How to Fight Fake News
Dr Victoria Baines, IT Livery Company Professor of IT
6th December 2022

“Fake news” seems to be everywhere we turn in contemporary society, but its rise in our collective consciousness is very recent. What is it? How can governments, platforms, and civil society combat harmful falsehoods online? And what can we do to protect ourselves and defend the truth?

First, a story…

On 4th March 2020, I was asked to appear on Newsnight, the BBC’s late night current affairs programme. The topic for discussion was the then-very-novel coronavirus, and it was to be my last in-person TV interview for some months. Racing up from a talk on cybersecurity at a school on the South Coast, I found myself backstage at Broadcasting House with a Professor of Virology from a London university, and someone who worked for a fact-checking organisation. My job was to explain what social media companies and search engines were doing to combat misinformation on COVID. When asked by the presenter how easy it was for platforms like Facebook to take down this sort of information, this is what I said:

> “From what I understand and from what we’ve seen, reporting and taking down the material when it’s been reported is just one of the measures that companies like Facebook, and I would say also Google with Google Search and with YouTube, are taking at the moment. They’re working with the World Health Organisation and with the NHS, so they have a hotline, if you like, from those official sources. But they’re also promoting those official sources. So, people logging into Facebook today and using Google Search today will have noticed that there are SOS alerts at the top of their newsfeeds, and on their search pages when they search for coronavirus. Then, behind the scenes there will also be another approach to try and investigate, identify and remove some possible sources of coordinated disinformation, which is something President Putin of course has alluded to today. It’s not impossible that hostile actors will be looking to sow disinformation deliberately.”

My intention was to reassure viewers that tech companies were taking steps to promote trustworthy information and counter misuse of their platforms. So far, so sensible. Barring a trip in a black London cab through the New Forest in the pitch dark in the middle of the night, this evening faded from my memory, overtaken by a global public health emergency and several lockdowns.

Until 20th August that year, when my brother informed me that a friend of his had spotted me in a documentary called *Plandemic: Indoctornation*. It turned out that what I had said on Newsnight five months earlier had been repurposed to support the theory that tech companies and governments were conspiring to bury the truth about COVID-19, and that Microsoft founder Bill Gates was attempting to microchip the world’s population. This is the clip they used:

> “They’re working with the World Health Organisation and with the NHS, so they have a hotline, if you like, from those official sources, but they’re also promoting those official sources.”

The sentence itself wasn’t fake, but it had been taken out of context and reproduced as evidence of something that isn’t true and I don’t believe. I found this somewhat ironic, given that the interview was itself about misinformation. It also illustrates the importance of context in many cases where we need to distinguish fact from fiction, and enforce against falsehood.

What exactly is Fake News?

'Fake news' is a catch-all term that is often used to describe false information. I and other researchers would argue that it is actually quite unhelpful, as it bundles together several different types of content and behaviour that have different aims, tactics and impacts, and may therefore require different countermeasures.

Firstly, there is **disinformation**. This is "the dissemination of deliberately false information, esp. when supplied by a government or its agent to a foreign power or to the media, with the intention of influencing the policies or opinions of those who receive it; false information so supplied" (*OED*). Disinformation is sometimes referred to as 'information warfare' or 'influence operations'. It denotes coordinated activity by, for example, Russian authorities seeking to influence people's thinking in the run-up to the US elections and the Brexit referendum in 2016. The coordinated aspect of this means that there are technical measures platforms can deploy to stop its spread online.

Misinformation is "wrong or misleading information" (*OED*) that is not strictly intended to mislead. It is the more accurate term for false information that is uncoordinated or spreads organically, and is more likely to be shared unwittingly, in good faith or out of fear. It is the massively distributed, social media era version of gossip or rumour. **Conspiracy theories**, including those about COVID's origins, vaccines, 5G mobile phone technology, and the Deep State, would fall into this category.

