





"Fake" news, disinformation, online abuse, and bots

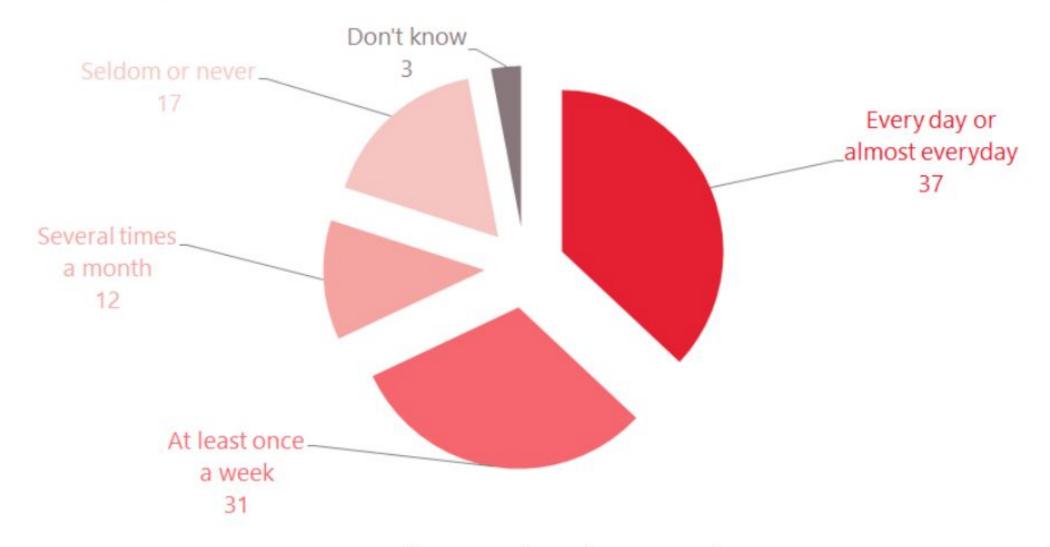
Kalina Bontcheva
The University of Sheffield



Is Online Misinformation a Big Problem for Citizens?

Q2 How often do you come across news or information that you believe misrepresent reality or is even false?

(% - EU)



Base: All Respondents (N=26,576)



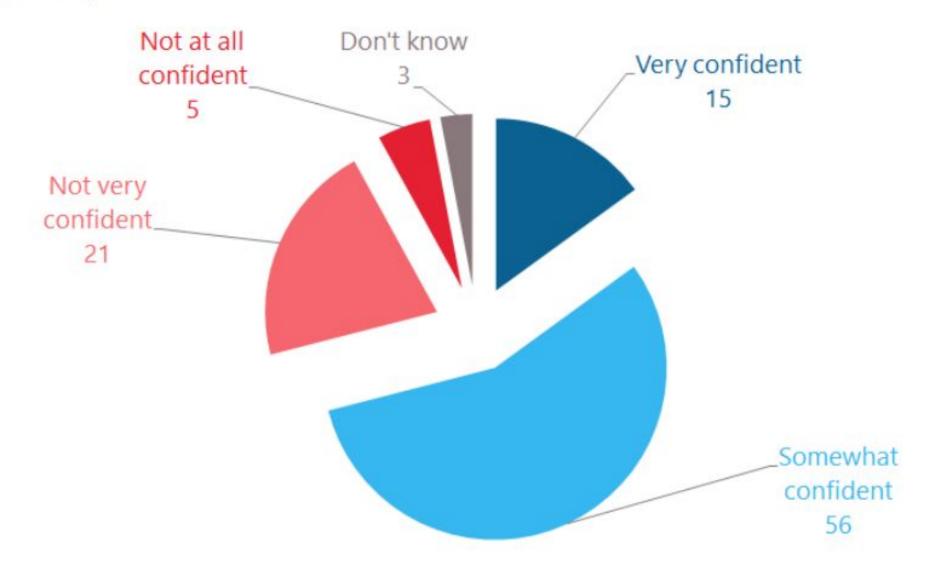


Source: Eurobarometer 2018

Do citizens know how to spot it?

Q3 How confident or not are you that you are able to identify news or information that misrepresent reality or is even false?

(% - EU)



Base: All Respondents (N=26,576)

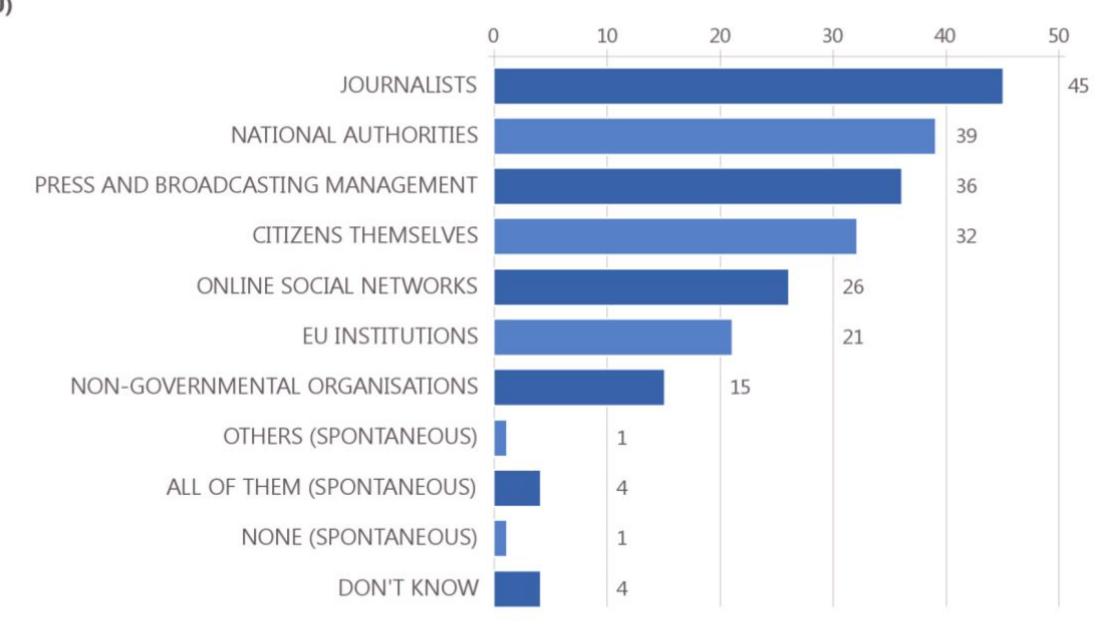




Who should act to prevent it?

