Written Representation 94

Name: Claire Wardle
Research Fellow at the Shorenstein Center for Media, Politics and Public Policy, Harvard Kennedy School and Executive Director of First Draft

Received: 1 Mar 2018

Evidence submitted to the Singapore Parliament Select Committee on Deliberate Online Falsehoods

by Dr. Claire Wardle, Research Fellow at the Shorenstein Center for Media, Politics and Public Policy, Harvard Kennedy School and Executive Director of First Draft[1]

Definitions

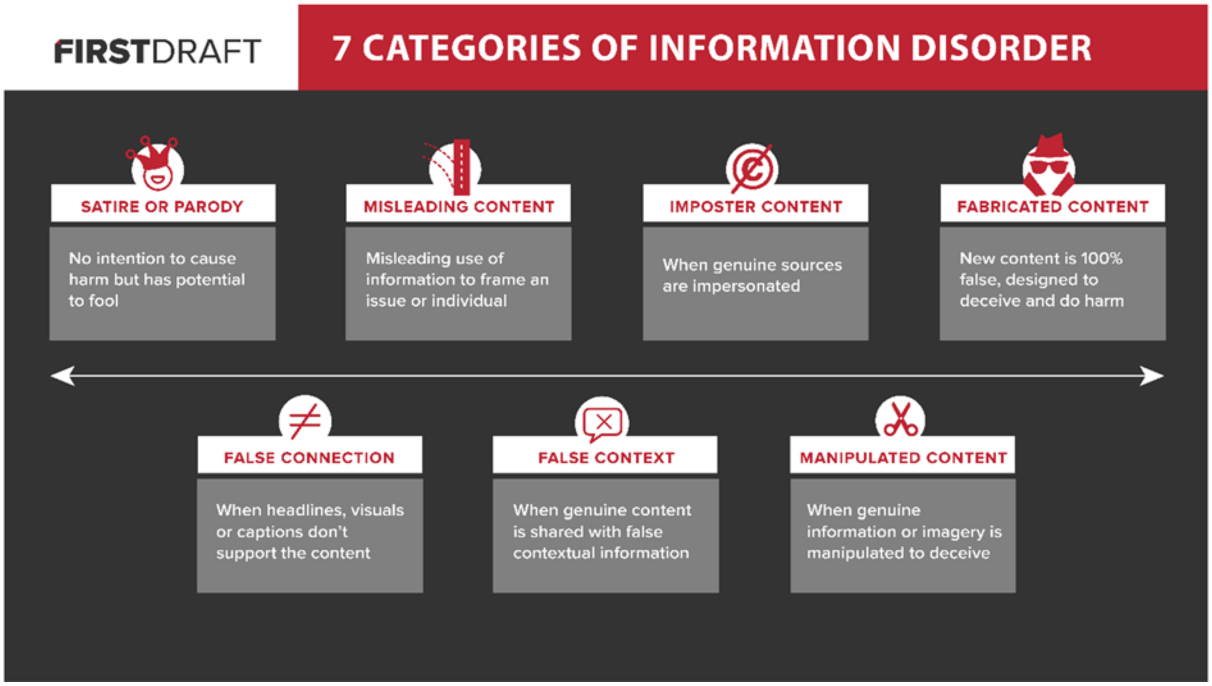
Language and terminology matter. The term "fake news" is woefully inadequate in capturing the complexity of what ought to be called "information disorder." For one, much of the content being debated isn't actually fake, but instead used out of context or manipulated.[2] Further, the ecosystem of polluted information extends far beyond content that mimics "news." The term "fake news" shouldn't be used.

However, clearly delineating what counts as information disorder is difficult. Legislators struggle with content that might be legal in other contexts—incitement to violence or hate speech—but nevertheless harms individuals, organizations, or even the democratic process. The definition of information disorder is not black and white; it's fluid.

In this diagram, I highlight seven types of information disorder, illustrating this fluid spectrum. Satire, the least problematic form of information disorder, sits on one end of the spectrum, while fabricated content, specifically content created to spread false information, sits on the other.

