderation Research Consortium dataset; partial data loss

Detect: identify agent types and coordinated behavior



IC-TH clusters IO agents from specific countries based solely on the timing of the cascades in which they participate; it identifies even individual “troll farms”.



Qualitative investigations uncovers strategies of Russian trolls farms:
C1: Russian news with patriotic framing;
C2: Regional and conservative news;
C3: tweet in English, #music, #usa, relationship advice



The technical detail:

Interval-censored Transformer Hawkes; Twitter Moderation Research Consortium dataset; partial data loss

Detect: #DeathTax coordinated campaign



Clear structure with two clusters:
disinformation (right) and debunking (left)

Disinformation cluster: tightly connected, coordinated and timed retweeting

◆ **Debunking cluster:** organic retweeting, reactionary, loosely connected, multiple communities



The technical detail:

Map information networks from social media; content, interactions, structure and diffusions analyse; social network analysis

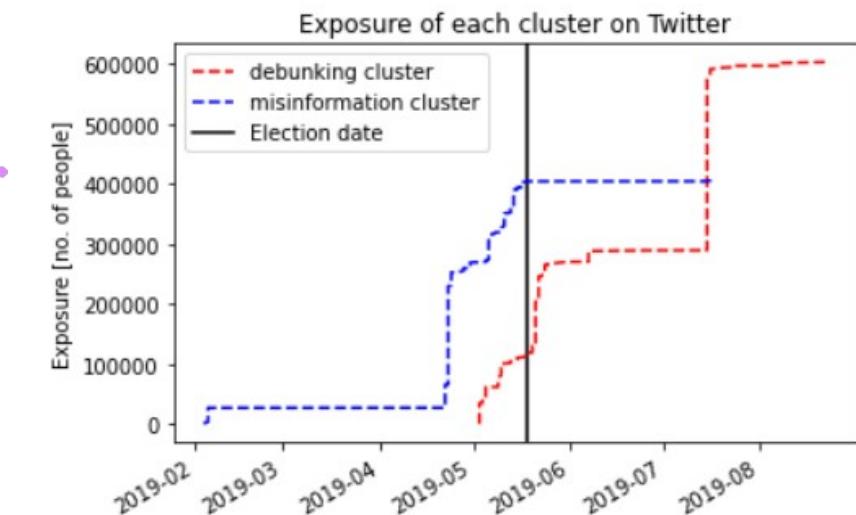
Detect: #DeathTax coordinated campaign



Clear structure with two clusters:
disinformation (right) and debunking (left)

Disinformation cluster: tightly connected, coordinated and timed retweeting

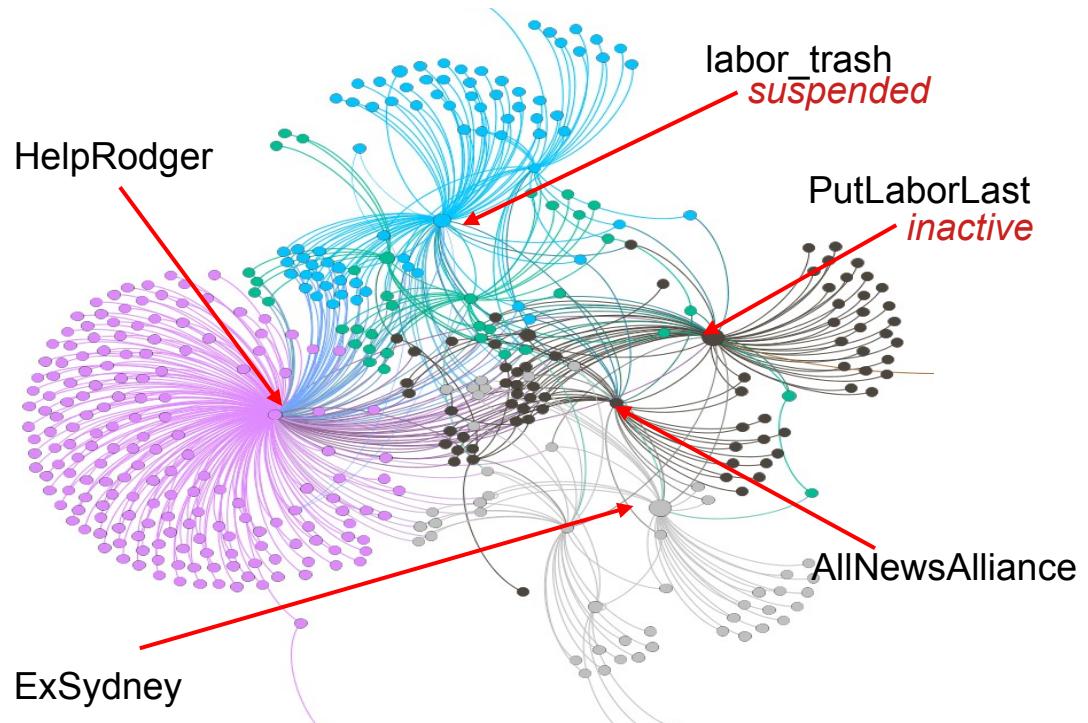
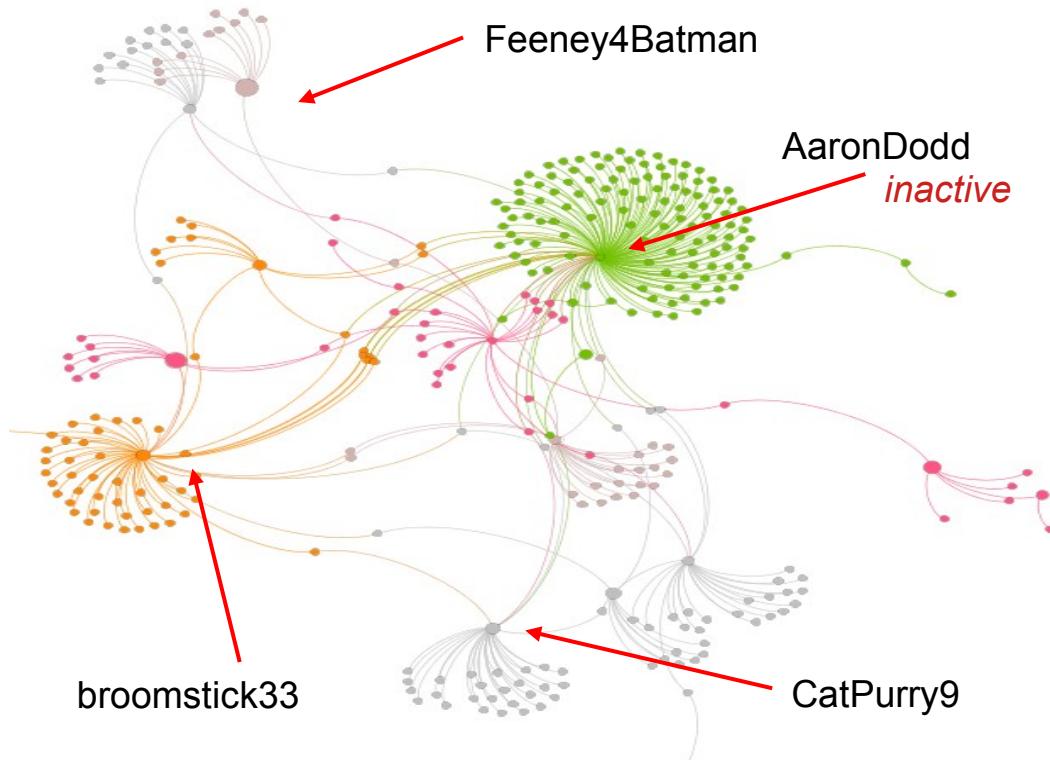
◦ **Debunking cluster:** organic retweeting, reactionary, loosely connected, multiple communities



The technical detail:

Map information networks from social media; content, interactions, structure and diffusions analyse; social network analysis

Detect: two clusters emerge



Observations:

- Two clusters – one misinformation and one debunking the misinformation
- Misinformation cluster (right) is strongly connected compared to debunking cluster (left), cross resharing



Detect: Detect Russian IO operatives and their narratives in Ukraine

Collect data on Telegram and Twitter, in three languages (Russian, Ukrainian and English)

Kherson counteroffensive (29/08/2022- 11/11/2022):

#Kherson / #Херсон
#Khersoncounteroffensive
#Khersonoffensive
#Enerhodar / #Енергодар
#Zaporizhia / #Запоріжжя / #Запорожье
#Voznesensk / #Вознесенськ / #Вознесенск
#Mykolaiv / #Миколаїв / #Николаев
#Melitopol / #Мелітополь / #Мелитополь