News or information that misrepresent reality or that are even false are called "fake news". Which of the following institutions and media actors should act to stop the spread of "fake news"? (MAX. 3 ANSWERS)

(% - EU)



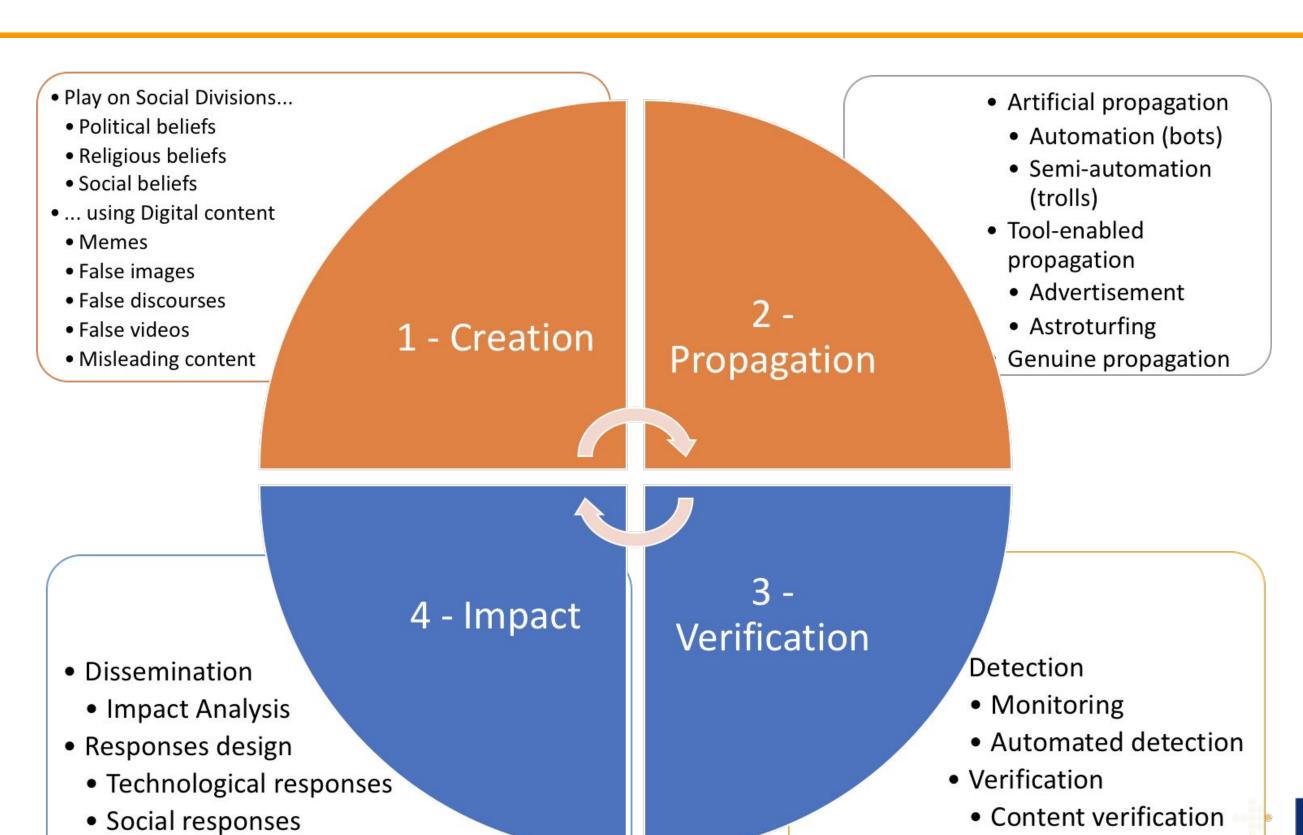
Base: All Respondents (N=26,576)





The Disinformation Lifecycle

Legal responses



Source: STOA report 2019

European Commission Information Society and Media

Source verification

The 6 Questions of Disinformation Analysis

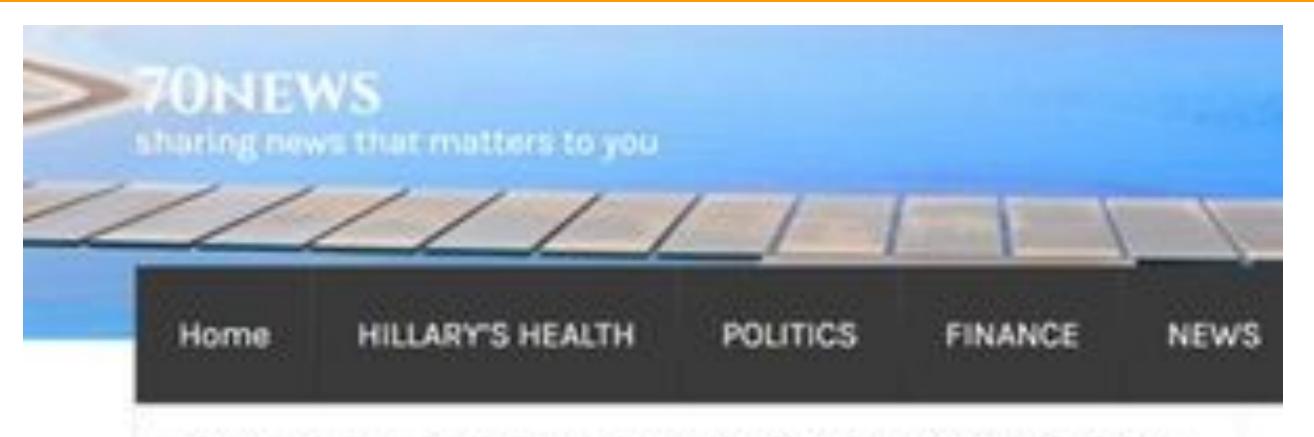
- What is being spread?
- Who is spreading it?
- When it spreads?
- Where it spreads?
- Why it spreads?
- How it spreads?



The 6 Questions of Disinformation Analysis

What is being spread?

Online Falsehoods - Examples



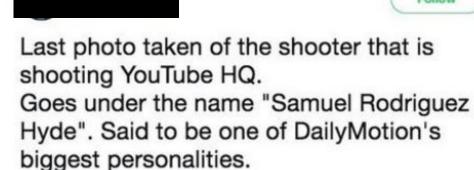
FINAL ELECTION 2016 NUMBERS: TRUMP WON BOTH POPULAR (62.9 M -62.2 M) AND ELECTORAL COLLEGE VOTES (306-232)...HEY CHANGE.ORG, SCRAP YOUR LOONY PETITION NOW!

POSTED ON NOVEMBER 12, 2016 LIPOATED ON NOVEMBER 13, 2016





Online Falsehoods - Examples



Motivations unclear.



3:15 PM - 3 Apr 2018

12 Retweets 24 Likes

NOTORIOUS MASS MURDERER AND WHITE SUPREMACIST, SAM HYDE CONFIRMED AS SHOOTER AT YOUTUBE HQ #YoutubeHQShooting #YouTube #Shooting #MarchForOurLives



3:13 PM - 3 Apr 2018









Online Falsehoods - Examples



The EU has failed us all





Don't you mean "Fake News"?