[1] A non-profit focused on experimental projects to fight disinformation (firstdraftnews.org)

[2] One example of this a photo of a Muslim woman walking past a victim during the Westminster bridge attack was amplified by a Russian bot. The photograph was genuine but the amplification techniques were part of a coordinated campaign to shape public opinion about the event http://www.telegraph.co.uk/news/2017/11/13/russian-bot-behind-false-claim-muslim-woman-ignored-victims/



FIRSTDRAFT

7 CATEGORIES OF INFORMATION DISORDER

SATIRE OR PARODY

No intention to cause harm but has potential to fool

MISLEADING CONTENT

Misleading use of information to frame an issue or individual

IMPOSTER CONTENT

When genuine sources are impersonated

FABRICATED CONTENT

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

1. Satire and Parody: Including satire here is perhaps surprising. However, people often don't realize that satire is actually satire, especially when they are reading on a social feed. In fact, in our Crosscheck project monitoring the French presidential election,[3] we found that people disseminate disinformation masquerading as satire, in order to avoid fact-checks.

2. False Connection: A false connection is when headlines, visuals or captions don't support an article's content. The most common example is clickbait headlines, which are becoming more popular.

[3] Smyrnaios, N., S. Chauvet and E. Marty (November 2017) The Impact of CrossCheck on Audiences and Journalists, First Draft, https://firstdraftnews.org/crosscheck-qualitative-research/

3. Misleading Content: Misleading content appears when information is used to inaccurately frame an issue or an individual. For example, someone may misguide their reader by cropping a photo or by choosing a quote or statistic to remove relevant context. Visuals like cropped photos are particularly effective, as our brains are less critical of visuals than they are of text.

4. False Context: Here, genuine content is circulated out of its original context, misleading the reader. Content exhibiting “false context” is one of the many reasons that the term “fake news” is so unhelpful.

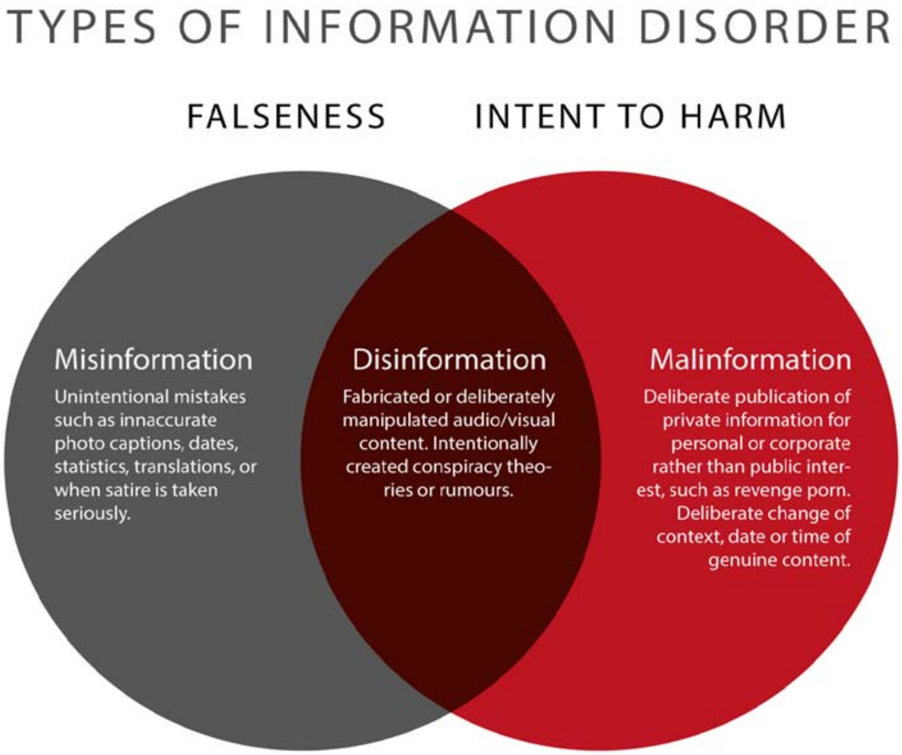
5. Imposter Content: Journalists often see their bylines alongside articles they did not write, and organizations’ logos are used in video and images they did not create.