Kharkiv counteroffensive (06/09/2022-02/10/2022):

#Kharkiv / #Харків / #Харьков
#Izium / #Ізюм / #Изюм
#Balakliia / #Балаклія / #Балаклия
#Kupiansk / #Купянськ / #Купянск
#Sviatohirsk / #Святогірськ / #Святогорск
#Svatove / #Сватове
#Kremenna / #Кремінна / #Кременная
#Luhansk / #Луганськ / #Луганск
#Shevchenkove / #Шевченкове
#Verbivka / #Вербівка
#Yakovenkove / #Яковенкове



Australian Government

Department of Defence

Defence Science and
Technology Group

NARRATE

Topics

Add new topic

ukraine

Labels
Facebook Pages
Settings
Logout

data collection

users

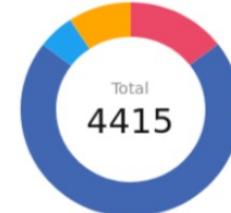
narratives

Saved views

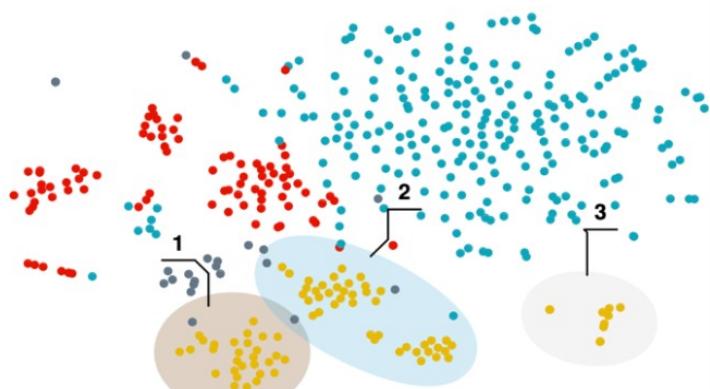
Filter Sort by Posting Date DESC

Summary

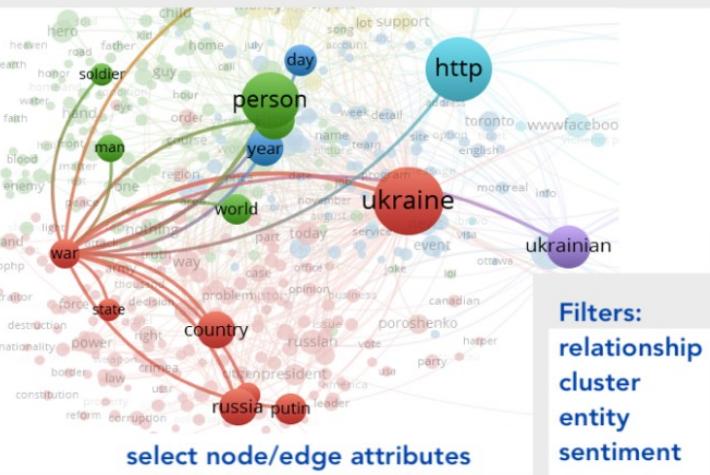
Hour Day Week All



Source	Posts	Value
Instagram	643	14.6%
Facebook	3124	70.8%
Twitter	239	5.4%
News	409	9.3%

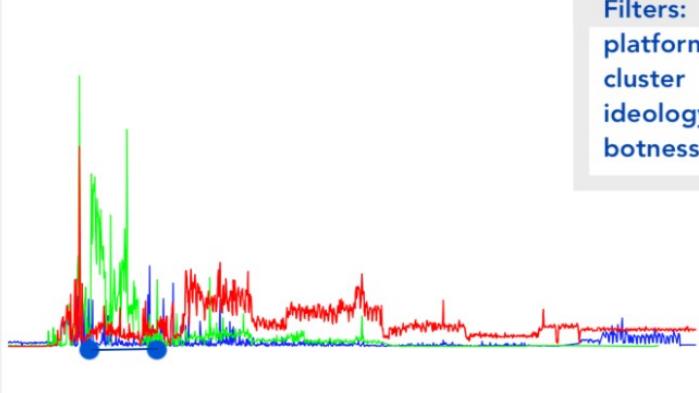


select cluster(s) to analyse



select node/edge attributes

Filters:
relationship
cluster
entity
sentiment



select a time interval

Natural language summary:

Narrative #1
Narrative #2
Narrative #3
...

CSV export:
Narrative #1
Narrative #2
Narrative #3
...

Action plan



**Response
level**



Objective



Approach

Monitor

Detect

Predict

Mitigate

How can we develop and deploy dashboards to monitor discussion on both the social media and traditional media outlets, in which the adversaries are most likely to deploy the influence operations?

Monitor discussions on social and traditional media

Characterising the dynamic interaction between traditional and social media ecosystems in the flow and spread of disinformation and problematic content.

Develop and deploy a "mission control" dashboard to retrieve content from a constantly updating list of traditional media and Internet sources.

How do we most effectively identify and triage information campaigns based on the characteristics of the message, how it spreads, who is communicating it, and where it is being communicated?

Detect adversarial information campaigns

Utilise information diffusion techniques to identify problematic content based on the way it moves through and across online channels

Deploy natural language processing techniques to automate the detection of problematic online messages based on the structure and content of the message

What factors accelerate and intensify the communication and reach of weaponized messages within and across online environments, and which factors lead to the most significant real-world harms?

Estimate the effectiveness of influence operations

Model the impact of networks and influencers on the virality and reach of problematic messages

Track the spread of problematic messages across and between online platforms and into the real-world

What are practical approaches that allow us to both pro-actively and re-actively limit the harms of problematic messaging, including identifying where, when and how counter-messaging should be deployed?

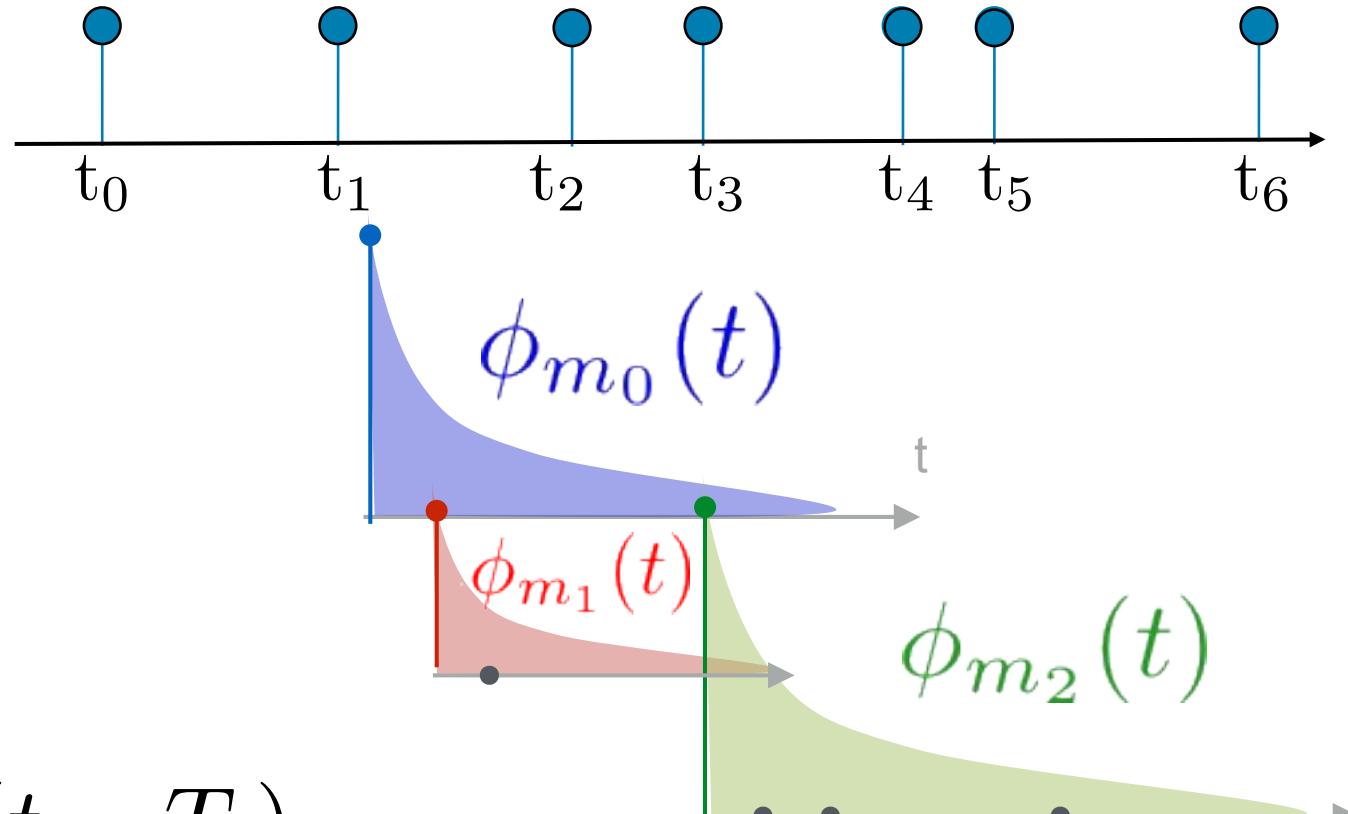
Design and apply countermeasures

Use natural language processing to automatically generate counter-messaging that is tuned for the platform and target group of interest