Imposter sites



Parody



Boeing CEO Admits Company Made Mistake By Including Automatic Self-Destruct Function On All 737 Max Planes

CHICAGO—Acknowledging that certain practical considerations had been overlooked in production, Boeing president and CEO Dennis Muilenburg admitted at a press conference Monday that the company had made a mistake by including an automatic self-destruct function on all 737 Max airplanes. "At the time, we thought that having a simple, one-step option for destroying...

Continue Reading







Definitions

- Information disorder theoretical framework (Wardle, 2017; Wardle & Derakshan, 2017)
- Three types of false / harmful information:
 - Mis-information: false information that is shared inadvertently, without meaning to cause harm.
 - Dis-information: intending to cause harm, by deliberately sharing false information.
 - Mal-information: genuine information or opinion shared to cause harm, e.g. hate speech, harassment.





TYPES OF INFORMATION DISORDER

FALSENESS

INTENT TO HARM

Misinformation

Unintentional mistakes such as innaccurate photo captions, dates, statistics, translations, or when satire is taken seriously.

Disinformation

Fabricated or deliberately manipulated audio/visual content. Intentionally created conspiracy theories or rumours.

Malinformation

Deliberate publication of private information for personal or corporate rather than public interest, such as revenge porn.

Deliberate change of context, date or time of genuine content.

Definitions (2)

 Currently the most widely agreed upon definition comes from the High Level Expert Group report:

"Disinformation....includes all forms of false, inaccurate, or misleading information designed, presented, and promoted to <u>intentionally</u> cause public harm or profit." (Buning et al, 2018).

Definitions (3)

- Often hard to distinguish mis- from dis-information
 - Intention of the information source or amplifier may not be easily discernible
 - Hard not only for algorithms, but also for human readers (Jack, 2017; Zubiaga et al, 2016)
- Mis- and dis-information are sometimes addressed as if they are interchangeable



Categories of Information Disorder

FIRSTDRAFT

7 COMMON FORMS OF INFORMATION DISORDER



No intention to cause harm but has potential to fool



Misleading use of information to frame an issue or individual



When genuine sources are impersonated



New content is 100% false, designed to deceive and do harm



FALSE CONNECTION

When headlines, visuals or captions don't support the content



FALSE CONTEXT

When genuine content is shared with false contextual information



MANIPULATED CONTENT

When genuine information or imagery is manipulated to deceive

7 Categories of Information Disorder. Credit: Claire Wardle, 2017.

Examples

Parody



Boeing CEO Admits Company Made Mistake By Including Automatic Self-Destruct Function On All 737 Max Planes

CHICAGO—Acknowledging that certain practical considerations had been overlooked in production, Boeing president and CEO Dennis Mullenburg admitted at a press conference Monday that the company had made a mistake by including an automatic self-destruct function on all 737 Max airplanes. "At the time, we thought that having a simple, one-step option for destroying..."

Misleading Content



 Decontextualised images/videos





3:13 PM - 3 Apr 20

17



0

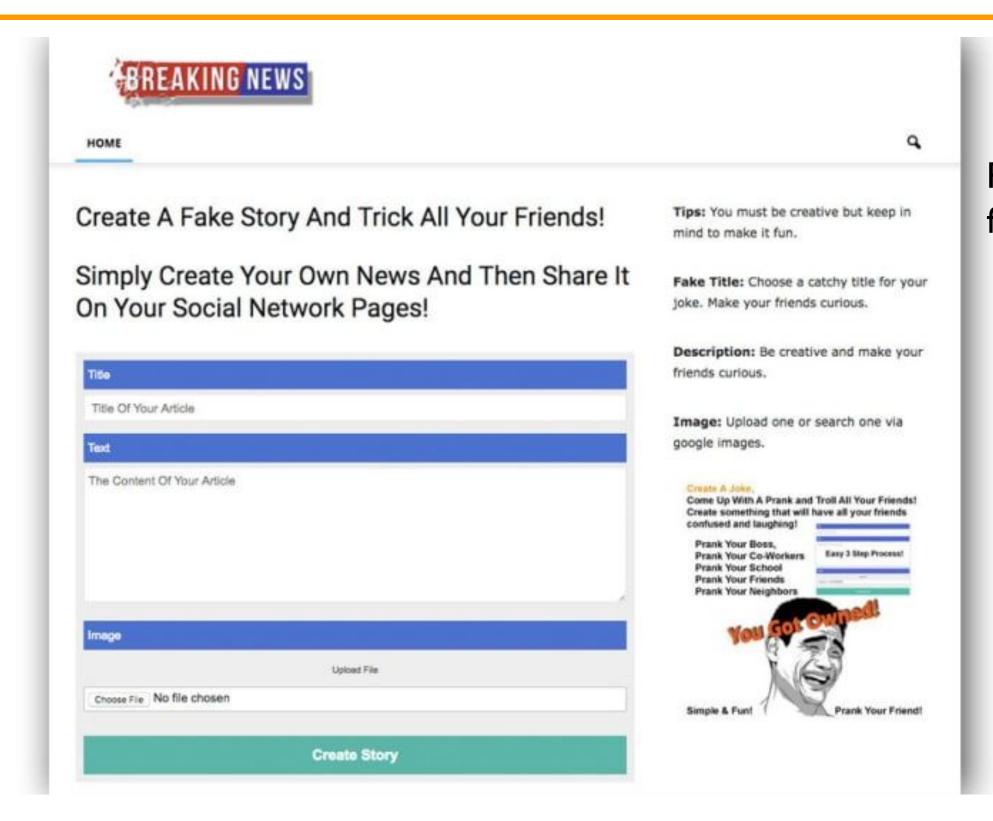


The 6 Questions of Disinformation Analysis

- What is being spread?
- Who is spreading it?
- When it spreads?
- Where it spreads?
- Why it spreads?
- How it spreads?