6. Manipulated Content: Manipulated content is when genuine content is manipulated, in Photoshop for example, to deceive.

7. Fabricated Content: Fabricated content can be textual or visual. For example, the fabricated ‘news site’ WTOE5 News published an article suggesting that the Pope had endorsed Donald Trump. Or, consider the visual example in which a graphic targeted at minority communities on social networks suggested that people could vote for Hillary Clinton via SMS.

Three Types of Information Disorder

These seven categories can be categorized into three camps, based on truthfulness and intention to harm. This diagram demonstrates that there are three types of information disorder. Content that is false but not intended to harm is called ***misinformation***. This can include satire, clickbait, or misleading quotes and images. Content that is false and intended to harm is considered ***disinformation*** and includes malicious lies, fabricated content, and manipulation campaigns. Finally, truthful information that is intended to harm is considered to be ***malinformation***[4].

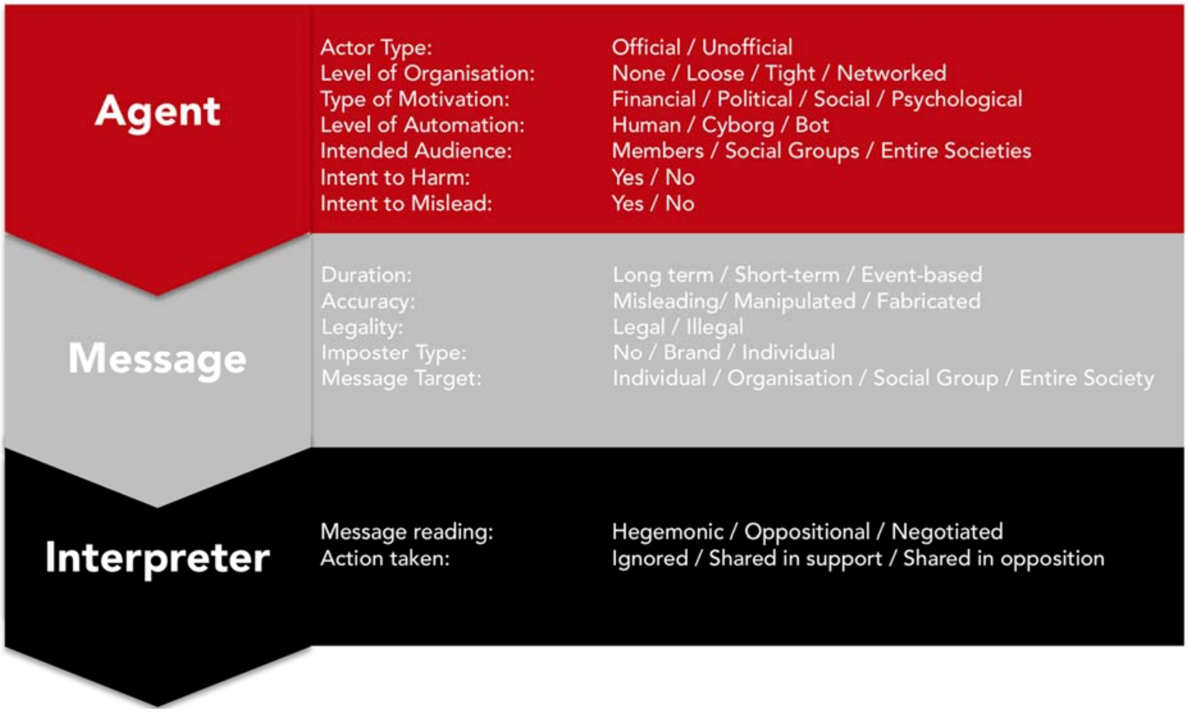

TYPES OF INFORMATION DISORDER
FALSENESS
INTENT TO HARM
Misinformation
Unintentional mistakes such as innaccurate photo captions, dates, statistics, translations, or when satire is taken seriously.
Disinformation
Fabricated or deliberately manipulated audio/visual content. Intentionally created conspiracy 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.