Identify key message inoculation points in social networks based on how information flows and gains velocity

Self-exciting (Hawkes) processes

[Hawkes, 1971]



Event intensity

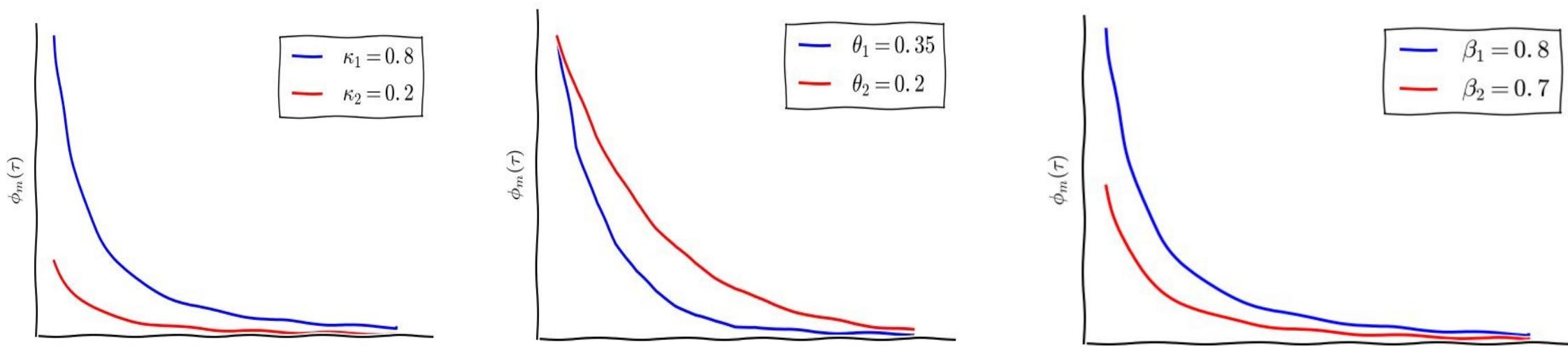
$$\lambda(t|\mathcal{H}_t) = \mu(t) + \sum_{i:t>T_i} \phi(t - T_i)$$

base intensity
(exogenous)

self-excitation
(endogenous)

The “social media” kernel

$$\phi_m(t - T_i) = \frac{\kappa m^\beta (t - T_i)^{-(1+\theta)}}{\text{the rate of 'daughter' events} \times \text{content virality} \times \text{user influence} \times \text{memory}}$$

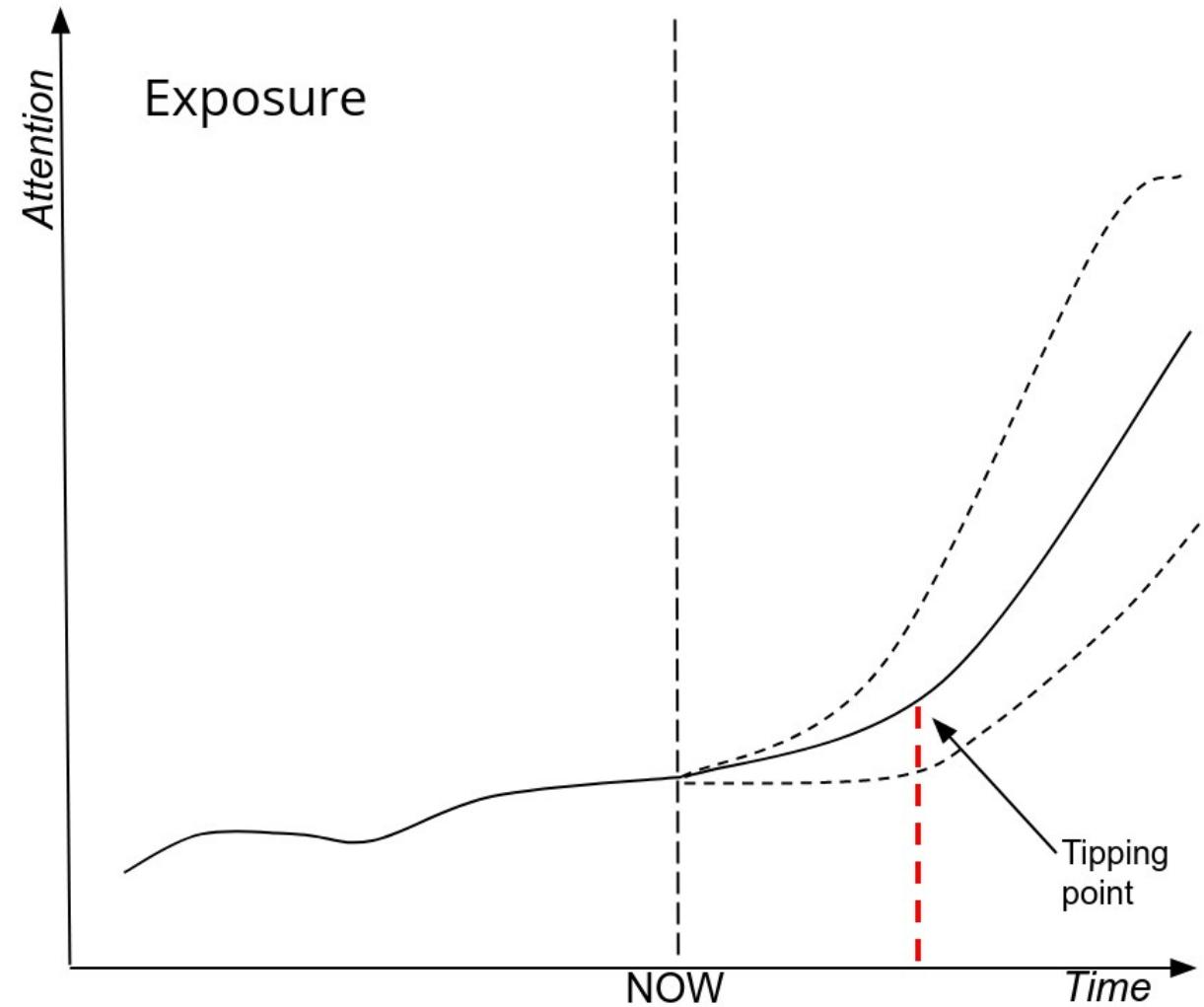


Predict: Future adoption and tipping points

Given an early observed spread, predict the future attention a narrative will gain

tipping points – the points at which the reach of IO increases exponentially

When to act? (hint: before tipping point)
What to counter? (hint: potentially popular IO)



Predict: Effectiveness of state-sponsored narratives

Purpose: Predicting the Effectiveness of State-sponsored Influence Operations: a Case Study of the Solomon Islands and Melanesia

Questions:

RQ1: Do Chinese IO operations in different regions of Melanesia consistently use particular narrative constructions and language? Do they seek to elicit specific reactions?

RQ2: How to effectively identify and triage information campaigns based on the characteristics of the message, how it spreads, who is communicating it, and where it is being communicated?

RQ3: What factors accelerate the reach of weaponized messages within and across online environments, and which factors lead to the most significant real-world harm?

