Automated tackling of disinformation

STUDY

Panel for the Future of Science and Technology

European Science-Media Hub

EPRS | European Parliamentary Research Service

Scientific Foresight Unit (STOA)

PE 624.278 – March 2019

EN

Automated tackling of disinformation

Major challenges ahead

This study maps and analyses current and future threats from online misinformation, alongside currently adopted socio-technical and legal approaches. The challenges of evaluating their effectiveness and practical adoption are also discussed. Drawing on and complementing existing literature, the study summarises and analyses the findings of relevant journalistic and scientific studies and policy reports in relation to detecting, containing and countering online disinformation and propaganda campaigns. It traces recent developments and trends and identifies significant new or emerging challenges. It also addresses potential policy implications for the EU of current socio-technical solutions.

AUTHORS

This study was written by Alexandre Alaphilippe, Alexis Gizikis and Clara Hanot of EU DisinfoLab, and Kalina Bontcheva of The University of Sheffield, at the request of the Panel for the Future of Science and Technology (STOA). It has been financed under the European Science and Media Hub budget and managed by the Scientific Foresight Unit within the Directorate-General for Parliamentary Research Services (EPRS) of the Secretariat of the European Parliament.

Acknowledgements

The authors wish to thank all respondents to the online survey, as well as first draft, WeVerify, InVID, PHEME, REVEAL, and all other initiatives that contributed materials to the study.

ADMINISTRATOR RESPONSIBLE

Mihalis Kritikos, Scientific Foresight Unit

To contact the publisher, please e-mail esmh@ep.europa.eu

LINGUISTIC VERSION

Original: EN

Manuscript completed in March 2019.

PE 624.278
ISBN: 978-92-846-3945-8
doi: 10.2861/368879
QA-04-19-194-EN-N

http://www.europarl.europa.eu/stoa (STOA website)
http://www.eprs.ep.parl.union.eu (ESMH website)
https://sciencemediahub.eu/ (EPRS website)
http://epthinktank.eu (blog)

Executive summary

The currently ubiquitous online mis- and disinformation poses serious threats to society, democracy, and business. This study first defines the technological, legal, societal, and ethical dimensions of this phenomenon and argues strongly in favour of adopting the terms misinformation, disinformation, and malinformation instead of the ill-defined "fake news".

Next, it discusses how social platforms, search engines, online advertising, and computer algorithms enable and facilitate the creation and spread of online misinformation. It also presents current understanding in why people believe false narratives, what motivates their sharing, and how they impact offline behaviour (e.g. voting).

Drawing on existing literature, the study also summarises state-of-the-art technological approaches to fighting online misinformation. It follows the AMI conceptual framework (Agent, Message, Interpreter), which considers the origin(s) and the impact of online disinformation alongside the veracity of individual messages.

This is complemented by a brief overview of self-regulation, co-regulation, and classic regulatory responses, as currently adopted by social platforms and EU countries. User privacy and access to data for independent scientific studies and development of effective technology solutions are also discussed. In addition, the study summarises civil society and other citizen-oriented approaches (e.g. media literacy).

We have included a roadmap of initiatives from key stakeholders in Europe and beyond, spanning the technological, legal, and social dimensions. It is complemented by three in-depth case studies on the utility of automated technology in detecting, analysing, and containing online disinformation.

The study concludes with the provision of policy options and makes reference to the stakeholders best placed to act upon these at national and European level. These include support for research and innovation on technological responses; improving transparency and accountability of platforms and political actors over content shared online; strengthening media and improving journalism standards; supporting a multi-stakeholder approach, involving also civil society.

Table of contents

Table of figures

Table of tables

1. Problem Definition and Scope

The past few years have heralded the age of ubiquitous mis- and disinformation - often referred to as fake news - which poses serious questions over the role of social media and the internet in modern democratic societies. Topics and examples abound, ranging from the UK "Brexit" referendum and the 2016 US presidential election to medical misinformation (e.g. miraculous cancer cures). While the strong presence of state-backed disinformation campaigns during the recent US, French, and Italian elections has now been established (Faris et al, 2017; Storyful, 2017; Allcott & Gentzkow, 2017), a complete understanding of their true reach and impact on voter behaviour and election outcomes is still lacking. This has led the EU High Level Expert Group (HLEG) on fake news and disinformation (Bunning, 2018) to conclude that: "Special attention should be paid to the threat represented by disinformation aimed at undermining the integrity of elections (local, national or EU elections)."

Moreover, the corrosive effect of online disinformation is much wider reaching than elections alone. Recent research reported on by the UK Royal Economic Society (RES, 2018) found that social media bots are influencing stock market performance, based on a study of more than 49 million tweets mentioning the names of 55 large FTSE100 companies. The findings also suggest that small investors are more likely to fall victim to bot tweets. Other targets have included women (e.g. the investigative journalist – Jessikka Aro and prominent female journalists and political leaders (RSF, 2018).

Therefore, online disinformation and propaganda have not only a societal and personal cost, but also can lead to significant economic losses. Forbes reported (Rapoza, 2018) that in $130 billion in stock value was lost when the Associated Press (AP) Twitter account was hacked and a post was made claiming there was an "explosion" that injured Barack Obama. Although stock prices recovered, this points to how disinformation on social media can impact high-frequency trading algorithms that rely on text to make investment calls. Similarly, the French construction company Vinci became a target in 2016 of a false press release (Agnew, 2016) which caused share price to go down by 19% (a loss of more than 6 billion euro) before recovering by close of trading.
Researchers have also started encountering the so called misinfodemics, where health-related misinformation is facilitating the spread of diseases such as Ebola and measles. (Gyenes & Mina, 2018).

> "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)

The reasons behind the sudden and ubiquitous rise in online mis- and disinformation are complex and lie at the intersection of:

1. **Online propaganda and for-profit disinformation sites:** State-backed (e.g. Russia Today), ideology-driven (e.g. misogynistic or Islamophobic), or for-profit clickbait websites and social media accounts are all engaged in spreading misinformation, often with the intent to deepen social division and/or influence key political

outcomes. However, they should not be regarded as the sole source of online disinformation[1].

2 **Post-truth politics**, where politicians, parties and governments frame key political issues in propaganda, instead of facts. Misleading claims are repeated, even when proven untrue by media or independent fact checkers. This has a highly corrosive effect on public trust.

3 **Partisan media**: Today's highly competitive online media landscape has resulted in poorer quality journalism and worsening opinion diversity, with misinformation, bias and factual inaccuracies routinely creeping in. Many outlets also resort to highly partisan reporting of key political events, which, when amplified through social media echo chambers, can have acrimonious and divisive effects (Moore & Ramsay, 2017).

4 **Polarised crowds:** As more citizens turned to online sources as their primary source of news, the social media platforms and their advertising and content recommendation algorithms have enabled the creation of partisan camps and polarised crowds, characterised by flame wars and biased content sharing, which in turn, reinforces their prior beliefs (typically referred to as *confirmation bias*).

5 **Technological affordances of advertising algorithms and social platforms**: technologies such as search engine optimisation, personalised social feeds, and micro-targeted advertising have been used widely to promote, spread, and monetise online misinformation and propaganda. Other contributing factors are anonymity of advertisers, lack of algorithmic transparency, and lack of oversight on third-party use of users private data.

1.1. Public Perception of the Problem

Now let us first examine how ubiquitous are mis- and disinformation online and to what extent citizens feel affected by the phenomenon and capable of identifying unreliable information online. Firstly, both scientists and journalists have now gathered wide-ranging, well documented evidence that online misinformation and disinformation affect all social media platforms and mobile applications, with the key ones being Twitter, Facebook, YouTube, Reddit, 4Chan/8Chan, WhatsApp, Discord, Telegram, etc. This is supported by the findings of the public consultation on fake news and online disinformation[2] that 74% of the respondents encountered such unreliable content primarily through social media and messaging apps. This aspect will be discussed further in section 2.

The agents behind the disinformation campaigns are diverse and include for-profit fake news sites, hyper-partisan sites, alt-right/alt-left communities, states (e.g. Russian or Chinese propaganda), politicians/parties, media organisations, fake think tanks, and sometimes individuals.

Moreover, online mis- and disinformation are used increasingly to manipulate public opinion, increase societal, religious, and cultural divisions, and influence elections.

This trend is accelerating fast, with the number of countries affected by organised social media manipulation rising from 28 in 2017 to 48 in 2018 (Bradshaw & Howard 2017; Bradshaw & Howard, 2018). These studies found evidence of political parties and candidates spreading disinformation through social media platforms, with the intent to manipulate voter opinions and thus influence election outcomes. Fake social media accounts, trolls, and bots are also used to artificially inflate popularity of certain content (e.g. make some hashtags trend on Twitter), promote extremist opinions, or distort conversations.

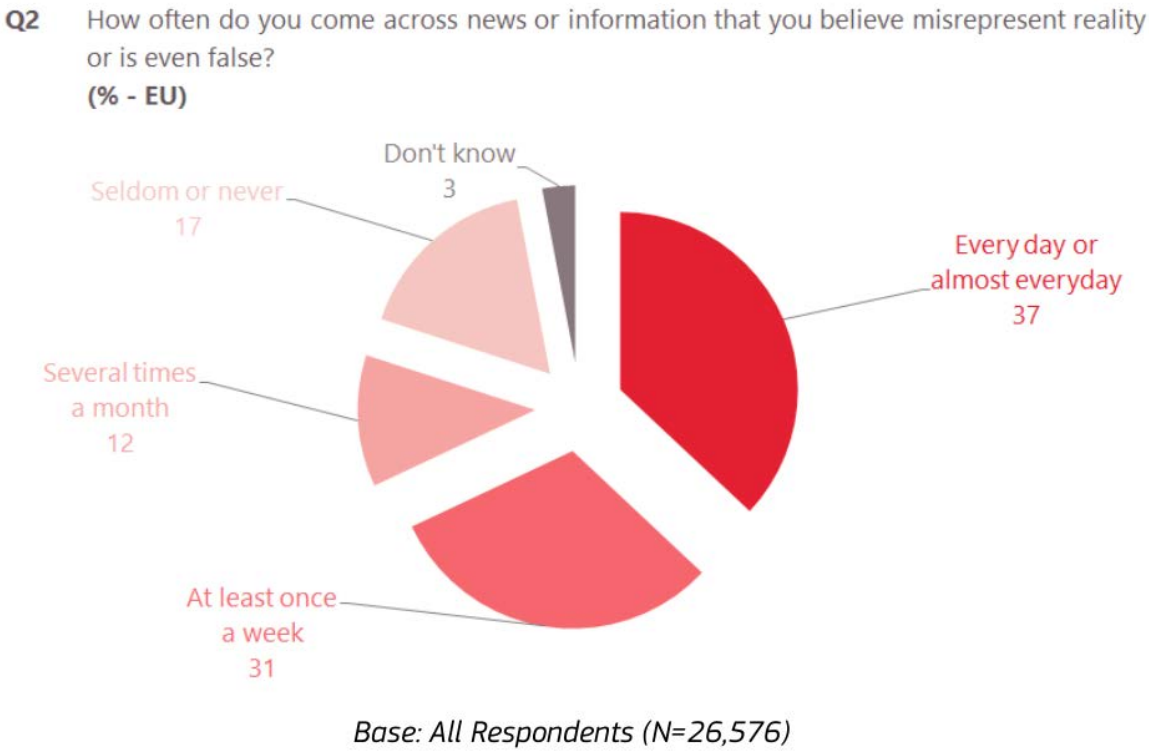
[1] https://www.nytimes.com/2018/01/25/opinion/russian-trolls-fake-news.html

[2] http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=51810

Amongst the 48 countries included in the 2018 study (Bradshaw & Howard, 2018), eight are from the European Union: Austria, the Czech Republic, Germany, Hungary, Italy, Netherlands, Poland and the United Kingdom. The actors involved in social media manipulation vary from country to country and include government organisations (CZ, DE, HU, and UK), parties and politicians (AT, CZ, DE, IT, NE, PL, UK), and private contractors (AT, PL, UK). The most widely used social media manipulation strategies use fake accounts and bots (AT, DE, HU, IT, NE, PL, UK), primarily to carry out attacks on the opposition, post distrating messages, or engage in trolling and harassment. This is achieved not only through posting replies or comments, but also through the creation of new content such as fake videos, blogs, memes, or websites. The study also observed an "*increasing use of paid advertisements and search engine optimization on a widening array of Internet platforms*".

A 2018 Eurobarometer survey (Eurobarometer, 2018) of over 26,000 EU citizens in the 28 EU member states found that over 83% considered online disinformation a threat to democracy, with consistent results across member states. At least half of the respondents also said that they encounter online disinformation and "fake news" at least once a week, with the highest number of reports (between 73% and 78%) coming from Spain, Hungary, Croatia, Poland, France, Greece, and Slovakia.

Figure 1: Eurobarometer 2018 Survey, Frequency of coming across to information misrepresenting reality or is even false

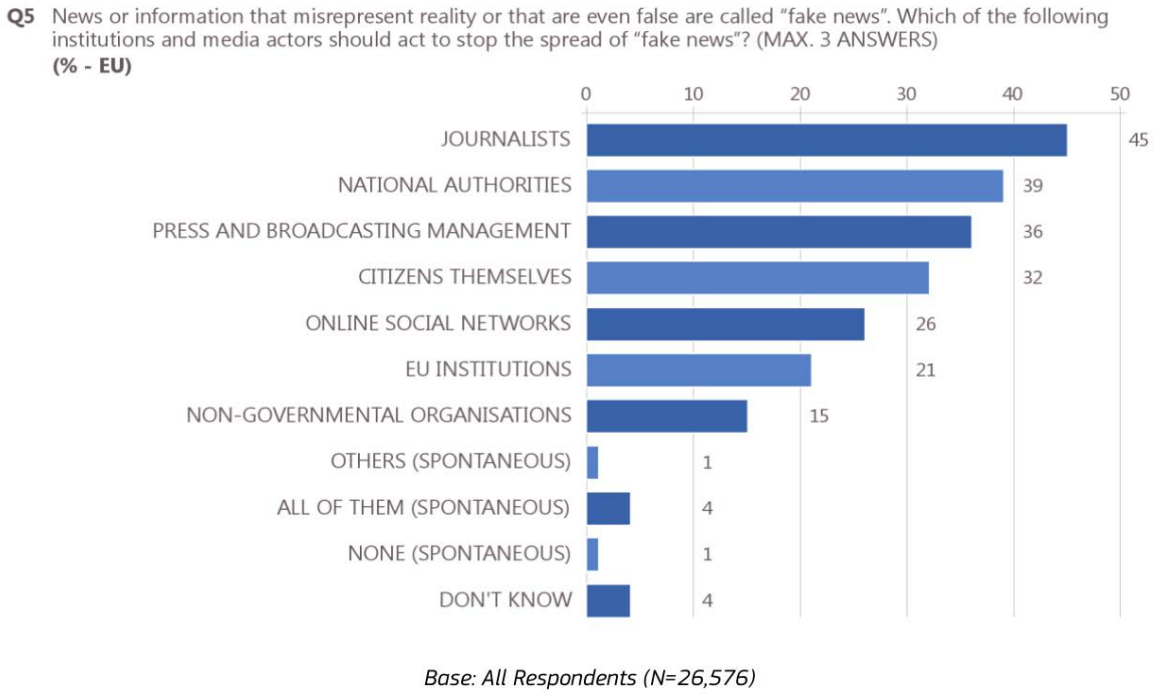


Source: Eurobarometer 2018

The Eurobarometer findings are in line with those of the public consultation on fake news and online disinformation (EC public consultation, 2018) where around 70% of respondents reported encountering online disinformation daily or weekly, with a further 12.5% - monthly. This consultation attracted responses from 2,784 individuals and 202 legal organisations and journalists, with most replies from Belgium, France, the United Kingdom, Italy and Spain, and high participation in Lithuania, Slovakia and Romania. The consultation findings are complementary to the Eurobarometer citizen survey, as the former involved a wide range of legal entities, including private

news media companies, civil society organisations, the online platforms themselves, research and academia, and national and local public authorities.

Figure 2: Eurobarometer 2018 Survey, Institutions and media actors that should act to stop the spread of "fake news"



Source: Eurobarometer 2018

The 2018 Eurobarometer study also asked citizens who are the actors that should be responsible for stopping the spread of online disinformation, with 45% naming journalists, 39% - national authorities, and 36% - the press and media (Eurobarometer, 2018). Interestingly, only just over a quarter of respondents (26%) named online social networks as having to be made responsible for stopping "fake news". These findings contrast to those of the Reuters News Institute (Newman *et al*, 2018), who reported in their own 2018 survey that: "Most respondents believe that publishers (75%) and platforms (71%) have the biggest responsibility to fix problems of fake and unreliable news." This is attributed to respondents being primarily concerned in this case with biased or inaccurate news from the mainstream media, as there is a potential lack of public awareness of the existence of other harmful kinds and sources of mis- and disinformation and their negative impact on democracy and society.

The recent 'Yellow Jackets' crisis in France highlighted (mis)information practices around demonstrations and social protesters. While there are concerns around social media manipulations, on-site interviews conducted by journalists from Franceinfo[3] highlighted the distrust protesters had towards traditional media organisations.

[3] https://mobile.francetvinfo.fr/economie/transports/gilets-jaunes/entre-medias-collabos-et-merde-de-facebook-comment-les-gilets-jaunes-s-informent-sur-les-ronds-points_3096513.html#xtref=acc_dir

An onsite interview from Le Monde[4] also highlighted certain lack of media literacy regarding online platforms, with citizens not diffrentiating clearly between news editors and social media platforms, as well as equating social platforms and the internet :

> *"It is not nice to spread false information because this is not why we pay for newspapers (…) Facebook and the Internet are the same, we pay for it every month, we should have real information on it."*

1.2. Definitions and Conceptual Frameworks

This report argues strongly in favour of adopting the terms *misinformation, disinformation,* and *malinformation* instead of the more ill-defined "fake news". This section defines each of these terms and introduces the First Draft *information disorder recent* theoretical frameworks (Wardle, 2017), which describes them in more detail. First, however, let us present the rationale behind this choice.

1.2.1. Terminology

The term "fake news" is increasingly regarded as inadequate both by scientists as it is too nebulous and imprecise. For instance, a recent study of thirty four academic papers (Tandoc et al, 2017) concluded that "fake news" encompasses a wide range of phenomena: news satire, news parody, fabrication, manipulation, advertising, and propaganda. Moreover, "fake news" is misleading, as it is also increasingly used by politicians "*to describe news organisations whose coverage they find disagreeable. In this way, it's becoming a mechanism by which the powerful can clamp down upon, restrict, undermine and circumvent the free press*." (Wardle & Derakhshan, 2017).

The term "fake news" has also been rejected by the High Level Expert Group (HLEG) appointed by the European Commission to advise on fake news and online disinformation (p.10; Buning et al, 2018), as well as the non-profit coalition First Draft News, which is dedicated to improving skills and standards in the reporting and sharing of online information (Wardle, 2017). The UK Parliamentary Inquiry into disinformation and "fake news" makes the following recommendation:

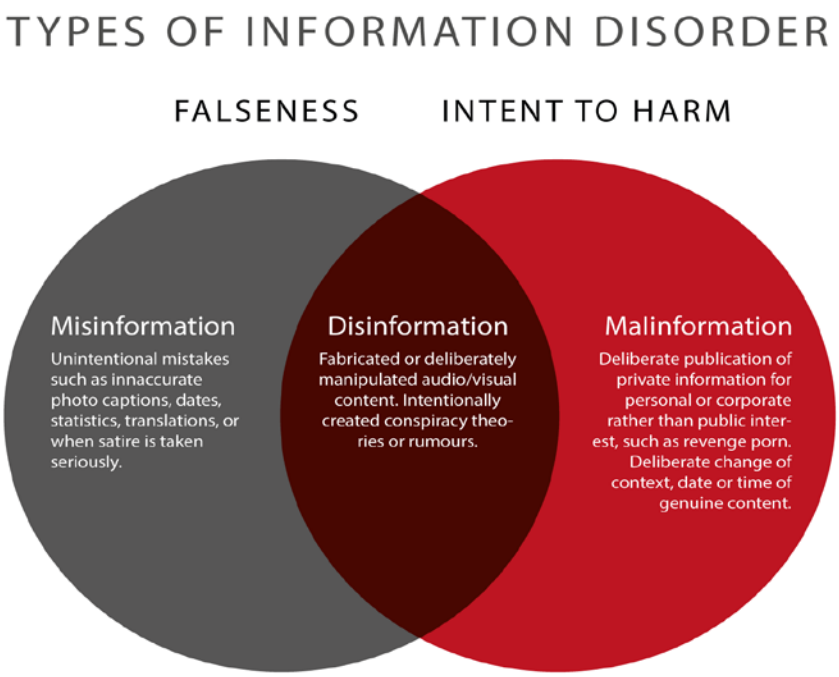
> *"The term 'fake news' is bandied around with no clear idea of what it means, or agreed definition. The term has taken on a variety of meanings, including a description of any statement that is not liked or agreed with by the reader. We recommend that the Government rejects the term 'fake news', and instead puts forward an agreed definition of the words 'misinformation' and 'disinformation'. With such a shared definition, and clear guidelines for companies, organisations, and the Government to follow, there will be a shared consistency of meaning across the platforms, which can be used as the basis of regulation and enforcement."*

Similarly, a French report (Jeangène Vilmer et al, 2018) by the French Policy Planning Staff (CAPS) of the Ministry for Europe and Foreign Affairs and the French Institute for Strategic Research (IRSEM) of the Ministry for the Armed Forces argues for "*substituting the vague and controversial notion of "fake news" for the more precise term, "information manipulation." The latter is understood as the intentional and massive dissemination of false or biased news for hostile political purposes.*"
Therefore, instead of "fake news" we adopt the *information disorder* theoretical framework (Wardle, 2017; Wardle & Derakshan, 2017), which defines three types of false and/or 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.

[4] https://www.lemonde.fr/politique/article/2018/12/13/gilets-jaunes-sur-les-ronds-points-la-chasse-a-l-info-et-la-tentation-du-complot_5396917_823448.html

Figure 3: Types of Information Disorder



Source: (Wardle & Derakshan, 2017)

Others have defined disinformation specifically in the context of elections, as "*content deliberately created with the intent to disrupt electoral processes*" (Giglietto, Iannelli, Rossi, & Valeriani, 2016). It must be noted, however, that this definition is potentially quite narrow.

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

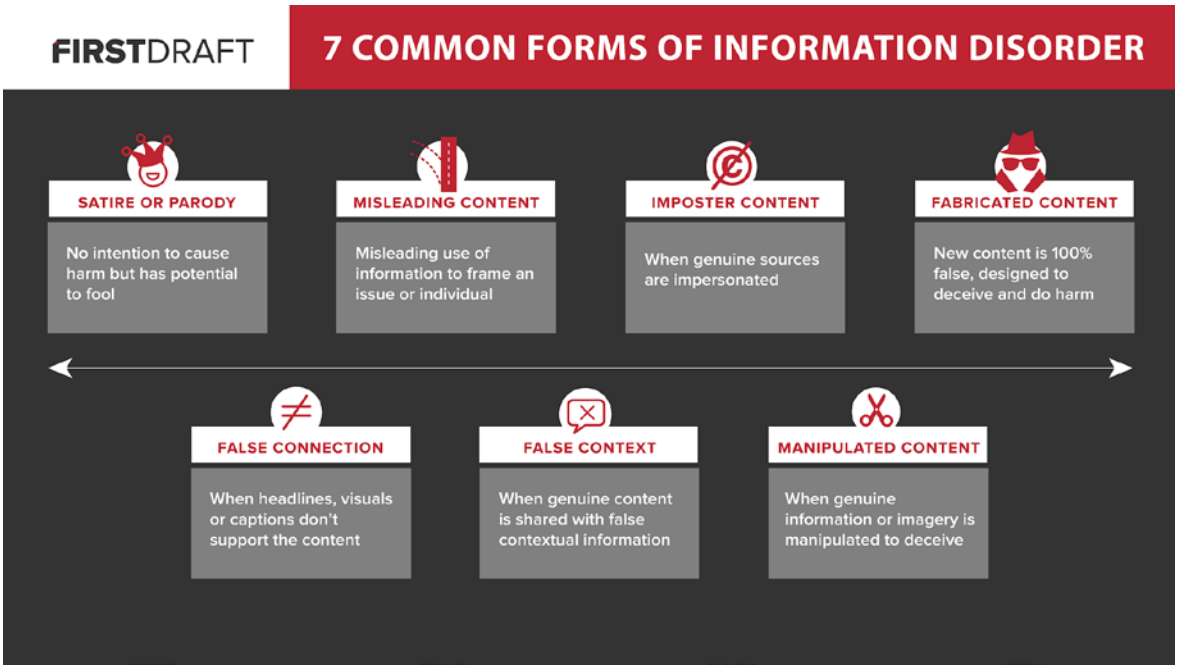
> *"Disinformation….includes all forms of false, inaccurate, or misleading information designed, presented, and promoted to intentionally cause public harm or profit." (Buning et al, 2018).*

In the context of online disinformation, this is typically amplified through social media and other online platforms.

It must also be noted that, in practice, it is often hard to distinguish mis- from dis-information, as the intention of the information source or amplifier may not be easily discernible, not only by algorithms, but also by human readers (Jack, 2017; Zubiaga et al, 2016). Therefore, mis- and dis-information are sometimes addressed as if they are interchangeable.
The *information disorder* framework also discusses a typology of the different kinds of mis- and dis-information that have been found to circulate online (Wardle, 2017):

- Satire or parody: no intention to cause harm but with potential to fool.
- Misleading content: misleading use of information to frame an issue or an individual.
- Imposter content: when genuine sources are impersonated.
- Fabricated content: news content is 100% false, designed to deceive and do harm.
- False connection: when headlines, visuals or captions do not 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.

Figure 4: Categories of Information Disorder



Source: Claire Wardle, 2017

Disinformation is also related to propaganda which "*...is neutrally defined as a systematic form of purposeful persuasion that attempts to* ***influence the emotions, attitudes, opinions, and actions of specified target audiences*** *for ideological, political or commercial purposes through the controlled transmission of* ***one-sided messages (which may or may not be factual)*** *via mass and direct media channels." (Nelson 1996: p232-233)*

Deceitful propaganda techniques (e.g., selective use of facts, unfair persuasion, appeal to fear), however, are employed much more widely, e.g. in anti-EU campaigns, post-truth politics[5], ideology-driven web sites (e.g., misogynistic or Islamophobic), and hyperpartisan media, often with the intent to **deepen social division**, **increase polarisation**, **influence public opinion**, or **impact key political outcomes**.

1.2.2. Propaganda techniques

What makes online disinformation even more challenging for citizens to identify and protect themselves against, is not just its ubiquitous online presence, but also the way it exploits propaganda techniques around **linguistic, cultural, and national differences, to create new social barriers and divisions, as well as causing financial and personal damages**.

Not only are disinformation and propaganda campaigns ubiquitous, but they are also **cross-border and cross-language**, often **using different persuasion techniques and sources**. For instance, in the UK, anti-immigrant narratives are mainly being pushed through foreign sources close to the US alt-right. At the same time, in France, pro-Kremlin disinformation communities have focused on protests and in Italy - on Nigerian immigration.

5 Post-truth was the word of 2016 by Oxford Dictionaries, defined as "relating to or denoting circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief".

Key **propaganda persuasion techniques** employed in making online disinformation more credible include (Faris *et al*, 2017):

- **Ad hominem** – statements attacking the opponent, rather than their argument.
- **Ad nauseam** – constant repetition of a slogan or an idea, to make it mainstream and accepted.
- **Cherry-picking** – facts and opinions are unequally represented.
- **Appeal to authority** – citing of prominent figures or sources, sometimes out of context.
- **Appeal to fear, anger, or prejudice** – building support through fear, anger, prejudice, or stereotyping.
- **Deception** – deceptive/misleading presentation of facts or viewpoints.
- **Humour** – used to discredit democratic values or politicians; involves structural manipulation of message content (NATO StratCom COE, 2017).