Authoring False Stories is Easy

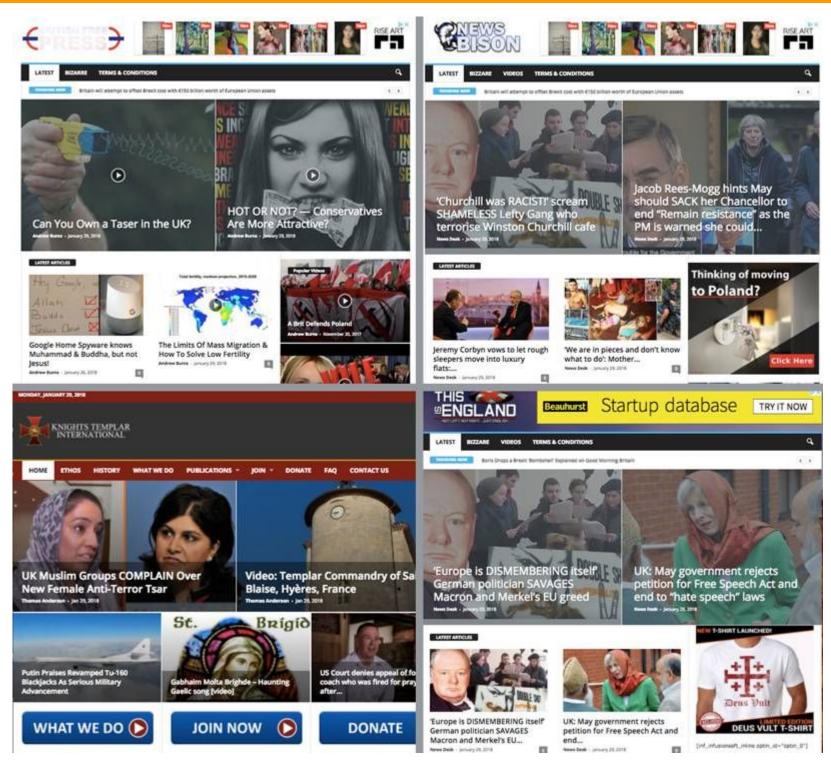


Play on emotions like fear and anger



Source: (Silverman, Lytvynenko & Pham, 2017)

Low credibility web site networks



- A network of websites that post disinformation or distorted, out-of-context news stories
- This example: a network of seemingly UK-based far-right news sites (operated from Eastern Europe)
- Shared and amplified through thirteen related Facebook pages
- 2.4 million likes (more than any UK Facebook political page)

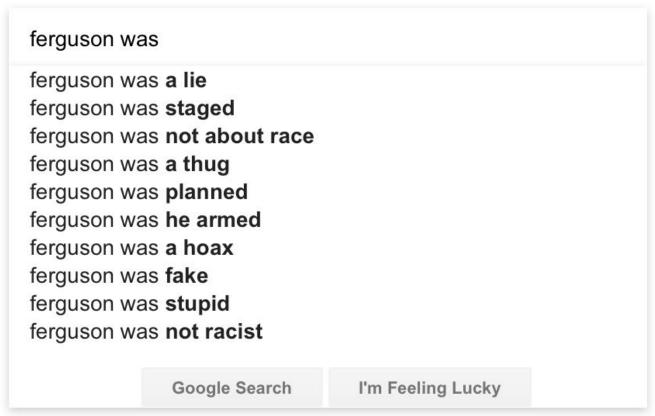




Source: (Reynolds, 2018)

Misinformation in search results



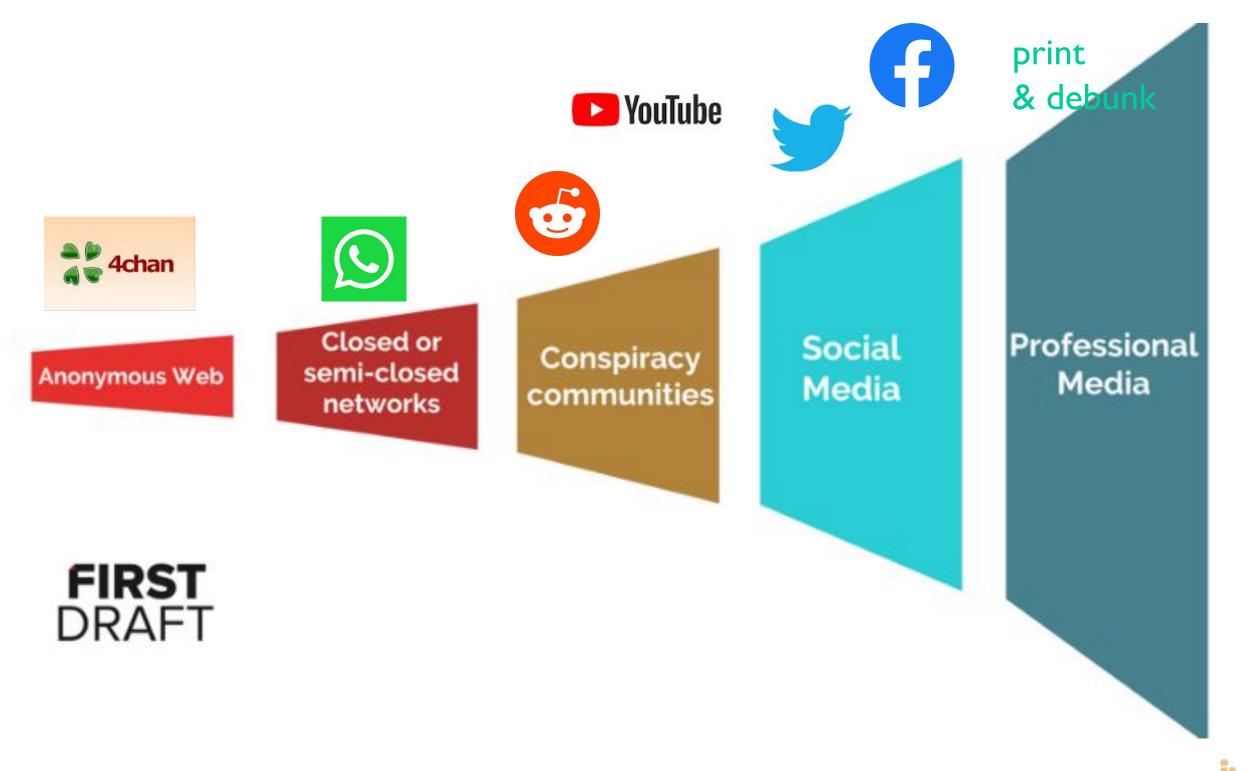


Report inappropriate predictions



Source: (Albright, 2018)

The Trumpet of Amplification





Closed networks



- WhatsApp groups are used to spread misinformation in many countries
- No API and messages are encrypted
- This example: Video produced in Pakistan (public service announcement), spread in India with false claims of child kidnapping
- Resulted in mob murders of innocent people
- Some steps from WhatsApp being implemented





The 6 Questions of Disinformation Analysis

- What is being spread?
- Who is spreading it?
- When it spreads?
- Where it spreads?
- Why it spreads?
- How it spreads?