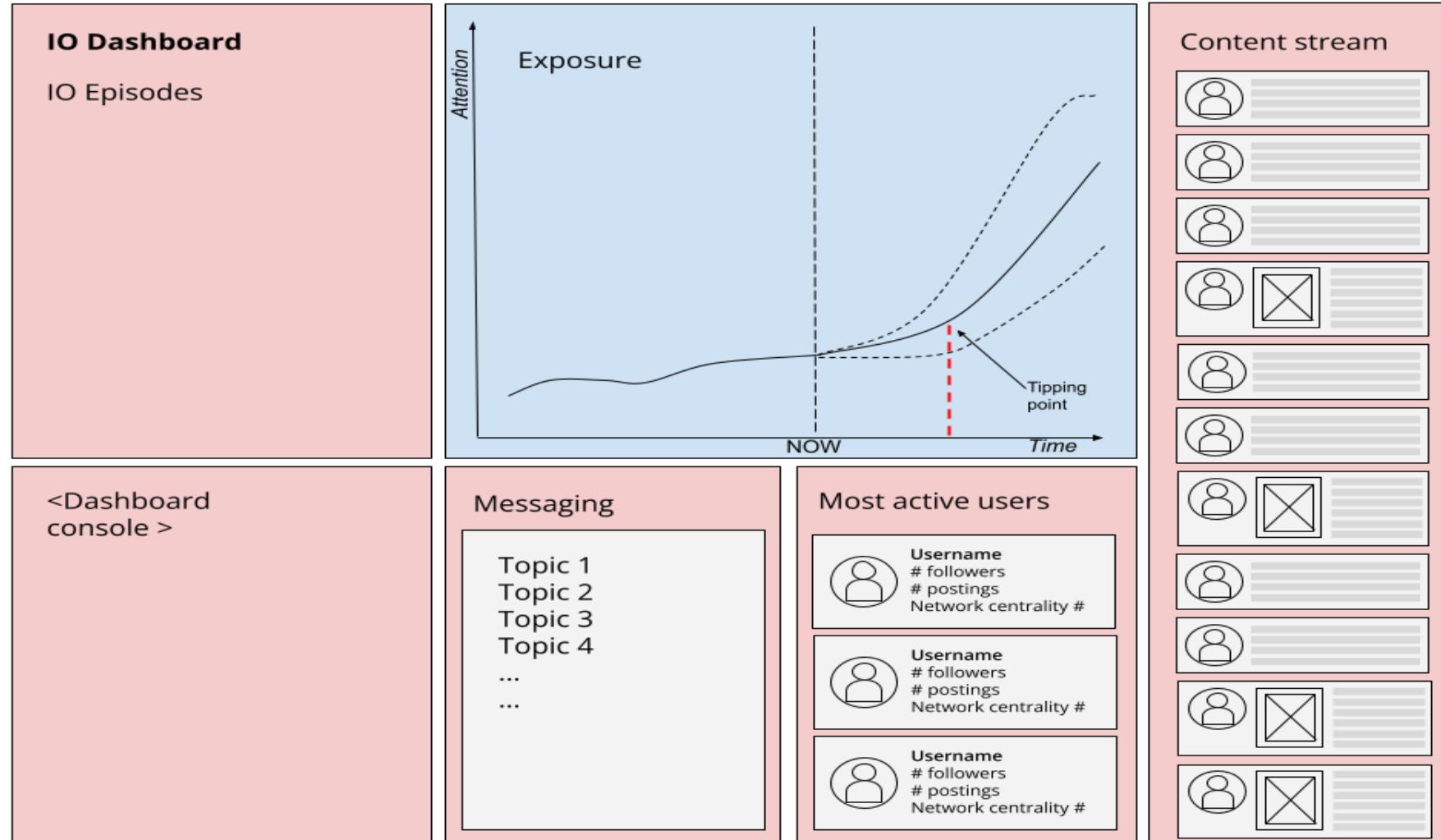
RQ4: How to forecast the future attention received by an IO? How to distinguish those IO likely to be effective? Is there a tipping point before which countermeasures are most effective?

Predict: Effectiveness of state-sponsored narratives

(some) Identified narratives:

1. Solomon Islands should become economically self-reliant and vibrant post-RAMSI, through relations with China.
2. The West excludes China from the South Pacific to perpetuate their domination.
3. **Australia and the US are colonialist bullies that threaten Solomon Islands sovereignty.**
4. **The West perpetuate a ‘Cold War mentality’.**
5. Westernised activists and civil society groups undermine the Solomon Islands government.
6. Australia is trying to interfere in Solomon Islands democracy.
7. The West instigated the November 2021 Honiara riots.
8. **Australia and the West will invade the Solomons because of its security relations with China.**

Predict: Effectiveness of state-sponsored narratives



Action plan



**Response
level**



Objective



Approach

Monitor

Detect

Predict

Mitigate

How can we develop and deploy dashboards to monitor discussion on both the social media and traditional media outlets, in which the adversaries are most likely to deploy the influence operations?

Monitor discussions on social and traditional media

Characterising the dynamic interaction between traditional and social media ecosystems in the flow and spread of disinformation and problematic content.

Develop and deploy a "mission control" dashboard to retrieve content from a constantly updating list of traditional media and Internet sources.

How do we most effectively identify and triage information campaigns based on the characteristics of the message, how it spreads, who is communicating it, and where it is being communicated?

Detect adversarial information campaigns

Utilise information diffusion techniques to identify problematic content based on the way it moves through and across online channels

Deploy natural language processing techniques to automate the detection of problematic online messages based on the structure and content of the message

What factors accelerate and intensify the communication and reach of weaponized messages within and across online environments, and which factors lead to the most significant real-world harms?

Estimate the effectiveness of influence operations

Model the impact of networks and influencers on the virality and reach of problematic messages

Track the spread of problematic messages across and between online platforms and into the real-world

What are practical approaches that allow us to both pro-actively and re-actively limit the harms of problematic messaging, including identifying where, when and how counter-messaging should be deployed?

Design and apply countermeasures

Use natural language processing to automatically generate counter-messaging that is tuned for the platform and target group of interest

Identify key message inoculation points in social networks based on how information flows and gains velocity

Mitigate: countering conspiracy theories, misinformation and state-sponsored IO



Questions:

RQ1: Are messages written in particular styles more effective in spreading?

RQ2: Do different audiences/segments of the population prefer specific styles?

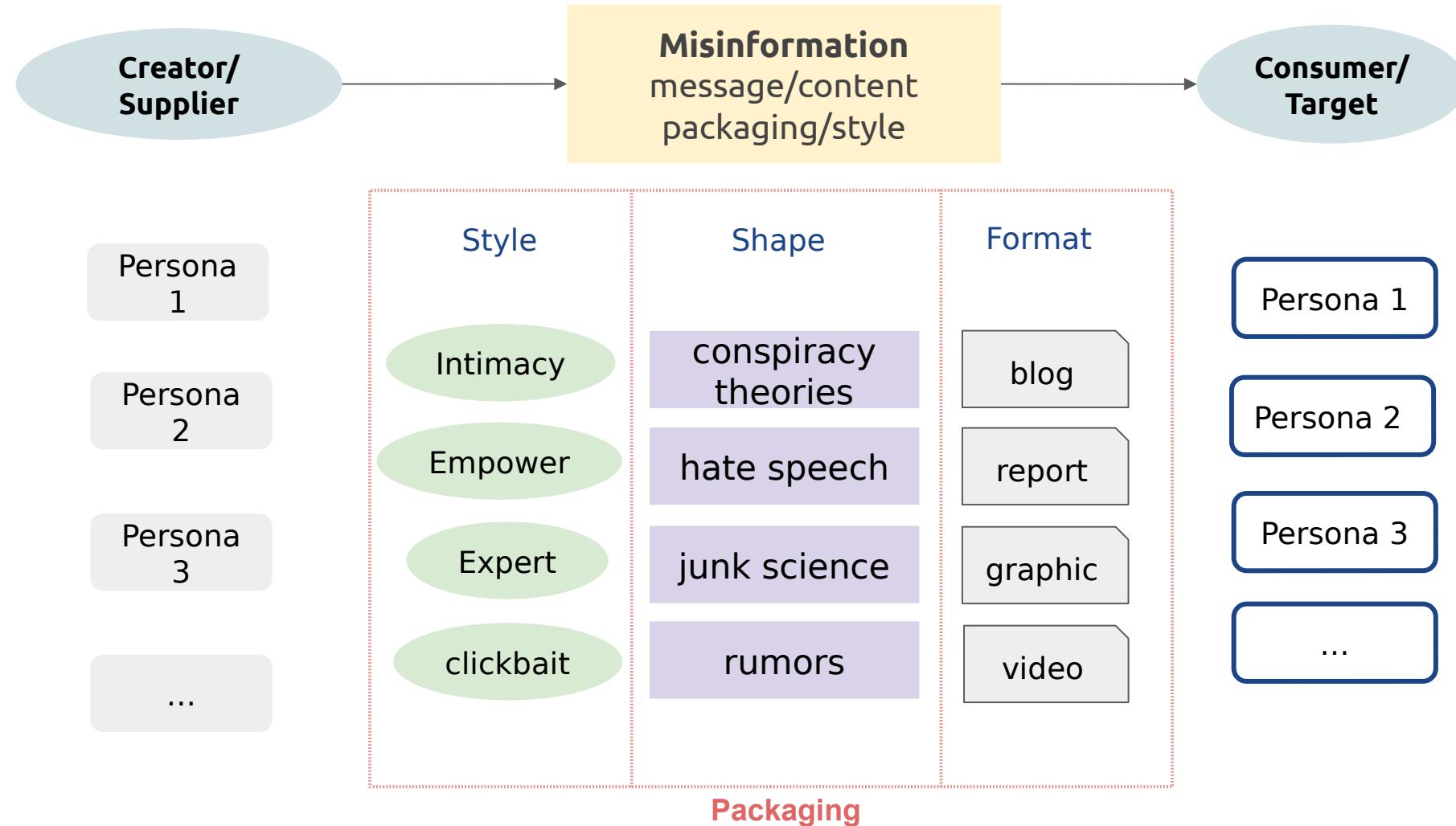
RQ3: Can we use Generative Language models (ChatGPT-like) to construct effective interventions “in the style of...”