The rationale behind employing propaganda techniques in online disinformation campaigns is to enhance the **credibility** of the message. It must be emphasized that the **credibility of a claim/content (i.e. message) is separate from its veracity**, since the former is about subjective perception of whether a claim is credible, whereas verification is about evidence-based, objective assessment.

Likewise, propaganda techniques are employed to enhance the credibility of an information source. For example, false amplifiers such as hyperpartisan media and sock puppets often repackage and/or republish content in an attempt to give it credibility and gain acceptance through familiarity. This can be achieved through retweet and mention patterns between such false amplifiers, e.g. in the 2016 US Presidential Election (Faris *et al*, 2017). Social scientists (Starbird, 2017), however, have recently discovered the much harder to detect practice of *information laundering*, where text is reused between sites to give the impression that a large number of independent sources are reporting in different ways on the same "facts". Section **Error! Reference source not found.** discusses this is more detail.
Therefore, it is important to include awareness of content and source credibility and propaganda persuasion techniques as part of media literacy initiatives, as well as fund quantitative large-scale studies of their use in disinformation campaigns.

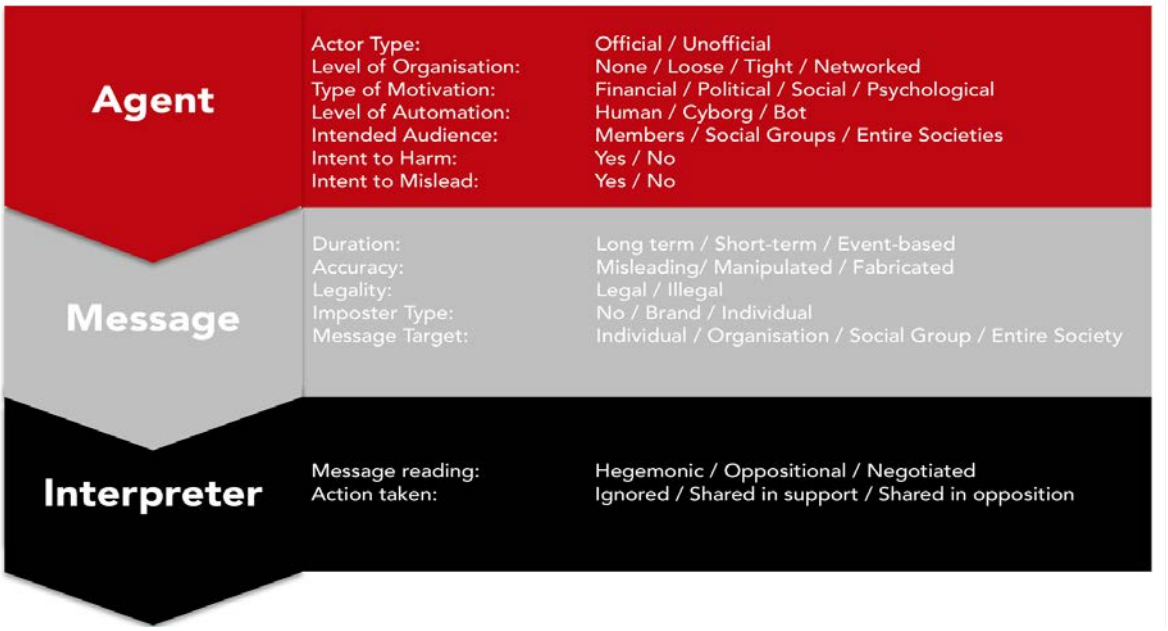
1.2.3. Conceptual Framework

In addition, as argued in (Wardle & Derakhshan, 2017) in order to understand online disinformation, its spread, and effects, it is necessary to adopt a more sophisticated conceptual framework, which distinguishes between:

- the **Agents** involved **-** who are the authors or distributors of disinformation and what is their motivation;
- the **Message** - the false content that is being spread, how it is expressed, and the techniques used to enhance its credibility;

- the **Interpreters** - who are those reading the disinformation and what are its effects on their beliefs and actions.

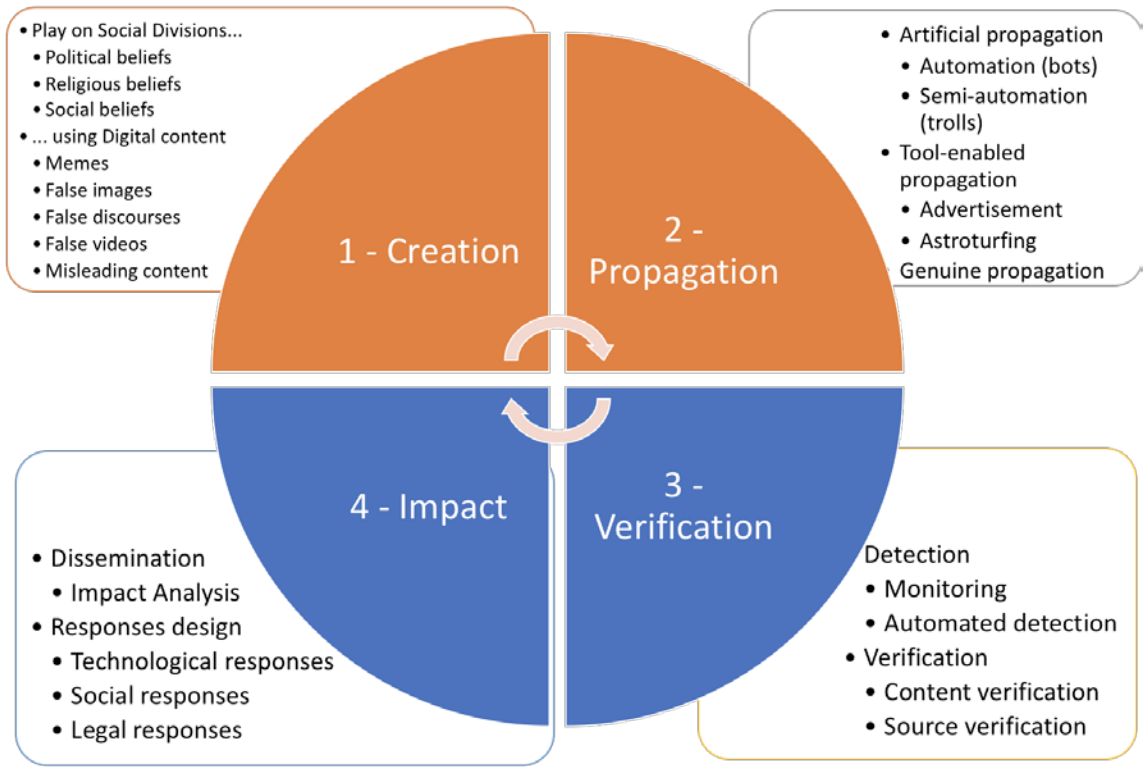
Figure 5: The Agent, Message, Interpreter Conceptual Framework



Source: (Wardle & Derakshan, 2017)

This report adopts this Agent, Message, Interpreter (AMI) conceptual framework, as it models not only message veracity and formulation, but also emphasizes the importance of understanding the origins of a given disinformation campaign, who are its key propagators, what is the extent of its reach, and whether they have impacted offline behaviour (e.g. voting).
In order to better understand the disinformation cycle, we summarised it in the following diagram. In orange, we can see how disinformation is created and amplified, in blue the efforts that needs to be designed to counter it.

Figure 6: Disinformation Lifecycle



Source: own elaboration

2. Social Platforms and Other Technological Factors Helping the Spread of Online Disinformation

This section will discuss in more depth the socio-technical causes of online disinformation, with focus specifically on the role of algorithms throughout the complete life-cycle. First, we discuss the role of social platforms and web search engines, as the vehicles for creation, publication, and monetisation of online disinformation. Next, we look into amplification strategies, including advertising, bots, and the reasons why some genuine users believe and share misinformation. The section concludes with a discussion of the newly emerging ways in which artificial intelligence and deep learning are being harnessed to produce highly credible online disinformation.

2.1. Social Platforms and Web Search Engines: Algorithms, Privacy, and Monetisation Models

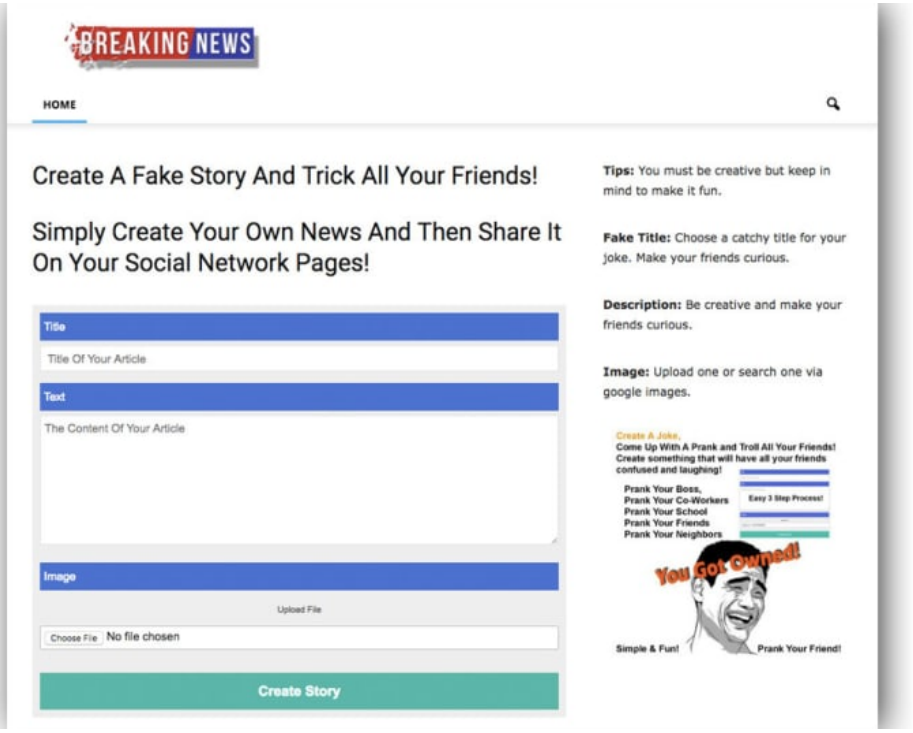
As numerous studies and polls have shown (e.g., Duggan, 2015; Newman, 2011; Ireton & Posetti, 2018; Newman *et al*, 2018), social platforms (e.g. Facebook, Twitter, Reddit) and web search engines (e.g. Google) are now firmly established as major sources of news and information online for an ever growing number of citizens worldwide[6].

[6] There are differences in the level of adoption by countriesand age groups (Newman *et al*, 2018; Mitchell *et al,* 2016).

Consequently, billions of users worldwide[7] have become targets of online disinformation and propaganda campaigns through these online platforms and technology.

When it comes to disinformation, the weakness of online sites and social platforms is also their main strength, i.e. the easy way in which their users can create, publish, share, and engage with online content (e.g. express likes, anger, sadness). As shown in Figure 7, the process of writing a false story, adding an image, and sharing it on Facebook is made even simpler through dedicated web sites. A typical strategy is to create a fake headline, coupled with misleading images or videos, to create an emotionally provocative story, which entices social media users to share and click, on platforms like Facebook (Silverman, 2017), Twitter, and WhatsApp (McLaughlin, 2018).

Figure 7: Web form for creating a false story and sharing it on Facebook



Source: (Silverman, Lytvynenko & Pham, 2017)

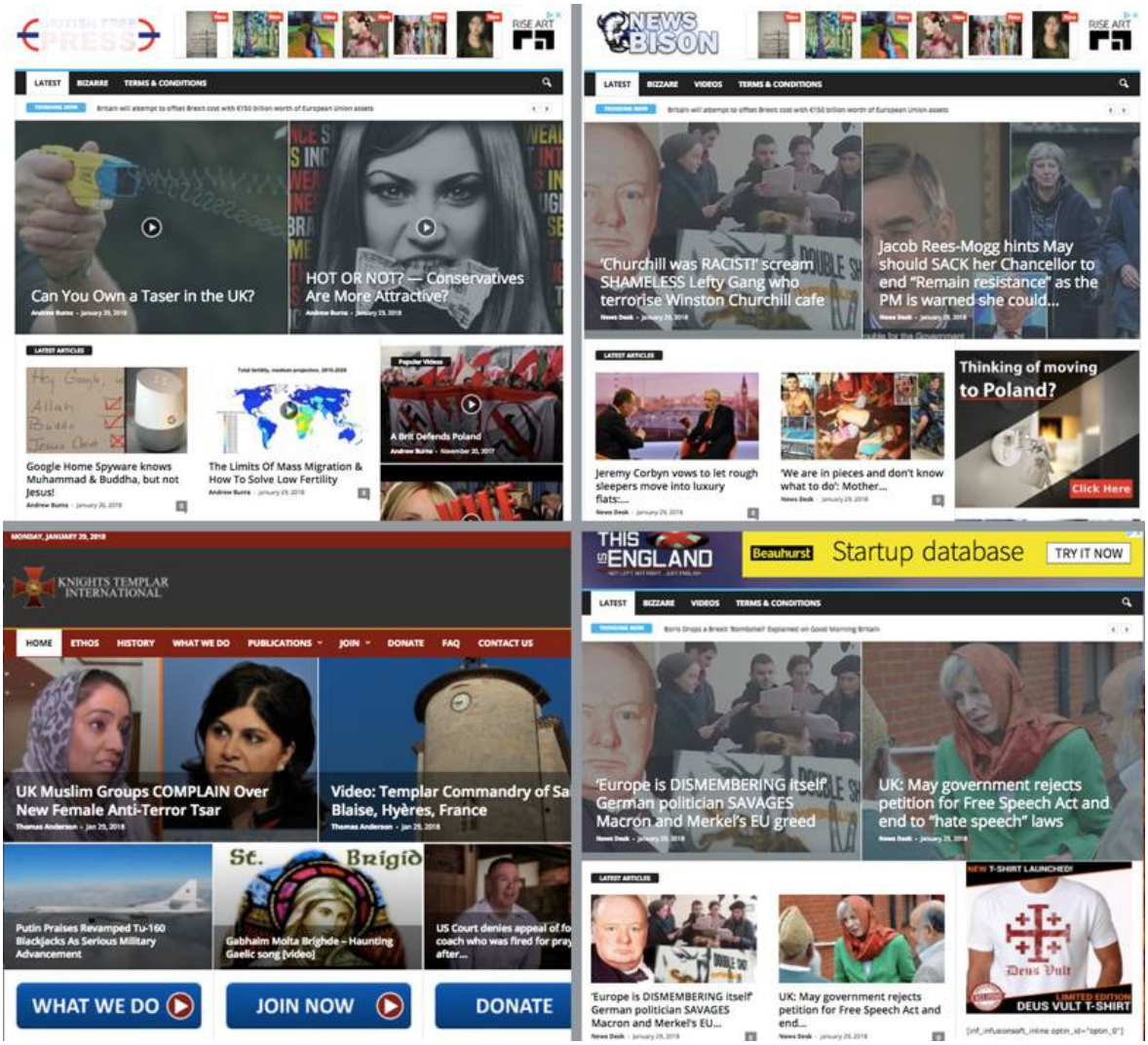
Many successful disinformation campaigns tend to harness the emotional power of videos by posting them first on YouTube (Albright, 2018a) and then sharing the content through Facebook and similar channels. Their reach and impact is then amplified by the platforms' trending topics algorithms, which frequently promote misinformation (e.g. conspiracy theories) during major events and crises (Lapowski, 2018).

In addition, successful propaganda and disinformation campaigns often leverage a network of websites that post disinformation or distorted, out-of-context news stories, designed specifically to invoke emotional response and online engagement. For example, Figure 8 (Reynolds, 2018) shows some examples from a network of far-right news sites (operated from Eastern Europe), which are then shared and amplified through thirteen related Facebook pages, targeting British Facebook users. As shown by (Reynolds, 2018), this social media strategy is typically extremely successful as, in this case, it attracted over 2.4 million Facebook likes (more than any other UK political page on

7 Facebook alone has over 2 billion montly active users in 2018: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

Facebook), despite the fact that the political group behind them has only around 1,000 members offline.

Figure 8: Examples from a network of far-right news sites shared and amplified through Facebook pages



Source: (Reynolds, 2018)

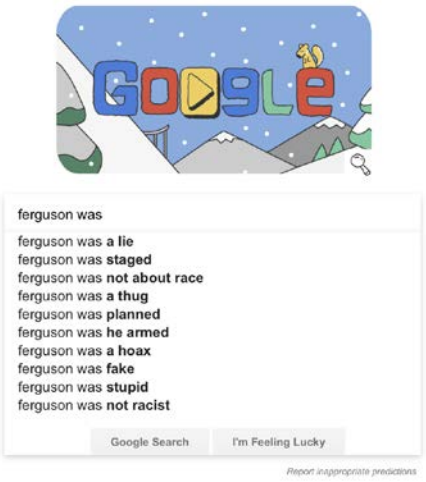
Following initial scepticism, the social platforms have now accepted the need to contain the spread of online disinformation and started implementing changes to their algorithms. Nevertheless, in 2017 the top fifty "fake stories" as identified by BuzzFeed (Silverman, Lytvynenko, & Pham, 2017) attracted over 23.5 million engagements (i.e. shares, likes, and comments), which is 2 million more than in 2016. One of the earliest measures, which Facebook implemented, for combating online disinformation was to rely on external fact checking organisations, such as Snopes, PolitiFact, and FactCheck.org. However, fact checking on its own is not sufficient, since debunking stories receive significantly less Facebook engagement and spread (just 0.5% for the top 50 stories above). In response, Facebook are now experimenting, amongst other new techniques, with making debunkings more prominent in the News Feed.

Another recent challenge is that disinformation methods are evolving rapidly, moving from one platform to another. Fake content campaigns are often prepared and trialled on less visible platforms such as 4chan or Discord, and then the successful ones are being spread on the

mainstream ones. According to Storyful, a US Irish-born social media news agency, recent use cases like the Parkland shooting in Florida in February 2018, have shown "the importance of understanding fringe social networks" and of corroborating claims and sources manipulation through multiple independent channels (Storyful, 2017). Identifying quickly how, when and why disinformation networks are created on underground networks was key to help quell misinformation before it spreads too widely during the French 2017 Presidential election, according to Benjamin Decker, Head of Research at Storyful.

Journalists and scientists are increasingly having to study disinformation on closed social platforms, i.e. those with end-to-end encryption such as Whatsapp, Telegram, Signal, and Discord. This is particularly difficult, as users are typically communicating anonymously through groups, where membership is heavily controlled. Even though the extent of disinformation spread on these platforms is currently hard to establish, its significant negative impact in terms of physical harm and violence have been documented. For instance, cyber-police in India has estimated that two dozen people have been killed since May 2018, as a result of false rumours spread via WhatsApp (McLaughlin, 2018).

Figure 9: Disinformation in Google Search Suggestions



Source : (Albright, 2018)

Last but not least, the proliferation of online mis- and disinformation is also affecting web search results, which further hinders users in finding and reading trustworthy online information. For example, as recently as February 2018, Google search suggestions were showing misleading or false information (Albright, 2018) especially in the top one to five suggestions, e.g., feminists are crazy, police are evil, Ferguson was a lie (see Figure 9 for other suggestions), Michael Brown was a thug.

Google has invested significantly in the past couple of years in preventing false content from appearing in its search algorithms through using fact checking and trust indicators (e.g. ClaimReview[8] markup), as well as promoting content from more authoritative publishers and direct feedback tools. It is also investing heavily in the Google Digital News Initiative[9], which has funded many media-led projects aimed at developing new technologies and tools for content verification and disinformation detection. While definitely being steps in the right direction, these measures are far from being sufficient, since online disinformation campaigns and their orchestrators are continuously evolving in sophistication. This is particularly evident during breaking news events, such as the February 2018 school shooting in Florida, when CNBC reported that Google News searches soon after the event were returning false reports, alongside legitimate content (Salinas, 2018). The problem is not confined to news search, with YouTube's top trending video on the shooting also being a conspiracy theory (Salinas, 2018). This is due to lack of human oversight over the algorithmically curated trending topics - a problem which is also faced by Facebook and its own trending news section.

[8] https://schema.org/ClaimReview

[9] https://newsinitiative.withgoogle.com/dnifund/

Next, let us examine the most commonly used mechanisms for creating, promoting, and spreading online disinformation.

2.1.1. Fake Profiles and Groups

Social platforms are plagued by fake profiles and groups, which are created to post online disinformation and amplify its spread and perceived importance (e.g. make it trend on Facebook or Twitter). The programming interfaces of Facebook, Twitter, Instagram and other platforms make it possible to do this fully automatically and at scale (Weedon, Nulan & Stamos, 2017). To give just one example, a Macedonian man created and ran in a coordinated fashion over 700 Facebook profiles, spreading online disinformation for monetary gain (Silverman, 2017).

Fake accounts can also be created and purchased to act as fake followers and thus artificially inflate the importance and popularity of genuine accounts (Confessore *et al*, 2018). At the same time, fake accounts created for other purposes often follow popular genuine accounts, in order to be perceived as credible by the social platform and avoid detection. Affected accounts have included media outlets, e.g. 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.

Groups, in particular, are playing an increasingly important role in disinformation campaigns. Some are dedicated to astroturfing, i.e. creating artificial appearance of "grass-roots" support (recall the far-right British Facebook pages discussed above) and are 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).

The way in which fake profiles and groups amplify artificially online disinformation is through frequent and coordinated content posting, commenting, sharing, and reactions (e.g. Facebook likes, Twitter favourites) (Weedon, Nulan & Stamos, 2017). Genuine groups can also be flooded with posts and shares from fake accounts, in order to divert the discussion focus and saw dissent.

2.1.2. Online Advertising and Clickbait

Online advertising has been used extensively to make revenues for junk news sites, as they receive payments when the adverts are shown alongside the fake content. Owners of such web sites have claimed to earn between $10,000 and $30,000 per month from online advertising. One such example is the CEO of a company called Disinfomedia, which operated a group of such misinformation web sites (Sydell, 2016).

Traffic is driven to these sites from social platforms through a combination of organic clicks attracted by *clickbait posts* and *promoted posts* (which themselves could be clickbait in nature).
A clickbait post is designed to provoke emotional response in its readers, e.g. anger, compassion, sadness, and thus stimulate further engagement by following the link to the webpage, which in turn generates ad views and revenues for the website owner. Clickbait typically omits key information about the linked content (Chakrabarti et al, 2017), in order to create a curiosity gap (Loewenstein, 1994) and thus entice users to click. The sensationalist and emotive nature of social media clickbait has been likened to tabloid journalism and found to provide an "alternative public sphere for users drifting away from traditional news" (Chakrabarti et al, 2017). Clickbait tweets have been found to retain their popularity for longer and attract more engagement, as compared to non-clickbait

tweets (Chakrabarti et al, 2017). These characteristics make them highly successful in propagating organically online mis- and disinformation through the social networks of genuine users, as well as being adopted in many highly-viewed adverts.
Promoted posts on Facebook and Twitter are marked as advertisements and can be reposted, liked, replied to, etc. as any normal post can. Advertisers are billed by the social platform based on the amount of engagement generated, e.g. likes, shares, clicks and views.
In many cases advertisers can choose which users will see the promoted post, based on information such as geographic location, gender, interests, device type, or other specific characteristics. When adverts are targeted at a very narrow set of users, with very specific profiles, the practice is called *micro-targeting* (see Section 2.1.3).

As users visit web sites and social media platforms, they are willingly or implicitly giving away invaluable **personal information**, e.g. their location, mobile device used, IP address, browsing history, social media engagements (e.g. likes and shares). Their social profiles are also data rich and include further personal data, including birthday, relationship status, family members, workplace, education history, etc. Moreover, users' online behaviour is continuously tracked through technology such as cookies, tracking scripts and images, display ads, and CSS/HTML code. All this data is what enables the automated profiling of users and the resulting micro-targeted delivery of personalised advertising and/or content.

Platforms try to control what is being promoted by prohibiting adverts for illegal goods and services including drugs; those containing hate speech, abusive content, and violence; or posts containing offensive or inflammatory content. Screening is partly algorithmic and partly based on user reports. The lack of transparency and regulation of online advertising on Google, Facebook, Twitter, and other online platforms has caused very significant concerns lately. Firstly, platforms need to **disrupt the advertising-based monetisation of online disinformation**, as recommended by the EU High Level Expert Group on fake news and online disinformation:

> *"Platforms should adapt their advertising policies, including adhering to "follow-the-money" principle, whilst preventing incentives that leads to disinformation, such as to discourage the dissemination and amplification of disinformation for profit. These policies must be based on clear, transparent, and non-discriminatory criteria" (HLEG report, 2018)*

The second key concern arises especially in relation to **political advertising** and its potential to unfairly influence voters. The issue of election interference through online advertising on these platforms came to the fore following investigations into Russian influence in the 2016 US presidential elections. For instance, Twitter reported that Russia Today and related account promoted just under 2,000 election-related tweets that violated the platform's advertising policies and thus generated around 53.5 million impression on U.S. based users (Edgett,2017). Similar Russia-led campaigns were carried out on Facebook, Google, YouTube, and other platforms. Russia, however, is not the only state trying to influence elections, with China being another recent example.

State-sponsored propaganda aside, a 2017-2018 UK Parliamentary Inquiry into disinformation and "fake news" uncovered serious **issues around digital campaigning by political organisations** during the 2016 UK EU membership referendum (DCMS report, 2018). Likewise, there are outstanding questions over the way the Trump presidential campaign used over 5.9 million Facebook adverts to achieve maximum impact and engagement (Frier, 2018).
Consequently, the EU HLEG report (2018), as well as UK and many other European policy makers (DCMS report, 2018) and independent fact checking organisations (FullFact, 2018) have strongly advocated that paid-for political advertising data must be made publicly accessible for search and include information about the advertising organisation, country of origin, and who are the ads targeted at.