The Three Elements of Information Disorder

The ecosystem of information disorder includes different actors, very different messaging formats, and wildly different audience interpretations. Thus, we need to separately examine the ‘elements’ of information disorder: the agents, messages, and interpreters. In this matrix we pose questions that need to be asked of each element. As we explain, the ‘agent’ who creates a fabricated message might be different from the agent who produces that message—who might also be different from the agent who distributes the message. We need to thoroughly understand not only who these agents are, but also what motivates them. Similarly, we must understand the types of messages distributed by agents, so can properly estimate their scale and properly address them. Finally, we need to more deeply understand how these messages are interpreted, what action are being taken by those who see them (eg. re-sharing to their networks with new comments),

[4] a term invented by myself and my co-author Hossein Derakshan to describe genuine information used to cause harm (for example revenge porn, the leaking of private emails, some forms of harassment)

and how various audiences ‘read’ these messages when they’re coming from trusted family members, friends or peers.



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

First, we need to examine the ‘agents’ - those who have the idea for the message. They might be an operative in the Russian government. They might be an individual who sees the opportunity for financial gain. Or they might be a Trump supporter who wants to publicly connect with other like-minded people to push a misleading narrative.

There are four motivations for creating misleading or inaccurate information: (i) financial, (ii) political (either geo-political or campaign politics), (iii) social (to connect with others like you) or (iv) psychological (to cause trouble or harm for the sake of it).

The types of actors vary widely. Actors can work on behalf of a state, or as part of a loose network of passionate supporters of a country, party, or cause. The target of the disinformation can be an individual, a cause, a party, a religion, or a country. Actors can program bots, or they can post as humans or cyborgs (humans who post so regularly they take on the characteristics of bots). Actors may intend to mislead and cause harm or they may not.

Types of messages vary widely as well. They may be legal or illegal; they can be individual messages or part of a longer term manipulation campaign; they can be slightly misleading with a kernel of truth or widely exaggerated and wholly inaccurate.

In addition to all the reasons already discussed, “fake news” is problematic term because it has confined the debate around it to the world of text. Because the debate focuses on fabricated news ‘sites,’ visual content, like images, visualizations, graphics, and videos, is rarely considered, even when it’s misleading, manipulated or fabricated. Technology companies have aimed their solutions at fabricated articles, mostly because text is easier to computationally analyze than visuals. However, visuals are often more persuasive than text[5], making them a more powerful vehicle for information disorder. In addition, in the past few months, we’ve seen that audio and video can be manipulated to falsify reality. [6]

Finally, messages are interpreted in a whole host of ways. Their interpretation depends on the source of the message, who created it, who shared the message, and how the message interacts with a reader’s existing beliefs. The impact of digital m/disinformation over social networks has not been studied nearly enough, making the current debate all the more challenging.

However, important work has certainly been done. Stuart Hall’s[7] seminal work, “Encoding/Decoding,” is still incredibly relevant. We ought to be asking the questions he poses: Are users accepting the messages as designed? Are they challenging certain parts of a message or dismissing it entirely? If we examine how messages are “re-

[5] Birdsell, D. S., & Groarke, L. (1996). Toward a theory of visual argument. Argumentation and Advocacy, 33(1), 1-10

[6] WNYC Radio Lab (July 27, 2017) Breaking News, http://www.radiolab.org/story/breaking-news/