Related concepts of **propaganda** – whether produced and distributed by another state or our own – and low quality **junk news** content also merit consideration as fake news. The phenomenon of clickbait, salacious online content that encourages users to follow often dubious web links, relies on our appetites for junk news and gossip. And of course for one former President of the US, 'fake news' has become short-hand for representations in the mainstream media that are simply unfavourable. Against the backdrop of political populism, concepts such as '**alternative facts**' have gained legitimacy, and there is justifiable concern that we are now living in a '**post-truth**' society, where objective fact is harder to identify and allegedly less important.

As well as presenting information that is simply untrue, the fake news ecosystem abounds in the practices of manipulation and distortion. My comments on Newsnight in 2020 are an example of precisely this: for most viewers of a quality news programme, public health organisations and tech companies working together to suppress misinformation and promote official advice can be a source of reassurance; but for those choosing to watch a movie that promotes conspiracy theories, it is evidence of their worst fears of collusion between governments and technology providers to suppress 'the truth.'

Disinformation and Democracy

When we hear of disinformation, our thoughts naturally turn to hostile state influence operations in another country. The Russian word *dezinformatsiya* denotes the established practice of injecting false information into foreign intelligence holdings, which over time has evolved into the injection of false information into public discourse, including on social media. In November 2016, Facebook founder Mark Zuckerberg dismissed as a "pretty crazy idea" the notion that fake news on social media influenced the outcome of the US election that year. A year later, the company revealed that it had found evidence of fake accounts using the platform to share and amplify data stolen from the Democratic National Committee's email accounts. A few months after that, the company's Chief Security Officer detailed how 470 fake accounts had spent around 100,000 USD on roughly 3000 advertisements on Facebook, concluding "Our analysis suggests these accounts and pages were affiliated with one another and likely operated out of Russia."[1]

Although these ads were served around the time of the US election, Facebook's analysis found that the vast majority focused instead on sowing discord in communities around subjects such as LGBTQ rights, race and ethnicity, immigration, and gun ownership. Subsequent research by Stewart, Arif & Starbird of government affiliated Russian troll accounts on Twitter also in the run-up to the 2016 election traced how disinformation on the topic of the Black Lives Matter movement was harnessed in an attempt to create greater division and polarisation in American communities online.[2] And in the UK, think tank Demos found that UK-related tweets from accounts affiliated with Russia's Internet Research Agency were notable for their anti-Islamic

[1] https://about.fb.com/news/2017/09/information-operations-update/

[2] https://faculty.washington.edu/kstarbi/examining-trolls-polarization.pdf

sentiment.[3] While there is some evidence to suggest that exposure to fake news did have an impact on voting decisions in the 2016 US election, the jury is still out on whether its influence was sufficient to determine the outcome of this election or of the Brexit referendum.[4]

So much, so Cold War spy thriller. Lesser known is the fact that Western governments also engage in influence operations on social media. Recent research by Graphika and the Stanford Internet Observatory found "an interconnected web of accounts on Twitter, Facebook, Instagram, and five other social media platforms that used deceptive tactics to promote pro-Western narratives in the Middle East and Central Asia."[5] On further investigation, Meta – the parent company of Facebook and Instagram – found links to individuals associated with the US military.[6]

Tempting as it is, it would therefore be a mistake to dismiss former US President Donald Trump's propensity to play fast and loose with the truth, and to seek to heroise himself as a character from Game of Thrones. By branding quality news outlets 'fake news', and by claiming to be both a speaker and arbiter of truth – even naming his proprietary social media app Truth Social – Trump consistently and dishonestly exploits the ability to challenge fact with so-called 'alternative facts'. He has also been known to perpetuate conspiracy theories in order to garner support from their believers, in particular the viral theory known as QAnon, the central belief of which is that Satan-worshipping Democrats who run a paedophile ring are trying to control US politics and media. With Trump's explicit and tacit encouragement of the movement to "Stop the Steal" of what he and his supporters considered to be a rigged 2020 election, QAnon believers were prominent among those who attacked the Capitol Building in Washington DC on 6th January, 2021.