In response, Google, Facebook, Twitter, and YouTube have started tightening their advertising policies, as well as implementing automated methods to screen out adverts that violate them. Unfortunately, these algorithms are yet to reach sufficient accuracy, with adverts promoting spam, misinformation, or phishing content still in circulation, e.g. (Silverman, 2018).
Google, Facebook, and Twitter have also started making initial steps towards transparency in online political advertising:

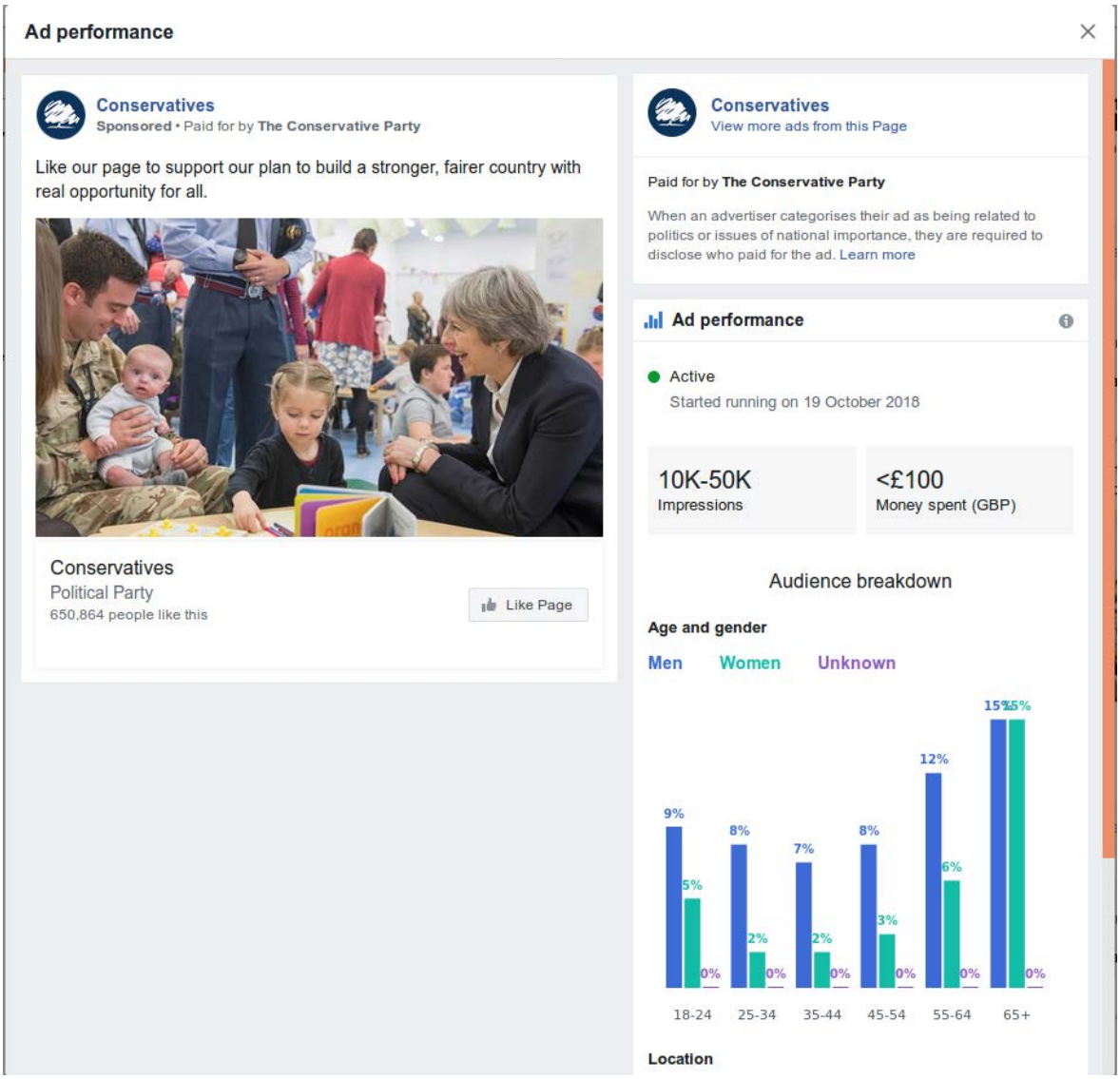
- **Google's Political Ads database**[10] is a public database of US political ads. It shows overview statistics by location (US states), top advertisers, and top keywords. There is also a search interface where adverts can be found by advertiser or candidate, dates it ran, amount spent, number of impressions, and type (text, image, video).
- **Twitter's Ad Transparency Center**[11] currently allows search by Twitter account, which then shows all promoted tweets by that account (if any) in the past seven days and the number of retweets and favourites for each. For US-registered political campaign accounts (e.g. @TheDemocrats) the search returns much more in-depth information, including the organisation behind this account and its spent so far, the age, gender, and location breakdown for all impressions of each political advertising tweet, and also details of the targeted audience. Twitter is also implementing public disclosure of issue ads (e.g. gun control) with exemption currently allowed for certain US-based media organisations. Concerns have been raised, however, that this might also hide from public scrutiny issue-based ads from heavily partisan sites with large online follower count (Romm, 2018).
- **Facebook's Ad Archive**[12] is currently the only one to offer public search over political and issue-based adverts from countries other than the US (currently Brazil and the UK). Ads are archived and accessible for up to seven years - the most comprehensive to date. Figure 10 shows an example advert and the available information about its performance and cost. The location information currently supported is very coarse: England, Wales, Scotland, and Northern Ireland. There is also a programming interface, providing automated access to this information.

[10] https://transparencyreport.google.com/political-ads/library

[11] https://ads.twitter.com/transparency

[12] https://www.facebook.com/ads/archive

Figure 10: Political advert by the UK Conservative Party



Source: Facebook Ads Archive website (accessed 24 October 2018)

These three initiatives represent promising first steps towards full online ad transparency. Firstly, they need to be expanded towards all EU countries and beyond and also evaluated independently how well they work in practice, as reports of outstanding problems are beginning to emerge (Lucas, 2018a).. Secondly, Google and Twitter need to implement APIs for automated access and analysis. In addition, the duration for which ads are being archived needs to be increased significantly, to match Facebook's seven year limit. Lastly, some policy makers and independent fact-checking organisations (FullFact, 2018) are advocating the need to go one step further and create a **public open data repository**, where political adverts are made available by the platforms in a unified machine readable format, including information about ad content, target audience criteria, audience reach and demographics, and amount spent. This is a necessary and important step, as it will provide an **independent mechanism for monitoring political advertising across all social platforms.**

2.1.3. Micro-Targeting and Third-Party Data Analysis of User Data

Social platforms, such as Facebook, Twitter, LinkedIn, and Pinterest, offer advertisers the option of creating the so called *dark ads*. These are online adverts that are visible only to the users that are being targeted (e.g. voters in a marginal UK constituency (Cadwalladr, 2017a)). Dark ads do not appear on the advertiser's timeline or in the feeds of the advertiser's followers. They have been used during political campaigns to spread misinformation, with the intent of influencing election outcomes. Moreover, due to their highly personalised nature, dark ads can be used to target susceptible users with misinformation which they are likely to believe is correct. As dark ads are hidden from view of other users, misinformation within cannot be discussed or counter-evidence posted by the user's friends.

Traditional social media ads, e.g. on Facebook, typically specify their target audience in very general way, e.g. age, gender, or areas of interest. Facebook (and other social platforms) also offer much finer-grained ad targeting (referred to as *micro-targeting*), based, for example, on job titles or demographic data. Facebook also offers the Lookalike tool, where the advertiser provides some seed accounts of the kinds of users being targeted by the ad and then Facebook's algorithms find similar Facebook users and show them the advert.

Dark adverts are effectively personalised adverts, seen only by their target audience. It is possible for an advertiser to run several micro-targeted ads at the same time, since they will be shown to different audiences. Dark ads are also used to test different versions of a Facebook post (organic or promoted), in order to optimise for engagement. For instance, during the 2016 US presidential campaigns Trump ran 5.9 million ads, to identify and then promote those variants that generated most Facebook engagement (Frier, 2018).

Since micro-targeted ads do not appear on the advertiser's Facebook page and in the feeds of its followers, they also help advertisers from cultivating a certain public image through their organic posts, while at the same time they can promote their hidden messages through hidden, micro-targeted adverts.

In the context of elections and politics, dark adverts have been employed extensively by political parties and candidates with the purpose of influencing voters. For instance, during the UK EU membership referendum the VoteLeave campaign used dark ads (Cadwalladr, 2018) containing misinformation regarding the weekly cost of Britain's EU membership and Turkey being set to join the EU (see Figure 11). Similarly, in 2017 the UK Conservative party sought to influence voters in marginal constituencies through micro-targeted ads containing hyperpartisan content (Cadwalladr, 2017b). This led the UK DCMS parliamentary inquiry to recommend that: "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)

Figure 11: A VoteLeave dark ad made public by Facebook as evidence to the UK DCMS parliamentary inquiry

Source: (Cadwalladr, 2018)

Such elaborate voter targeting requires large amounts of fine-grained, personal data to work effectively. Social media profiles contain a wealth of such information, including likes, posts, comments, and profile data. Based on such personal data, researchers have developed algorithms that can infer demographic information about users where that's missing, including gender, age, location, and political orientation (Rao *et al,* 2010; *Kosinski et* al, 2013; Colleoni *et al*, 2014). Machine learning models have also been developed for the automatic detection of personality traits based on Twitter (Golbeck *et al,* 2011) or Facebook (Golbeck *et al*, 2011a; Kosinski *et al*, 2013) posts and profile information. The methods score users using the so called OCEAN model (Openness, Conscientiousness, Extroversion, Ageeableness, and Neuroticism).

In 2015, the Ted Cruz presidential campaign hired Cambridge Analytica[13], which used automatically derived personality models in order to micro-target US voters with personalised campaign messages (Davies, 2015). However, in order to obtain the best results, such models need to be trained on the profile data and posts of tens of thousands of users, for which personality traits are established through psychometric tests. Cambridge Analytica hired a UK-based company (Global

13 A detailed write-up of the complex relationships between Cambridge Analytica, SCL, GSR, Aggregate IQ and US and UK politicians and political campaigns is beyond the scope of this report. We refer the interested reader to Chapter 3 of the Interim Report of the UK DCMS inquiry into fake news and disinformation (DCMS report, 2018).

Science Research – GSR) led by Alexandr Kogan, to collect personal Facebook data on millions of US voters and train such personality models (Davies, 2015).

At the time, GSR was able to collect the personal data of millions of US Facebook users, without their consent, thanks to Facebook's policy of allowing Facebook applications to access personal user information (including that of their network of friends), irrespective of the user's privacy settings (i.e. even information that was set to private) (Soltani, 2018). Facebook direct messages were also collected, for the users who gave Kogan's personality quiz access to their Facebook mailbox.[14] The data was then made available to Cambridge Analytica, which was also employed by Donald Trump's presidential campaign (Kang & Frenkel, 2018). In April 2018, Mark Zuckerberg admitted publicly that up to 87 million users may have been affected by this privacy data breach (Kang & Frenkel, 2018). However, as early as 2011 Facebook were made aware by the FTC and Irish data protection commissioner that Facebook app developers can access private Facebook user data without consent.[15] Moreover, the FTC also found Facebook's app review and oversight programme lacking (Soltani, 2018). According to Richard Allan (Facebook Vice President of Policy Solutions), at the time Facebook chose not to act since there was no legal obligation and there was no precendent with abuse of private user data (Allan, 2018).

The Facebook app associated with Cambridge Analytica and GSR is not the only Facebook app to obtain personal user data without express consent. Despite Facebook making changes in 2014 to its app privacy policies (Facebook, 2014) to prevent such occurrences, the New York Times reported that device makers (Apple, Samsung, and numerous of others) had access to Facebook data from user profiles and friend networks without express user consent, some as late as 2018 (Dance *et al*, 2018). This demonstrates why a purely self-regulation approach to user privacy and advertising transparency is ineffective in practice.

In addition to the privacy concerns raised by unauthorised access to Facebook profile data by third-party apps, there are also other related important privacy concerns:

- The models derived from personal data of millions of users are as, if not more, important than the raw data itself, as demonstrated by the reluctance of Cambridge Analytica to delete these throughout the 2016 US elections and beyond.16
- **Non-user data**: Facebook stores also some data about people who are not Facebook users (Allan, 2018). These include information, such as email addresses, uploaded by Facebook users for their contacts. There is also data collected from visits to web sites that host Facebook plugins. Belgium has flagged this as a privacy invasion case, which is still ongoing (Allen, 2018).
- **Data sharing between WhatsApp and Facebook**: Following the acquisition of WhatsApp by Facebook, concerns have been raised on privacy implications of data sharing between the two social apps and GDPR compliance (Denham, 2018).

As a result, the issue of ensuring effective personal data protection on social platforms has now emerged as a priority for national governments and policy makers worldwide. It is also an area where the benefit of international cooperation is becoming recognised, with the UK Information Commissioner recently stating that "we must have the tools to work together, to collaborate and co-operate, because this is truly a global issue. That is a strong recommendation." (Denham, 2018)

14 https://www.theguardian.com/uk-news/2018/apr/13/revealed-aleksandr-kogan-collected-facebook-users-direct-messages

15 Q4177 DCMS evidence http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/digital-culture-media-and-sport-committee/disinformation-and-fake-news/oral/92923.html

16 https://www.theguardian.com/uk-news/2018/may/06/cambridge-analytica-kept-facebook-data-models-through-us-election

She then also went on to add: "The message I heard from Facebook this morning was that unless there is a legal order compelling a change in their business model and their practice, they are not going to do it."

In conclusion, it must also be noted that due to these recently uncovered data-driven campaign practices, political parties and candidates also need to respect privacy laws, give people control over how their information is used, and be made accountable, in much the same way as social media platforms (Denham, 2018).

2.2. Genuine Amplifiers: Online News Consumption Habits, Confirmation Bias, and Polarisation

Recent research established that the main amplifiers behind viral misinformation and propaganda are human users (Vosoughi et al, 2018). The key questions then are why do people fall for online misinfomation, what motivates them to share it, and what is the impact of online misinformation on their offline behaviour (e.g. does it affect their voting in elections).

Due to the recent nature of large-scale online misinformation "wildfires", answers to these key questions are currently incomplete and, in some cases, somewhat contradictory. This has motivated policy makers and independent experts to recommend that governments need to invest into further research on these topics (HLEG report, 2018; DCMS report, 2018; NED, 2018), including not just data science approaches, but also ethnographic studies (NED, 2018).

Nevertheless, key concepts and current findings are presented here, as they are necessary for the understanding of existing socio-technical approaches to fighting online misinformation.

People's exposure to online misinformation has grown thanks to their use of the internet and social platforms as their news source (Ofcom, 2018; Messing & Westwood, 2014). In the UK, for example, 84% of young adults under 24 and 73% of adults from ethnic minority groups get their news from online and social sources. People often orient towards and read content and new sources which are aligned with their political and other views - a practice referred to as *confirmation bias*. Moreover, an individual's online news and information sharing and commenting behaviour is influenced by the behaviour of their online social connections, as they tend to be usually like-minded people - referred to as *homophily*.

Taken together, confirmation bias and homophily lead to the creation of *online echo chambers* (Quattrociocchi et al. 2016), where people are exposed to and share with their connections predominantly information conforming to their pre-existing beliefs and thus lack exposure to diverse or opposing perspectives and opinions (Garrett, 2013). According to the latest study by Gallup and the Knight Foundation[17], people generally share information that they trust and do so primarily for social or personal reasons.

Unsurprisingly, therefore researchers have found that "social networks and search engines are associated with an increase in the mean ideological distance between individuals" (Flaxman *et al,* 2018), i.e. lead to polarisation. These findings hold not only for the online US political landscape, but also for other countries, e.g. Turkey and Russia (Kelly & François, 2018).

Experimental research has also shown that when polarised online communities are exposed to misinformation which conforms to their preferred narratives, it is believed and shared (Quattrociocchi et al. 2016). Consequently, when such users and communities are exposed to

[17] https://www.knightfoundation.org/reports/in-the-internet-we-trust-the-impact-of-engaging-with-news-articles

debunks or opposing opinions, these either have little effect or strengthen their pre-existing beliefs and misconceptions.

There are also concerns over the role of the platform's automatic content recommendation algorithms in creating so called *filter bubbles*, as they tend to promote posts that are similar to those liked or shared by the user. A Facebook study of 10 million US users (Bakshy *et al,* 2015), however, has reported that algorithms play a smaller role in limiting user exposure to opposing views and more diverse content, than did the users' own homophily and echo chamber conformance.

Another contributing factor to the viral spread of misinformation by genuine users (as opposed to bots and fake accounts) is information overload and the limited attention that users have (Qiu *et al*, 2017), when they engage with social media feeds on mobile devices.

The extent to which confirmation bias, online echo chambers, filter bubbles, and (mis)information overload impact online and offline user behaviour and beliefs is still being studied. As already discussed, many researchers consider these phenomena as limiting exposure to diverse information and opinions. Others, however, have argued that there is sufficient diversity in the users' social networks thanks to diversity in real-world connections (Bakshy *et al*, 2015) and that this diversity ensures that users are still exposed to some opposing viewpoints. Given the importance of social endorsements and this diversity in social ties, researchers (Messing & Westwood, 2014) have also argued that these can reduce the dominance of partisan selective exposure . Other recent research (Shore *et al,* In press) has shown that the average user tends to share links to more moderate news content. These findings, however, do not contradict previous research on echo chambers and polarisation, since (Shore et al, In press) have also found evidence of the existence of a number of highly popular and very active users who "do exhibit cross-sectional evidence of polarization and are responsible for the majority of tweets received overall due to their popularity and activity".

The social platforms are also aware of this problem and have started experimenting with changing their algorithms to reduce the impact of misinformation and polarising content. Facebook, for example, will be prioritising content from friends and family over articles from the News Feed (Bickert, 2018). Twitter is also taking steps to expose its users to opposing political views (Bail, 2018) in order to try reducing polarisation. As with debunks, however, there are concerns that it may actually have the opposite effect of entrenching hyperpartisan views (Bail, 2018).

Therefore, the question still remains open as to what is the best strategy to reduce people's susceptibility to misinformation and the likelihood of its amplification through organic sharing. Strategic communications researchers (Pamment et al, 2018) have argued for 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." Such an audience-sensitive approach also means that instead of simply presenting the truth or opposing views, it may be more productive to encourage debate and critical reflection instead (Harford, 2018; Pamment *et al*, 2018), facilitated by real-life social connections (e.g. friends, colleagues) (Harford, 2018).

2.3. Fake Amplifiers: Social Bots, Cyborgs, and Trolls

Social bots are 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. Political bots have been used to try and influence democratic elections and referenda in countries worldwide, including the US

(most notably in the 2016 presidential elections), the UK (the 2016 EU membership referendum), France (the 2017 French general election), and Russia (in 2012).
Sockpuppets are fake accounts that pretend to be ordinary human users and aim to connect to and influence real social network users. Sockpuppets are typically human controlled, but can also employ some automation, in which case they are referred to as cyborgs. Politically oriented sockpuppets, especially those controlled by governments or affiliated organisations are known as **trolls**.

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), which flood the platform with manipulated content thus making high quality information hard to find. Automated accounts and trolls also engage in astroturfing - a propaganda technique which "creates the illusion of a spontaneous, popular movement on the internet started by a fake grassroots organization." (RSF, 2018).

Another kind of false amplification is the use of **fake followers** to inflate artificially the perceived influence of a social media account. A recent New York Times investigation (Confessore *et al*, 2018) has uncovered companies that sell fake social media followers and retweets. Devumi, for example, was shown to operate 3.5 million fake accounts, which were used to generate over 200 million fake followers. The investigation also uncovered evidence of social media identity theft, where names, photos and personal details of real social media users were used to create more authentic-looking fake accounts. The problem is not confined to Twitter, with Facebook also acknowledging that up to 60 million automated accounts may exist on its platform (Confessore *et al,* 2018). The business model of artificial amplification and its use by governments, politicians, celebrities, businesses, and many other kinds of social media users are due to the *attention-centric* nature of social platforms, where quantity (e.g. number of followers, post engagements) is more important than quality.

Fake accounts can also follow high profile social accounts, in an attempt to gain credibility, attract attention, and motivate them to engage with and repost their content. For instance, (Campoy, 2018) have reported that over 60% of the 55 million followers of Donald Trump's Twitter account are fake. Moreover, Russia-linked political bots not only amplified heavily his messages in the run up to the 2016 presidential elections (De Vynck & Wang, 2018), but Trump himself has sometimes retweeted and replied to fake accounts (Bertrand, 2017).

Another example of artificial amplification and attempts to skew democratic processes are **fake comments** on government consultations or phantom signatures on online petitions. For example, a Wall Street Journal investigation uncovered a large number of fake comments on federal regulations on controversial issues, such as net neutrality (Grimaldi & Overberg, 2017). Estimates vary as to the number of fake comments on net neutrality alone, with some putting it at hundreds of thousands, while others - at 1.3 million (Kao, 2017).

As demonstrated here, the scale and ubiquity of artificial amplification are such that urgent measures need to be taken by platforms, policy makers, researchers, journalists, and civil organisations. These are discussed in detail in section 3.2.

2.4. Artificial Intelligence, Synthetic Media, and "Deepfakes"

In the past several years, advances in artificial intelligence and the availability of affordable high performance computing have led to the emergence of synthetic (i.e. computer-generated) images, audio, and video. The so called "**deepfakes"**, in particular, are synthetic videos and images generated using deep neural network models, which "look and sound like a real person saying something that that person has never said." (Lucas, 2018)

"Deepfakes" can potentially cause significant harm, as they look credible; are harder for citizens to verify; and can give the impression, e.g. that a politician has said or done something that they did not. Pornography-oriented deepfakes are also offensive to the target and can be used as part of online abuse campaigns, e.g. against journalists.
Until this technology came into existence, generating even simple synthetic media content required significant expertise and effort. Since off-the-shelf, easy-to-use tools for the creation of deep fakes have become widely available, they have been used to perform face swapping in images and video, lip-syncing, speech modification, or image alterations (Hui, 2018).

For example, FaceSwap[18] (and now to a lesser degree FakeApp, as it is no longer so easily available for download) have been used to swap in the faces of celebrities (e.g. Gal Gadot[19]) or other targeted women (Curtis, 2018) in porn videos, or put famous actors in movies they never played (e.g. Nicolas Cage[20]). Another example is a video of President Obama who is lip-synced to an actor (see Figure 12). Adobe has recently unveiled Voco[21] - an application that allows an audio recording to be altered with words that were never spoken by the person, who was recorded even though they sound as if they did. Deepfake face-swap has recently become even easier thanks to Aphrodite.ai[22], which requires only a desktop/laptop computer, web camera, and a modern web browser to run. The heavy computation is carried out using cloud computing.

Figure 12: Video demonstrating a lip-syncing deepfake

Source : YouTube

Social platforms, and Reddit in particular, have helped accelerate the pace of deepfake creation by hosting online communities that exchange know-how, code, and the resulting deepfakes themselves. The original r/Deepfakes subreddit that gave rise to the FaceSwap repository has been banned by Reddit for violating their terms of service. However, another related community is r/SFWdeepfakes[23], which as of 16 Nov 2018 had 3,200 subscribers.
The ability to generate highly believable synthetic media poses important legal challenges (e.g. are deepfakes illegal, who owns the copyright, is this parody), which are yet to be addressed (Ellis, 2018; Beres&Gilmer, 2018). Even if deepfakes are legal, there are ethical aspects that have to be considered, including lack of consent and private-only use (Zucconi, 2018). Audio manipulation

Figure 13: Example of NVIDIA technology that modifies weather conditions automatically

Source : The Verge

[18] https://github.com/deepfakes/faceswap

[19] https://motherboard.vice.com/en_us/article/gydydm/gal-gadot-fake-ai-porn

[20] https://www.youtube.com/watch?v=Ex83dhTn0IU

[21] https://www.bbc.co.uk/news/technology-37899902

[22] http://www.aphrodite.ai - created by a Hong Kong based company

[23] https://www.reddit.com/r/SFWdeepfakes/

software like Adobe Voco could also pose cybersecurity risks, e.g. in cases where people being identified from their voiceprint[24].

Researchers and technology companies have also been working on tools and AI algorithms for detecting synthetically generated content. So far these have been successful at detecting existing fakes, thanks to some of their imperfections. For instance, the GIF-hosting platform Gfycat is using a combination of face recognition and AI-based video matching to identify deepfake content (Matsakis, 2018). While technology can help in some cases, social platforms need to take responsibility to prevent and remove synthetic pornographic content in an efficient and transparent manner (Curtis, 2018).

Journalists and media also need to have robust processes in place for identifying deepfake video footage, in order to ensure high quality news reporting. Recently the Wall Street Journal started training its journalists to recognise deepfakes via a combination of methods, including examining the source who shared it, the way it was obtained, the image and video metadata for evidence of tampering (e.g. via the InVID plugin, see section 7.1), slowing down and examining the footage for unnatural movement and fuzziness, blurriness compared to the non-facial areas of the video, change of skin tone near the edge of the face, etc.

While fairly easy to detect at present, synthetic media is evolving rapidly and becoming more believable and harder to detect automatically or by humans, due to new developments in deep learning that allow the face swap algorithm to learn from an AI-based detection algorithm. This has led to journalists, researchers, and policy makers becoming increasingly concerned that synthetic media will soon become the latest, and possibly the most dangerous, weapon in the arsenal of viral online disinformation campaigns.

The recent success and wide reach of spreading misinformation through political memes (e.g. some of them published by Russian trolls (Tunikova, 2018)) has demonstrated that misinformation in images is extremely successful at going viral (Ng, 2018), while being significantly harder to detect by automatic means or by fact-checkers. Compared to memes, misinformation in deepfakes can be even more damaging, as videos can be even more believable and subversive.

At the same time, humans have been shown to struggle distinguishing between real and fake videos (Rössler *et al*, 2018) at present, which has led also to calls for media literacy training (Witness & First Draft, 2018), as "shared learning on how to detect synthetic media that brings together existing practices from manual and automatic forensics analysis with human rights, Open Source Intelligence (OSINT) and journalistic Practitioners."

Ultimately, the biggest threat from deepfakes is that they can undermine further public trust in news media and politicians, as well as spread virally highly believable misinformation. In other words, the existence of believable deepfakes can make it hard to prove that "something real is real" (Witness&First Draft, 2018). To help combat this, Witness and First Draft recommend significant foundational research and development of techniques to "track authenticity, integrity, provenance and digital edits of images, audio and video from capture to sharing to ongoing use."

2.5. Outlook

This section introduced the socio-technical enablers and actors of online disinformation campaigns, as they are currently understood. Over the past few years, the tools and methods used to spread

24 https://www.bbc.co.uk/news/technology-37899902

online propaganda and disinformation have become more and more sophisticated, and now often span multiple social platforms, types of content (text, images, videos), and types of amplifiers.

This makes effective detection and response particularly challenging, since "the public manifestations of communication do not tell the full story of the intent and the coordinated efforts behind those communications." (Storyful, 2017)

Moreover, as can be seen from the rapid emergence and improvement in quality of deepfakes, cutting edge technology can quickly be harnessed into producing a new generation of online dis- and malinformation. This applies not only to the creation of more sophisticated types of fake content, but also to its more efficient artificial amplification. In particular, disinformation bots can soon become more believable too, thanks to rapidly improving AI chatbot technology.

These evolving technological complexity and effectiveness will bring about new ethical, legal, and societal challenges, which in turn will require new policy discussions and socio-technical solutions to online propaganda and disinformation campaigns to be developed continuously. Next we will discuss the state-of-the-art in these areas.

3. Technological approaches to fighting disinformation

This section focuses on algorithms and technology tools for (semi-)automated detection of online disinformation campaigns. All modalities of disinformation are considered: text, images, and videos. The complete lifecycle of disinformation detection and analysis is addressed: content and source validity analysis; organic and artificial network spread; measuring impact on citizen beliefs and actions; and debunking methods.

As discussed in Section 1, there are seven types of mis- and disinformation: satire/parody; misleading content, imposter content, fabricated content, false connection, false context, and manipulated content (Wardle, 2017). Amongst these, state-of-the-art verification tools and methods largely focus on identifying manipulated content. Detection of satire, imposter, and fabricated content have also been studied, in particular hoaxes, fake news, and conspiracy theories. For instance, recent computational research on disinformation detection in Wikipedia (Kumar et al, 2016) showed that some hoaxes remained undetected for long periods and were widely cited across the web. They also showed that humans were significantly worse at detecting hoax articles than the machine learning algorithm. Researchers have also studied network-based visualisations of claim and misinformation spread, e.g. the Hoaxy system (Shao & al, 2016).

Following the AMI model (Agent, Message, Interpreter), described in section 1, studies and technology needs go beyond individual pieces of content (i.e. the message), and consider the actors and the impact of disinformation campaigns on their recipients. Initially this was addressed by journalists and political science researchers at the micro level, who focused on specific disinformation campaigns. For instance, the Donbass conflict in Ukraine exemplifies how information warfare can exploit multiple media to spread propaganda in conflict zones, especially through Russian state TV (Khaldarova & Pantti, 2016). The 2016 United States presidential elections showed the importance of disinformation campaigns on democratic processes, while also highlighting the difficulty in providing accurate assessment of their impact on voting behaviour (Allcott & Gentzkow, 2017). Even in the so-called "swing states" - states that played a deciding role in the result of the elections - where misinformation and fake news was most shared, an impact assessment has been difficult to produce (Howard, Kollanyi, Bradshaw & Neudert, 2018).

The majority of academic research is focused on methods for macro-level analysis. Broadly speaking, we can identify three classes of approaches. The first set of approaches focused on investigating the role of echo chambers - questioning the influence of social media platforms and online news sites and their influence in creating partisanship echo chambers. For instance, the role of misinformation content in generating homogenous and polarised echo chambers (Del Vicario et al., 2016), as well as the role of confirmation bias (Quattrociocchi, Scala & Sunstein, 2016), have been demonstrated on Facebook. The influence of recently created alt-right media in the Republican partisanship echo chambers in the US has also been analysed, with Democrat echo chambers relying more on mainstream media and traditional networks like CNN (Faris et al., 2017).