Agent, Message, Interpreter (AMI) model

Actor Type: Official / Unofficial Level of Organisation: None / Loose / Tight / Networked Type of Motivation: Financial / Political / Social / Psychological Agent Level of Automation: Human / Cyborg / Bot Intended Audience: Members / Social Groups / Entire Societies Intent to Harm: Yes / No Yes / No Intent to Mislead: Long term / Short-term / Event-based Duration: Misleading/Manipulated/Fabricated Accuracy: Legal / Illegal Legality: No / Brand / Individual Message Imposter Type: Individual / Organisation / Social Group / Entire Society Message Target: Hegemonic / Oppositional / Negotiated Message reading: Interpreter Ignored / Shared in support / Shared in opposition Action taken:

Fake Profiles

- A Macedonian man created and ran in a coordinated fashion over 700 Facebook profiles, spreading online disinformation for monetary gain
- IRA operated cyborg accounts spread misinformation during the 2016 US elections
 - Jenna Abrams widely considered a real American; posted on diverse topics
- Fake accounts can be purchased to inflate follower counts



Fake Accounts (2)

- Fake accounts try to gain credibility by following genuine accounts
 - the USA Today Facebook page lost around 9 million followers when the platform detected and suspended a large coordinated network of fake accounts (Silverman, 2017)
 - Politician's Twitter accounts are another example, with a recent study estimating as many as 60% of Donald Trump's followers being suspected fake accounts (Campoy, 2018), 43.8% - for Hillary Clinton, and 40.8% for Barack Obama.



Advertising and Clickbait

- Online advertising used extensively to make money from junk news sites
 - Payments when the adverts are shown alongside the false content
 - Often employ clickbait to attract users
- A clickbait post is designed to provoke emotional response in its readers, e.g. anger, compassion, sadness, and thus stimulate further engagement



Misinformation in Political Advertising Online

- Russia Today & related acc promoted just under 2,000 election-related tweets
 - generated around 53.5 million impression on U.S. based users (Edgett, 2017)
- Issues around digital campaigning by political organisations during the 2016 UK EU membership referendum (DCMS report, 2018)
- Questions also around the Trump presidential campaign use of over 5.9 million Facebook adverts (Frier, 2018)



Micro-targetted or dark ads

- online adverts that are visible only to the users that are being targeted
 - e.g. voters in a marginal UK constituency (Cadwalladr, 2017a)
- Do not appear on the advertiser's timeline or in the feeds of the advertiser's followers
- Micro-targeting fine-grained ad targeting,
 based on job titles or demographic data
 - Cambridge Analytica (!)



Micro-targetted or dark ads

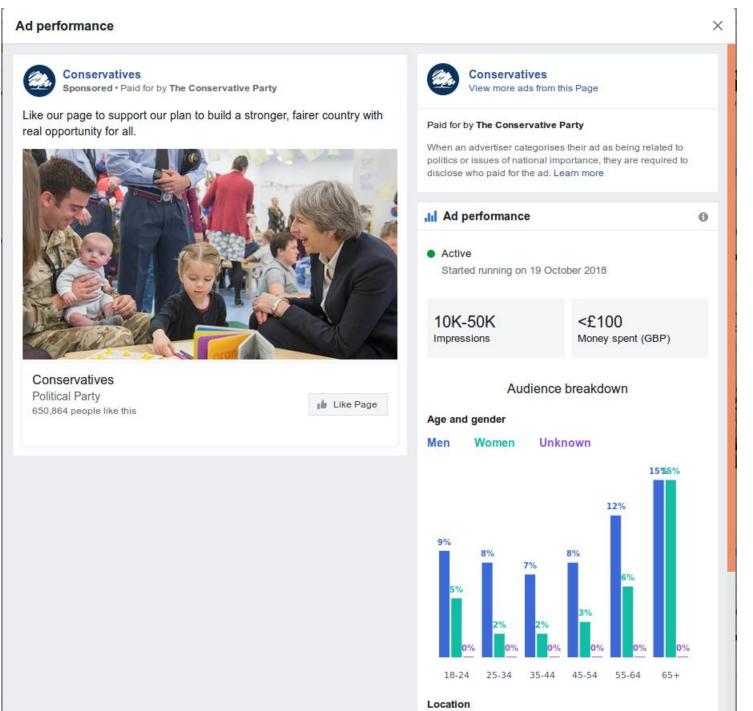
- Used to spread misinformation during election campaigns
 - A VoteLeave dark ad made public by Facebook as evidence to the UK DCMS parliamentary inquiry
- Effective personal data protection on social platforms priority for national governments and policy makers (<u>non-users also affected</u>)





Transparency of Online Advertising

- Google political ads database
 - https://transparencyrepor t.google.com/political-ads /library
- Twitter's Ad Transparency
 Center
 - https://ads.twitter.com/tran sparency
- Facebook Ad Archive
 - https://www.facebook.com/ ads/archive







Transparency of Online Advertising (2)

- Some of these lack automated APIs
 - manual analysis does not scale
- Calls for public open data repository (most notably FullFact)
- Independent mechanism for monitoring political advertising across all social platforms
- "There should be a ban on micro-targeted political advertising to lookalikes online, and a minimum limit for the number of voters sent individual political messages should be agreed, at a national level." (DCMS report, 2018)







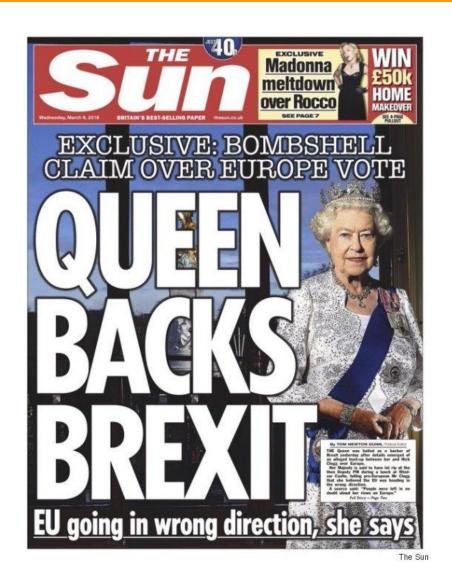
Impact of Inaccurate Claims by Politicians



- £350m false claim 10.2 times more tweets than the 3,200 tweets by the Russia-linked accounts suspended by Twitter
 - More than 1,500 tweets from different voters
 - I am with @Vote leave because we should stop sending £350 million per week to Brussels, and spend our money on our NHS instead.
 - I just voted to leave the EU by postal vote! Stop sending our tax money to Europe, spend it on the NHS instead! #VoteLeave #EUreferendum
 - Ipsos Mori (22/06/2016) for 9% the NHS was the most important issue in the campaign
 - Ipsos Mori over half of the UK population believed this claim to be correct



Mainstream Media



- Mainstream media also sometimes publish inaccurate, misleading, or distorted information
- Sometimes it is intentional
- Sometimes they are misled themselves
 - Pressure of 24 news cycle
- The Queen Backs Brexit example was then picked up and spread by the Russian troll accounts
- UK Independent Press Standards
 Organisation upheld the complaint



Euromyths

Aviation - 'Open Skies' treaty threatens right agains global warming, Feb 2006

```
В
```

```
Baby rattles - EC to ban noisy toys, Jan 1996
```