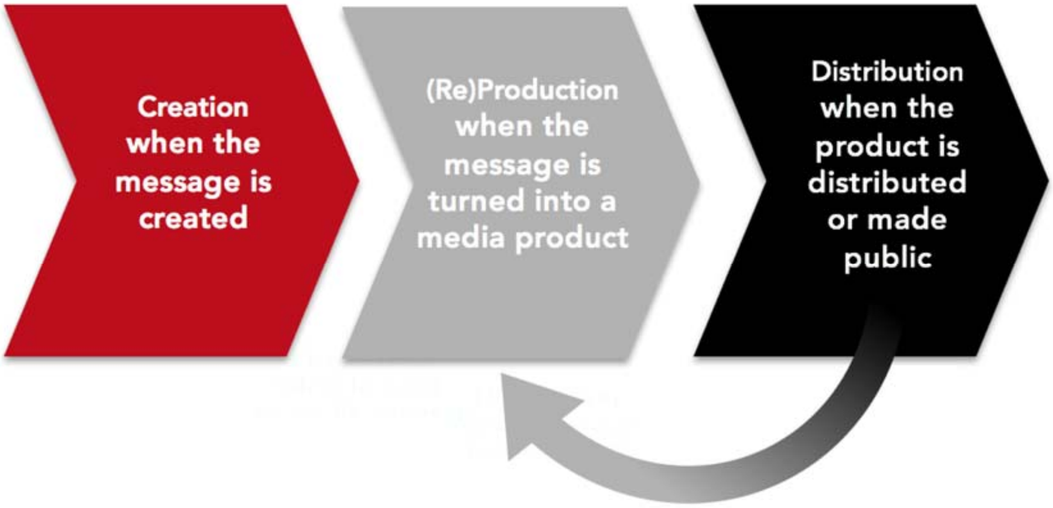
[7] Hall, S. (1980). Encoding/decoding. Culture, media, language: Working Papers in Cultural Studies, 1972-79, pp. 128-138.

shared," we gain insight into how people make sense of particular messages. More analysis of this kind is needed to understand how messages are shared, online and off, and how messages are interpreted, particularly when the messages are mediated by a recipient's trusted peers.

Communication research, such as the two-sept flow model first discussed by sociologist Paul Lazarsfeld in 1944, reminds us that understanding a message's impact means focusing on moore than just the number of people who clicked on any given link. People interact with information in a complex set of ways—their beliefs and attitudes are shaped by opinion leaders, the mass media, and, in the era of social media, often by just an individual journalist, a friend, or even a commenter online. Measuring information disorder is perhaps our most significant challenge.