The second strand of research focused on detecting fake amplifiers of false narratives. The role of automated or semi-automated accounts (bots) in amplifying false narratives has been demonstrated especially during the US elections and the Brexit referendum (Howard & Kollanyi, 2016; Gorrell et al, 2018). Even though Facebook and Twitter themselves have invested significant effort in identifying and suspending such bot accounts (Pickles, 2018; Bickert, 2018), it is still necessary to have independent researchers and journalists studying these, as there are still cases which are being missed by Twitter and Facebook's algorithmic approaches (Phillips & Balls, 2017). While the use of disinformation and propaganda techniques in election campaigns is not new, we are still lacking the human skills and the technology needed for analysing the large volumes of

ubiquitous online disinformation, which frequently spans countries, languages, cultures, and online social platforms.

The third strand of work is on combining content analysis with network analysis through the use of semantic tools and machine learning (Conroy, Rubin & Chen, 2015), which highlighted the necessity to have a combined machine/human approach and to fuse different techniques to assess veracity of information.

Once identified, another key challenge is containment of disinformation spread in social networks. Computer science models have focused on identifying the key nodes that need to be "decontaminated" (Nguyen & al. 2012), e.g. using epidemiological models (Tambuscio & al. 2015) or nodes that can be recruited to spread debunking information through the network (Budak & al. 2011). However, most of these models fail to account for the effect of partisan nodes and alternative media, as well as lack empirical validation on real social network data. In addition, recent journalism research has found that exposing victims of disinformation to factual, non-partisan debunking may change their knowledge, but not their beliefs (Swire & al, 2017).

Despite these efforts, there is an urgent need to address these diverse major challenges and develop a new generation of content verification and disinformation analysis tools. This has also been recognised by the pan-European High Level Expert Group (HLEG) on Fake News and Online Disinformation. In particular, their report (Bunning et al, 2018) recommends to "develop tools for empowering users and journalists to tackle disinformation and foster a positive engagement with fast-evolving information technologies".

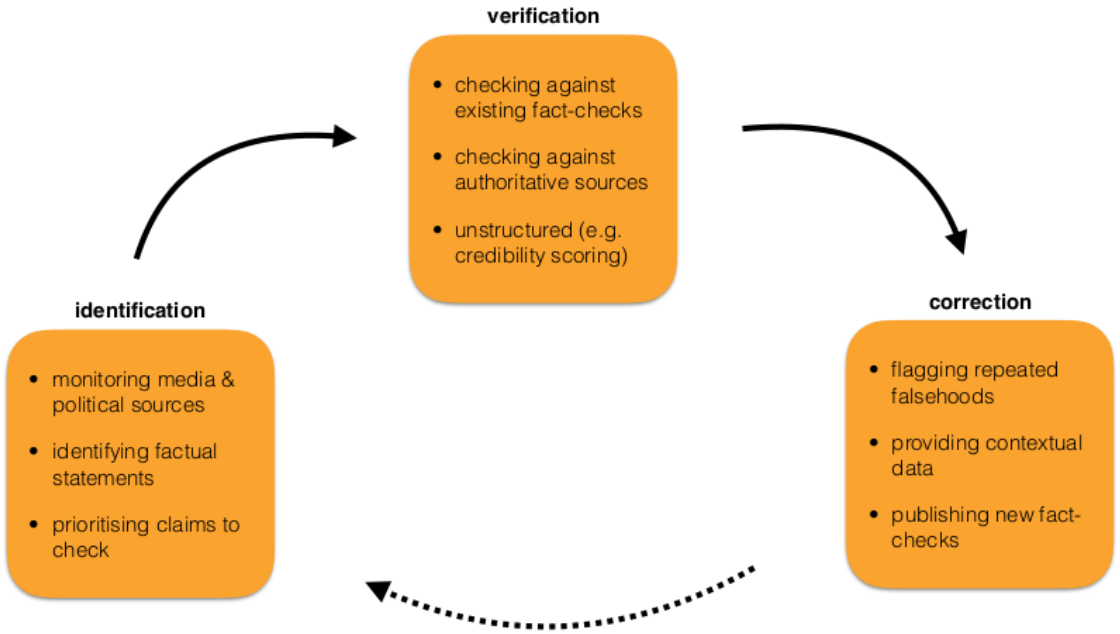
The rest of this section goes back to discussing state-of-the-art technology solutions for each of the key disinformation analysis tasks, before concluding with a discussion on ethics, which is often overlooked especially by companies.

3.1. Fact checking and Content Verification

In order to counter subjectivity, post-truth politics, disinformation, and propaganda, many media organisations and non-partisan institutions worldwide have started fact checking initiatives – 114 in total, according to Poynter (Mantzarlis, 2017A). These mostly focus on exposing disinformation in political discourse, but generally aim at encouraging people to pursue accuracy and veracity of information (e.g. Politifact, FullFact.org, Snopes). A study by the American Press Institute (API,2015) has shown that even politically literate consumers benefit from fact-checking as they increase their knowledge of the subject. Independent act checking of claims made by politicians and news is also hugely important, as it can also act as a deterrent to political parties and news media organisations, as well as provide unbiased evidence to policy makers and government organisations as to how objective are politicans and media.

Consequently, fact-checking is regarded by policy makers as a key element of a successful multi-dimensional approach to limiting the spread and influence of online disinformation, e.g. the EU High Level Expert Group on online disinformation (HLEG report, 2018) and the UK DCMS inquiry (DCMS, 2018).

Figure 14: Fact checking workflow



Source: (Graves, 2018)

Professional fact checking is a time-consuming process that consists of numerous complex steps (see Figure 14 for details). Therefore, manual fact-checking cannot cover a significant proportion of the claims being propagated via social media channels.

A number of automated fact-checking tools are being developed in response by fact-checking organisations and startup companies, e.g. FullFact[25], Duke University's Reporters Lab[26], Factmata[27], Chequado[28], ContentCheck[29]. The aim is to assist the human fact-checkers in tasks, such as automatic detection of factual claims made by politicians and other prominent figures in TV transcripts and online news, e.g. Full Fact's Live tool[30], Duke's Tech&Check which uses Claimbuster (Funke, 2018). Other automation tools offer tracking mentions of already known false claims, e.g. Full Fact's Trend tool and automatic checking of simple numeric claims against authoritative databases, e.g. Full Fact Live. Due to the subjective and difficult nature of some fact-checks, some tools, e.g. Duke's Tech & Check also send automatically identified factual claims to independent fact-checking organisations, such as PolitiFact and The Washington Post Fact Checker (Funke, 2018). Complementary to these are efforts like Storyzy[31], which is offering a continuously updated

[25] https://fullfact.org/

[26] https://reporterslab.org/

[27] https://factmata.com/

[28] https://chequeado.com/

[29] ContentCheck https://team.inria.fr/cedar/contentcheck/

[30] FullFact's Live and Trend tools https://fullfact.org/automated; Live was developed originally for English but is now also adapted to Spanish and Portuguese: https://www.poynter.org/news/these-fact-checkers-teamed-across-atlantic-cover-presidential-debate-real-time

[31] http://storyzy.com/about

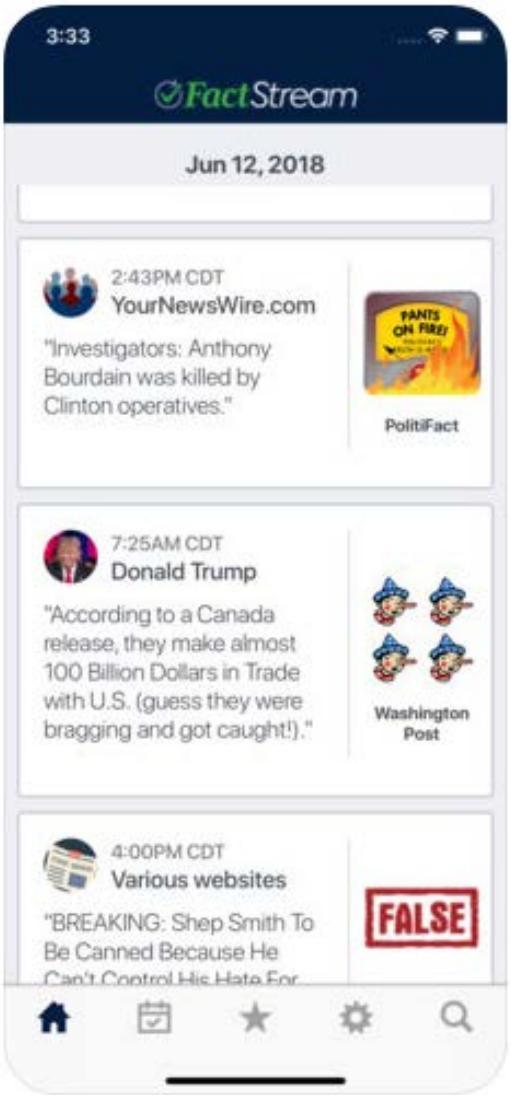
database of fake news sites and video channels, and WeVerify[32], which is building a blockchain database of known false claims and fake content. Automated Fact Checking tools, e.g. Full Fact Live, Duke's Tech&Check also check incoming claims against existing fact-checks (see Figure 15), stored either in internal databases and/or assembled automatically based on trustworthy, publicly shared fact-checked claims tagged with the open Claim Review[33] standard schema.

Automated fact-checking methods based on Natural Language Processing (NLP) and Artificial Intelligence (AI) techniques are also being researched. One of the seminal approaches (Vlachos and Riedel, 2015) focused on identifying simple statistical claims (e.g. "the population of UK is 60 million people") and checking their validity against a structured database. While the accuracy of these methods is improving continuously, thanks to the creation of large datasets of validity-annotated textual claims (Thorne, Vlachos, et al, 2018), it is still considered insufficient for practical use (FullFact, 2016). However, as more and more human-verified claims are shared openly in machine readable formats, e.g. Claim Review, these will help NLP and AI fact checking algorithms reach maturity.

For the time being, as noted by a recent Reuters Institute report on Automated Fact-Checking (Graves, 2018): "*Both researchers and practitioners agree that the real promise of AFC technologies for now lies in tools to assist fact-checkers to identify and investigate claims, and to deliver their conclusions as effectively as possible*."

Perhaps the biggest challenge in fact-checking is **efficient correction,** i.e. bringing fact-checks to the attention of citizens (see Figure 15). FactStream[34] is a mobile phone app, developed by Duke Reporters' Lab, which alerts American users of the latest fact-checks from the three largest U.S. fact-checking organizations – the Washington Post, PolitiFact and FactCheck.org. Each claim is clearly displayed together with its source, truth value (true, mostly true, half true, false, pants on fire), and the fact-checking organisation that verified it.

Figure 15: FactStream mobile phone app



Source : FactStream

Fact-checking is also being harnessed by the social media platforms, in order to help curtail the viral spread of disinformation. Facebook, for example, is letting users flag suspect stories, which are then verified by a network of trusted fact-checking organisations. According to M. Bickert from Facebook (DCMS HC 363, 2018), the platform's

[32] https://twitter.com/WeV3rify/status/1044876853729796099

[33] https://schema.org/ClaimReview

[34] https://itunes.apple.com/us/app/factstream/id1327422405

algorithms then reduce by 80% the visibility of stories deemed to be false by the fact checkers. While initially Facebook's authorised fact-checking organisations were all non-partisan, recently the platform approved also the conservative Weekly Standard, which has led to a controversy over what is fake versus clickbait and the role of partisan organisations in fact-checking (Newton, 2018). There are also reports of conflicting priorities between Facebook and the independent fact-checking organisations, e.g. as the latter cannot choose what stories to fact-check and then feed them to Facebook, but rather have to fact-check the stories selected in a somewhat non-transparent way by Facebook (Annany, 2018). Lastly, evidence is emerging that Facebook may have over-stated their fact-checking efforts and, in practice, is sending just one post per day to independent fact-checkers (Salinas, 2018).

Outstanding technological challenges (apart from improved NLP and AI algorithms) include adapting existing mostly English-centric AFC tools to all EU languages, obtaining and sharing reliable, machine-readable datasets, more effort dedicated to source-checking, i.e. verifying the trustworthiness of the information source. As argued by Full Fact (Babakar & Moy, 2016), these need to be complemented by community-wide effort to bring together and coordinate AFC tool development to avoid duplication and speed up progress. An essential part of this is the creation of open standards, open source algorithms, and open evaluation datasets (FullFact, 2016).

Content Verification is complementary to fact-checking and is concerned with verifying whether an image, video, or a meme have been tampered with or promote false information. Some of the best known tools have focused on crowdsourced verification (e.g. CheckDesk, Veri.ly), citizen journalism (e.g. Citizen Desk), repositories of checked facts/rumours (e.g. Emergent, FactCheck). Currently, the most successful verification platforms and products include SAM[35], Citizen Desk[36], Verily[37], Check[38], and Truly Media[39].

There are also some browser tools and plugins aimed at journalists, e.g., the InVID[40] and Frame by Frame[41] (video verification plugins), Video Vault[42] (video archiving and reverse image search), RevEye[43] (reverse image search), Jeffrey's Image Metadata Viewer[44] (image verification), NewsCheck[45] (verification checklist). Plugins offering web content and social media monitoring are Storyful's Multisearch[46] plug-in for searching Twitter, Vine, YouTube, Tumblr, Instagram and Spokeo, with results show in separate tabs, without cross-media or social network analysis; and Distill[47] which monitors web pages.

[35] https://www.samdesk.io/

[36] https://www.superdesk.org/

[37] https://veri.ly/

[38] https://meedan.com/en/check/

[39] http://www.truly.media/

[40] https://www.invid-project.eu/tools-and-services/invid-verification-plugin/

[41] https://chrome.google.com/webstore/detail/frame-by-frame-for-youtub/elkadbdicdciddfkdpmaolomehalghio

[42] https://www.bravenewtech.org/

[43] https://chrome.google.com/webstore/detail/reveye-reverse-image-sear/keaaclcjhehbbapnphnmpiklalfhelgf

[44] http://exif.regex.info/exif.cgi

[45] https://firstdraftnews.org/launching-new-chrome-extension-newscheck/

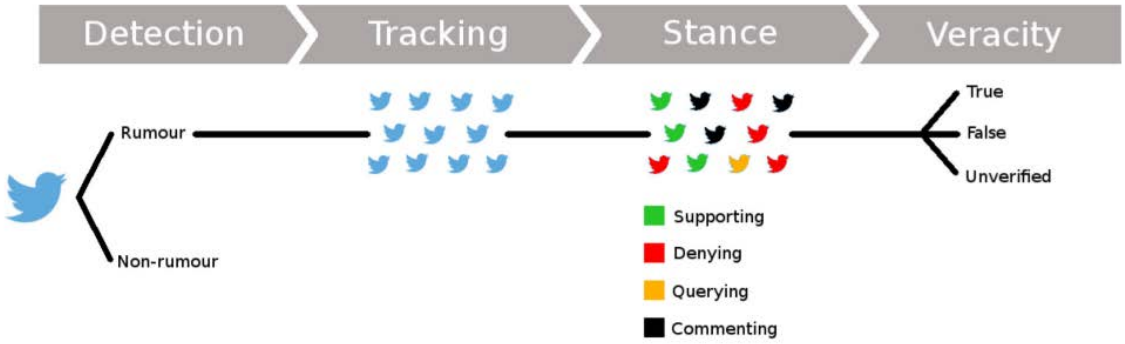
[46] https://chrome.google.com/webstore/detail/storyful-multisearch/hkglibabhninbjmaccpajiakojeacnaf

[47] https://chrome.google.com/webstore/detail/distill-web-monitor/inlikjemeeknofckkjolnjbpehgadgge

With respect to algorithmic analysis, there are tools for photo, image, and video forensics, e.g. Forensically[48], FotoForensics[49], the Image Verification Assistant[50] developed in the REVEAL FP7 EU project, and the InVID video and image verification plugin[51]. The functionalities currently being offered are based on algorithms that highlight tampered areas, metadata categorization and analysis, and near-duplicate retrieval based on keyframe matching through reverse image search (typically through Google). According to (Zampoglou et al, 2016), the REVEAL Image Verification Assistant is the broadest amongst image verification tools, featuring seven state-of-the-art image analysis algorithms and some basic metadata analysis tools.

Some of the most accurate tools tend to combine metadata, social interactions, visual cues, the profile of the source (i.e. originating agent), and other contextual information surrounding an image or video, to assist users with the content verification task. Two of the most widely used such tools are the InVID plugin (Teyssou et al., 2017) and the Amnesty International Youtube Data Viewer. The InVID video verification plugin is discussed in depth in section 7.1. The YouTube Data Viewer extracts metadata listing and offers image-based similarity search using keyframes.

Figure 16: Rumour Analysis Workflow



Source: (PHEME project)

Rumour analysis is another hotly researched topic related to fact checking and content verification (Zubiaga et al, 2018). Following the conceptual framework proposed by the PHEME project[52] (see Figure 16), rumour analysis can be broken down into the following four stages, where each stage can be carried out either manually or (semi-)automatically:

- **Rumour detection** is concerned with whether a piece of information constitutes a rumour. A typical input to a rumour detection component is a stream of social media posts. An AI algorithms is trained to classify each post as a rumour or a non-rumour. This component is useful for identifying emerging rumours; however, it is not necessary when dealing with rumours that are known a priori (e.g. already identified by fact checkers).
- **Rumour tracking** collects and filters posts discussing the already identified rumour, using the rumour's originating post or keywords as the starting point.
- **Rumour stance classification** is concerned with classifying each post about a rumour, according to the reaction it has elicited in the post's author. The stance being expressed is either supporting (i.e. agreement with the rumour's statement), denying, querying, or

[48] https://29a.ch/photo-forensics/

[49] http://fotoforensics.com/

[50] http://reveal-mklab.iti.gr/reveal/

[51] https://www.invid-project.eu/tools-and-services/invid-verification-plugin/

[52] http://pheme.eu

commenting. Different algorithms for automatic stance classification have been studied (Zubiaga *et al*, 2018), with some of the best reported results (Aker, Derczynski & Bontcheva, 2017) achieved by a model combining linguistic, message-based, and topic-based features.

- Finally, **rumour veracity** is determined (true, false, or unverified). This is a complex decision which often requires expert human judgement (Zubiaga & Ji, 2014). Researchers have also experimented with building AI-based veracity classification models (Dungs *et al,* 2018; Zubiaga *et al,* 2018), with the best ones currently achieving around 80% overall score (Dungs *et al*, 2018).

One area in urgent need of further research and practical tool development is **meme tracking and verification.** On one hand, memes have been shown to feature prominently in propaganda and disinformation campaigns (Haddow, 2016). On the other, research on their veracity, origin and spread is still very much in its infancy (Zannettou *et al,* 2018).

In conclusion, while existing technology for content verification and fact checking can assist journalists, media, and social media platforms to flag or filter false content more effectively, it fails to address another key challenge – that of analysing the mechanisms and scale of online disinformation and propaganda campaigns, their sources and goals, and their effects on citizens, business, and society. The next few sections discuss these in more detail.

3.2. Detecting Computational Amplification and Fake Accounts

Numerous studies of computational amplification of disinformation and propaganda has focused on detecting bot and sockpuppet accounts, clickbait, and astroturfing.

The importance of research into computational amplification was brought to public and political attention, as a result of studies trying to establish the role of bots and sockpuppets in the 2016 US presidential elections, the UK EU membership referendum, and the 2017 French presidential elections. Political bots, in particular, have been shown to try influencing political opinion, e.g. attack political leaders or journalists, even though current evidence seems to indicate bots do not change voter intent (Howard et al, 2018b). Research has shown that during the elections a large number of (coordinated) bots and sockpuppet accounts were used for spreading disinformation and political rumours, as well as promoting news from Russian sources (primarily Russia Today and Sputnik) (Phillips & Ball, 2017; Howard et al, 2018; Gorrell et al, 2018; Howard & Kollanyi, 2016).

Consequently, both Twitter and Facebook carried out significant internal research, which in turn resulted in updated platform policies and algorithms for detecting fake accounts. The actual bot and sockpuppet detection algorithms implemented by the platforms themselves are not publicly available, and thus it is not possible to have independent scrutiny of their accuracy and possible flaws. Based on publicly available policy documents, however, we do know that Twitter has recently started using new features some of which were proven as important in academic research first. Examples include whether an account uses a stock or stolen avatar photo, stolen or copied profile text, or misleading profile location (Harvey & Roth, 2018). In comparison, Facebook has fewer automated bot accounts (due to its more closed API), but needs to identify more sock puppet and impersonation accounts instead. Identifying these automatically is much harder (and sometimes impossible) than finding bots, due to the more authentic human-driven behaviour (Weedon et al, 2017).

State-of-the-art research on bot detection methods (Varol & al, 2017; Woolley & Howard 2016; Cresci & al. 2016) use predominantly social behaviour features (e.g. tweet frequency, hashtag use,

following large number of accounts while being followed by just a few). There are also approaches that detect bots based on the high correlations in activities between them (Chavoshi *et al,* 2018).
A widely used Twitter bot detection service is Botometer[53] (previously BotOrNot), which is provided free of charge by Indiana University. Users can check the bot likeliness score of a given Twitter account, based on its user profile information, friends, and followers. Usage is subject to Twitter authentication and rate limiting. The service can also be accessed programmatically as well. The accuracy of this service needs to be established through independent evaluation.

In general, as research-based methods can only use the publicly disclosed data about Twitter accounts, there are also concerns[54] regarding how accurate can they be, given that human curators can often struggle to identify bots from public Twitter profiles alone. This is set to become even harder, as more sophisticated bots are starting to emerge.

So far, research has focused primarily on Twitter bots, due to Facebook API restrictions. Existing methods from academic research are yet to reach very high accuracy, as they operate only over publicly accessible account data (e.g. account description, profile photo). The social media platforms often make use of additional account-related information, including IP addresses, sign-in details, email account, browser cache, which all make the task somewhat easier. Nevertheless, the platforms also find bot detection a very challenging task, due to the sheer scale of data and activity that take place daily. The short lifespan of political bots and other fake accounts and the fast emergence of new ones is a key challenge.

Quantitative findings of both academic researchers and social platforms alike have so far indicated that:

- **The volume of both fake accounts and their artificially amplified posts is significant.** For instance, Twitter are currently detecting and investigating 9.4 million suspect accounts each week (Harvey & Roth, 2018). In a retrospective analysis of all 2016 US election tweets, they established that Russian-linked, automated accounts generated 1.4 million tweets, but these were less than one percent (0.74%) of all election-related posts (Edgett, 2017).
- **Artificially amplified large volume does not necessarily lead to high engagement.** "In the aggregate, automated, Russian-linked, election-related Tweets consistently underperformed in terms of impressions relative to their volume on the platform". (Edgett, 2017)
- Automated accounts are being used to accelerate the spread of true and false content in equal measure (Vosoughi et al, 2018).
- Automated accounts however play crucial role in the early stages, to help misinformation become viral (Shao *et al*, 2018).

The key enabler behind all this work **are datasets** of proven bots and sock puppet accounts and all their social media data (e.g. posts, social profile, shares and likes). This "gold data" is necessary for the training of the machine learning algorithms for bot and sock puppet detection. Until recently, such datasets were created by academics (e.g. the DARPA Twitter Bot Challenge (Subrahmanian et al., 2016) and the Bot Repository[55]). However as part of their transparency drive, Twitter have now released a very large dataset[56] of more than 10 million Tweets and 2 million images and videos posted by 3,841 Russia-affiliated and 770 other accounts, potentially originating in Iran. This is a very

[53] https://botometer.iuni.iu.edu

[54] https://twitter.com/yoyoel/status/1058471828706930688

[55] https://botometer.iuni.iu.edu/bot-repository/

[56] https://about.twitter.com/en_us/values/elections-integrity.html#data

helpful first step and further such datasets and transparency initiatives are needed, especially for accounts operating in languages other than English.
Such **shared datasets are very important for independent research**, e.g. quantifying the impact of bots in diffusing political messages and influencing democratic processes. One of the recommendations of the EU High Level Expert Group on disinformation is to create a network of independent European Centres for (academic) research on disinformation (HLEG, 2018). One of the key functions of this network will be to provide "a safe space for accessing and analysing platforms' data and for a better understanding of the functioning of algorithms".

3.3. Detecting Mis- and Disinformation Campaigns

Journalists and researchers are increasingly faced with the challenge of verifying not only the authenticity of online content (e.g. a viral news story or a video), but also digging deeper, to find the originating source of viral disinformation, track its spread across online communities and networks, and establish the likely impact on citizens.

The technological challenges in carrying out large-scale studies of partisanship, propaganda, and disinformation campaigns are still very significant, however, as demonstrated by two of the largest studies of the 2016 U.S. Presidential Election (Faris et al, 2017) and the 2016 UK EU membership referendum (Moore&Ramsay, 2017). The former analysed both web and news media content, as well as their shares on Facebook and Twitter, whereas the latter focused purely on the role of UK media. Due to the large scale of these studies coupled with the lack of suitable easy-to-use tools for analysing propaganda campaigns, researchers had to employ primarily human coding and qualitative analysis, coupled with basic keyword matching and link mapping.

As the largest study comparing the spread and impact of true and false news has also shown (Vosoughi et al, 2018), it is not sufficient to only limit artificial amplification through bot detection:

> *"Contrary to conventional wisdom, robots accelerated the spread of true and false news at the same rate, implying that false news spreads more than the truth because humans, not robots, are more likely to spread it."*

In other words, it is necessary to study online disinformation and propaganda campaigns in their entirety. For this purpose, we advocate the adoption of the AMI model (Agent-Message-Interpreter) conceptual framework (defined in Section 1.2). Following this model, all three key elements of disinformation and propaganda campaigns need to be examined (Wardle&Derakhshan, 2017): the trustworthiness of the agents (original source, amplifiers, and the connections between them), the credibility and accuracy of the message, and the effects on the interpreters' beliefs (i.e. the targeted audience).

3.3.1. Agents: Source Trustworthiness and Information Laundering

A key aspect of the AMI framework is analysing the originating agents of the disinformation campaigns, the other key agents involved, and the explicit or implicit network connections between them. An essential aspect of that is the trustworthiness and credibility of these disinformation agents. This is being referred to as source checking in (Wardle & Derakhshan, 2017), who argue that it is hugely important, while currently overlooked, especially in terms of assistance from automated tools and approaches.

Journalism research (Lacy & Rosenteil, 2015) has proposed several metrics for assessing the quality of news and online media, such as partisan bias, structural bias, topical bias, and source transparency. However, there are currently no automated methods for calculating these. Such content-based source trustworthiness indicators complement the currently better understood indicators from bot detection research, discussed in Section 3.1 above.

Academic research has started proposing network models for trust modelling, e.g., for Twitter user networks (Wang, 2010), but these are still very much experimental. They typically take a naïve PageRank style score and propagate trust between nodes in the network automatically. To manage trust based on content requires some authoritative or near-authoritative sources. (Levien & Aiken, 1998) calculate global reputation for each actor in a network, and also allow the listing of certain nodes as 'bad', thus cutting out unreliable parts of the network.

There is also a strong need for further research around detection of hyper-partisan and biased sources. Journalism research (Lacy & Rosenteil, 2015) proposes a model of structural bias based on whether the source provides more coverage of certain viewpoints but not the opposing ones. For each source, topical diversity measures also need to be considered, where some automation assistance can be provided by automatic topic detection tools. There is also need to model topical diversity over time, in order to capture sock puppet and disinformation sites that do not post between major events or switch language or country focus over time. All these trust indicators are complementary to the network-based ones described above.

A promising new initiative is the Global Disinformation Index[57], which is building a real-time rating system of news source reliability worldwide. The aim is to have a neutral, independent, and transparent source of probability risk ratings (similar to those of bonds in finance), based on AI algorithms that identify junk domains automatically (Melford et al, 2018). This will be complemented by programmatic access (API) for advertisers, wishing to have access to an always up-to-date dynamic blacklist of untrustworthy sites.

The Journalism Trust Initiative (JTI)[58] is a complementary, journalism-led initiative, has been launched by Reporters Without Borders (RSF), Agence France Presse (AFP), the European Broadcasting Union (EBU) and the Global Editors Network (GEN). It aims to create within the next 12 to 18 months a voluntary standard of media credibility to assist consumers, advertisers, and online platforms with identifying and promoting credible information.