Bagpipes – EU safety directives to ban playgrounds, trapeze artists and bagpipes, Aug 2005

Bagpipes, noise levels - Bagpipes to be quietened by an edict from Brussels, Apr 2008

Bailout - EU demanding £6bn from Britain for a new bailout, Sep 2013

Bailout, emergency funding - UK called upon to provide more emergency funding, Jun 2016

Balloons – EU bans children from blowing up balloons, Oct 2011

Bananas - Curved or small bananas to be banned by Brussels, Sep 1994

Bananas - Curved bananas, Mar 1998

Bananas, funding - £3billion for bananas, Feb 2013

Bank accounts, ID requirements when opening – Children to be prevented from opening bank due to ID laws, Apr 1995

Barmaids - EU health directive to prevent barmaids from showing cleavage, Nov 2005

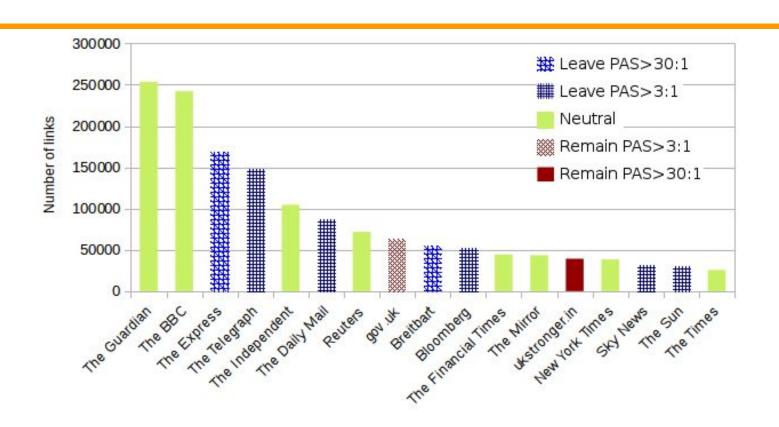
Barometers - Preassure rises on Britain's barometer makers, Mar 2006





Media Role in the Brexit Online Debate





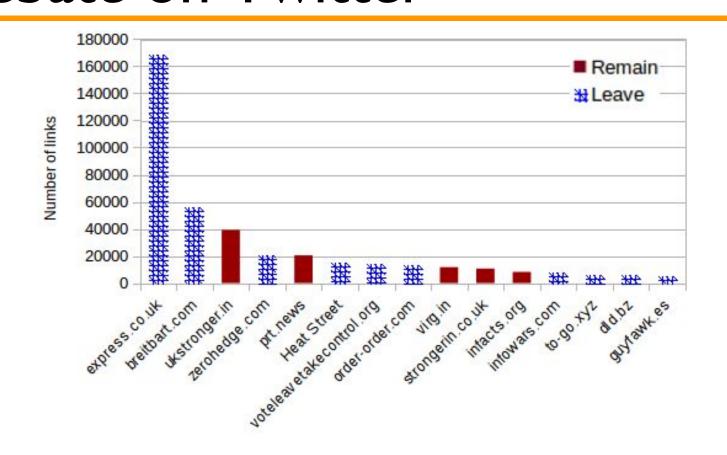
- We studied how many leavers and how many remainers linked to a domain
- The audience ratio between partisan groups has been called "Partisanship Attention Score" (PAS)
- Partisan media not as influential as in US elections





The Influential Partisan Sites in the Brexit debate on Twitter





- Express dominated (over 160,000 links)
- Breitbart second most linked to (almost 40,000 links)
- The remain partisan sites were campaign, not media sites

http://eprints.whiterose.ac.uk/136940/1/gorrell-influencers-brexit.pdf





Misinfodemics

 "We know that memes—whether about cute animals or health-related misinformation—spread like viruses: mutating, shifting, and adapting rapidly until one idea finds an optimal form and spreads quickly. What we have yet to develop are effective ways to identify, test, and vaccinate against these misinfo-memes. One of the great challenges ahead is identifying a memetic theory of disease that takes into account how digital virality and its surprising, unexpected spread can in turn have real-world public-health effects."

Source: (Gyenes & Mina, 2018)



Anti-vaccine Misinformation & Measles Outbreaks









False Amplifiers: Bots



Am J Public Health. 2018 October; 108(10): 1378-1384.

Published online 2018 October. doi: 10.2105/AJPH.2018.304567

PMCID: PMC6137759

PMID: 30138075

Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate

- Bots that spread malware and unsolicited content disseminated anti-vaccine messages
- Russian trolls promoted discord
- Accounts masquerading as legitimate users create false equivalency, eroding public consensus on vaccination





False Amplifiers: Definitions

Social bots

- programs "capable of automating tasks such as retweets, likes, and followers. They are used to disseminate disinformation on a massive scale, but also to launch cyber-attacks against media organizations and to intimidate and harass journalists." (RSF, 2018)
- A political bot is a social bot designed to promote political content.
- Sockpuppets/cyborgs are fake accounts that pretend to be human users; aim to connect and influence
- Trolls politically oriented sockpuppets/cyborgs



False Amplifiers: Impact

- Social bots, cyborgs, and trolls have all been employed as fake amplifiers in online misinformation and propaganda campaigns (Gorwa & Guilbeault, 2018)
- In Mexico, for instance, it is estimated that 18% of Twitter traffic is generated by bots (RSF, 2018)
- Flood the platform with manipulated content thus making high quality information hard to find
- Fake comments on government consultations or phantom signatures on online petitions



False Amplifiers: Fake Groups

- Astroturfing, i.e. creating artificial appearance of "grass-roots" support (recall the far-right British Facebook pages discussed above)
- Initially seeded with fake accounts, before drawing in genuine users.
- Other fake groups are created to "spread sensationalistic or heavily biased news or headlines, often distorting facts to fit a narrative" (Weedon, Nulan & Stamos, 2017).
- The credibility of fake Facebook groups can be enhanced through fake verification check marks (Silverman, 2017).