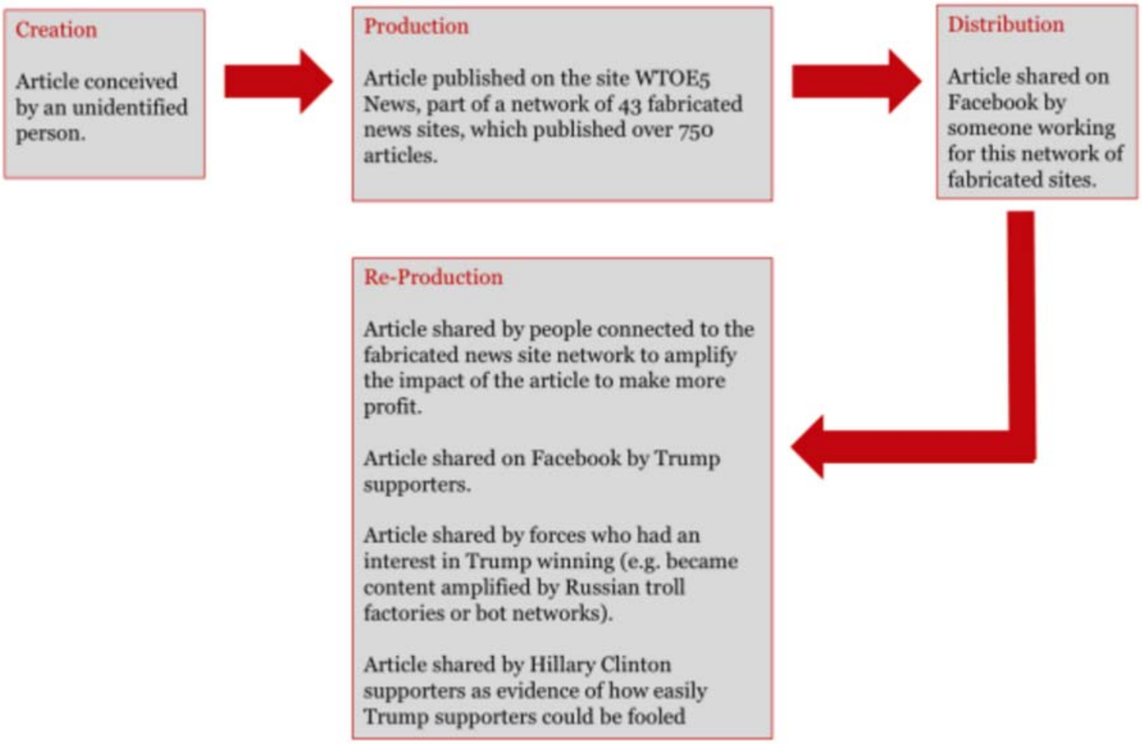
Three Phases of Information Disorder

Finally, we need to think about the different ***phases*** of information disorder: creation, production, distribution, and, frequently, re-production. Across these different phases, there are often different agents, and the message itself often evolves. Those that interpret the original message become agents themselves when they re-share with their own communities.


Creation when the message is created
(Re)Production when the message is turned into a media product
Distribution when the product is distributed or made public

To examine how the phases of creation, production and distribution help us understand information disorder, let’s use the article ‘Pope Francis Shocks World, Endorses Donald Trump for President, Releases Statement’ published on the self-proclaimed fantasy news site WTOE 5 in July 2016. For an in-depth analysis of this article and the network of sites connected to it, we would recommend reading ‘The True Story Behind the Biggest Fake News Hit of The Election’ from Buzzfeed.[8]

If we think about the three phases in this example, we can see how different agents changed the impact of this content.



Creation

Article conceived by an unidentified person.

Production

Article published on the site WTOE5 News, part of a network of 43 fabricated news sites, which published over 750 articles.

Distribution

Article shared on Facebook by someone working for this network of fabricated sites.

Re-Production

Article shared by people connected to the fabricated news site network to amplify the impact of the article to make more profit.

Article shared on Facebook by Trump supporters.

Article shared by forces who had an interest in Trump winning (e.g. became content amplified by Russian troll factories or bot networks).

Article shared by Hillary Clinton supporters as evidence of how easily Trump supporters could be fooled

The role of the mainstream media as agents in amplifying (intentionally or not) fabricated or misleading content is crucial to understanding information disorder. Fact-checking has always been fundamental to quality journalism, but the techniques used by hoaxers and

[8] Craig Silverman (Dec 2016) The True Story Behind the Biggest Fake News Hit Of The Election, Buzzfeed, https://www.buzzfeed.com/craigsilverman/the-stran

those attempting to disseminate dis-information have never been as sophisticated as they are now. With newsrooms increasingly relying on the social web for story ideas and content, forensic verification skills and the ability to identify networks of fabricated news websites and bots is more important than ever before.

Politicians, as well as their supporters, have appropriated the term “fake news” all around the world to describe news organizations whose coverage they find disagreeable. In this way, the term has become a mechanism by which the powerful can clamp down upon and restrict free speech, as well as undermine and circumvent the free press. The inquiries currently taking place in Europe, both this one, as well as the EU Commission’s High Level Group on ‘Fake News’ (of which I am a member) are being watched incredibly closely by governments around the world. Recommendations suggested in Europe could become blueprints for regimes where protections for free speech and independent media do not exist. Already, German laws against hate speech on social media have been used by other countries to clamp down on free speech.