Disinformation agents are often not acting independently, even though this could be hard to establish sometimes. In order to give the impression that a large number of independent sources are reporting in different ways on the same 'facts', some disinformation sites and/or sock puppet accounts reuse and republish content, in a practice known as **information laundering** (Starbird, 2017). Journalists currently lack easy-to-use tools that show which alternative media sites or social network accounts have reused content from another. This is important, since hyperpartisan media and sock puppets are repackaging and/or republishing content in an attempt to give it credibility and gain acceptance through familiarity. So far, research has focused primarily on studying retweet and mention patterns between such false amplifiers, e.g. in the 2016 US Presidential Election (Faris et al, 2017), but technology for much more in-depth analysis is needed.

3.3.2. Message Credibility: Beyond Fact Checking and Content Verification

Propaganda and disinformation campaigns, with their focus on purposeful persuasion, one-sided messages, and influencing emotions and attitudes of its target audiences, are significantly more challenging to detect, analyse, and respond to effectively. Consequently, the required methodological and technological advances need to go beyond the fact checking and content verification work discussed in Section 3.1.

[57] https://www.disinformationindex.com

[58] https://rsf.org/en/news/rsf-and-its-partners-unveil-journalism-trust-initiative-combat-disinformation

In particular, the credibility of a claim/content is separate from its veracity (established through fact checking or content verification), since the former is about subjective perception of whether a message is correct, whereas verification is about evidence-based, objective assessment. Researchers (Zubiaga & Ji, 2014), for example, have found that non-expert crowd workers tend to judge inaccurate information as true. Their judgements were based on credibility perceptions, as they lacked the necessary expertise in cross-checking facts against external sources and other content verification techniques.

Manual content analysis in journalism (Lacy & Rosenteil, 2015) has considered partisan balance (the space given to sources of a particular orientation), structural balance (space allocated to opposing views), source diversity, source transparency, breadth of covered topics. In addition, other manipulation techniques are biased selection of topics and/or headlines, manipulative choice of words, defamation, discrediting, and mockery (SRF,2016).

The journalist-led Credibility Coalition[59] has proposed a number of content-based indicators for message credibility and annotated manually a small dataset of highly shared articles to validate the indicator definitions (Zhang et al, 2018). These indicators are the starting point for new web schema standards, or extensions to existing schemas (most importantly, schema.org), to be developed by the newly established W3C Credible Web Community Group[60]. Given the vast scale of online and social content, new tools for their automatic generation need to be developed at the same time. The indicators most correlated with disinformation were found to be clickbait title and some logical fallacies. Hedging, quotes from experts and citations of organisations and studies were also found to be important, in line also with the manual content analysis research discussed above (Lacy & Rosenteil, 2015).

While there are automated tools available for tasks like emotion detection and opinion mining, there is very little research on methods for automated credibility analysis of online disinformation. Linguistic cues, for example, have been used in analysing framing bias (introduced by subjective words and phrases) and epistemological bias (realised by lexical and grammatical linguistic features) (Recasens et al, 2013). Linguistic analysis has also shown that propaganda often targets morals and authority, and contains bias markers (Rashkin et al, 2017). There is a need to go beyond detecting deceptive framing, and address a wider range of techniques including sexist framing, appeal to fear and anger, and partisan imbalance in quoted information sources. There is also need to measure source credibility assessment, as well as transforming this experimental research into practically useful tools. Based on this data, machine learning algorithms for rumour stance detection were also developed in the PHEME project (Aker *et al*, 2017).

3.3.3. Interpreters

Interpreters are the recipients of the disinformation message, who themselves, if convinced, can act like agents by spreading the message further through their social networks. The largest present study of fake news spread on Twitter (Vosoghi et al, 2018) has concluded that: "…human behavior contributes more to the differential spread of falsity and truth than automated robots do. This implies that misinformation containment policies should also emphasize behavioural interventions, like labelling and incentives to dissuade the spread of misinformation, rather than focusing exclusively on curtailing bots."

Unfortunately, as argued also in (Vosoghi et al, 2018), our current understanding of the real impact of disinformation on the beliefs of exposed individuals and their subsequent actions is currently very

[59] https://credibilitycoalition.org/

[60] https://www.w3.org/community/credibility/

limited. Prior studies have typically measured impact based purely on quantitative metrics, such as number of retweets or replies.
As detailed in (Wardle & Derakhshan, 2017), however, exposed individuals can respond in one of the following ways: ignore the message, share in support, or share in opposition. The last two types of shares need to be distinguished, since supportive shares indicate that the interpreter believes the disinformation message (hegemonic response in (Wardle & Derakhshan, 2017)), whereas sharing in opposition is indicative of an oppositional stance. The PHEME[61] project analysed manually annotated rumours with respect to the stance taken by their interpreters: accepting, denying, questioning, or commenting (Zubiaga *et al*, 2016), which showed that disinformation is questioned more, attracts more affirmations and denials, and results in deeper conversation threads. Based on this data, machine learning algorithms for rumour stance detection were also developed in the PHEME project (Aker *et al*, 2017).

3.4. Malinformation: Hate Speech, Online Abuse, Trolling

Online abuse, hate speech, and trolling are online behaviours where an individual or a group is being attacked or discriminated against, on the basis of their political orientation, race, religion, ethnic origin, gender orientation, sex, or other characteristics. Researchers have differentiated online abusive language alongside two dimensions: explicit vs implicit and generic vs directed to a specific individual (Waseem *et al,* 2017).

The need to address abusive online language and behaviour is urgent, as recent research has shown: "social media has not only become a fertile soil for the spread of hateful ideas but also motivates real-life action" (Müller & Schwarz, 2018). In this particular case, researchers (Müller & Schwarz, 2018) studied the way in which right-wing anti-refugee Facebook posts are correlated with anti-refugee incidents in Germany. There are also many well-documented examples where journalists and media outlets are being targeted via online harassment campaigns, which often cause significant online and offline damage (RSF, 2018). Likewise, politicians (Gorrell *et al,* 2018; Phillips, 2017) and women (RSF, 2018; Zuckerberg, 2018) are also major targets.

Online harassment campaigns are often based on disinformation and employ the same amplification mechanisms. While Reporters without Borders have investigated these specifically in the journalism and media context (RSF, 2018), the same tactics are used also in other cases of directed online abuse:

1. **Disinformation**: journalistic or other legitimate content posted by the abuse target on social networks is drowned in a flood of online disinformation;
2. **Amplification**: impact of the disinformation is enhanced artificially by bots, ads, and paid-for trolls. Online harassment is devastating because the trolls writing the hate messages have an advantage over the journalists they target – virality. The threats, insults, and false information act as clickbait, creating outrage for those living within what the Internet activist Eli Pariser calls "filter bubbles." (RSF, 2018)
3. **Intimidation**: the targeted individual is subjected to a high volume of personal online abuse, threats, and disinformation, in order to discredit them and reduce them to silence.

Online harassment is most widely studied on Facebook and Twitter, but increasingly Reddit, 4Chan, YouTube, and WhatsApp are also employed. The actors behind these campaigns can be classified along several dimensions: financially or ideologically motivated; automated vs human; individual vs organised. Some political trolling campaigns can also have different degrees of state involvement: state-executed, state-directed, state-incited, and state-endorsed (Nyst & Monaco, 2018).

[61] http://pheme.eu

Establishing reliably the actors responsible for a given online harassment campaign can be very difficult, as they "*are designed to appear spontaneous and organic, camouflaged by the chaotic ephemera*" (Nyst & Monaco, 2018).

RSF (RSF, 2018) and others have documented aggressive cyber-harassment campaigns in many countries worldwide, including countries like Germany, France, and the UK who have already enacted or are in the process of defining legislative controls (see section 4.3).

Social platforms are aware of the problem and have started implementing semi-automated solutions, which are needed in order to screen efficiently the large number of posts and comments received daily. Facebook, for example, have implemented machine learning algorithms that can identify the most straightforward cases of abusive language (Allan, 2017). The core of the decisions, however, are made by humans, i.e. the platform relies on its users to report content, which is then reviewed by a human curation team. In 2017, the latter consisted of 4,500 Facebook employees worldwide, with their number set to increase by another 3,000 (Allan, 2017).

Currently, there are no publicly available figures on the accuracy of the automatic hate speech detection algorithms used by social platforms. In academic research, state-of-the-art methods currently have 65-70% precision (Wulczyn, Thain, and Dixon 2017). Even though the social platforms have access to additional, non-public information about the post (e.g. its originating IP address), the algorithms are still not sufficiently accurate to be used in a fully automated manner. For instance, recently Facebook's hate speech detection algorithms were triggered by part of the US Declaration of Independence, which resulted in the post being automatically withheld from initial publication (MacGuill, 2018).

However, there is even a more significant challenge, yet unsolved - that of having an agreed upon definition of what constitutes hate speech. For instance, when humans are asked to annotate racial slurs, they have been found to agree with each other in only 69% of the cases (Bartlett *et al*, 2017). The task of distinguishing polite from impolite tweets has been found easier for humans, with agreement ranging from 80% to 95% depending on the language of the tweet (Theocharis *et al*, 2016).

These findings demonstrate that dealing with online abuse is a very complex task, where even human curators can disagree. This has led to calls for platforms to not only act swiftly on reported content (Nyst&Monaco, 2018; RSF, 2018), but also to provide transparency of their curation practices and implement safeguards to protect freedom of speech (RSF, 2018).

3.5. Accuracy and Effectiveness

Social platforms and researchers are actively developing methods based on machine learning algorithms, in order to identify automatically disinformation on social media platforms. Given the extremely large volume of social media posts, key questions are:

1. Can disinformation be identified in real time?
2. How accurate/effective are these methods?
3. Should they be adopted by the social media platforms without human oversight?

The very short answer is: Yes, some of it can, although we are still far from developing highly effective socio-technical methods, which can be adopted in practice without any human oversight. When it comes to containing the spread of disinformation, we should be mindful of the problems which such technology could introduce:

- **Non-trivial scalability**: While some algorithms work in near real time on specific datasets such as tweets about the UK EU membership (Brexit) referendum - applying them across all

posts on all topics as a social platform would need to do, for example, is very far from trivial. Just to give a sense of the scale involved - prior to 23 June 2016 (referendum day) there were approximately 50 Brexit-related tweets per second or fewer. Twitter, however, would need to screen more than 6,000 tweets per second, triage them based on their topic, and then screen them for misinfomation. This is a non-trivial software engineering, computational, and algorithmic challenge.

- **Algorithms make mistakes**, so while 90 per cent accuracy intuitively sounds very promising, we must not forget the errors - 10 per cent in this case, or double that at 80 per cent algorithm accuracy. At 6,000 tweets per second, 10 per cent amounts to 600 wrongly labeled tweets per second rising to 1,200 for lower accuracy algorithms. To make matters worse, automatic disinformation analysis often combines more than one algorithm – e.g. first to determine which story a post refers to and second - whether this is likely true, false, or uncertain. Unfortunately, when algorithms are executed in a sequence, errors have a cumulative effect. The scale of potential errors therefore means that algorithms require human oversight and robust appeal procedures, in order to protect the fundamental human rights of freedom of expression and privacy.
- **Algorithmic mistakes can be very costly**: broadly speaking algorithms make two kinds of errors: (1) *false negatives* in which disinformation is wrongly labelled as true or bot accounts are wrongly identified as human; and (2) *false positives, when* correct information is wrongly labelled as disinformation or genuine users are wrongly identified as bots. False negatives are a problem on social platforms, because the high volume and velocity of social posts (e.g. 6,000 tweets per second on average) still leaves a lot of disinformation "in the wild". If we draw an analogy with email spam - even though most of it is filtered out automatically, we are still receiving a significant proportion of spam messages. False positives, on the other hand, pose an even more significant problem, as falsely removing genuine messages is effectively censorship through artificial intelligence. Facebook, for example, has a growing problem with some users having their accounts wrongly suspended.

Therefore, the best way forward is to implement human-in-the-loop solutions, where people are assisted by machine learning and AI methods, but not replaced, as accuracy is still not sufficiently highly and we need mechanisms to alleviate the danger of opaque algorithms censoring people.

Even when algorithms are not given free reign, there is still the question of how effective they are and thus do they really save time for journalists, social platform employees, citizens, and other stakeholders engaged in large-scale disinformation analysis. Typically the effectiveness of academic methods is evaluated against state-of-the-art baselines and performance is reported on relatively small-scale benchmark datasets (e.g. the rumours annotated in the PHEME project by SwissInfo journalists (Zubiaga *et al*, 2016)).

While useful for comparing algorithms against each other to establish the best one, such evaluations are rather limited in practical utility when it comes to predicting performance should they be adopted by social media platforms and search engines.

An important characteristic of evaluation datasets that must always be considered, is that they are a limited snapshot of the language and strategies of misinformation, as they were at the time when this dataset was created. Likewise, technological solutions trained on past data would be limited to the kinds of misinformation topics and strategies that existed at the time when the training data was created. However, as new kinds of misinformation start appearing over time, these can well be missed by automated tools that have not been re-trained or otherwise updated on more recent datasets, thus leading to significantly worse performance than originally reported. For example, in the case of recognising names of people, organisations, and locations automatically, this

phenomenon of performance decay is referred to as entity drift (Decrzynski, Bontcheva & Roberts, 2016).
Likewise, evaluation datasets that are several years old may not measure accurately the performance of automatic tools on current kinds of misinformation. Therefore, it is advisable, when using such older datasets, to supplement them with some newly annotated content, in order to ensure longitudinal document diversity and thus mitigate for temporal drift (Derczynski, Bontcheva & Roberts, 2016).

Given these and other limitations of evaluation datasets (e.g. their limited size), performance evaluation metrics and datasets only really enable the relative comparison of different automated tools on the same data.

Ultimately, technology validation is therefore best carried out through use in practical settings. Until such prolonged use demonstrates that automatic methods have matured to the point where they are capable of identifying and analysing misinformation at scale, in real-time, and with very high accuracy, they should be used only in a semi-automated, assistive fashion.

With respect to evaluating the effectiveness and accuracy of the bot detection solutions, hate speech identification, and other techniques already implemented by the social platforms, there is an even bigger challenge. These are all proprietary algorithms, the details of which are not disclosed in public and neither are any independent quantitative evaluations, if such exist. The only information available is from blog posts and press releases of the companies themselves, but, for example in the case of bots and fake account detection, what is disclosed is the number of accounts being suspended. While this is a good start, it is not sufficient as there needs to be also information on the number of wrongly suspended accounts and, where possible from a formal evaluation, also an estimate of the likely percentage of missed bot/fake accounts.

This is one of the reasons why this study recommends the establishment of ongoing, long-term cooperation and data sharing between scientists, journalists, platforms, and other stakeholders. This would enable the independent study, evaluation of effectiveness and oversight of algorithms. While there are already some bilateral collaborations between individual research labs and some social platforms/apps, these are not sufficient since any publications arising from such sponsored interactions raise the potential issue of conflict of interest and biased results.

Policy makers are also becoming aware of the strong need for algorithmic transparency and auditing algorithms deployed on social and other online platforms, not only for the reasons already discussed, but also to ensure their compliance with human rights legislation (Denham, 2018). On 27 November 2018, the UK Information Commissioner testified in front of a joint parliamentary committee of nine countries:

> *"Under the GDPR, there are enhanced requirements for explainability in algorithms in the context of personal data. If a company or an organisation is making a decision that is basically done by machines, it has to be explainable and it is challengeable. As another part of the work we do, we are going to be auditing algorithms for issues of fairness, data protection, bias—all those issues. So we are going to be in the space of looking at algorithms and looking at transparency... "*

Overall, while it is currently impossible to judge how well the existing algorithmic approaches of social platforms and web search engines work, there are continuously emergeing reports of:

- wrongly suspended accounts;
- failure to sufficiently prevent the spread of hate and online abuse with well documented negative real-world consequences, e.g. the role of Facebook hate content in the Rohingya genocide in Myanmar;
- increased exposure to junk news during the 2018 midterm elections as compared to 2016;
- recurrent trending of misinformation during shootings and crises.

Alltogether, this demonstrates that current automated solutions are not sufficiently effective.

In conclusion, there is also strong need to look beyond "simply" evaluating the accuracy and effectiveness of misinformation technology and also consider how succeptible to abuse are current algorithms:

> *"I have been trying to promote the idea of "abusability testing". Platforms invest resources into seeing how their platforms can be abused to harm consumers. I think that smart policy would incentivise that kind of investment, as we have seen that kind of incentivising around cyber security in the last 10 years."* (Soltani, 27 Nov 2018)

4. Legal responses

Under public pressure and media articles pointing the dangers of disinformation on our societies, and most notably the Cambridge Analytica scandal (see section 2.1.3 for details) lawmakers have become aware of the problem and started to draft regulatory responses.

In this section, we describe how different lawmakers have been approaching the regulation issues, the main paths adopted, along with the risks and opportunities that each is bringing. The challenges of user privacy and access to data for scientific studies and development of effective technology solutions to fake news is also discussed.

4.1. Self regulation

In this section, we describe approaches that go along with the self regulation theory. The self-regulation approach consists relies on stakeholders to conduct internal and holistic change, sometimes through incentive or discouragement (fiscal, reputation, etc.). On this issue, self-regulation could be seen as the measures that social media platforms like Facebook or Twitter decided to implement voluntarily in reaction to the issue.

Our definition of self-regulation initiatives can be summarised as the implementation of third-party partnerships to develop new services or extended capacities for the selected partners. For instance, Twitter, through its Healthy conversations initiative[62] announced by Jack Dorsey, Twitter CEO on March 1st 2018[63], granted two academic partnerships centred on defining indicators for healthy digital conversations through studies on uncivil discourse and echo chamber effects. This initiative aims at working with researchers on developing new conversation metrics that could help to better measure the disinformation issue.

Facebook implemented a similar approach with an initiative called Social Media impact on elections[64]. This action is designed to build a group of scholars that will define their own research agenda, solicit call of proposals for independent research and peer-review scholars that will participate to these calls. This approach could be defined as the setup of an independent research board.

Facebook also implemented a partnership with other stakeholders as fact-checkers. Through these partnerships like the one in Philippines[65], Facebook is giving special access and tools to fact-checkers so they can expand their work directly on the social platform.

However, all research partnerships initiatives should not be considered as self-regulation. For example WhatsApp, a Facebook-owned company, announced WhatsApp Research awards for social science and misinformation[66], an initiative focused on core research, not directly linked to WhatsApp products. Awardees from this initiative can submit research proposals that cannot be implemented through the social platform as WhatsApp would not give access or data or the capacity to test it on the platform to awardees. Therefore, we do not consider this initiative as self-regulation.

62 https://blog.twitter.com/official/en_us/topics/company/2018/measuring_healthy_conversation.html

63 https://twitter.com/jack/status/969234275420655616

64 https://newsroom.fb.com/news/2018/04/new-elections-initiative/

65 https://newsroom.fb.com/news/h/fact-checking-philippines/

66 https://www.whatsapp.com/research/awards/

It is interesting to remark that due to the high scrutiny under these platforms operate, they do tend to emphasise the independence of the third-parties they're collaborating with. The social media impact on elections initiative from Facebook has been running with 7 different foundations[67]. Solicitations of independent research under this program are managed by the Social Science Research Council, an independent and recognised body. They also emphasise the advanced selection process they implemented like Twitter:

> *"In the first round, each proposal was reviewed and scored by two different people. The 50 proposals that made it to the next round were each reviewed by a total of four different people, and then the 16 semi-finalists were further evaluated by a small committee of subject-matter experts. Finally, we interviewed a handful of finalists by video conference before making the final selection. By the end, we had written more than 350 reviews in total and examined the finalists in depth; as a result, we're confident in the process that led us to the two extraordinary partners we've selected."*

4.1.1. Risks and opportunities of self regulation

We see two main opportunities and three risks in this approach. First, there is a capacity for independent researchers to have access to data to better understand and seize the disinformation issue. This is clearly needed, especially at times where APIs are questioned after cyberattacks or piracy acts like Cambridge Analytica or the recent Google+ hijacking. Then, these third-parties can also rely on the platform to effectively implement new techniques, new methodologies and evaluate their efficiency. This is a crucial step to bridge academic research and social use in order to design better innovative solutions.

On the other hand, we raise the issue of independence of the third-parties both in their selection and in their actions. One risk is that social platforms may prioritise addressing certain issues which may not necessarily be the most important ones from disinformation containment perspective. There is also a risk third-parties could become fully dependent of the access and tools given to them. This could create distortion towards the research scope and evaluation of the implemented measures. Eventually this approach also raises the question of the cross platform expertise. As every platform is launching its own partnership, there is a risk expertise could not be fully shared and that a cross-platform vision could not be developed.

4.2. Co-regulation

Regulation of disinformation at a European level belongs to a co-regulation approach. From a structured dialogue instituted by the state or authority, this type of self-regulation is based on a compromise found between the actors, with implementation being monitored by the authority.

4.2.1. European Commission approach

In April 2017, Vice-President Andrus Ansip in charge of the completion of the Digital Single Market described fake news as "a serious problem"[68]. With the communication "tackling online disinformation: a European approach"[69], the European Commission chose to engage with stakeholders in order to define an action plan to tackle the spread of disinformation in Europe. At the end of 2017, the European Commission announced the creation of a high level expert group[70], composed of representatives of the civil society, social media platforms, news media organisations,

67 Laura and John Arnold Foundation, Democracy Fund, the William and Flora Hewlett Foundation, the John S. and James L. Knight Foundation, the Charles Koch Foundation, the Omidyar Network, and the Alfred P. Sloan Foundation

68 https://ec.europa.eu/commission/commissioners/2014-2019/ansip/announcements/statement-vice-president-ansip-european-parliament-strasbourg-plenary-debate-hate-speech-populism_en

69 https://ec.europa.eu/digital-single-market/en/news/communication-tackling-online-disinformation-european-approach

70 https://ec.europa.eu/digital-single-market/en/news/call-applications-selection-members-high-level-group-fake-news

journalists and academia. The group delivered its report in March 2018. As well as a public consultation and a multi-stakeholders conference, the report from the expert group served for the drafting of a self-regulatory code of practice released at the end of September 2018. Alongside the code of conduct, the European Commission has launched an independent network of fact-checkers and an online platform on disinformation, as well as measures aiming at enhancing media literacy. The monitoring of the progress made will examine the need for further actions.

The European level certainly is the best scale of regulatory intervention. The code of practice agreed with online platforms seems like a good starting point and the consultation of experts in the high level group certainly brought significant knowledge to the topic. But in the meantime, we can regret the lack of implications from NGO and civil society besides journalists in the expert group. As the commitments from the platforms rely on a voluntary basis, there is a need to monitor their effective implementation and decide for further actions accordingly.

4.2.2. Belgian platform

In May 2018, the Belgian minister in charge of digital agenda launched the consultation of an expert group. This expert group, composed of academics, media representatives and NGOs, made a series of recommendations in order to tackle the spreading of fake news and disinformation online. Parallel to the expert group, a citizen consultation platform stopfakenews.be was launched to gather citizen opinion on disinformation. In its conclusions[71], the expert group suggests not to legislate in a repressive manner but to foster research, media literacy and an enhanced and ongoing dialogue with the platforms. The next step of this process has been announced by Charles Michel, Belgian Prime Minister on Monday, October 8th, through the creation of a fund dedicated to fund civil society initiatives related to fact-checking and disinformation[72].

One particularly strong aspects is that the Belgian initiative involves a wide range of actors and the civil society as a whole in this process. It is a good starting point for fostering research and cross-expertise dialogue on that topic before considering harder regulation. The creation of fund dedicated to support civil-society initiatives certainly will strengthen the disinformation-fighting community structure and build stronger initiatives. Their initiatives implementation will be monitored closely.

4.2.3. Denmark

In September 2018, the Danish government has launched an action plan of 11 initiatives[73] in order to strengthen safeguards against influence on Danish democracy and society. In particular, an intergovernmental task force has been set up in order to coordinate and develop countermeasures against influence campaigns. Hence, the Ministry of Foreign Affairs, has launched a strengthened monitoring of disinformation. Particular attention will be payed to the spreading of misinformation in the context of the upcoming Danish parliamentary elections. Also, the government intends to initiate a dialogue with the media and the platforms to find models of cooperation on countering potential foreign attempts to influence the upcoming parliamentary elections.

The Danish initiatives on fighting disinformation mostly rely on governmental forces and the Ministry of Foreign affairs to safeguard the Danish democracy from foreign influence. If such action is key to the country cyber security, we can only welcome the complementary dialogue with platforms and the media. Otherwise, state-based only initiatives could be considered as counter

71 https://www.decroo.belgium.be/fr/un-groupe-d%E2%80%99experts-formule-des-recommandations-pour-lutter-contre-les-fake-news

72 https://decroo.belgium.be/nl/fact-checking-fonds-de-steigers-strijd-tegen-fake-news

73 http://um.dk/en/news/NewsDisplayPage/?newsID=1DF5ADBB-D1DF-402B-B9AC-57FD4485FFA4

propaganda and will thus not resolve the problem of disinformation, while damaging the government's reputation.

4.2.4. Opportunities and risks of co-regulation

Co-regulation opens up the prospect of cross expertise collaboration by bringing very different stakeholders involved at different moments of the disinformation chain. The diversity of expertise supports new ideas and solution designs, especially if the organisations involved have not previously worked together. In that specific case, we consider the involvement of civil society and academics in this field vas a welcome and necessary new development which is likely to bring significant benefits.

The drawback of cross-expertise collaboration is the difficulty to mix methodologies and cultural habits. Therefore, solution design and implementation can be harder to realise. For instance, peer review methodology from academics might not be applicable in fact-checking organisations. Thus, a significant effort is needed from all stakeholders to focus their interests on a reduced number of activities and leads. Moreover, evaluation of co-regulation is also difficult as indicators need to be built in advance and be easily measurable.

4.3. Classic regulation

Making an observation that self-regulation only has failed, several countries chose to take stricter actions and draft legislation forcing the social media networks to take responsibility for the content published on their platforms.

4.3.1. German regulation

The Act to Improve Enforcement of the Law in Social Networks NetzDG[74] entered into force on 1 October 2017. It covers defamation, dissemination of propaganda material, incitement to commit serious violent offense endangering the state, public incitement to crime, incitement to hatred and the distribution of pornography. The Act applies to social media networks that have two or more million registered users in Germany. The Act set up a procedure for online platforms to remove "obviously illegal" posts within 24 hours or risk fines of up to €50 million. The Act makes the platforms accountable for the monitoring and removal of content. In particular, the handling of complaints shall be monitored via monthly checks by the social network's management.

The German law might be considered as the most accomplished regulatory effort to tackle harmful content online, but its implementation has been difficult due to a number of omissions in the text, thus creating a regulatory uncertainty. In particular, the German legislator did not give a sufficiently strict definition of how to recognise manifestly unlawful content. It is therefore unclear how social media platforms will take down illegal content within the intended timeframe. Moreover, there is no appeal procedure for the user whose content has been removed. It must also be noted that the Act does not have enforceable mechanisms to combat misinformation.

4.3.2. French regulation

In January 2018, the French president Emmanuel Macron announced a law[75] aiming to prevent the spread of false information and propaganda during an election period. The draft law would apply during the three months preceding an election. France already has a law against the spread of false information, but according to the government this would be insufficient to quickly take down

74 https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf

75 http://www.assemblee-nationale.fr/15/propositions/pion0799.asp

information online, in particular in "the recent electoral climate", referring in particular to the US elections and UK referendum.

The law would give the Audiovisual Council (the French broadcasting regulator), the authority to block foreign state-controlled broadcasters that publish false information. Users would have the possibility to denounce potential false stories, and social media platforms would have to justify that they take action to remove them. Also, electoral candidates would call on a judge to take down a reported false story within 48 hours.

After a difficult debate between the Assemblée Nationale and Senate, a final text has been adopted in November 2018. Nevertheless, the law proposal has been largely criticised, as some consider it would threaten freedom of speech and have perverse effects on the democratic debate. In practice, the judge might not be able to make a decision in 48h, as we know that fact-checking is a longer process. Besides, with the majority of viral misinformation spreading within a matter of hours, rather than days, it is unclear how effective this approach would be even in cases where content is removed within the statutory 48h window.