Genuine Amplifiers

- The main amplifiers behind viral misinformation and propaganda are genuine human users (Vosoughi et al, 2018)
- Confirmation bias
 - o reading news that conforms to the individual's political views
- Homophily
 - Individual's information sharing and commenting behaviour is influenced by the behaviour of their online social connections
- Confirmation bias and homophily lead to the creation of online echo chambers (Quattrociocchi et al. 2016)
- Polarisation
 - "social networks and search engines are associated with an increase in the mean ideological distance between individuals" (Flaxman et al, 2018)





Genuine Amplifiers (2)

- The main amplifiers behind viral misinformation and propaganda are genuine human users (Vosoughi et al, 2018)
- Confirmation bias
 - reading news that conforms to the individual's political views
- Homophily
 - Individual's information sharing and commenting behaviour is influenced by the behaviour of their online social connections
- Confirmation bias and homophily lead to the creation of online echo chambers (Quattrociocchi et al. 2016)
- Polarisation
 - "social networks and search engines are associated with an increase in the mean ideological distance between individuals" (Flaxman et al, 2018)
- Polarised communities believe and share misinformation which conforms to their preferred narratives (Quattrociocchi et al. 2016)

Genuine Amplifiers (3)

- What is the best strategy to reduce people's susceptibility to misinformation and the likelihood of its amplification?
 - Strategic communications researchers (Pamment et al, 2018) recommend presenting corrective information in ways that consider how and why the false story seemed credible.
 - What are the audience's dispositions?
 - Who do/don't they trust?
 - What aspects of the truth are they least/most likely to resist?
 - Question the frame, not just the content.
- Encourage debate and critical reflection (Harford, 2018;
 Pamment et al, 2018)



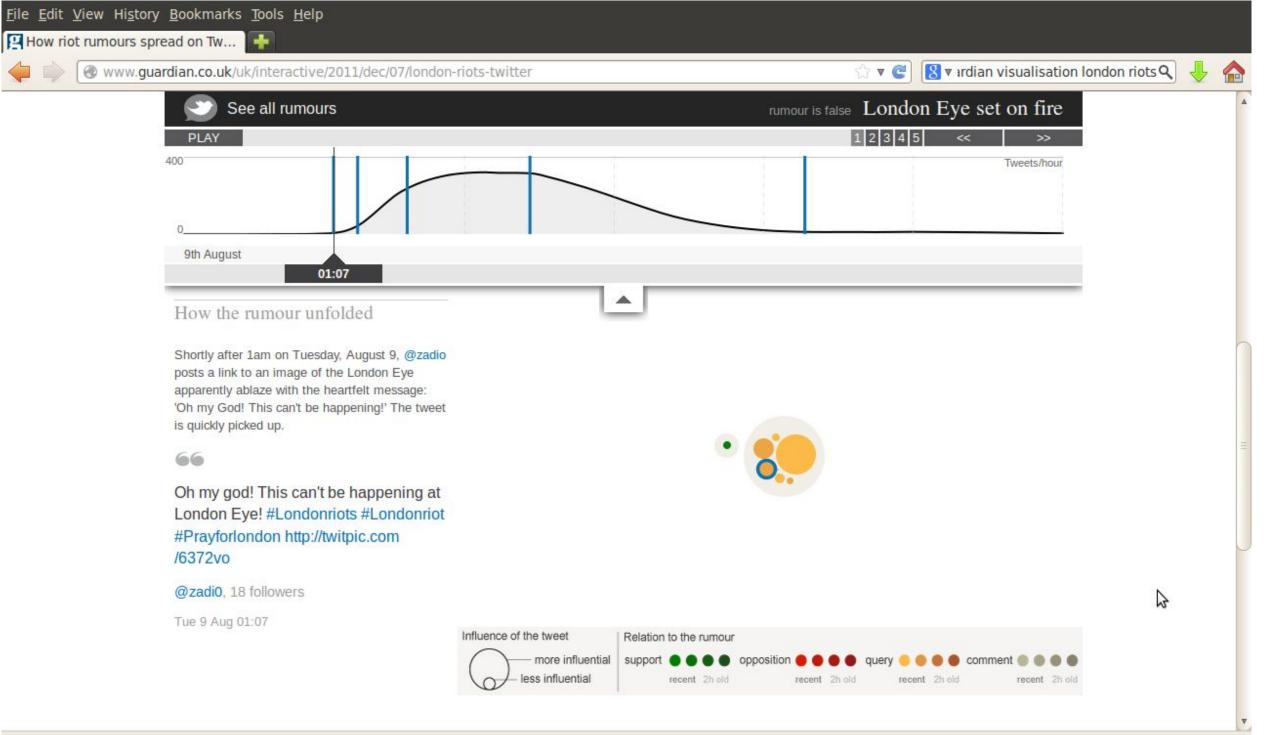
The 6 Questions of Disinformation Analysis

- What is being spread?
- Who is spreading it?
- When it spreads?
- Where it spreads?
- Why it spreads?
- How it spreads?



What about online rumours?





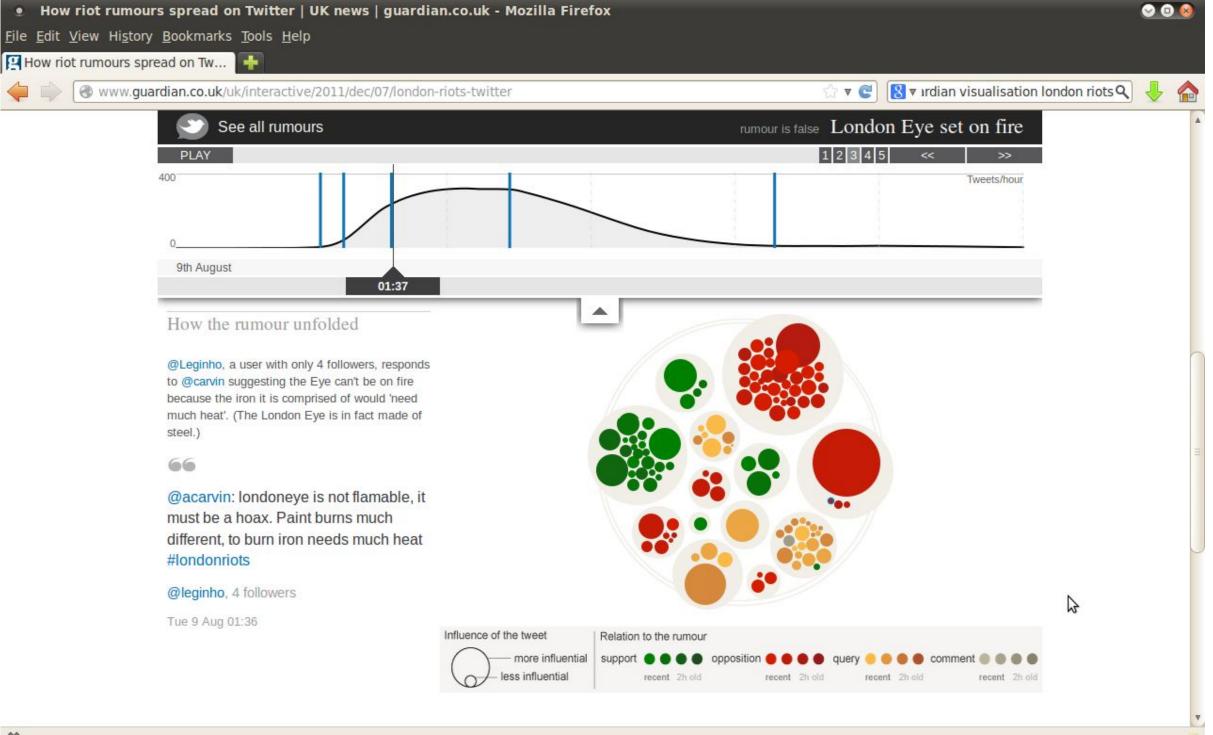




European Commission Information Society and Media

PHEME: ...Veracity, and Spread









PHEME: Analysing Online Rumours



- Rumour is "a circulating story of questionable veracity, which is apparently credible but hard to verify, and produces sufficient skepticism and/or anxiety"
- Memes are thematic motifs that spread through social media in ways analogous to genetic traits
- We coined the term phemes to add truthfulness and deception to the mix
- Named after ancient Greek Pheme, "embodiment of fame and notoriety, her favour being notability, her wrath being scandalous rumours"