RQ4: Can we test the interventions in simulated environment before deploying them?



Prof Hany Farid
UC Berkeley

Mitigate: misinformation as a content consumption problem



Mitigate: personas prefer specific writing styles



Jennifer Matthews

age: 25-40
residence: Australia
education: BA
occupation: Formerly a communications specialist for a non-profit, but retired after the birth of her first child
marital status: Married with two children (a newborn and a 2 year old)



"A mother's work is never done"

Jennifer wakes up at 6am each day, often after a broken night's sleep due to her newborn still waking up. She prepares breakfast for her husband and does some yoga, and misses the morning class she used to attend. Jennifer's peace ends when the children wake up, and she spends most of the day juggling their needs. If she can successfully get them to have a nap, she checks Facebook and Instagram to try and keep up with her friend's lives and stay in touch. She might also browse Pinterest, saving pictures to boards about yoga, healthy and quick meals, motherhood advice, and parenting tips. Twice a week, she has an afternoon walk with the kids to the nearby supermarket for groceries. In the evening, she makes dinner for her husband. He enthusiastically talks about work, but she feels like she has very little to say about her life; and it's not very interesting to him.

Intimacy + Junk Science

Barrier to gov comms: formal, masculine style

Mitigate: personas prefer specific writing styles



Jennifer Matthews

age: 25-40

residence: Australia

education: BA

occupation: Formerly a communications specialist for a non-profit, but retired after the birth of her first child

marital status: Married with two children (a newborn and a 2 year old)



"A mother's work is never done"

Jennifer wakes up at 6am each day, often after a broken night's sleep due to her newborn still waking up. She prepares breakfast for her husband and does some yoga, and misses the morning class she used to attend. Jennifer's peace ends when the children wake up, and she spends most of the day juggling their needs. If she can successfully get them to have a nap, she checks Facebook and Instagram to try and keep up with her friend's lives and stay in touch. She might also browse Pinterest, saving pictures to boards about yoga, healthy and quick meals, motherhood advice, and parenting tips. Twice a week, she has an afternoon walk with the kids to the nearby supermarket for groceries. In the evening, she makes dinner for her husband. He enthusiastically talks about work, but she feels like she has very little to say about her life and it's not very interesting to him.

Intimacy + Junk Science

Barrier to gov comms: formal, masculine style

Aaron Turner

age: 21

residence: Australia

education: Law student (final year)

occupation: Summer law clerk

marital status: Single



"I object!"

Aaron wakes at around 8am and makes a mad rush for either university or his summer job, skipping breakfast. He is enthusiastic about both and works hard at each. By lunch time, he is feeling exhausted and deserving of a "treat", often buying a burger or pizza meal--and after all, he skipped breakfast, which makes up for the extra calories. He works intently through the afternoon, and leaves work feeling exhausted again. He often tells himself that he's too tired to hit the gym now, but he'll go tomorrow. On Fridays, he spends the night with his Church youth group, but it's increasingly bittersweet, as he will have to leave it after his upcoming birthday. On nights when he doesn't have Church, he often winds up having a few beers and scrolling through Reddit, where he can vent his frustration about still being single and find other interesting blogs to explore.

Empower + Hate Speech

Barrier to gov comms: authoritative tone

Mitigate: personas prefer specific writing styles



Australian Government

Jennifer Matthews

age: 25-40

residence: Australia

education: BA

occupation: Formerly a communications specialist for a non-profit, but retired after the birth of her first child

marital status: Married with two children (a newborn and a 2 year old)



"A mother's work is never done"

Jennifer wakes up at 6am each day, often after a broken night's sleep due to her newborn still waking up. She prepares breakfast for her husband and does some yoga, and misses the morning class she used to attend. Jennifer's peace ends when the children wake up, and she spends most of the day juggling their needs. If she can successfully get them to have a nap, she checks Facebook and Instagram to try and keep up with her friend's lives and stay in touch. She might also browse Pinterest, saving pictures to boards about yoga, healthy and quick meals, motherhood advice, and parenting tips. Twice a week, she has an afternoon walk with the kids to the nearby supermarket for groceries. In the evening, she makes dinner for her husband. He enthusiastically talks about work, but she feels like she has very little to say about her life and it's not very interesting to him.

Intimacy + Junk Science

Barrier to gov comms: formal, masculine style

Aaron Turner

age: 21

residence: Australia

education: Law student (final year)

occupation: Summer law clerk

marital status: Single



"I object!"

Aaron wakes at around 8am and makes a mad rush for either university or his summer job, skipping breakfast. He is enthusiastic about both and works hard at each. By lunch time, he is feeling exhausted and deserving of a "treat", often buying a burger or pizza meal--and after all, he skipped breakfast, which makes up for the extra calories. He works intently through the afternoon, and leaves work feeling exhausted again. He often tells himself that he's too tired to hit the gym now, but he'll go tomorrow. On Fridays, he spends the night with his Church youth group, but it's increasingly bittersweet, as he will have to leave it after his upcoming birthday. On nights when he doesn't have Church, he often winds up having a few beers and scrolling through Reddit, where he can vent his frustration about still being single and find other interesting blogs to explore.

Patrick Bear

age: 67

residence: Suburban NSW

education: Bachelor of Education

occupation: Retired primary school teacher

marital status: Married for 45 years, 3 children, 4 grandchildren



"Kids these days"

Patrick wakes early and takes a walk around the neighbourhood, before helping his wife up and organising breakfast for her. She's become quite frail, and he has to help her get dressed and walk. After, they often sit outside in the garden and he reads to her. He organises lunch and they both have a short rest, before the grandkids come over after school until their parents finish work. These frantic few hours are their favourite part of the day, even if it gives them a bit of a headache. He helps the kids with homework, and then takes care of chores as his wife supervises them watching TV or spending time in the garden. When the grandkids have been retrieved by their parents he organises dinner, and as his wife watches late night TV, he logs on to the Facebook his daughter signed him up for. He tries to keep up with the posts there, and also finds interesting news items about issues he never hears about on TV. His wife thinks the internet is silly, but he's excited about the new things he can learn there. As he gets ready for bed, he often finds himself still thinking about the things he's read.

Expert + Conspiracy

Barrier to gov comms: belief in own expertise, difficulty in understanding long complex text

Empower + Hate Speech

Barrier to gov comms: authoritative tone

Mitigate: run “packaged” ads campaigns



Instruction set:

Write a 100 word social media post that would appeal to Jennifer, accounting for:

- the post comes from a fellow mother who understands Jennifer's fears
- she reassures her that vaccines are safe for her children.
- she vaccinated her children against smallpox the other day, and 2 weeks later all the children in the neighborhood got it, except hers.
- write it as if it is send on a mum's Facebook group.

A synthetically-generated social media post , campaigning for vaccines targeting the “Jennifer” persona.

Mitigate: run “packaged” ads campaigns



Generated post:

Hey mamas,

I just wanted to share something with you all that happened recently. I vaccinated my kids against smallpox a couple of weeks ago and, unfortunately, all the kids in our neighborhood got it except mine. It's a scary thing when there's a new vaccine out there, but I just wanted to let you all know that vaccines are safe and effective. They've been tested a ton, and they work to protect our kids from some really dangerous diseases. So if you're still on the fence about vaccinating your kiddos, I'm here to tell you - do it!