4.3.3. UK regulation

Amid an investigation of Russia's reported use of social media to spread disinformation about the Brexit referendum in 2016, the British government set up a national security communications unit with the task of "combating disinformation by state actors and others." After the "Cambridge Analytica" inquiry, the Home Office and the Department for Digital, Culture, Media and Sport (DCMS) started an 18 months investigation. The DCMS report (DCMS report, 2018), published in July 2018, recommends social media companies to take more responsibility for the content published on their platforms. Following this investigation and report, the lawmakers plan to announce a new legislation tackling "social harms" by winter 2018. In the meantime, the DCMS committee has continued its investigations on the topic, most notably with the hearing of the Facebook Vice President of Policy Solutions on 27 November 2018 (Allan, 2018), that was attended by parliamentarians from 8 countries around the world.

The new proposals will likely include the introduction of a mandatory code of practice for social media platforms and strict new rules such as "takedown times" forcing websites to remove illegal hate speech within a set timeframe or face penalties. This could be implemented by a regulation authority similar to Ofcom, the UK communications regulator. Another option suggested by the LSE Commission on Truth, Trust and Technology composed of MPs, academic and industry leaders[76], would be the creation of an independent platform agency that could provide a permanent forum for monitoring and reviewing the behaviour of online sites, and produce an annual review of "the state of disinformation".

4.3.4. Risks and opportunities of regulation

A classic regulation approach on disinformation brings many risks. First, territoriality is one of the main concerns as disinformation content can be created, distributed and amplified from outside national territory. Moreover, the main stakeholders like the social media platforms are very often international corporations difficult to sanction. Thus, it is very difficult to enforce controls on external stakeholders.

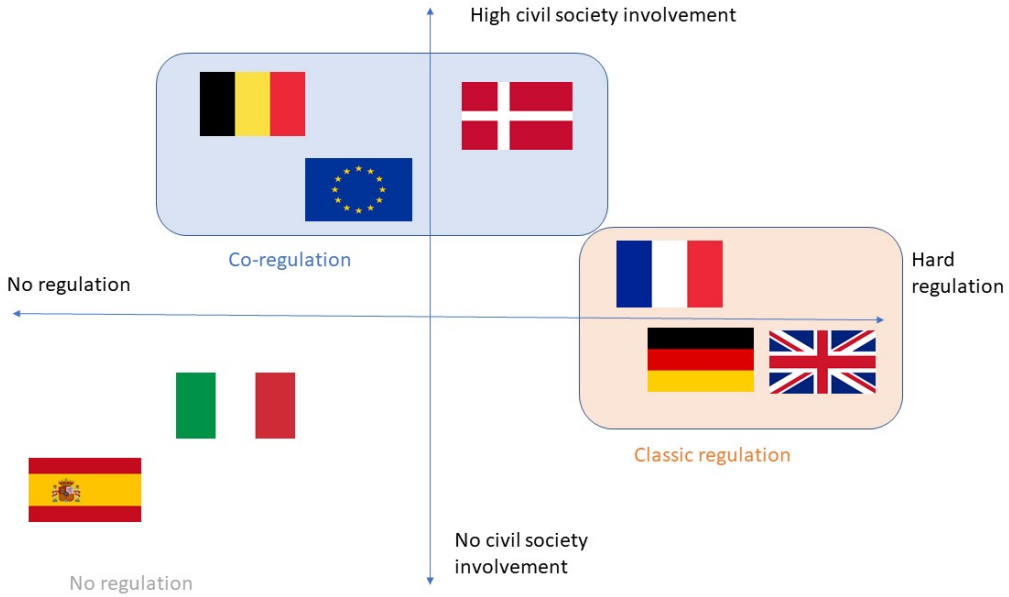
Moreover, we do notice there are derivative risks on freedom of expression as most of the regulations trying to address disinformation issue are labelling it as "false information" or "non-illegal content". The very complex nature of disinformation, as exposed in our problem definition,

[76] http://blogs.lse.ac.uk/mediapolicyproject/2018/11/19/tackling-the-information-crisis-a-new-report-from-lse/

creates a regulation scope where censorship could be reinforced. Thus, classic regulation empowers the judicial actors to rule on the truthiness of an information. For instance, the distorted use of the term "Fake news" to qualify political opponents discourse in non-democratic regimes has increased with regulation in Egypt, which opened complaints against journalists and bloggers[77].

4.4. Compared approach on regulation

Figure 17: Classification of national regulations



Source: own elaboration

Figure 17 shows the national initiatives classified regarding the type of regulation, ranging from no regulation to hard regulation, relative to the civil society involvement. Spain has not yet implemented any specific regulation on disinformation. In Italy the Postal Police[78] has been charged to issue warning on false reports. Citizens can also alert the police in case they spot wrong information. Then we see the co-regulation initiatives implemented by the European Union, Belgium and Denmark, which foster greater civil society involvement. Then, France, UK and Germany have adopted strict regulations, giving much power to the judiciary, and less cooperation with the civil society.

4.5. National and International Collaborations

Since online misinformation affects pretty much every country in the world and every social platform, there is a strong argument for multiplying and amplifying the effects of individual, organisational, and national initiatives and research through the establishment of cross-disciplinary international collaborations.
Due to the global nature of tech companies, regulation should involve international cooperation. In this respect, the European Commission initiative will lay the foundations for further work.

77 https://www.theguardian.com/global-development/2018/jul/27/fake-news-becomes-tool-of-repression-after-egypt-passes-new-law

78 https://www.bloomberg.com/news/articles/2018-06-07/who-you-gonna-call-postal-police-is-italy-s-fake-news-fix

One already existing successful European initiative is the East Stratcom Task Force and the related EUvsDisInfo website[79]. The latter has a rather narrow remit with its focus on pro-Kremlin disinformation. Also at EU level, the recent establishment of the independent High Level Expert Group (HLEG) on Online Disinformation is also a promising step towards a Europe-wide, collaborative approach (HLEG report, 2018).

However, as the Cambridge Analytica scandal has shown[80], there are also other political actors promoting disinformation and online propaganda and some of them also operate in different countries.

As representative bodies, national parliaments as well as the European parliament have felt the urge to work together on questioning the accountability of platforms. Such cooperation was particularly visible during the auditions of Facebook by the European Parliament, and more recently the British parliament, together with nine parliaments gathered in a "Grand committee" questioning the company's accountability on the "Cambridge Analytica" scandal and its aftermath. Following this audition, parliamentarians from across the world signed a declaration on the "Principles of the Law Governing the Internet"[81], stating that "The internet is global and law relating to it must derive from globally agreed principles". G7 countries are considering a coordinated action on disinformation, also following reflexions in international fora such as the OCDE and the Internet Governance Forum of the United Nations.

This has also motivated a number of prime ministers, presidents, and Nobel peace prize winner to issue a joint statement[82] on 14 November 2018, calling for a sustained and coordinated approach:

> *"We must act now to protect our access to independent, pluralistic, facts-based information, which is essential for people to freely form their opinions and play an active and constructive role in democratic debates. "*

The need for stronger international collaboration is also echoed in some of the recommendations of the UK DCMS Parliamentary Inquiry into online disinformation (DCMS report, 2018) and the UK Government's response (DCMS HC 1630, 2018).

[79] https://euvsdisinfo.eu/

[80] https://www.newyorker.com/news/news-desk/new-evidence-emerges-of-steve-bannon-and-cambridge-analyticas-role-in-brexit

[81] https://www.parliament.uk/business/committees/committees-a-z/commons-select/digital-culture-media-and-sport-committee/news/declaration-internet-17-19/

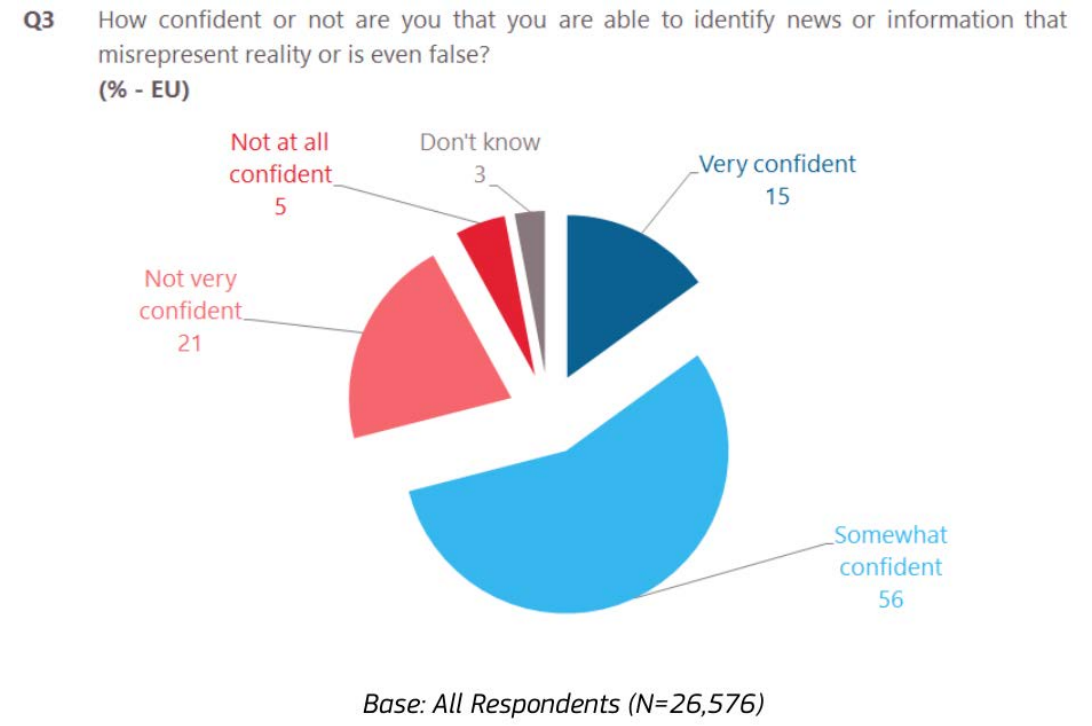
[82] https://www.thestar.com/opinion/contributors/2018/11/14/democracies-must-take-action-against-threats-to-freedom-of-expression.html

5. Social and Collaborative Approaches

A 2018 Eurobarometer survey (Eurobarometer, 2018) of 26,576 EU citizens in the 28 EU member states established that only 15% of the respondents felt very confident in identifying false news content. The first step of the verification process is the ability to distinguish factual from opinion statements, where a Pew Research Center study[83] has shown that (on average) only 25% of Americans are able to recognise factual news statements (Gottfried & Grieco, 2018), with the number rising to 33% for younger Americans.

This lack of knowledge, coupled with the fact that legal and technological approaches cannot fully eradicate online propaganda and disinformation, has underlined the importance of finding ways to turn citizens away from being genuine disinformation amplifiers and into engaged disinformation filters.

Figure 18: A Eurobarometer on fake news and disinformation



Source: (Eurobarometer, 2018)

5.1. Media Literacy

In that context, media literacy and critical thinking are widely regarded as the key skills required by citizens to engage more effectively with online information and social platforms. According to Wikipedia[84], media literacy "encompasses the practices that allow people to access, critically evaluate, and create media".

[83] The online quiz used by this study is available at: http://www.pewresearch.org/quiz/news-statements-quiz/

[84] https://en.wikipedia.org/wiki/Media_literacy

As only a fraction of the EU population is "born digital" and have developed such advanced media literacy skills, the public consultation[85] on fake news and online disinformation concluded that is it necessary to establish and strengthen "efforts in increasing media literacy at all levels, from school pupils to adult audience, and among actors, from end-users to journalists".

Gamification, i.e. teaching through participation in a game, is an engaging way for people (not just school children) to gain knowledge and experience. The Drog initiative[86] has brought together academics, journalists, and media experts to build an online game - Bad News, which aims to educate people about the various tactics employed in online propaganda and disinformation campaigns, by offering a hands-on experience in how they work in practice (see Figure 19, left). Another similar educational game is Fakey[87] (see Figure 19, right) by the University of Indiana. It asks players to share or like credible articles and report for fact-checking suspicious ones. The game shows headlines, images, and the first couple of sentences of each article, in a simulation of a standard social media timeline. Article source information is only revealed if the player presses the Hint button.

Figure 19: Disinformation tactics taught in the Bad News game (left) and the rules of the Fakey game (right)

Source: Left Right

There are also more traditional, school-based approaches to media literacy, which are targeting pre-teens and teens, just as they start taking interest in social media, news, and politics. Lie Detectors[88] is one such non-profit initiative in Belgium and Germany, which puts journalists in the classrooms to interact directly with pupils and teach about news literacy and news verification practices. There is a similar initiative in France led by journalists from Le Monde[89]. In the UK, the BBC has been running media literacy training in schools[90] and is now rolling out this programme in India, Kenya,

85 https://ec.europa.eu/digital-single-market/en/news/synopsis-report-public-consultation-fake-news-and-online-disinformation

86 https://aboutbadnews.com/

87 https://fakey.iuni.iu.edu/

88 https://lie-detectors.org/

89 https://www.lemonde.fr/les-decodeurs/article/2017/02/02/le-monde-s-engage-dans-l-education-a-l-information_5073215_4355770.html

90 https://www.bbc.co.uk/newsround/39032291, https://www.bbc.co.uk/cbbc/quizzes/real-or-fake-news-quiz

and more globally through its Beyond Fake News project (BBC News, 2018). In Italy, politicians have spearheaded a program[91] to teach media literacy to high school children. For older teens (15-18 years old), the International Factchecking Network has produced a role-playing card game[92] (in English, Italian, Portuguese, and Spanish), with students playing newsroom journalists covering a controversial referendum, marred by online propaganda and disinformation.

With recent studies showing that the older generation (above 50 years old) has lower than average ability to recognise factual information (Gottfried & Grieco, 2018) and remember already debunked false claims (Mantzarlis, 2017), one very significant challenge is how best to deliver media literacy information to that generation. The previously discussed approaches are not suitable, as older people are much less likely to use online games and rely significantly less on social media platforms as their source of news (Ofcom 2018). One promising way is to deliver special programmes through mainstream TV channels, similar to the BBC's Beyond Fake News project, which has just launched an entire series of documentaries, special reports and features across the BBC's networks in Africa, India, Asia Pacific, Europe, and the Americas and is delivered via TV, radio, and online.

Online materials on verification and fact checking, aimed at the general public are also increasingly becoming available (e.g. Edutopia[93], a New York Times lesson plan[94]), some of them also in languages other than English. To make the process easier to understand, the International Fact Checking Network has produced a 7-step fact checking cartoon[95], currently available in English, French, Italian, Portuguese, and Swahili. The UK independent fact-checking charity Full Fact has produced a similar, 10-step misinformation detection toolkit[96], as well as offering a collection of children-oriented literacy materials. The main challenge at the moment is how to help the general public (especially those holding polarized views) to discover these and invest the time to learn and practice in mindful social media engagement behaviour.

5.2. From Amplifiers to Filters: Citizens' Role in Disinformation Containment

In addition to utilising AI algorithms to flag false content, social platforms have also empowered their users with the ability to flag suspicious content, which is then checked manually by the platform's content moderators. This is another reason why raising citizen awareness of online disinformation and knowledge of how to recognise and contain it are even more important.

Recent research has demonstrated that some citizens, when made aware of misinformation, can alter their online behaviour and help reduce further misinformation spread by refraining from liking or re-sharing. For example, when users encounter a false post on Facebook and are shown related stories that correct this misinformation (e.g. from fact-checkers), their belief in the false post is significantly reduced (Bode & Vraga, 2015). This is the advanced warning technique from social and behaviour science research (Cook *et al, 2015)*. Initiatives such as Politifact's Truth-o-meter also help, as they increase transparency and help voters see which politicians make trustworthy statements.

While changing user online behaviour can significantly reduce organic amplification of disinformation, the current challenge is in finding the best way to make such a behaviour change

91 https://www.pri.org/stories/2017-10-31/italian-politician-wants-kids-become-fake-news-hunters

92 https://factcheckingday.com/lesson-plan

93 https://www.edutopia.org/blogs/tag/media-literacy

94 https://www.nytimes.com/2017/01/19/learning/lesson-plans/evaluating-sources-in-a-post-truth-world-ideas-for-teaching-and-learning-about-fake-news.html

95 https://factcheckingday.com/articles/24/this-cartoon-has-7-tips-for-fact-checking-online-information

96 https://fullfact.org/toolkit/

take place. This is particularly difficult at present, as we are still lacking sufficient understanding of "...what demographics are most vulnerable or most likely to be targeted, why they are receptive to disinformation, and the mechanics of how disinformation spreads within their networks online and offline is key to finding effective solutions in the long term." (Monaco, 2018)

Further research and algorithm development on these topics are very important, for two reasons. Firstly, debunking or labelling misinformation is not effective for highly polarised users who have already strong belief in misinformation by encountering it many previous times - the familiarity effect (Swire *et al*, 2017b). Psychology and social science research on debunking and misinformation has also demonstrated that "persistence [of beliefs in misinformation] was stronger and the debunking effect was weaker when audiences generated reasons in support of the initial misinformation." (Chan *et al*, 2017).

Secondly, such better demographic modelling would also help with formulating more effective strategies to change online citizen behaviour. Researchers, for example, have found that longer, more explanatory debunks are more effective in changing the beliefs of some users (Chan *et al,* 2017; Swire *et al*, 2017b). However, at the same time the same longer debunks amplify the misinformation-persistence effect in other users (Chan *et al*, 2017).

These findings demonstrate the need for a more personalised approach towards persuading social platform users to reduce their belief in and exposure to online misinformation. One promising avenue is to use micro-targeting to amplify the effects of debunking, e.g. by suggesting debunks from sources that are better aligned with the user's political views. Bots can likewise help to scale up such personalised debunking efforts. This approach will be experimented with in the new WeVerify[97] H2020 research initiative. If successful, social platforms themselves could adopt such an approach, based on the fine-grained user profiling information that they already have.
Citizens can also be effective in mitigating online abuse. For example, as part of the TrollBusters initiative, online communities of users have helped counter trolls by posting positive messages that drown out the abusive content from the troll account(s) (RSF, 2018).

Civil society organisations and citizens also play an important role in checking continuously that social platforms protect freedom of expression and implement robust mechanisms for appealing wrongly suspended accounts and deleted posts. The Santa Clara Principles on Transparency and Accountability in Content Moderation[98] call on social platforms to report regularly on how many posts and accounts are suspended/deleted and the reasons behind these actions, as well as provide a timely appeal mechanism where decisions are re-reviewed by a human moderator. As detailed in a recent civil petition towards Facebook[99], this is essential for protecting freedom of expression and absolutely necessary since the platform's algorithms and moderators do make mistakes and given the number of users and daily posts on the platform, these do amount to a significant problem.

Recent victims have been a Danish politician[100] and the Philadelphia Museum of Modern Art, when art photos posted by these accounts were censored by Facebook's nudity detection algorithms. What these cases also demonstrate is that current AI algorithms are still very limited in their "understanding" of culture and context and thus, it is essential that social platforms do not implement fully automated processes, without humans in the loop.

[97] https://www.weverify.eu

[98] https://santaclaraprinciples.org/

[99] https://www.hrw.org/news/2018/11/14/open-letter-mark-zuckerberg

[100] https://www.bbc.co.uk/news/blogs-news-from-elsewhere-35221329

5.3. Journalist-Oriented Initiatives

Journalists, researchers, and some organisations have created a number of initiatives and resources, aimed at improving news quality and fighting online misinformation.

The recent Yellow Jacket movement in France is considered by fact-checkers like Guillaume Daudin[101] to have helped enlarge the fact-checker's audience and credibility. Debunking posted by Daudin's team seems to have been shared more than the initial hoax[102], although no impact studies have been conducted to assess the audiences reached by the debunking story and whether they overlap with those spreading the hoax.

Firstly, a growing amount of verification literacy materials and programmes are being created, e.g. the UNESCO Handbook for Journalism Education and Training (Ireton & Posetti, 2018) and the learning module on the history of disinformation (Posetti & Matthews, 2018). The First Draft initiative[103] also provides courses for journalists in verifying media, websites, visual memes, and manipulated videos.

A second kind of useful shared resources for journalists are aimed at helping with accurate reporting, e.g. by providing a trustworthy resource of the latest research on key news topics[104], latest advice on media engagement strategies[105], a centralised resource of public government data[106] or thoroughly fact-checked information and statistics on economy, health, immigration, etc[107]. Many of these resources, however, are currently country- and language-specific and are designed purely for human consumption. Their usefulness in fact-checking and content verification can be improved further, if they are also equipped with programming interfaces and support data interchange standards.

There is also now awareness that journalists can benefit from working more closely with researchers, in order to benefit from the latest advances in psychology, social science, and data science (Lazer *et al*, 2017). There is also scope for learning from experts in information operations and strategic communications (Jeangène Vilmer *et al*, 2018), e.g. around the best debunking strategies for countering misinformation.

Since media itself and journalists are often targets of online disinformation campaigns and online abuse, there are now also social initiatives aimed at tracking press freedom and journalist abuse. One such example is the U.S. Press Freedom Tracker, which is a nonpartisan online resource supported by more than 20 press freedom organisations, including Reporters Without Borders (RSF 2017)

Trusted journalists and media are also in the position to carry out new research into the origins and mechanisms for spreading online misinformation, e.g. through data analysis and interviews. For instance, the BBC's Beyond Fake News project has released new research into misinformation on encrypted messaging apps in India, Kenya, and Nigeria (BBC News, 2018). Unlike Twitter, Facebook, YouTube, and Reddit, information on such closed apps is not available to independent researchers

[101] https://www.poynter.org/fact-checking/2018/the-yellow-vest-protests-showcases-the-enduring-reach-of-misinformation-and-the-desire-for-fact-checking/

[102] https://twitter.com/Jacques_Pezet/status/1065653483108515844

[103] https://firstdraftnews.org/en/education/learn/

[104] https://journalistsresource.org/

[105] https://mediaengagement.org/

[106] https://datausa.io/

[107] https://fullfact.org/finder/

through public APIs, which is currently a major limiting factor in academic research on online misinformation.

6. Initiatives Mapping

This section first discusses the online survey of disinformation initiatives that was carried out as part of this study. Next, Section **Error! Reference source not found.** presents a summary of EU and national technological, legal, and social initiatives for fighting online misinformation.

6.1. Survey of initiatives

During this study an online survey was used to gather the opinions and experience of different initiatives about disinformation. The survey run from August to October 2018 and received 16 responses. The survey questionnaire is available in the Annexes. The purpose of the survey was to collect information from different initiatives, which has been incorporated in the different sections of this report. It also aimed to collect the views of the initiatives about disinformation, obstacles they experience, collaboration efforts, legislation and policy measures against it. This section describes the initiatives views.

6.1.1. Initiative obstacles to achieve their objectives

Obstacles to data access

The first obstacle reported most frequently is about access to data from social media networks. Data access is currently reported as either impossible or with problems. Some social media platforms lack data access schemes, while others offer limited access and, in some cases, very expensive. Access to historical data is needed in addition to real time data, as sometimes the relevance of some data becomes clear only after the event or it could be removed. When content is removed, it is unclear if the platform or the user removed it and it is impossible to determine the original piece of information. In general, the responses report a lack of political willingness from the platforms to help media or research activities on content verification.

Obstacles about resources

Human, financial and infrastructure resources to make new and keep current initiatives operational were also identified as an obstacle in the responses. Financial resources are linked to having appropriate funding for personnel, debunking activities and the tools needed to support these activities. Since debunking disinformation is an area with limited commercial potential, securing funding for these activities from the private sector is difficult and public funds are needed.

An additional resource related obstacle extends to the lack of resources dedicated to fact-checking, as well as, the interest and time/effort availability of journalists as main actors in the debunking processes. Although in theory everyone agrees that fact-checking is important, in practice, some media organisations devote minimal to zero resources to verify news before publishing.

Although technology can strongly support content verification, humans are still required and - at least for the near future – have the role and responsibility to decide about information being true or false and the many in-between results of partially true or false. Usually, tools do not give a final verdict as to whether a claim is real or not but help users by providing useful relevant information, which they need to analyse and interpret to complete their assessment.

Lastly, infrastructure resources are related to the need for having available high-performance computational power and high storage space to store and analyse the large scale of the available data.

Obstacles to diffusion of debunking disinformation

One of the responses quoted the nature of the disinformation spread as being extremely effective. The overwhelming consequences it has on raising a political divide and a disputed political climate can consequently swing votes in favour of populism. This facet of disinformation was reported as the “catalyst on a minefield of social issues”. The lack of an effective diffusion strategy or process of

debunking results to large audiences was reported. This aspect was also mentioned in the 3rd case study of this report about disinformation during the French elections in 2018.

Technological obstacles
Reported obstacles also extend to weaknesses found in available technologies with some technologies still requiring further research. While the visual content found online, images and video, are of low quality, it makes the work of technological solutions more difficult.
Frequently, the available tools may not scale well on cases with large data sets and the process of data analysis in the background is not always clear to end users, e.g. how algorithms calculate a particular score. Technical skills may be required to install the newly developed tools and their components, being the outputs from research efforts and often, the results of these tools are hard to communicate to non-experts.

Legal obstacles
Legislation imposed actions do not seem to introduce significant obstacles, other than compliance with GDPR.

6.1.2. Collaboration amongst the initiatives

The responses received show ongoing collaboration efforts between some of the initiatives. Overall, initiatives are aware of each other and participate in each other's networks. Collaboration amongst research projects, academia, civil society organisations, human rights organisations, media literacy associations and teams of social media networks working on fact checking was reported.

A lot of the existing initiatives are collaborative efforts building on the outputs and experience of previous collaborations. Research project partners continue their work on new projects and in some cases, consortium partners have formed private partnerships, e.g. Truly Media. Initiatives have also funded the work of other initiatives, e.g. the development of the resource "A Field Guide to Fake News" by the Public Data Lab was funded by First Draft.

6.1.3. Legislation as a measure to fight disinformation

The initiatives that responded to the survey do not show a uniform opinion on whether legislation is an effective measure to fight disinformation, proving the complexity of the question. Some responses report that legislation is already available but not enforced rigorously.

One approach to legislation against fake news that was reported is to make accountable the responsible entities, when there is proof of disinformation beyond reasonable doubt. A different approach is to regulate the channels and web platforms that act as a tool to spread disinformation, in the same way media organisations are regulated, resulting in an approach of unified handling of mass media and the web.

A legislative approach to disinformation should be able to adapt to the acceleration of the high-speed societies, e.g. very quick bailiff's reports. On a similar note, a reason why legislation may not be effective is also the high-speed that technological tools progress and their use is changing, introducing new ways that disinformation might appear and spread.

It was reported that the emphasis of legislation should be more on fighting the quick spreading characteristic of disinformation, "virality", than the actual disinformation content. Focusing on the disinformation content and the publisher introduces a risk of restricting or censoring free speech. This topic is the main point of those who believe that legislation is not the right measure to fight disinformation and stresses the importance of adopting a very careful and moderate

implementation. A false positive result in this case, may end up being used to censor legitimate content and news.
Those not supporting legislation as a solution think that any legislation that targets the content and content providers will not be effective at addressing the problem. Instead it will cause more harm than good. Similarly, putting the onus on social media platforms to remove content that is "fake" could lead to collateral censorship, where platforms over-censor to avoid liability. This could have a chilling effect on freedom of speech and degrade the overall value of these platforms for legitimate democratic participation and deliberation.

A response that does not support legislative actions, describes that democracy must also tolerate false propositions. Adequate provisions should be made to educate citizens and ensure tools are available for self-education or self-training.

6.1.4. Policy actions as a measure to fight disinformation

The most reported policy action by respondents is about media literacy and public awareness. Policy should focus on empowering civil society, journalists and activists to enhance their media literacy and learn ways to spot disinformation and avoid spreading it further.

Citizens should be aware about the impact of disinformation, about its scale in terms of how often it appears and how important it is, about the critical thinking they need to exercise and about the tools and resources available to them that can help verify information. Media literacy and awareness is reported as an essential skill that should start early in classrooms and be carried out throughout school education, even primary education. Collecting and exposing to the broadest public what is happening in the disinformation ecosystem will result in more actors establishing common ground and can help bridge the gaps between academia and civil society.