Regulating Falsehood

Some governments hold that the solution is for online platforms to identify and remove everything that isn't true. In 2019, Singapore passed the Protection from Online Falsehoods and Manipulation Act, which prohibits the communication of a false statement of fact that is likely to

(i) be prejudicial to the security of Singapore or any part of Singapore;
(ii) be prejudicial to public health, public safety, public tranquillity or public finances;
(iii) be prejudicial to the friendly relations of Singapore with other countries;
(iv) influence the outcome of an election to the office of President, a general election of Members of Parliament, a by-election of a Member of Parliament, or a referendum;
(v) incite feelings of enmity, hatred or ill-will between different groups of persons; or
(vi) diminish public confidence in the performance of any duty or function of, or in the exercise of any power by, the Government, an Organ of State, a statutory board, or a part of the Government, an Organ of State or a statutory board.

But is removal of false or fake content really always the best option? One concern is that by putting the onus on platforms to remove content, people will be less exposed to untrue narratives, and will therefore be less able to distinguish fact from fiction, or to challenge falsehood. Their critical thinking may in fact deteriorate if it is not used. Alternative responses are for platforms to 'downrank' low quality or junk news sources, effectively making them less visible by pushing them further down users' newsfeeds. Some large social media platforms already do this. They also apply labels to identify paid-for advertising, especially in relation to political campaigns, and work with fact-checkers so that they can apply labels to posts that have been found to be inaccurate or misleading.

"Who will guard the guards themselves?" (Juv.6.347f.)

Focus on platforms' removal of content at scale presupposes that those in power always post objective facts online and do not engage in influence operations, which unfortunately is not the case. Some governments use disinformation tactics against their own citizens. In my first lecture in this series, *Who Owns the Internet?*,

[3] https://demos.co.uk/press-release/new-demos-analysis-finds-russian-influence-operations-on-twitter-targeted-at-uk-were-most-visible-when-discussing-islam/
[4] https://theconversation.com/trump-may-owe-his-2016-victory-to-fake-news-new-study-suggests-91538
[5] https://stacks.stanford.edu/file/druid:nj914nx9540/unheard-voice-tt.pdf
[6] https://about.fb.com/news/2022/11/metas-adversarial-threat-report-q3-2022/?utm_source=substack&utm_medium=email

I discussed how Russia, China, Iran and North Korea greatly restrict access to the internet and promote pro-government content online and in the mainstream media. There have been numerous instances since the invasion of Ukraine of Russian state media putting rather more than a positive spin on their military's performance. Perhaps the most amusing Russian disinformation tactic in recent years has been the tendency to use video game footage in coverage of military operations. In 2017, the Ministry of Defence published a scene from the game *AC-130 Gunship Simulator: Special Ops Squadron* as "irrefutable proof that the US provides cover for ISIS combat troops". And in 2018, Russian state TV used footage from the game *Arma 3* in a news item about a Russian soldier killed in Syria in 2016.

When a source that we trust starts to manipulate the narrative, whether that's a public figure, political party, or a media outlet, this can be harder to spot and we may require specialist assistance. Political debates in the UK are now routinely fact-checked by academics, non-profits, and mainstream media outlets alike. Ahead of the US mid-term elections in 2018, right-wing media organisations pushed false narratives about the Hungarian born billionaire George Soros, a leading Democrat donor. These included accusations that he had funded the migrant caravan then heading towards the US, that he had been a Nazi SS officer, and anti-Semitic conspiracy theories apparently endorsed by celebrities and the family of the then President. Around the same time, automated 'bot' accounts on Twitter posing as Democrat supporters began to discourage people from voting in the mid-term elections. On this occasion, Twitter removed an estimated 10,000 bot accounts.

Another tactic is to encourage real account holders to copy and paste a message that seems authentic, complete with grammatical errors, but which on closer inspection is anything but. So-called 'keyboard armies' are paid to do this. Platforms are able to use tools that identify and remove the same false content being shared by large numbers of users. In both cases – whether bots or real people acting just like them – tech comapnies look for inauthentic behaviour *in addition to* seeking to verify the accuracy of the content shared.

The Death of Satire?

For thousands of years, societies have turned to satire and parody to voice their disapproval of or criticise politicians. Mocking public figures reminds us and them that they are only human and that – in democracies, at least – they are in power only because we have elected them. Satire has an important social function, and it is not always comfortable reading, listening or viewing.