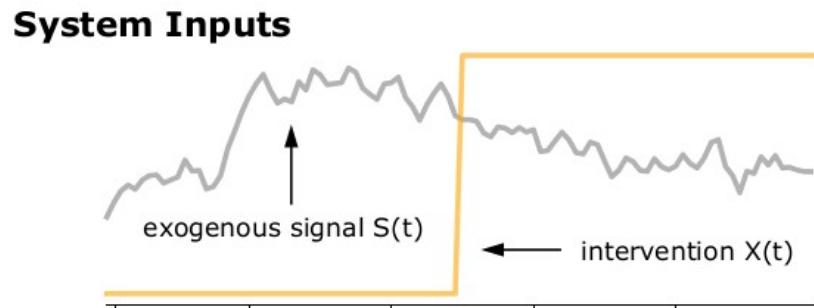
Let's keep our babies healthy and safe together!

#MomSquad #VaccineSafety #SmallpoxPrevention
#HealthyKids.

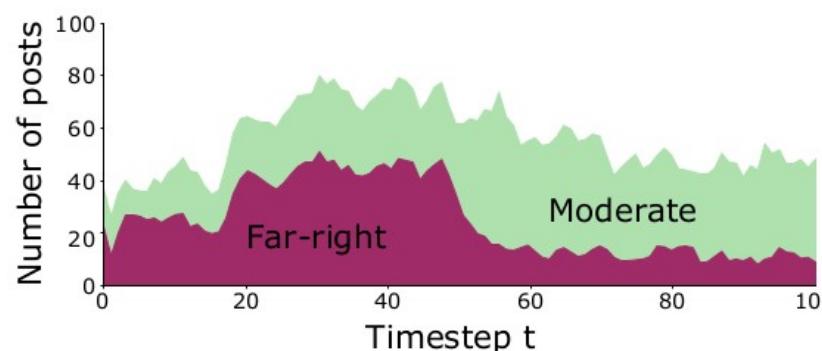
Love,
One of your fellow moms.

A synthetically-generated social media post , campaigning for vaccines targeting the “Jennifer” persona.

Mitigate: synthetic testing of interventions

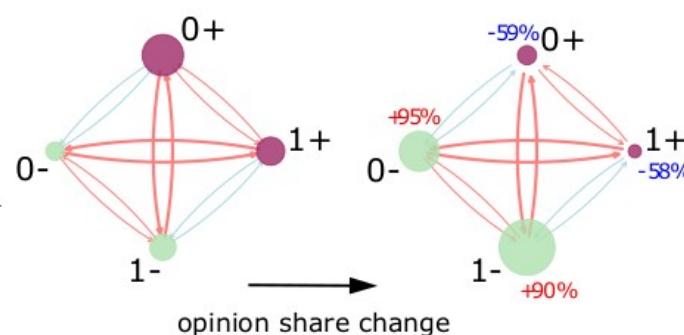


Tier 1
(attention volumes)

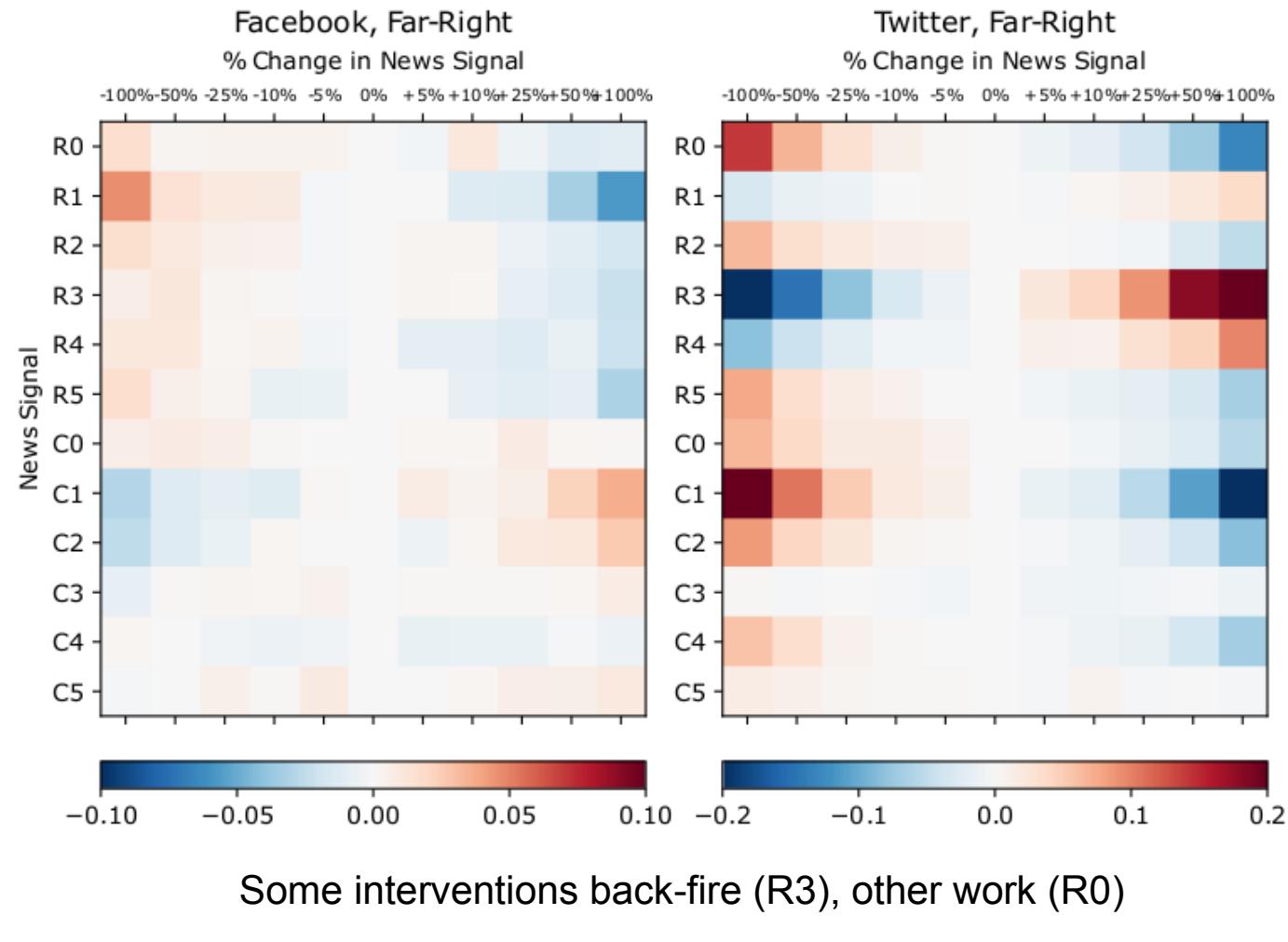
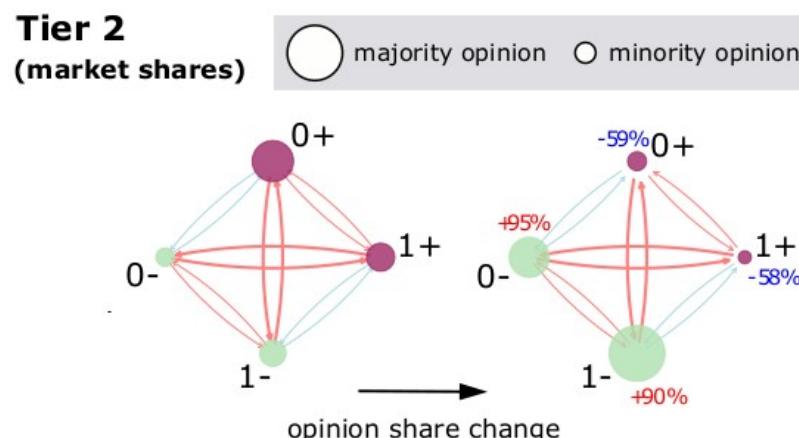
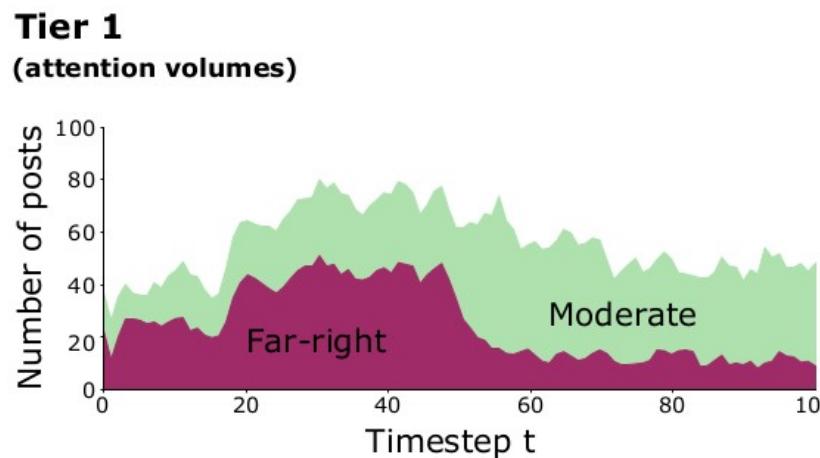
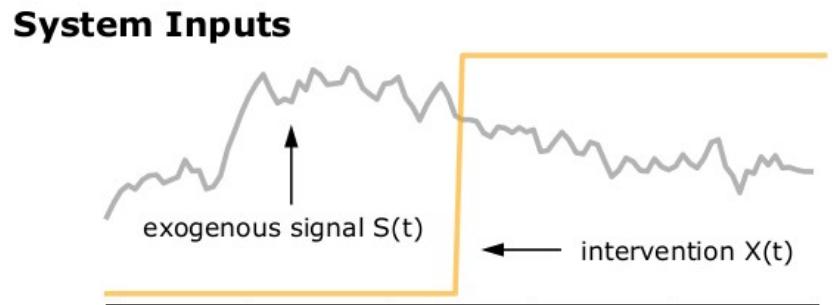