Example Rumours and Events



• Events:

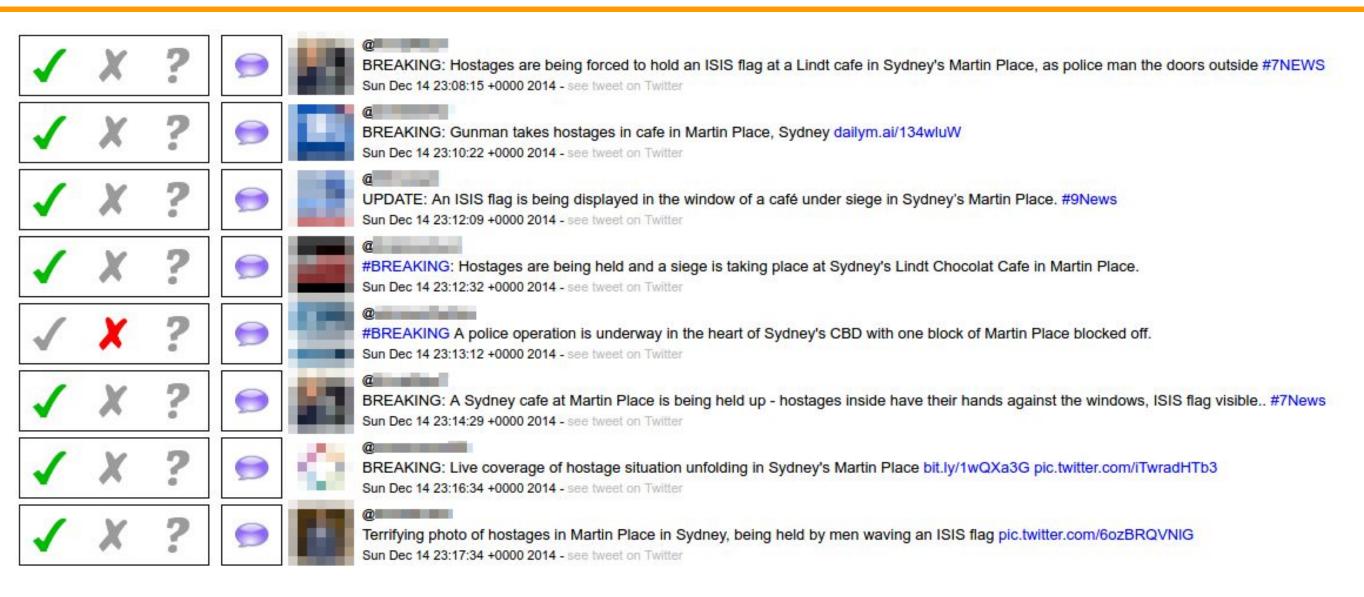
- Ferguson unrest
- Ottawa shooting
- Sydney seige
- Charlie Hebdo shooting
- German Wings crash
- Specific rumours:
 - Putin missing
 - Prince concert
 - Michael Essien
 - Gurlit collection







Rumourous Thread - Example

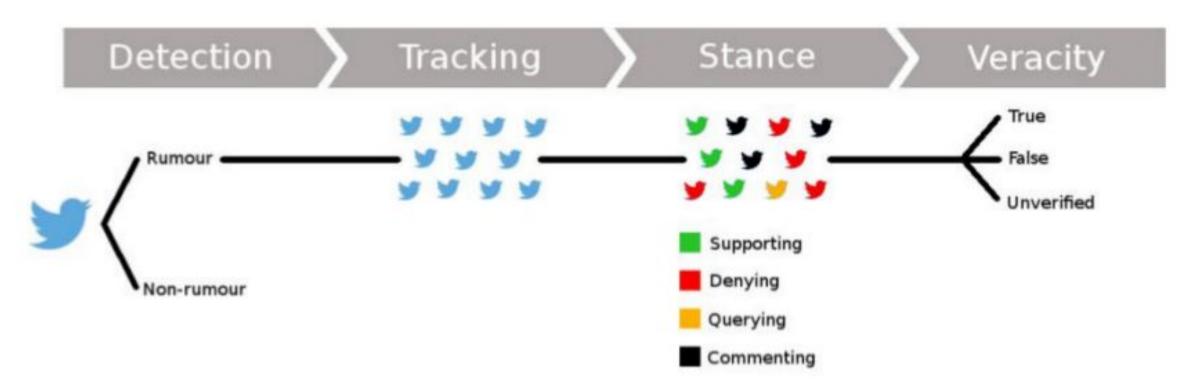






PHEME: Analysing Rumours





Source: (PHEME project)





PHEME: Rumour Stance Observations



- Supporting tweets are more likely to include links.
 i.e. provide evidence
- Looking at the temporal dimension, S/D/Q tend to occur in early stages of a rumour, and then mostly comments later
- We've looked at persistence, finding that users supporting a rumour tend to post more tweets to argue their beliefs





The 6 Questions of Disinformation Analysis

- What is being spread?
- Who is spreading it?
- When it spreads?
- Where it spreads?
- Why it spreads?
- How it spreads?



Other Challenges

- Deep fakes
 - Synthetic videos and images "look and sound like a real person saying something that that person has never said." (Lucas, 2018)
- Preserve important social media content for future studies
- Establish policies for ethical, privacy-preserving research and data analytics
- More funding for inter-disciplinary research
- Measure the effectiveness of technological solutions implemented by social media platforms
- Strengthening media and improving journalism and political campaigning standards



Thank you!

Questions?

Details in this STOA report:

Alaphilippe, A., Bontcheva, K., Gizikis, A. Automated tackling of disinformation: Major challenges ahead.

http://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_STU(2019)624278

email: K.Bontcheva@sheffield.ac.uk

twitter: @kbontcheva



Acknowledgements

Work partially support by funding from the European Union's Horizon 2020 research and innovation programme WeVerify(825297), SoBigData(654024), COMRADES (687847).

WeVerify: https://weverify.eu/

COMRADES: https://www.comrades-project.eu/

SoBigData: http://sobigdata.eu/index