Scale

As just explained, this ecosystem includes different actors, very different messaging formats, and diverse interpretations by audiences. This includes, but is certainly not limited to:

- Websites created to deliberately spread disinformation;
- Inaccurate posts on public social media, forums and message boards (Facebook, Twitter, Reddit, 4Chan, Gab etc.);
- Inaccurate information shared on closed messaging apps such as WhatsApp, Facebook, Messenger, Telegram or Discord;
- Visual posts on social media sites (Instagram, YouTube, Pinterest) and closed messaging apps (including inaccurate photographs, videos, memes, and data visualizations that have been manipulated or fabricated);
- Inaccurate information published via so-called ‘dark posts’ on social networks that micro-target updates to certain users;[9]

[9] These posts are described as dark as the post is not visible on the organisation’s public profile

- Text, image and video results on search platforms (e.g. Google, Bing, YouTube)
- Inaccurate comments or content published on consumer review sites (e.g. Amazon, TripAdvisor)
- Manufactured signatures on online petitions (e.g. Change.org)
- Offline events created online, for example the creation of Facebook ‘Events’ pages designed to encourage passionate supporters from either side of a controversial topic to take their protests to the streets.[10]

Key Challenges

- Agents of disinformation understand that to most effectively influence public opinion they require the amplification of rumours and fabricated content by platforms, politicians and the mainstream media. They target platforms by deliberately ‘gaming’ trending or search results, trick politicians by targeting them directly via social media by posing as concerned citizens or by hoping hashtag campaigns will gain traction, and they hoax the mainstream media through sophisticated disinformation campaigns. They use many forms of manufactured amplification, from automated bot networks to groups of people paid to act as bots (cyborgs); they repost or re-share technology, moving content across networks automatically; and they astroturf using people who are paid or who are motivated by ideology to game online petitions or forums. Currently, most journalists and politicians, as well as technology companies and newsrooms, are neither equipped to handle nor sufficiently aware of these threats.
- Today, people’s brains are overloaded by more information than ever before. In this environment, posts delivered via smartphone alerts or social streams often look identical. The heuristics humans require to judge credibility are absent. Mental shortcuts therefore become more powerful. People often judge a piece of content’s credibility based on the friend who shared the information. Or, often, they use what’s known as the ‘familiarity heuristic’—if they have heard the same information

[10] An example highlighted in the US was a pretext orchestrated entirely by Russia http://money.cnn.com/2018/01/26/media/russia-trolls-facebook-events/index.html

before they are more likely to believe it.[11] These challenges are leading people to share inaccurate information, thereby amplifying disinformation.

- The lack of transparency around promoted content on social platforms and search companies makes it impossible to know which campaigns have been undertaken on a platform, and who paid for those campaigns in real-time. In this environment, it is much easier for systematic manipulation to occur.
- People are predisposed to seek out, consume and engage with information that supports their worldview. Social algorithms are designed to encourage this behaviour.

  Online disinformation does not function in a vacuum. It moves across platforms, it can flow to and from mainstream media, and it certainly flows between people in conversations offline. Unfortunately, much of the data used to understand this phenomenon is text based (because of access to available API's, etc.). And because research on images, videos, and offline conversations is difficult to undertake, it is difficult to understand the entire ecosystem of information disorder.

Proposed interventions

1) Research and Data

- Currently the technology companies themselves are the only ones who can view the scale of the problem. Even Twitter, which is researched most often because of its open and accessible API, is hard to study, because it claims that it 'cleans' feeds algorithmically, something not visible via the API. Without any external access to this data, there is no way to independently audit the scale of the problem, to understand how and when users have interacted with disinformation, and to understand how disinformation moves across platforms. In addition, the lack of data means that when initiatives such as the Facebook's Third Party Fact Check project or Google's 'fact-check' tags are implemented, there is no independent method for assessing their impact.

[11] Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. Journal of Pragmatics, 59, 210-220.

- There is a need for funding and co-ordination of an international research agenda for monitoring the scale and impact of disinformation. Academic researchers from many different disciplines (including digital anthropology, computer science, cognitive psychology, political communication, neuro-science) need to be incentivized to work collaboratively on understanding the phenomenon. In addition, the methodologies need to be shared and standardized to allow longitudinal and cross-border studies to take place.
- Advanced computational technology is needed. Specifically, we need technology that can automate the detection and categorization of different types of visual and text based disinformation created and uploaded to the (social) web, in multiple languages across many countries, every day. The development of this technology should be incentivized, and engineers working for the large technology companies, as well as those working independently at startups or within research centers, should be given reasons to work together on these challenges.