Tier 2
(market shares)

○ majority opinion ○ minority opinion



Mitigate: synthetic testing of interventions



Action plan in Countering Foreign interference and state backed IO



Monitor

Response
level



Objective



Approach

Detect

How do we most effectively identify and triage information campaigns based on the characteristics of the message, how it spreads, who is communicating it, and where it is being communicated?

Predict

What factors accelerate and intensify the communication and reach of weaponized messages within and across online environments, and which factors lead to the most significant real-world harms?

Mitigate

What are practical approaches that allow us to both proactively and reactively limit the harms of problematic messaging, including identifying where, when and how counter-messaging should be deployed?

Monitor discussions on social and traditional media

Detect adversarial information campaigns

Estimate the effectiveness of influence operations

Design and apply countermeasures

Characterising the dynamic interaction between traditional and social media ecosystems in the flow and spread of disinformation and problematic content.

Develop and deploy a "mission control" dashboard to retrieve content from a constantly updating list of traditional media and Internet sources.

Utilise information diffusion techniques to identify problematic content based on the way it moves through and across online channels

Deploy natural language processing techniques to automate the detection of problematic online messages based on the structure and content of the message

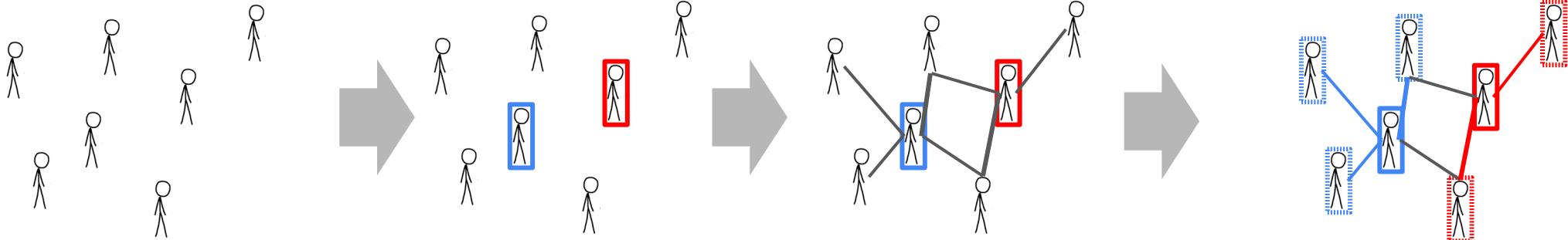
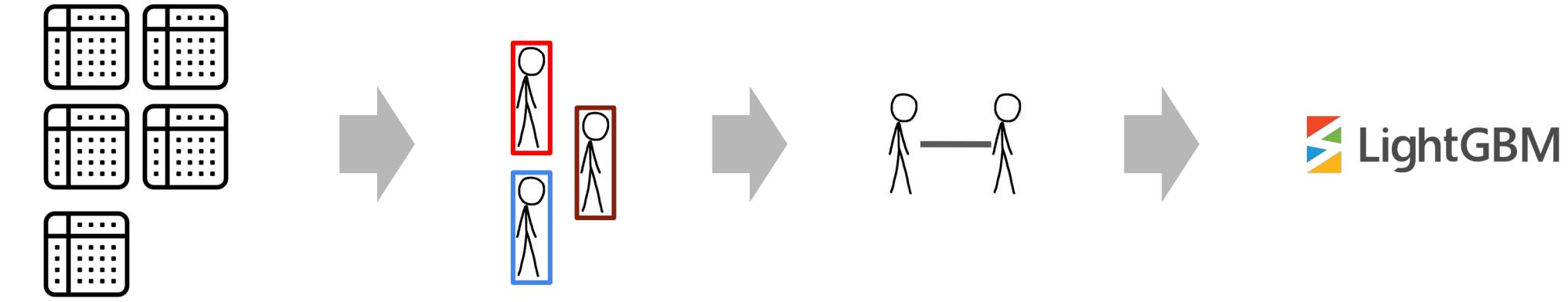
Model the impact of networks and influencers on the virality and reach of problematic messages

Track the spread of problematic messages across and between online platforms and into the real-world

Use natural language processing to automatically generate counter-messaging that is tuned for the platform and target group of interest

Identify key message inoculation points in social networks based on how information flows and gains velocity

Detect: far-right and extremist ideology detection



No data labelling required for new datasets

Seed users: use fringe news sharing to label a small subset of users

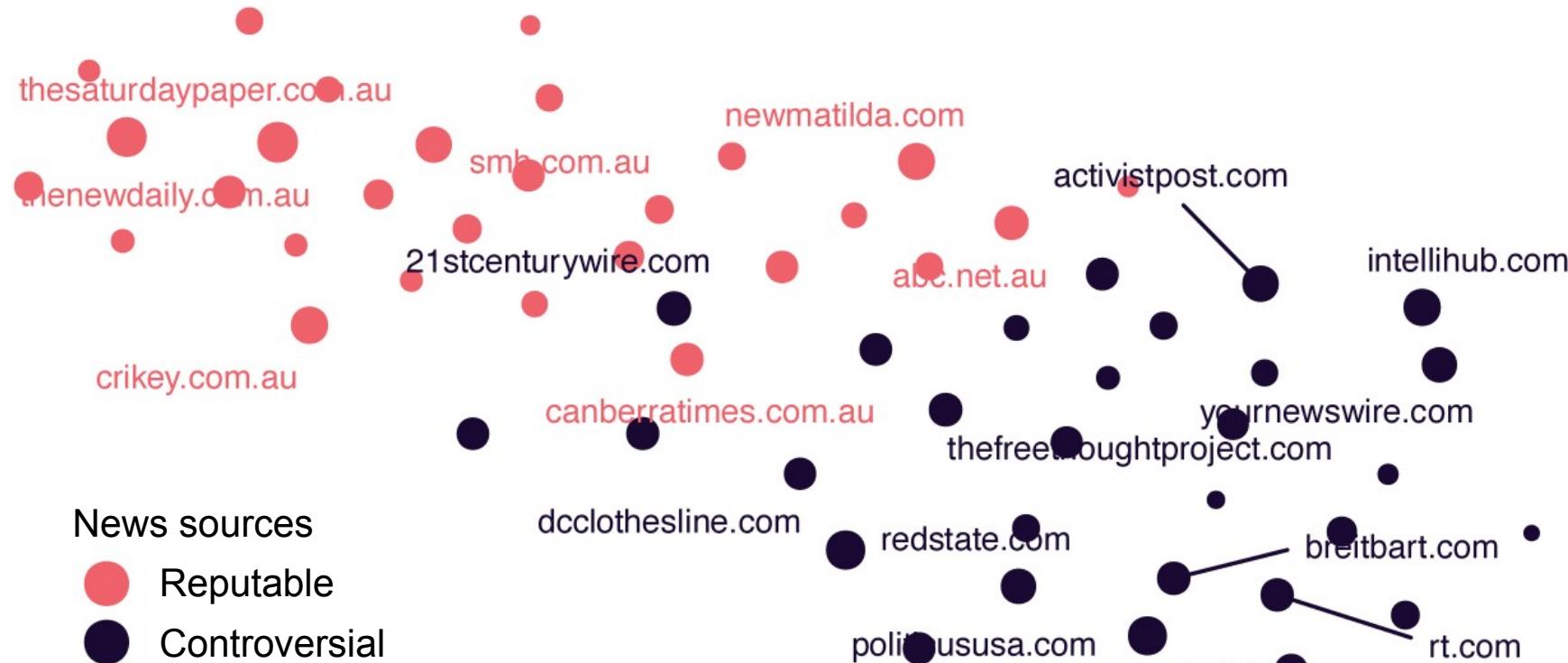
Text-based homophily quantification – “what you write is who you are”