One of the most famous examples of satire that disturbs and disconcerts is Jonathan Swift's *A Modest Proposal*. Published in 1729, it takes the form of a political essay to make the straight-faced suggestion that the impoverished Irish sell their own children to be eaten by the rich. It is a work of considerable grotesquery and rhetorical sophistication, which is often taken to be a critique not only of the contemporary economic policy, but of the dehumanisation of the Irish people. As a very bright Stanford Computer Science student pointed out during one of my lectures, it would likely fall foul today of some governments' expectations for suppressing content that sows community discord or diminishes public confidence in those in power.

Indeed, satire is of concern to regulators in other countries, including the UK and the US. In a 2018 report, the House of Commons' Digital, Culture, Media and Sport Committee of MPs included satire and parody in its definition of fake news, because they may "unintentionally fool readers."[7] In the US, satirical news website *The Onion* recently asked the Supreme Court to look favourably on a case brought by a man who had been charged with disrupting a public service for setting up a spoof police page on Facebook. In a beautifully written document that is itself satirical in tone and references luminaries from the literary tradition including Swift and the Roman poet Horace, The Onion set out to argue that i) Parody functions by tricking people into thinking that it is real ii) Because parody mimics the real thing, it has the unique capacity to critique the real thing, iii) a reasonable reader does not need a disclaimer to know that parody is parody, and iv) it should be obvious that parodists cannot be prosecuted for telling a joke with a straight face.

According to The Onion's brief, the US appeals court's decision to back the police "suggests that parodists are in the clear only if they pop the balloon in advance by warning their audience that their parody is not true. But some forms of comedy don't work unless the comedian is able to tell the joke with a straight face." Marking satire and parody precisely as such – deadening the effect and missing the point – is something we are already seeing more of.[8]

[7] https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/363/363.pdf
[8] https://help.twitter.com/en/rules-and-policies/parody-account-policy

The Rise and Rise of Junk News

The inclusion of satire and parody in definitions of fake news assumes that members of the public are unable to distinguish between fact and fiction, authoritative news reporting and parody. Quite apart from insulting the intelligence of citizens, this assessment also assumes that humans can't spot context or filter out junk content but machines can, when in fact the opposite is more likely.

In the UK there is a tradition of quality journalism, and a parallel tradition of journalism whose content readers have become accustomed to take with a pinch of salt. We appreciate headlines like "Freddie Starr ate my hamster" (*The Sun*, March 12, 1986) not as unadulterated records of objective fact, but as content that has been embellished to elicit sensation and heighten our engagement. Phrases like, "sources close to The Palace", or "a source close to the government" are tantamount to an admission that a story has been fabricated or is at least based largely on rumour. Readers are in on the joke. That is not to say that no one believes anything they read in a tabloid newspaper, but that readers are able to use their judgement to separate serious reporting from junk news.

In July of this year, amendments were added to the UK Online Safety Bill to "impose a duty on Category 1 companies [those with the largest audiences] to safeguard all journalistic content shared on their platform (including news publishers' journalistic content)."[9] What that means in effect is that a social media post making inflammatory claims about migrants could be liable to removal if posted by an ordinary user, but protected if published by a newspaper, radio or TV station.

The New Defenders: Fact-checkers

The big US platforms work with specialist fact-checking organisations around the world to verify the accuracy of online content. Such is the perceived trustworthiness of fact-checking services that during the 2019 UK general election, the Conservative party changed its Twitter account display name to 'factcheckUK', retaining the blue 'tick' for a verified account that it had previously obtained, and with more than a passing similarity both to Channel 4's 'factcheck' service and non-profit Full Fact. Although denounced by these organisations and roundly criticised by media outlets, there was no official sanction for this imposter tactic, nor for apparent manipulation of videos of opposition politicians, Keir Starmer and Jess Phillips among them.