2) Data and Information Literacy

- Traditional news literacy—learning how to differentiate opinion from hard news for example—needs to be expanded. Programs and curricula should include discussions of how to override confirmation bias and 'tribal identifications.' They should include training on how to be sceptical of information which produces an emotional response, particularly images. To ensure these programs include all elements of digital and information literacy, they should include modules on how to critically assess statistical and quantitative statements in the media.[12] In addition, they should include training on how to understand the power of algorithms on search results and social feeds, and the power of artificial intelligence to detect patterns, sort information and automate the creation of new forms of information[13.]

[12] Written evidence submitted by the Royal Statistical Society to the UK Parliamentary Inquiry on Fake News, https://www.parliament.uk/business/committees/committees-a-z/commons-select/culture-media-and-sport-committee/inquiries/parliament-2015/inquiry2/publications/

[13] Written evidence submitted by the UCL Knowledge Lab to the UK Parliamentary Inquiry on Fake News http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/culture-media-and-sport-committee/fake-news/written/48571.html

3) Providing motivations for stronger media, including collaborations

- A priority should be strengthening non-partisan media through additional investment and training opportunities. The impact of the digital revolution on the news media is well known, but the impacts of information disorder are still not fully understood. They include:
    - the devastation of the local news environment has meant that people no longer have access to high quality news about their local communities;
    - the shrinking of newsroom resources means that there are fewer editors catching honest mistakes or poor reporting;
    - the cutting back of newsroom training initiatives means that few journalists have been trained on how to verify content sourced via the social web, leaving journalists vulnerable to deliberate attempts to hoax reporters;
    - an increasingly competitive news environment, paired with the fact that social feeds, rather than news websites, are often people's direction connection to news, has led to the rise of clickbait headlines and sensational, image dependent, and emotionally driven coverage.
- Agents of disinformation know that fact-checks, if framed incorrectly, can provide oxygen to rumours and fabricated content. This means that newsrooms should be given a reason to work collaboratively--to ensure that manipulation tactics are being flagged and shared), to ensure best practices for debunking are instituted across newsrooms, and to prevent duplication of effort (25 newsrooms shouldn't be verifying the same meme independently).

4) Increased Transparency by the technology companies

- One the most challenging aspects of information disorder is that the techniques used by purveyors of quality information, such as targeting ads at specific audience, are the techniques used in disinformation campaigns. Any form of paid promotion needs to be flagged at the point of consumption. Digital and non-digital ads should be held to the same standard of transparency about who paid for the advertisement. This transparency should be required of all forms of digital advertising--disinformation campaigns often target cultural issues, rather than overt political candidates or policy issues,
- Over the past 18 months, journalists have investigated and shown that problematic

content is not being removed. Further, reporting on targeted advertising and algorithmic quirks on Facebook, Twitter, Instagram and YouTube have demonstrated that the platforms need to be undertaking much more frequent and in-depth investigations on their own technology or should allow independent auditing. Crucially, any proactive changes based on these investigations, such as changing how inauthentic accounts are being categorized or how certain types of content are being down-ranked, need to be transparent.

Conclusion

Current debates on this issue are focused disproportionately on the US, political disinformation, the Facebook newsfeed, and Twitter bots. In fact, this problem of information disorder is global, and includes powerful disinformation related to science, health, religion and ethnicity. In certain places it is leading to protests and violence, and people are losing their lives because of decisions based on inaccurate information. Perhaps most challenging is that increasingly people are receiving information via closed messaging apps, like WhatsApp, or other popular apps like LINE, KakaoTalk, WeChat, Viber or Telegram. When information (often in visual form) is sent and received in these environments, it is impossible to know what is being shared, which makes any fact-checking or debunking initiatives impossible.

One key element of any discussion about interventions is a need to recognize that the techniques and tactics used by those creating disinformation (either for financial or political advantage) are escalating daily. As manipulation technology, especially that powered by AI, becomes more sophisticated, as personal data sets grow, as virtual reality becomes more common, any intervention needs to be able to quickly respond to not just current trends, but also new developments.