Use advanced Machine Learning to label all users

The technical detail:

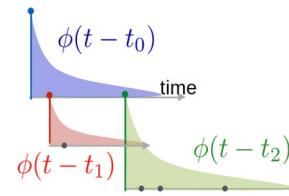
Ideology proxies; homophily lenses (text, follower, URLs); automatic user labelling

Detect: separating controversial from reputable



Reputable and controversial sources are separable based solely on how their information spreads

Detect controversial news without content analysis



evently

https://www.behavioral-ds.science/theme1_content/evently/



The technical detail:

Mathematical generative modelling; Hawkes processes; joint modelling

React: Identify influential inauthentic users (bots)



Identify users engaged in influence operations

Estimate their impact on the wider community



birdspotter

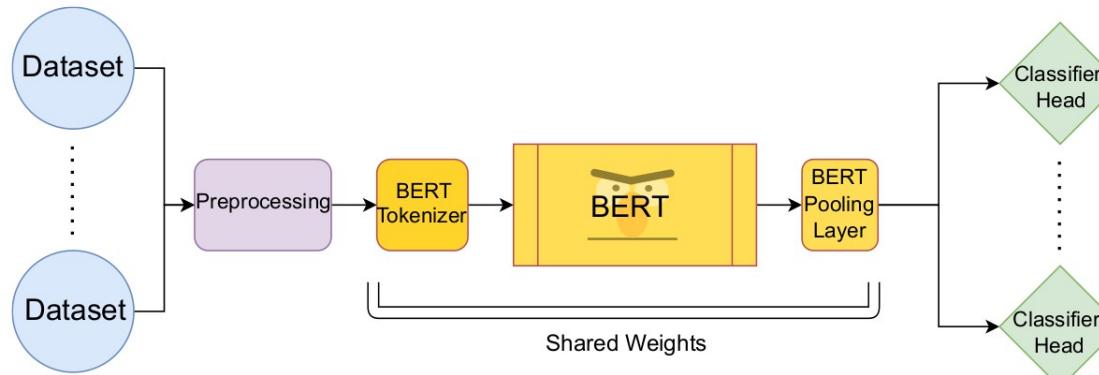
https://www.behavioral-ds.science/theme2_content/birdspotter/



The technical detail:

Influence estimation using stochastic modelling; content-free analysis

React: Detecting Hate Speech in Unseen Domains



Novel learning paradigm to leverage many disparate datasets to learn a single hate speech representation

Improved performances over the state-of-the-art, generalizable to novel datasets.

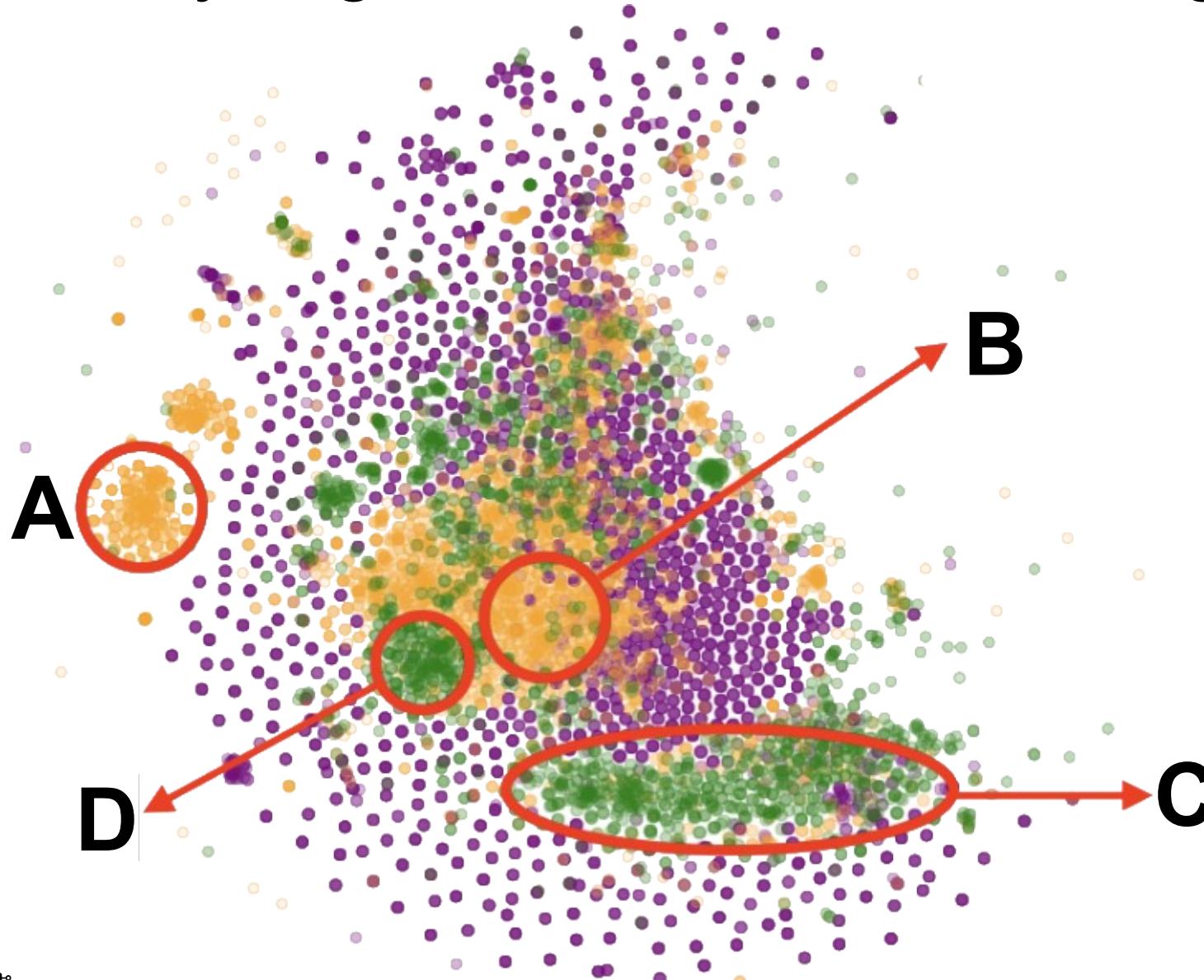
Model		Testing Dataset									# Wins
		DAVIDSON	WASEEM	REDDIT	GAB	FOX	STORM- FRONT	MANDL	HATEVAL	PUBFIGS-L	
MTL	MTL-NCH	0.6822	0.3801	0.8456	0.8738	0.6150	0.6826	0.5312	0.6449	0.6175	6
	MTL-MV	0.6455	0.4048	0.8263	0.8660	0.6030	0.6771	0.4834	0.6315	0.6231	1
BERT baseline trained on:	DAVIDSON		0.5556	0.5914	0.6731	0.4932	0.4597	0.5690	0.5414	0.5469	0
	WASEEM	0.6136		0.6000	0.6427	0.5519	0.5356	0.5099	0.5784	0.5611	0
	REDDIT	0.6135	0.4957		0.8083	0.5229	0.5559	0.4900	0.5741	0.5402	0
	GAB	0.5720	0.4595	0.8375		0.5075	0.5645	0.4277	0.5664	0.5185	0
	FOX	0.4285	0.4249	0.4234	0.4651		0.3865	0.4159	0.4490	0.3926	0
	STORMFRONT	0.4533	0.5467	0.5822	0.6487	0.5740		0.5104	0.5664	0.5659	0
	MANDL	0.3336	0.4822	0.4066	0.4582	0.4010	0.3518		0.4546	0.3633	0
	HATEVAL	0.5849	0.5824	0.5700	0.5796	0.5532	0.5466	0.5348		0.5432	0
	PUBFIGS-L	0.6351	0.6048	0.5970	0.6600	0.5546	0.5249	0.5963	0.5858		2



The technical detail:

Transfer learning; language models fine-tuning;

Analysing coordinated troll strategies



(yellow) right trolls: focused MAGA
(magenta) left trolls: surround discussion
(green) news trolls: selective highlighting



The technical detail:

Semantic edit distance; dimensionality reduction; Twitter trolls