The suggestions for policy action to promote investment in media literacy are followed by suggestions for policy actions to achieve improved technology for fighting disinformation. Funding for research efforts is quoted but also pressure on web platforms as a leverage to establish a more effective collaboration with independent research efforts and allow independent auditors to analyse the results that algorithms produce for different users.

The legislative approach is also reported in the responses, raising the concerns described in the previous section and firstly suggesting enforcement of existing rules and regulations more strictly. Quick and speedy rulings, e.g. court capacity, were mentioned in addition to making legal issues clearer to society at a greater level. Legislation is considered important but, in all cases, it should be applied very carefully in order to protect free speech. The role of governments was also described as not being directly involved but encouraging the private industry to establish standards and certification.

Policy actions suggested involved efforts to improve transparency, e.g. around political advertisements, publishing newsfeed statistics, publishing take-down statistics, improved standards for content moderation, implementing an appeals process for users, and hiring diverse and globally representative moderators. Policy actions don't necessarily directly address disinformation but deal with some of the larger systemic issues that are the root cause of why disinformation spreads quickly on social media platforms and aim to introduce the right measure to reduce this phenomenon and ecosystem.

A European observatory of disinformation is suggested in the responses without a clear description of its objectives. Considering the responses in other questions, it can be assumed that its role can be related to monitoring the disinformation phenomenon, raise awareness and drive policy actions.

6.2. Roadmap of initiatives

As part of this study, over 55 EU member state and 15 EU-level initiatives towards fighting online propaganda and disinformation were identified. The list is being continuously updated and is available on demand, as new ones emerge. A current snapshot appears in Appendix II and III.

Amongst the 15 EU initiatives, 8 are EU-funded research and innovation projects, developing technological solutions to online misinformation; 4 are government and policy initiatives; and the remaining ones are academic research and non-profit initiatives.

Amongst national initiatives from EU member states, the largest number are in the UK, France, Italy, Germany and Belgium. Many of these came in preparation for and following elections and referenda in these countries between 2016 and 2018. In terms of approaches, 20% of the initiatives are legal ones; 53% – social; and 27% – technological.

These figures indicate that further initiatives are needed, in order to develop new multilingual technology for identification and containment of online disinformation, especially since current technological approaches are aimed almost exclusively at English-language content.

7. Case studies

This section introduces three key case studies on:

1 the utility of automated technology in detecting
Case Study 1: The InVID verification plugin

2 **social media content analysis**
Case Study 2: Mis- and disinformation about the Brexit referendum

3 **identifying and containing online disinformation while it still spreads**
Case Study 3: Mis- and Disinformation during the French elections #MacronLeaks

7.1. Case Study 1: The InVID verification plugin

7.1.1. What is the InVID plugin?

The InVID plugin (InVID Project, Online 1) is a browser extension aiming to "debunk fake news and to verify videos and images". It can be added to Google Chrome and Mozilla Firefox on different operating systems and it offers several tools allowing end users to verify content on their own. The tools provided allow users to:

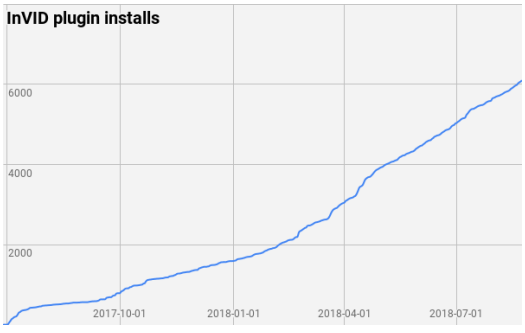
- get contextual information from videos, read video and image metadata and copyright information
- perform reverse image search on different search engines
- fragment videos into keyframes
- enhance and explore keyframes and images through a magnifying lens
- query Twitter more efficiently through time intervals and many other filters
- apply forensic filters on still images
- check video rights (on YouTube, Twitter or Facebook)

7.1.2. Who is using the InVID plugin?

The InVID plugin has been developed in the context of the EU funded project InVID and has been designed as a verification "Swiss army knife". It aims to help journalists save time and be more efficient in their fact-checking and debunking tasks on social networks, especially when verifying videos and images.

Journalists are one of the main user groups of InVID plugin, including the BBC social media, European Broadcasting Union, German and French TV, New York Times and others, who use InVID daily to debunk fake news and verify user generated videos or images. Other user groups include human rights activists, such as the Office of the United Nations High Commissioner for Human Rights, Amnesty International, media literacy scholars and emergency response professionals related to the European 112 emergency number.

Figure 20: InVID browser extension user growth



Source : InVID consortium

InVID's user base is growing very rapidly and the plugin has attracted a lot of attention. On 2 October 2018, the plugin surpassed 6000 users on Google Chrome, while the InVID project website reports on 13 July 2018 (InVID Project, 2018) that the InVID

extension surpassed 4,000 installations, resulting in almost 2,000 new users in a period of two and a half months. Figure 20 shows the user growth of the InVID browser extension users over time.

7.1.3. How is the InVID being used?

The tools offered by the plugin can be used in several ways to verify content. InVID will extract location information and the most pertinent comments from a video file, while it will also extract the keyframes and allow the user to do a reverse image search or check for tweets including the video. Observing the timeline of tweets and further filtering the reverse image search results by using for example the Google search tools to search results over a given time period, helps ensure that the content from the video has not been available before the event that the video claims to cover.

InVID also enhances the advanced search of Twitter and allows users to search for keywords or hashtags over a given time period, specified up to the minute. Similarly, to the previously described functionality, the Twitter search offered by InVID helps verify content by allowing easy access to the timestamps of previous uses of the content under verification.

The InVID plugin offers four more tools helping content verification. The image magnifier lens allows to zoom or apply a magnifying lens on the image, a feature helpful to discover implicit knowledge such as written words, signs, banners, etc. The metadata tool displays the metadata found on the image or video file, including geocoordinates, if available, a feature helpful if the original metadata shows traces of software edition or to confirm the location of an event thanks to the geocoordinates and the automated positioning on a Google map with the possibility to visually check the surroundings with Google Street View. The video rights tool retrieves metadata about video usage rights from YouTube, Facebook or Twitter, describing how the video can be used, a feature that helps complying with the platforms licences. Lastly, the InVID plugin provides access to an image forensic service developed in the Reveal EU funded project. This service offers an image verification tool by using a series of image tampering detection algorithms, aiming to identify images that have been forged or "photoshopped".

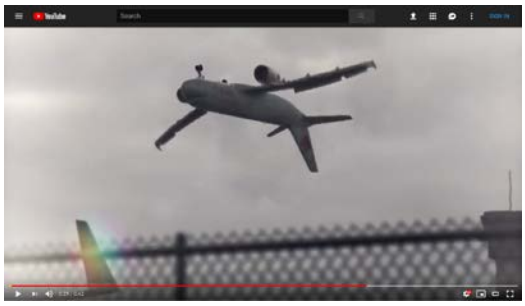
Overall, the InVID extension offers tools that provide easy and quick access to available technologies and tools, such as reverse image search, advanced search in social media, video analysis and makes metadata of the image and video files directly visible. These features are easily accessible to the user at the "click of a button", eliminating the need for series of manipulations to use these tools and speed up the verification process. The contextual menu allows to launch InVID tools or image reverse search in one click while the launcher helps users to find direct links to image and video content from other platforms like Instagram, Vimeo, Liveleak to later use the keyframes fragmentation service on videos or the magnifier lens and forensic reveal service for still images.

7.1.4. Using the InVID plugin to verify video footage

Mangkhut was a powerful typhoon that struck the Philippines and Hong Kong in September 2018. A video posted and shared online appeared to show an airplane doing a 360-degree turn, at very low altitude, before landing. The video then showed the passengers being evacuated from the emergency slides. It was shared numerous times claiming the airplane landing during the typhoon was in Shenzhen, near Hong Kong in China. AFP (Agence France Presse) reports that in one occasion the video was watched more than 11 million times.

Doing a reverse image search of the keyframes using the InVID plugin revealed that the footage was already available before the typhoon. The review of the comments in the video post, identified a link to an identical video from 2017. AFP contacted the video producer, who confirmed that the first part of the video, with the airplane doing the 360-degree turn, was done by CGI (Computer Generated Imagery) by a film production company. The second part of the video with the evacuation process was from another past incident of a damaged plane landing in Shenzhen. This example shows how the plugin helped to quickly identify previous online content with the same footage and lead the verification team to confirm that the video was a stitch of two previous videos.

Figure 21: Keyframe from fake video showin an airplane doing a 360-degree turn



Source : (MeniThings 2017)

The example is described as explained on a post (AFP 2018) of the "Fact Check" blog of AFP. The blog contains the detailed story of the video verification process, in addition to many other disinformation cases, which the InVID plugin helped verify. Figure 21 shows a key frame from the YouTube video by the original video producer (MeniThings 2017).

7.1.5. Technical dependencies and limitations

The functionality of image similarity search using the search engines of Google, Yandex, Baidu, Tineye, Bing, and Karmadecay (for Reddit) have proved to offer good accuracy. However, the success of identifying images and videos previously used directly depend on the exhaustivity of the indexing done by the search engines. if more fake images are indexed than the original, it may become difficult to retrieve the original image or video, especially over time.
Forensic analysis depends on the rate of compression of video files and poor quality of content spreading on social networks. More compression reduces the video quality and hence makes forensic analysis more difficult. Media forensic still needs more research and development, involving software makers.

Social network analysis relies on getting full or large access to the main social media platforms. In the survey conducted during this study, it was reported that some limitations are imposed by the platforms in the verification process due to privacy concerns and lack of political willingness to help third parties like media or research projects to verify the content spreading on their servers.
In addition to the InVID verification plugin, the InVID project has also developed a discovery dashboard of newsworthy user generated videos, an integrated video verification application (involving some of the main services of the plugin and more) and a mobile application for collecting user generated videos from the audience into a verification workflow.

7.2. Case Study 2: Disinformation during the 2016 UK EU membership referendum

7.2.1. Introduction

This study was carried out by researchers from the University of Sheffield, working closely with journalists from BuzzFeed UK. It was carried out retrospectively, with the data analysis taking place in November and December 2017. The initial aim was to quantify the role of Russia-linked Twitter accounts in the run up to the 2016 UK EU membership referendum. This led to the discovery of 45 linked suspected bot accounts that had been missed by Twitter, thus demonstrating the added

value of independent research. The study then investigated the spread of two key false claims made by politicians during the referendum campaign and demonstrated their much more significant impact.

7.2.2. Description of the dataset

The dataset includes around 17.5 million tweets between 30 March and 23 June 2016 inclusive (British EU referendum day). The highest volume was 2 million tweets on 23 June 2016 (excluding 3,300 lost due to rate limiting), with just over 1.5 million in poll opening times. Of the 2 million, 57% were retweets and 5% - replies. June 22nd was second highest, with 1.3 million tweets. The 17.5 million tweets were authored by just over 2 million distinct Twitter users (2,016,896).

Table 1: Description of dataset used to quantify the role of Russia-linked Twitter accounts in the run up to the 2016 UK EU membership referendum

Metric	Value
Tweets	~ 17.5 M
Distinct Twitter users	> 2 M (2,016,896)
Date period	Between 30 March and 23 June 2016 (inclusive)
Keyword/hashtag sought	votein, yestoeu, leaveeu, beleave, EU referendum, voteremain, bremain, no2eu, betteroffout, strongerin, euref, betteroffin, eureferendum, yes2eu, voteleave, voteout, notoeu, eureform, ukineu, britainout, brexit, leadnotleave

Some of the analysis focused on a subset of these, covering the month up to and including June 23rd. Within that period, there were just over 13.2 million tweets, from which:

- 4.5 million were original tweets (4,594,948)
- 7.7 million retweets (7,767,726)
- and 850 thousand replies (858,492)

These were sent by just over 1.8 million distinct users.
Data collection continued after the referendum, amassing over 93 million tweets between 24 June 2016 and 31 March 2017, when article 50 was triggered. A further 102 million tweets has been collected between 31 March 2017 and 18 December 2017. So far, the bulk of the analysis has focused on the pre-referendum tweets.

7.2.3. Russian involvement in social media during the referendum

The study analysed the activity of the accounts that were identified by Twitter as being associated with Russia in front of the US Congress in the fall of 2017 (Recode 2017). In addition, journalists from BuzzFeed UK and the team from the University of Sheffield used the retweet network to identify another 45 suspicious accounts, subsequently suspended by Twitter (Phillips & Balls, 2017). The study looked at tweets posted by these accounts one month before the referendum, and did not uncover evidence of significant influence of the Russian trolls, as compared to the overall number of tweets on the referendum. Both the Russia-linked ads and Twitter accounts did not have major influence.

These accounts produced 3,200 tweets in the data set analysed and just over 800 of those —about 26%— came from the new 45 accounts that were identified. However, the 45 new accounts were tweeting in German, so even though they produced 800 tweets, their impact on the British voter is not very likely to have been significant.
The accounts that tweeted on 23 June 2016 (the day of the British referendum on EU membership) were different from those that tweeted before or after, with all tweets posted in German. Their behaviour is also different - with mostly retweets on referendum day by a tight network of anti-

Merkel accounts, often within seconds of each other. The findings are in line with those of Professor Cram from the University of Edinburgh, as reported in the Guardian (Booth et al., 2017).
Similar to those identified by Twitter, the newly discovered accounts were largely ineffective in skewing public debate. They attracted few likes and retweets – the most successful message in the sample got 15 retweets.

An important distinction that needs to be made is between Russia-influenced accounts that used advertising on one hand, and the Russia-related bots found by Twitter and other researchers. The Twitter sockpuppet/bot accounts generally pretended to be authentic people (mostly American, some German) and would not resort to advertising, but instead tried to go viral or gain prominence through interactions on social media. An example of one such successful account/cyborg is Jenn_Abrams, illustrating how Russian talking points can seep into American mainstream media without even a single dollar spent on advertising (Collins & Cox 2017, O'Sullivan 2017, The Guardian Pass notes 2017).

7.2.4. Russia-sponsored media activity in social media

The study went on to analyse the influence of Russia-sponsored media and its Twitter posts. It first looked at the 6 Russia Today promoted tweets, as published by RT (Russia Today 2017). The 3 tweets before the referendum attracted 53 likes and 52 retweets, showing only a small scale activity.
The study continued by analysing all tweets posted one month before 23 June 2016, which were either authored by Russia Today or Sputnik, or were retweets of these. This gives an indication of how much activity and engagement there was around these accounts. Table 2Table 1 puts the activity in context, comparing it to the equivalent statistics of the two main pro-leave and pro-remain Twitter accounts.

Table 2: Russia-sponsored media activity in Twitter in the run up to the 2016 UK EU membership referendum

Account	Original tweets	Retweets by others	Retweets by this account	Replies by account	Total tweets
@RT_com - General Russia Today	39	2,080	62	0	**2,181**
@RTUKnews	78	2,547	28	1	**2,654**
@SputnikInt	148	1,810	3	2	**1,963**
@SputnikNewsUK	87	206	8	4	**305**
TOTAL	**352**	**6,643**	**101**	**7**	**7,103**
@Vote_leave	**2,313**	**231,243**	**1,399**	**11**	**234,966**
@StrongerIn	**2,462**	**132,201**	**910**	**7**	**135,580**

The analysis of the accounts which retweeted RT_com and RTUKnews identified the account with the most retweets (75 retweets of Russia Today tweets) was a self-declared US-based account that retweets Alex Jones from infowars, RT_com, China Xynhua News, Al Jazeera, and an Iranian news account. This account (still live as of 27 Nov 2018) joined in Feb 2009 and as of 15 December 2017 had 1.09 million tweets, i.e. an average of more than 300 tweets per day, indicating it is a highly automated account. It has more than 4K followers, but follows only 33 accounts. Two of the next most active retweeters are a deleted and a suspended account, as well as two accounts that both stopped tweeting on 18 Sep 2016.

For the two Sputnik accounts, the top retweeter made 65 retweets. It declares itself as Ireland based, has 63.7K tweets and 19.6K likes; many self-authored tweets; last active on 2 May 2017; account created on May 2015; with 87 tweets a day on average, which possibly indicates an automated account. It also retweeted Russia Today 15 times. The next two Sputnik retweeters with 61 and 59 retweets respectively, are accounts with a high average post-per-day rate (350 and 1,000 respectively) and over 11k and 2k followers respectively. Lastly, four of the top 10 accounts have been suspended or deleted.

In conclusion, as these numbers demonstrate and similar to the IRA-linked troll accounts, Russia Today and Sputnik also had insignificant impact in social media conversations related to the referendum. As Table 2 demonstrates, just the two flagship referendum campaign accounts combined (Vote_Leave and StrongerIn) generated over 363,000 retweets or just under 55 times the retweets of RT and Sputnik.

7.2.5. Impact of Russia-linked misinformation vs impact of false claims made by politicians during the referendum campaign

A study (Moore & Ramsay 2017) of the news coverage of the British EU Referendum campaign established that the economy was the most covered issue, and in particular, the remain claim that Brexit would cost households £4,300 per year by 2030 and the leave campaign's claim that the EU cost the UK £350 million each week. Therefore, the research focused on these two key claims and the relevant tweets about them.

With respect to the disputed £4,300 claim, 2,404 posts were identified in the dataset, including tweets, retweets, and replies referring to this claim. For the £350 million a week disputed claim - there are 32,755 pre-referendum posts in the dataset, including tweets, retweets, and replies. This is 4.6 times the 7,103 posts related to Russia Today and Sputnik and 10.2 times more than the 3,200 tweets by the Russia-linked accounts suspended by Twitter.
In particular, there are more than 1,500 tweets from different voters, with one of these wordings:

> *"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"*

Many of those tweets have themselves received over a hundred likes and retweets each. This false claim is being regarded by media as one of the key ones behind the success of VoteLeave (Kirk 2017). Despite these findings, it is not possible to measure exactly the impact of this misinformation on voting behaviour based on tweets alone.

In conclusion, while it is important to quantify the potential impact of Russian misinformation, the much wider range of misinformation that is posted on Twitter, Facebook and other platforms and its overall impact needs to be considered as well. Another related question is studying the role played by hyperpartisan and mainstream media sites during the referendum campaign.

7.3. Case Study 3: Mis- and Disinformation during the French elections #MacronLeaks

7.3.1. How #MacronLeaks started?

A few hours before the closing of the official campaigns for the presidential elections in France on 7 May 2017, the hashtag #MacronLeaks started to spread on social networks. This hashtag was used to disseminate a leak of emails from the Macron campaign. Due to the timing of the event, the leak looked very suspicious as some excerpts from the emails referred to alleged discussions regarding

Mr Macron's team openly buying drugs[108]. The hashtag first appeared on Friday 5 May 2017 at 8.43 pm CEST and the period of silence started at midnight of the same day, allowing very little time to respond. The following timeline presents how the hashtag emerged and how Nicolas Vanderbiest revealed its creation and the communities involved.

- **8.35 PM CEST, Friday, 5 May 2017**
  A message was posted on the board "politically incorrect" of the anonymous forum 4chan with links to a file sharing website including numerous documents leaked form the team of the presidential candidate Emmanuel Macron.
- **8:49 PM CEST, Friday, 5 May 2017**
  First tweet using #MacronLeaks from Jack Posobiec
  https://twitter.com/JackPosobiec/status/860567072010620929
- **9:31 PM CEST, Friday, 5 May 2017**
  Wikileaks announce they're examining the leak
  https://twitter.com/wikileaks/status/860577607670276096
- **11:01 PM CEST, Friday, 5 May 2017**
  First tweet from Nicolas Vanderbiest acknowledging there is something going on
  https://twitter.com/Nico_VanderB/status/860600388415873025
- **11:11 PM CEST, Friday, 5 May 2017**
  Nicolas Vanderbiest exposes pro-Trump/US alt-right accounts that are spreading the hashtag
  https://twitter.com/Nico_VanderB/status/860602824375365632
- **11:34 PM CEST, Friday, 5 May 2017**
  First map showing the spread, and the role of already identified French far-right activists into the spread
  https://twitter.com/Nico_VanderB/status/860608583217668101
- **11:41 PM CEST, Friday, 5 May 2017**
  Nicolas Vanderbiest exposes Florian Philippot (head of Marine Le Pen's campaign) and active officials from Front National of leaking fake news during the campaign. This tweet will be retweeted by the Macron Press team at 11:57 pm, 3 minutes before the end of the campaign.
  https://twitter.com/Nico_VanderB/status/860610487742074880
- **12:28 AM CEST, Saturday, 6 May 2017**
  Nicolas Vanderbiest exposes role and accountability of WikiLeaks and officials from the Front National in the spread of the #MacronLeaks
  https://twitter.com/Nico_VanderB/status/860622206841311232
- **12:40 AM CEST, Saturday, 6 May 2017**
  Nicolas Vanderbiest overlaps databases of people already tweeting about the MacronBahamas fake news, launched by Jack Posobiec two days before #MacronLeaks, and shows the same communities are spreading the rumours.
  https://twitter.com/Nico_VanderB/status/860625172344315904

These studies showed there were social media nests ready to spread fake news and rumours, these social media nests being populated by activists aggregated around certain candidates and foreign media outlets.

7.3.2. Sourcing #MacronLeaks

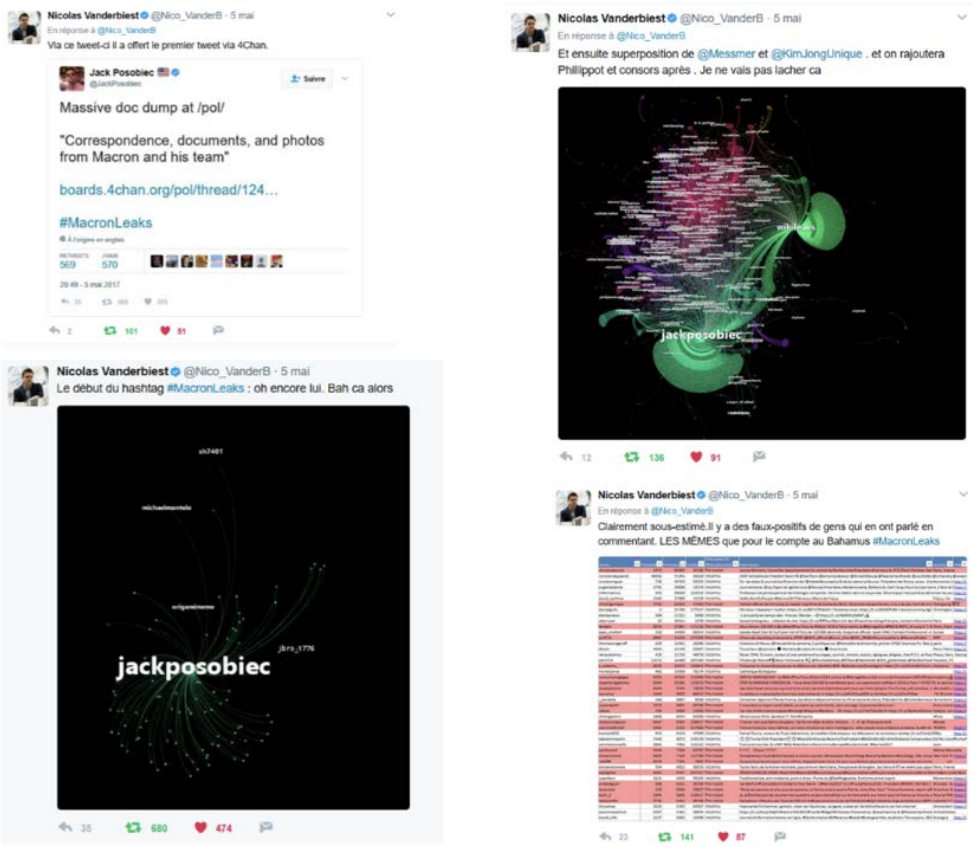
After the emergence of #MacronLeaks, many analysts focused on checking if the revealed information is related to fraud or misconduct by Emmanuel Macron and his team. Nicolas Vanderbiest checked who were the people spreading the hashtag and the information. The

108 https://www.liberation.fr/politiques/2017/05/09/macronleaks-la-grande-foire-aux-fake-news_1567868&sa=D&ust=1541153862613000&usg=AFQjCNGwcEymzrW7r33jepBZzDQ5Q9041Q

methodology used by Nicolas Vanderbiest is “Sourcing”, i.e. the ability to quickly go back to source of the information and look at who started to spread the information, and who relayed it.
While sourcing #MacronLeaks, Nicolas Vanderbiest used the pro-Kremlin media outlets analysis (Nicolas Vanderbiest Online 1) that he published before the elections. The previous analysis helped in quickly making a match between some of those who were sharing the hashtag and those already identified to have previously spread misinformation in France.

Nicolas Vanderbiest believes that sourcing is not sufficient without rapidly communicating the findings, in order to restrict the disinformation spread while following it in real time while it occurs. This was done very rapidly by Nicolas Vanderbiest. Figure 22 shows a summary of the evidences he could publish on his Twitter account between 11 pm and 11:57 PM on May 5th.

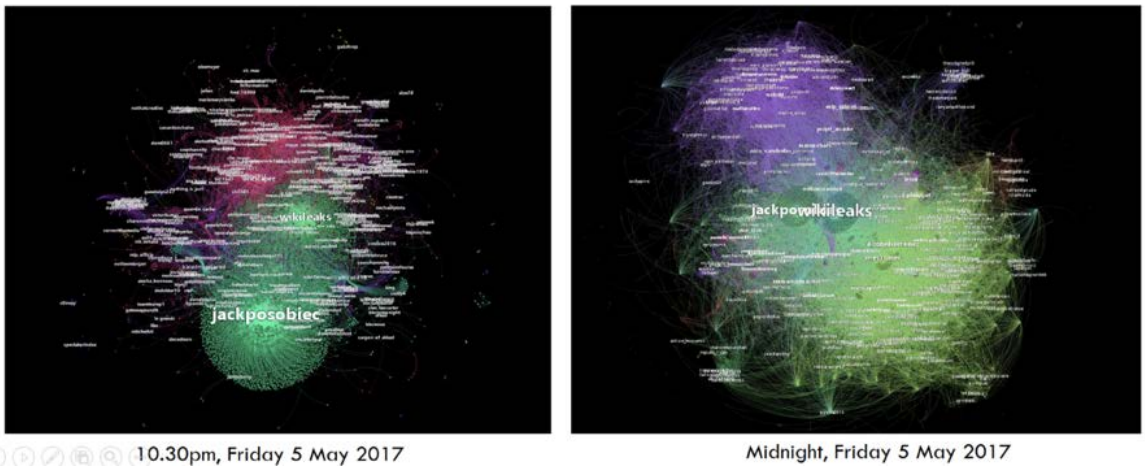
Figure 22: Summary of evidences published on Twitter about #MacronLeaks



Source: Nicolas Vanderbiest

The end result of sourcing is a map (Figure 23) depicting the Twitter involvement of different communities on the topic being analysed.

Figure 23: Map depicting Twitter activityrelated to #MacronLeaks



Source: Nicolas Vanderbiest

At Midnight, the community in purple is the community fighting the #MacronLeaks and created mostly by the quick sourcing done by Nicolas Vanderbiest. Sourcing enables people who want to fight the misinformation the ability to start verifying the information very quickly and before it has been widely spread.

7.3.3. How is sourcing different from fact checking?

In contrast to fact checking, sourcing does not focus on verifying content. It focuses on the content's online path. The aim of sourcing is to act as an early warning mechanism notifying the emergence of potential disinformation and giving time to fact checking to react quickly. In the experience of Nicolas Vanderbiest, this approach has been proven very helpful during the French elections and other events.

By focusing on communities instead of the content, sourcing allows to detect weak signals in rising stories. Sourcing is a live operational task during a crisis and produces results faster than fact checking. It alerts journalists and thus making an immediate impact on protecting "mainstream" communities from disinformation campaigns.

Sourcing is also different to quantitative content analysis and is not affected by the volume of information because it concentrates on the information spread and interactions between individuals constituting a community.