Here, too, there are technical measures that can and should be deployed by platforms. A name change for the social media account of a political party or politician is unusual, and should in most cases be flagged as suspicious – not least because it may indicate that the account has been hacked. Increasingly, tech companies use tools such as Microsoft's Video Authenticator that allow them to identify when a video has been artificially manipulated. Major journalistic outlets including the BBC, the New York Times and Reuters have partnered with Microsoft and Meta in the Trusted News Initiative. As well as sharing reports of disinformation in real-time for rapid response, the partners are working on Project Origin to create a common standard for establishing whether a piece of video content is authentic.

The advent of deepfakes has arguably made this need more pressing. Deepfakes are fake videos generated by machine learning. They are becoming more sophisticated and more convincing, and are finally being used in anger in political contexts. Recent research indicates that humans and computers are equally able to spot deepfakes, but humans significantly outperform leading deepfake detection technology when it comes to videos of well-known political leaders.[10] This in turn suggests that the most promising model for deepfake identification is human-machine collaboration. We can't sit back and let computers do all the thinking for us just yet.

What You Can Do

In cybersecurity, we divide successful security measures into people, process, and technology. They are the cornerstones of how we combat cyber-attacks and cybercrimes. Countering fake news, be this coordinated

[9] https://www.gov.uk/government/publications/fact-sheet-on-enhanced-protections-for-journalism-within-the-online-safety-bill/fact-sheet-on-enhanced-protections-for-journalism-within-the-online-safety-bill

[10] https://www.pnas.org/doi/10.1073/pnas.2110013119

disinformation by hostile states, junk news, misinformation shared out of fear or anxiety, or good old-fashioned propaganda, requires the same tripartite strategy.

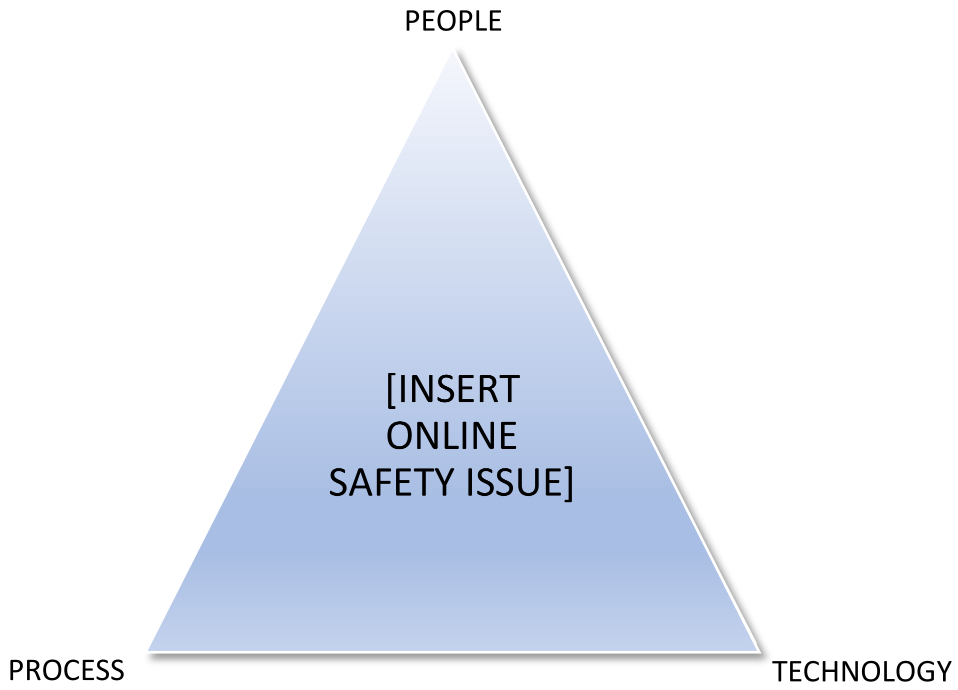

Fig. 1 The holistic approach to online safety issues

Society certainly requires regulations and rules to help distinguish what is prohibited from what is permissible speech and content, along with procedures for its suppression. Technical are necessary to identify bad actors, bots, inauthentic content and manipulated videos on a massive scale. But however stringent these rules, and however sophisticated these tools may become, strengthening the human response will always be just as important, because influence operations seek to exploit our susceptibility – to fear, to the next big scoop, the inside story, and the confidence of friendship. Those of us on social media are on the front line of the fight against fake news, because every time we share content with our family, our friends, and our professional networks, we risk being an unwitting agent of misinformation. And for that very reason, by using our critical thinking we can defend ourselves and those we know. We can even play an active role in protecting our national security.

Media literacy for all ages is crucial to fighting fake news. Research into the sharing of fake news during the 2016 US presidential campaign found that Facebook users over the age of 65 shared nearly seven times as many articles from fake news domains as those aged 18-29.[11] Many of the fact-checking organisations and quality news outlets around the world have built searchable databases to check the veracity of a news item or other piece of online content. A number have also produced training packages, educational videos, lesson plans and infographics to help people discern fact from fiction.

You shouldn't believe everything you see online, just as you've long been advised not to believe everything you read in the newspapers or see on TV. Take a few seconds to:

- Question the content: use search engines (including Google's reverse image search[12]) and fact-checking sites.
- Question the intent of the person or organisation posting.
- Question and check the source, even if it's someone you trust. Don't assume that they have done their own fact-checking.
- Where you can, let your friends and family know when they have shared something that you know to be untrue or manipulated.

[11] https://www.science.org/doi/full/10.1126/sciadv.aau4586

[12] https://www.google.com/imghp?hl=en

When we refuse to be taken in by fake, manipulated or junk content online, we protect everyone else from influence operations, and from misleading information that could harm their health or exploit their fears.

Resources

The Reporters' Lab at Duke University has identified nearly 400 fact-checking organisations in over 100 countries. You can search their database for the service in your country and/or language - https://reporterslab.org/fact-checking/

Notable English language fact-checking services include Snopes - https://www.snopes.com/. As well as checking political claims, Snopes has a large database of searchable checks on urban legends.

A number of mainstream media outlets operate fact-checking services, among them the BBC's Reality Check, with global coverage - https://www.bbc.co.uk/news/reality_check

Full Fact is a UK-based charity with a searchable database of fact checks on claims made by politicians, public institutions and journalists, and viral content online. They follow up on false claims by asking the people and organisations who have made them to correct the record. https://fullfact.org/facts/

The Institute for Strategic Dialogue, Google/YouTube and Parent Zone have produced the *Be Internet Citizens* toolkits for teachers and community educators. These help educate young people to tell fact from fiction and spot bias in media content. https://internetcitizens.withyoutube.com/#about

Newseum, the American museum of news and journalism, has an extensive online education offering that includes lesson plans for teachers and training for members of the public on recognising fake and junk news, countering propaganda, and when to share something online. https://newseumed.org/

EUvsDISINFO is an initiative of the EU's European External Action Service to counter Russian disinformation. In addition to online training and a searchable database on its website, the team also issues fact-checking updates via its Facebook and Twitter accounts. https://euvsdisinfo.eu/learn/

Further Reading

Centre for Data Ethics & Innovation (2021) *The role of AI in addressing misinformation on social media platforms* - https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1008700/Misinformation_forum_write_up__August_2021__-_web_accessible.pdf

Demos (2019) *Warring Songs: Information Operations in the Digital Age* - https://demos.co.uk/wp-content/uploads/2019/05/Warring-Songs-final-1.pdf

Graphika (2020) *Secondary Infektion* - https://secondaryinfektion.org/downloads/secondary-infektion-report.pdf

Graphika & Stanford Internet Observatory (2022) *Unheard Voice: Evaluating five years of pro-Western covert influence operations* - https://stacks.stanford.edu/file/druid:nj914nx9540/unheard-voice-tt.pdf

Groh, Epstein, Firestone & Picard (2021) "Deepfake detection by human crowds, machines, and machine-informed crowds", *PNAS 119.1*. https://www.pnas.org/doi/10.1073/pnas.2110013119#sec-2

Meta *Adversarial Threat Report, Third Quarter 2022* - https://about.fb.com/news/2022/11/metas-adversarial-threat-report-q3-2022/?utm_source=substack&utm_medium=email