7.3.4. How sourcing identifies content of potential disinformation?

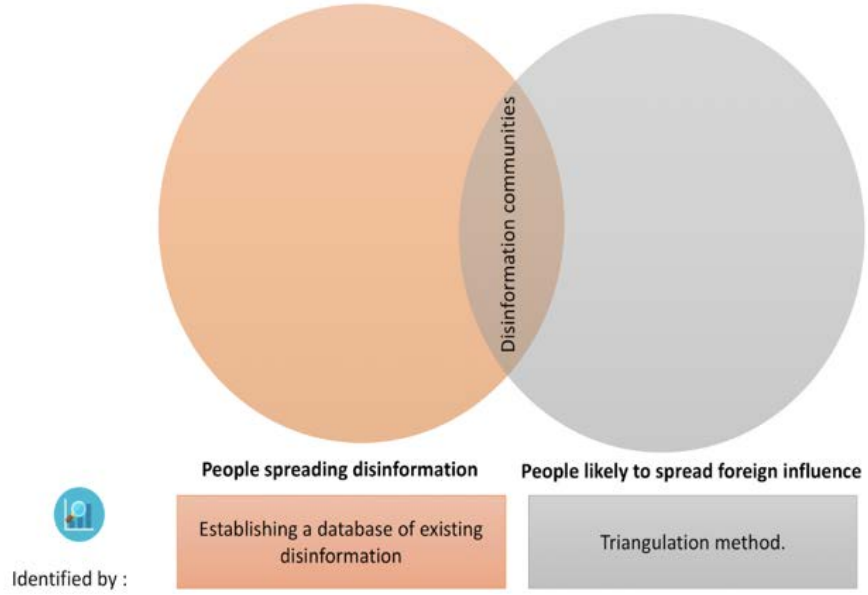
There are two ways to quickly identify content of potential disinformation, by narrative and by spotting accurate communities.

Identifying disinformation by narrative is based on the observation that most of the time, the narratives used are recurrent. Nicolas Vanderbiest identified that during the French presidential campaign, most of the false information concerned Emmanuel Macron employed the same narratives. These narratives are usually monitored, leading the monitoring process to the challenge of being able to filter out the "noise", i.e. content not being part of a disinformative narrative, e.g. users countering it while using the same keywords. On very popular topics like the French elections involving millions of online posts, it is highly difficult and special expertise is required to identify and associate keywords with disinformation in order to filter out the noise. For instance, if you monitor the content with the keyword "Macron", it will be difficult to spot very small things or weak signals

and filter the pointless discussions. Focusing on the main disinformation spreaders, the launchers and suspicious accounts disinformation detection will be quicker.
To identify disinformation by spotting accurate communities, it is necessary to identify the communities likely to pass on disinformation. To do that, Nicolas Vanderbiest follows two methods:

- **Triangulation** It is the starting point when not enough preliminary information is available about accounts spreading disinformation. First, a list of existing media or other accounts spreading disinformation, for instance listed by other partners, are identified and their audience is extracted in order to identify recurring accounts, forming a network of potential disinformation agents.
- **A disinformation database** A disinformation database is established by identifying each piece of disinformation and analysing the network of accounts that have been spreading this. The database of disinformation is then searched for potential patterns of disinformation spread, e.g. accounts present in several cases of disinformation spread. This process leads the analysis to spotting recurring users who are most likely doing it intentionally.

Figure 24: The role of the triangulation and the disinformation database in sourcing



Source: Nicolas Vanderbiest

The methods have been used by Nicolas Vanderbiest with good results during the French and Italian elections. By understanding communities regularly spreading misinformation and establishing patterns of misinformation spread, an early warning system is achieved able to score misinformation content. Networks that have been exposed in spreading "fake" content in the past could be monitored to spot new disinformation content and send early alerts to fact-checkers and journalists. This early warning system saves some time for journalists before the disinformation spreads and reaches the public, letting them gather enough information and facts to be ready for any spread outside of fringe communities.

7.3.5. Why and how is sourcing useful?

Fake news debunkers and fact-checkers focus on disproving information either by opposing factually correct information, or by conducting forensic analysis to prove manipulation, and then attempting to dilute the fake narrative with content debunking it. While crucial and highly effective, this process takes time and effort, requiring manual research or the use of forensic analysis tools. On

the example of #MacronLeaks, which was released only hours before the 48h electoral silence period, any traditional research would bring results long after the silence period was put in place, making it ineffective.
Sourcing on the other hand proved, within just an hour and a half, that the information was originating from highly untrustworthy sources and serial spreaders of fake news - alt-right MAGA & 4Chan activists, and then amplified by bots & Russian echo chambers. Nicolas Vanderbiest managed to publish results before midnight, giving civil society organisations and journalists enough time to share it and prevent a potential disaster.
The goal of Nicolas Vanderbiest was not to counter the attack within the community spreading the narrative. The impact of debunking with contrasting information has proven to be low (Zollo et al. 2017). The goal is to prevent journalists and other main influencers to become victims of disinformation that could seed doubt just before the elections.
Inherent to a massive operation like this is the impossibility to check and verify all content involved, especially in a matter of hours before an election. Therefore, the sourcing capability is a crucial complement to efforts of fact-checkers and journalists, buying time for them while information is verified.

The sourcing method of Nicolas Vanderbiest is not the whole solution. It offers a way to understand the communities spreading disinformation, the processes behind it, and provides a tool for live verification of information spread through social media. For the process to have impact it requires reaching an active audience as well as cooperating with traditional fact-checkers and journalists to definitively debunk a piece of disinformation. This is the beginning of building a complex, self-sustainable, systemic approach, which could shield societies from malicious external influence, considering the need for it to be constantly adapted to the ever-developing methodology of disinformation campaigns.

8. Policy options

This concluding section provides policy options on actions that could be taken and the stakeholders best placed to act upon these at national and European level. The likely effects of these actions will also be discussed, in as much detail as currently feasible. We have observed through this study and the case studies that solutions to counter disinformation require very different expertise and relied on the intersection of technological innovations and the involvement of civil society. Therefore the policy options are similarly organised: ones that are focusing on the solutions and tools that could be implemented, the others are looking at how to address the plethora of stakeholders involved.

8.1. Option 1: Enable research and innovation on technological responses

Many of the tech responses described in this report are relying on the new capabilities that machine learning and artificial intelligence are poised to bring. To enable such new research and innovation, a common data framework around disinformation should be put in place.

8.1.1. Preserving important social media content for future studies

The intersection of automated or semi-automated content, political propaganda, misinformation and disinformation is a key area in need of further investigation, but for which, scientists often lack the much needed data, while the data keepers lack the necessary transparency, motivation to investigate these issues, and willingness to create open and unbiased algorithms. The development of non-transparent monitoring systems within the large corporations that own and run social media platforms raises questions about their willingness and capacity to tackle the issue, when acting on their own. For instance, the recent New York Times investigation into Facebook (Frenkel *et al,* 2018), uncovered that prior to the Nov 2016 US presidential election, Facebook's senior management were already aware of Russian hacker activity targeting people involved in the campaigns. Another example is when journalists from BuzzFeed UK worked with academics to uncover 45 suspected Russia-linked Twitter accounts, that had been missed by the platform's own investigation (Phillips & Ball, 2017).

Data is essential for training machine learning and AI algorithms. In a data-driven environment, the main question is not only if the data is collected but whether if it is accessible, how and when. As disinformation techniques are evolving very quickly, researchers need to have uninterrupted access to large amounts of data, in order to evaluate, train and re-design methodologies.

When it comes to understanding online disinformation and its impact on society, there are still many outstanding questions that need to be researched and effective socio-technical solutions developed. Related to this, is the challenge of open and repeatable science on social media data, as again many of the posts in current datasets available for training and evaluation have been deleted or are not available. This causes a problem as algorithms do not have sufficient data to improve and neither can scientists determine easily whether a new method is really outperforming the state-of-the-art. Similar issues arise when trying to study online abuse of and from political stakeholders, as posts and accounts are suspended at a very high rate, or as dark adverts are used as a strategic campaign tool. As a first step to overcome at least partially this problem, some media organisations have started publishing social media datasets of high historical value, such as the deleted tweets of the Russia-linked troll accounts[109].

[109] https://www.nbcnews.com/tech/social-media/now-available-more-200-000-deleted-russian-troll-tweets-n844731

Governments and policy makers are thus in a position to help establish this much needed cooperation between social platforms and scientists, **promote the definition of policies for ethical, privacy-preserving research and data analytics** over social media data, and also to ensure the archiving and preservation of social media content of key historical value.

This could for instance, concern all publications and advertisements by political parties and official candidates **in a public depository that would be accessible for researchers and trusted third-parties**. Smart use of blockchain technology would also allow to make **content falsification** almost impossible by proving the origination of content. This technology would be particularly useful for videos or audio content, that can be falsified through deepfake technology. For example, a crowdsourced rating system could enable verified content to prevail unverified news.

8.1.2. Fund open-source and multidisciplinary research on automated methods for disinformation detection

As discussed already, there are emerging technologies for **veracity checking and verification of social media content** (going beyond images/video forensics). These include tools developed in several European projects (e.g. PHEME, WeVerify, InVID), tools assisting crowdsourced verification (e.g. CheckDesk, Veri.ly), citizen journalism (e.g. Citizen Desk), and repositories of checked facts/rumours (e.g. Emergent, FactCheck). However, many of those tools require further improvements. Whether disinformation is genuine, algorithmic, or paid, the corresponding detection algorithms should be developed further, to achieve accuracy comparable to that of email spam filter technology. New research should also focus on designing and piloting social and cognitive approaches to countering belief in and spread of misinformation. This research should be made open source, to enable open science and verifiable algorithms, as well as to enable companies and platforms to experiment easily with the new technology.

As already discussed in Section 2.3, artificial intelligence technologies make it possible to generate "deepfakes" which are even more difficult to identify and fact-check, especially as images and video are typically perceived as more credible. Journalists, researchers, platforms and technologist need to build a common understanding of those threats. A shared approach is required to address the human rights, journalistic, and technological challenges raised by this phenomenon. Collaborative research would enable the creation of tools such as reverse video search and prepare facing upcoming risks and scenarios.

8.1.3. Measure the effectiveness of technological solutions implemented by social media platforms and news media organisations

As discussed in Section 3, social platforms have started already to implement technologies and human curation workflows aimed at detecting bots and fake accounts, identifying hate speech, and flagging content that has been disputed by independent fact-checking organisations. As yet, however, there is no mechanism in place to verify independently the effectiveness of these counter-misinformation approaches, while at the same time evidence is emerging that they are not as effective as envisaged (e.g., reports that independent fact-checkers are receiving at most one article per day from Facebook (Salinas, 2018A), allegation of Russia-linked right wing propaganda still trying to influence the 2018 US midterm elections (Price, 2018)).

Consequently, we argue that objective and large-scale technology effectiveness evaluations are best carried out as a cooperation between academic researchers, journalists, and the platforms themselves. In order to prevent possible bias arising from exclusive sponsorships or data sharing agreements, the establishment of independent, multidisciplinary research centres on disinformation, as also advocated by the EU High Level Expert Group is proposed (HLEG report, 2018).

8.1.4. Outcomes for this option: Ethical implications of tech solutions

As private initiatives as well as public-funded projects already exist, these should deliver in the short to medium term new automated methods for disinformation detection. The European Union has several budgetary levers to help foster the needed innovation on science and technology. In particular, Horizon Europe (e.g. the new Digital Europe Program) could be used to fund collaborative research projects on disinformation detection. The creation of a shared data repository and infrastructure for large-scale multi-disciplinary research would require more significant investment. Such a flagship project, however, would have significant impact on regulation and research in the digital age.

At the same time, this new technology and innovation would only be possible within a strong legal framework, in particular regarding data protection. Moreover, the prospect of fully automated disinformation monitoring raises legitimate fears over censorship and infringement of freedom of speech. Therefore, algorithm transparency is a necessary condition to ensure that algorithmic choices are verifiable and bias-free. Algorithms do make mistakes, and as they are processing a large amount of data, these can have a cumulative effect. For example, wrongly suspending a social media account is a form of censorship and quick redress and independent human review mechanisms need to be put in place. At the same time, full algorithmic transparency carries out the danger of weakening their effectiveness, as it would allow actors who publish and disseminate disinformation to exploit weaknesses and devise evasion strategies.

8.2. Option 2: Improve the legal framework for transparency and accountability of platforms and political actors for content shared online

8.2.1. Build a transnational legal framework and support strong privacy protection

In order to build a coherent legal framework and avoid what could be considered as a "balkanisation" of the digital space[110], with different set of rules applicable at national and regional level, a coherent global framework of regulation should be put in place. in order to design the best possible approaches to overseeing or regulating global tech companies, legislators from multiple countries have already started to cooperate. One example is the "Grand Committee on Disinformation" which comprised parliamentarians from nine countries tasked with gathering evidence from Facebook in November 2018. Regulation on the minimum set of obligations on dataset access for researchers on disinformation and the public data repository recommended in option 1 would be more effective if applicable globally. European actions and reflections in international fora such as the OECD and G7 certainly should be prioritised.

In a data-driven environment, the collection of data raises questions of privacy protection. While some datasets need to be made available for analysis and algorithm training, this should not be at the cost of privacy. Scandals on the misuse of personal data for micro-targeted political advertising have shown the need for a robust policy framework protecting citizens from the misuse of their personal data. The general data protection regulation sets out a clear framework in this respect. In order to build a comprehensive understanding of the issues at stake, the European Data Protection Supervisor (EDPS, 2018) has recommended a structured dialogue between regulators (data protection, consumer protection, competition and electoral regulation) both at national and European levels.

[110] https://www.politico.eu/article/internet-governance-facebook-google-splinternet-europe-net-neutrality-data-protection-privacy-united-states-u-s/

Nevertheless, the capacity to use and share datasets for research on disinformation now raises new challenges, especially since this data is essential for developing and assessing the effectiveness of new technologies. Therefore, further discussions are required on the best ways to protect citizens' privacy, while making it possible for researchers to investigate the manipulation and diffusion of content online, within a well-defined framework of ethical and legal protocols and robust research methodologies.

Another important requirement towards such privacy-preserving legal framework is oriented towards transparency of algorithms on social platforms and provision of privacy-compliant access to data for journalists and researchers. With respect to preventing data misuse in online advertising, in particular, the EU High Level Expert Group on disinformation (HLEG report, 2018) has recommended that platforms "should ensure transparency and public accountability with regard to the processing of users' data for advertisement placements, with due respect to privacy, freedom of expression and media pluralism." There is also a strong argument for working with the online advertising industry to establish an ethics code and best practice guidelines.

There are also complementary technological developments, aimed at addressing the challenge of protecting user data. For instance, the Inrupt startup (led by the inventor of the World Wide Web - Sir Tim Berners-Lee) is developing a decentralised web platform, where users have a personal online datastore and control how applications access and use that (Brooker, 2018). If successful, this new individual data ownership model aligns well with the argument that "citizens do not lose control over the precious resource (data) and infrastructure (artificial intelligence), around which most of the future political and economic institutions will be built." (Morozov, 2018)

8.2.2. User-centric moderation and fiduciary responsibilities of social platforms

When it comes to content regulation, there is a strong argument for holding social media companies accountable for moderating the information shared on their online platforms. Moderation, however, should not compromise freedom of speech, privacy, transparency and accountability (RSF, 2018). As a first step towards implementing user-centric moderation practices, researchers have identified a set of principles known as the Santa Clara principles[111]. In particular:

- Companies should publish the numbers of posts removed and accounts permanently or temporarily suspended due to violations of their content guidelines.
- Companies should provide notice to each user whose content is taken down or account is suspended about the reason for the removal or suspension.
- Companies should provide a meaningful opportunity for timely appeal of any content removal or account suspension.

These recommendations are the necessary safeguards to the risks previously identified of tech-only solutions. Putting the user back at the centre of the moderation process, companies should provide citizens and regulators with transparent lists of the content removed and suspended accounts, as well as notice and justifications for moderation. Finally, the user should be able to appeal the moderation decision. In the spirit of these recommendations, Facebook announced the creation of an independent oversight group to adjudicate appeals on content moderation issues (Zuckerberg, 2018a).

To fully establish their liability, some advocate for social platforms to be considered publishers and be held accountable for the content shared on their online apps. As they operate more as a channel than as a media, such qualification can seem inadequate, with a risk of giving tech companies the possibility to limit freedom of speech. In this debate, there is a strong argument to create a new legal

[111] https://santaclaraprinciples.org/

category that would establish a clear liability for the tech companies to act against harmful and illegal content and protect users personal data.Going even further,digital media companies could become "information fiduciaries" - a new legal category, granting such companies specific role and obligations as facilitators of free-speech and collectors of personal data. (Balkin, 2019) The idea (similar to the obligations of doctors and lawyers) is to make user care and confidentiality the primary responsibility of social platforms, with loyalty towards end users not being compromised for business reasons. There is also an argument that fiduciary obligations exist irrespective of the contractual information in the platforms' terms of service. As the European GDPR framework sets strong privacy protection, it could be investigated how such fiduciary obligations could also be applied to information shared online.

8.2.3. Strengthening trust in public institutions and political discourse online

Accountability over online content is particularly sensitive when it comes to political discourse and political campaigns. Today, political advertising uses micro-targeting based on sensitive personal data, in order to reach key voters, a practice that can threaten the very essence of the democratic debate.

Election law should be updated in order to regulate political campaigns in the online public sphere, ensure transparency, and prevent voter manipulation. The UK independent fact-checking charity FullFact has made a list of recommendations (Staines & Moy, 2018), which is more detailed and wider in scope than the recommendations of the High Level Expert Group (HLEG report, 2018). First, it is crucial that voters can identify political advertising as campaign material and that claims made in political adverts are verified as correct prior to the advert being published. Secondly, the "imprint rule" which requires printed campaign materials to state who is promoting them, should be extended to apply to online advertising. Thirdly, it recommends a regulatory body for political advertising and the creation of an up-to-date centralised public database of political ads, also available in a machine-readable format. Lastly, social platforms and advertising networks need to provide government bodies, such as national regulatory authorities, with real-time access to political advertising information.

Related to that are short-term actions that platforms can take, in order to curtail the use of bots for automated amplification of disinformation and other kinds of information manipulation. Firstly, platforms need to continue investing in developing better algorithms for discovery and suspension of bot networks. This is likely to have a positive impact, especially with respect to preventing low-credibility content and disinformation from going viral (Shao *et al*, 2018). The second short-term change could be for platforms to require automated accounts to declare themselves as such to the platform users (Nyst & Monaco, 2018). In this way, users will be able to identify easily artificially amplified content and take this into consideration when deciding whether to trust the information source and its content.

These options go significantly beyond the partial voluntary steps already undertaken by Twitter and Facebook in this regard.

8.2.4. Outcomes: a human rights approach to tech solutions

When considering regulatory responses to disinformation, there is a major risk of endangering free speech by trying to control what should be a "truthful" discourse. The Special Rapporteur's 2018 report to the United Nations Human Rights Council advocates for smart regulation focused on ensuring transparency and mediation to enable the public to make choices on how to engage online. (UNHRC, 2018)

Another important area where international human rights law could be developed is towards recognising (state-sponsored) trolling attacks as infringing the right of the targeted person to freedom of speech and access to information online (Nyst & Monaco, 2018).

8.3. Option 3: Strengthening Media and Improving Journalism and Political Campaigning Standards

8.3.1. Support and promote high quality journalism and political campaign standards

Ahead of the UK EU membership referendum, several British tabloids published inaccurate information on migration, terrorism and border control (Lythgoe & Dixon, 2016). This was then amplified by politicians and on social media, thus manipulating voter perceptions. The problem was compounded further by inaccurate statements made by prominent UK politicians and used in emotive political advertising.

In order to safeguard public trust in media, strong press standards should be defined and followed, supplemented by transparency on political advertising and political campaign standards.
Many mainstream media are already carrying out key fact-checking and media literacy activities. However, these should be widened and adopted by all media. In addition, media should commit to respect high ethical standards, for example refraining from us of clickbait headlines on social media. In addition, as disinformation can sometimes spread locally, sometimes in a regional dialect, local reporting should also be supported with government subsidies.

8.3.2. Promote Fact Checking Efforts

National fact-checking initiatives should be promoted, as a collaboration between different media organisations, journalists and independent fact-checkers. Publicly funded media organisations should be supported and their content promoted on social platforms. In the meantime, politicians should be committed not to undermine the reputation of media and journalists.

Fact-checking can also be crowd-sourced with citizens flagging suspicious information which is then checked independently. Trust in public media organisations can also help with getting access to disinformation circulating on closed messaging platforms. For example, the BBC research project "Beyond Fake News"[112] relied on readers who gave access to their encrypted messaging apps in Kenya, Nigeria and India, thus enabling new research on disinformation spread through these platforms.

To become an efficient source of reliable information for citizens, fact-checking efforts could rely on automated methods for checking against statistical sources or a shared database of already debunked misinformation. In addition, the outcomes of new research (as described in Option 1) need to be made understandable by non-specialists and disseminated to journalists and public organisations, in order to inform and strengthen their ability to detect and debunk disinformation. There is also ample scope for collaborative projects between journalists, media and researchers focused on content verification and fact checking.

[112] https://www.bbc.co.uk/news/topics/cjxv13v27dyt/fake-news

8.3.3. Outcome: fact-checking on its own is not enough to combat disinformation

Fact-checking is a very time-consuming process, which unfortunately cannot be automated fully, due to the present limitations of AI algorithms but also the need for transparent decision making and the right to appeal. At present, it can take days to weeks to investigate whether a claim can be considered correct and comes from a trustworthy source. At the same time, hyperpartisan media, politicians, and online communities are rejecting the accuracy of reporting and the impartiality of mainstream media, since they consider them part of the establishment. This is not helped by the fact that a significant amount of unverified information and factually incorrect statements are published by traditional media and advocated by politicians. Such misinformation is then shared widely through social networks, where it can be amplified or presented out of its original context by online media, whose business model relies primarily on click-through traffic and online advertising. Altogether, this begs the questions if there is room for a sustainable business model of fact-checking.

8.4. Option 4: Interdisciplinary approaches and localised involvement from civil society

When it comes to fighting online disinformation, technology can be of significant help, but one should not forget that the phenomenon takes roots in and exploits social cracks. This study has described how multiple stakeholders with different backgrounds are working on the digital misinformation and disinformation ecosystem. Taken in isolation, these different approaches can address particular aspects of the issue, but, as argued also by the EU High Level Expert Group (HLEG report, 2018), its complex, multi-dimensional nature can only be tackled fully through a multi-stakeholder approach.

8.4.1. Support interdisciplinary approaches and invest in platforms for independent evidence-based research

Analysing the impact of social media on society and democracy involves a variety of actors from different backgrounds, including data scientists, artificial intelligence researchers, political and social scientists, as well as journalists. They need to work together alongside social media platforms and policy makers to gain full understanding of the mechanisms behind viral disinformation and the most effective ways to contain and prevent it. As argued already, journalists and scientists need access to public social media posts for research and experimentation purposes.
In order to prevent potential bias and conflict of interest, exclusive partnerships between social platforms and designated science labs need to be avoided. Instead, data and collaborations need to be made available to all stakeholders.

Thus, the creation, at the European and national levels, of independent platforms that facilitate collaborative evidence-based research and help promote best practice in detection and prevention of online disinformation is proposed. The platforms could also act as initiative incubators. They could be designed as independent consultative bodies, similar to the Digital National Council in France[113], combined with national or European research institutes that bring together scientists from multiple organisations and disciplines[114].

[113] The Digital National Council (Conseil National du Numérique) is a government-supported independent body, producing recommendations on digital economy regulations. Its board is mainly composed by digital entrepreneurs. https://cnnumerique.fr/

[114] The UK Alan Turing Institute was established in 2015 as a national institute for data science and artificial intelligence. Since late 2017 it brings together scientists from nine UK universities. https://www.turing.ac.uk/

8.4.2. Empower civil society to multiply efforts

As already discussed, social media companies have acknowledged the benefits of working alongside journalists and scientists, in order to gain better understanding of online disinformation and benefit from the latest advances in Artificial Intelligence research. For instance, Twitter have now released two large collections of tweets from already-suspended Russian and Iranian bot and troll accounts.[115] Nevertheless, such private company initiatives should not be the only ones that we depend on. Civil society can and should also play the role of a counter balance and an independent stakeholder, working alongside and in cooperation with private companies and platforms to flag and debunk misinformation. Therefore, funds should be secured to support civil society initiatives. In particular, knowledge sharing should be encouraged, thus building cross-expertise and avoiding duplicate projects by different organisations. Since disinformation campaigns can also cross linguistic, cultural, and national boundaries, there is also a strong argument for encouraging European-level civil society initiatives.

8.4.3. Promoting Media Literacy and Critical Thinking for Citizens

Media literacy and critical thinking certainly are key to fighting disinformation. From children over-exposed to online information, to elders needing to use new technologies, the audience spans all social and age groups. At present, media literacy education is most often delivered in face-to-face sessions by NGOs or widely respected media organisations (e.g. the BBC, Le Monde). To reach a wider audience across all social and age groups, such initiatives could be scaled-up in online format and promoted at a larger scale, for example via social networks and in schools. As a concrete example, the online game "Get bad news"[116] puts the player in the shoes of a fake news tycoon to teach citizens how disinformation is created and spread. The biggest challenge at present is how to design effective media literacy interventions for the elderly, who are much less aware of the danger of misinformation on the internet, as well as the best ways to protect their privacy online.

8.4.4. Outcomes for this option: the challenge of scaling up the action and overcoming cognitive bias

Media literacy education is still delivered on a very small, localised scale, and it will take some time to scale that up on European level. It is best delivered through initiatives independent of the state, as it could be misconstrued as yet another form of government propaganda. Most importantly, citizens should be able to understand the rationale behind media literacy education and the source of the educational materials. Where possible, the design and delivery of media literacy education should be aligned with the knowledge and beliefs (both political and others) of its target demographic. This is similar to the challenges faced by fact checkers and platforms on the best way to present information, which could reinforce, rather than weaken belief in misinformation, especially in highly polarised communities.

115 https://about.twitter.com/en_us/values/elections-integrity.html#data

116 https://getbadnews.com/#intro

REFERENCES

AFP, (2018), No, this is not footage of a plane caught in a typhoon in China, https://factcheck.afp.com/no-not-footage-plane-caught-typhoon-china, 20 September 2018, visited on 17 October 2018.

Agnew, H. (2016) Fake Vinci press release sends shares down 19%. Financial Times, 22 November 2016, https://www.ft.com/content/ba95f3fe-b0ce-11e6-9c37-5787335499a0, visited 26 November 2018.

Aker, A., Derczynski, L., & Bontcheva, K. (2017). Simple open stance classification for rumour analysis. arXiv:1708.05286 (2017). https://arxiv.org/abs/1708.05286

Albright, J. (2018). #NotOKGoogle search suggestions: 2018 edition. Medium. 21 February 2018. , visited on 20 October 2018.

Albright, J. (2018a). Untrue-Tube: Monetizing Misery and Disinformation. Medium. 25 Feb 2018. , visited on 3 Dec 2018.

Allan, R. (2017). Hard Questions: Who Should Decide What Is Hate Speech in an Online Global Community? Facebook. https://newsroom.fb.com/news/2017/06/hard-questions-hate-speech/

Allan, R. (2018) Oral evidence: Disinformation and 'fake news', HC 363. Questions 4131 – 4273. http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/digital-culture-media-and-sport-committee/disinformation-and-fake-news/oral/92923.html

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. Journal of Economic Perspectives, 31(2), 211-236.

American Press Institute. (2015) New studies on political fact-checking: Growing, influential; but less popular among GOP readers, 22 April 2015. https://www.americanpressinstitute.org/fact-checking-project/new-research-on-political-fact-checking-growing-and-influential-but-partisanship-is-a-factor/, visited on 31 October 2018.

Annany, M. (2018) The partnership press: lessons for platform-publisher collaborations as Facebook and news outlets team to fight misinformation. Tow Center for Digital Journalism, 4 April 2018. https://www.cjr.org/tow_center_reports/partnership-press-facebook-news-outlets-team-fight-misinformation.php, visited on 31 October 2018.

Babakar, M. & Moy, W. (2016) The State of Automated Factchecking. Full Fact report. 17 Aug 2016. https://fullfact.org/blog/2016/aug/automated-factchecking/, visited on 31 October 2018.

Bail, C. A. (2018). Twitter's Flawed Solution to Political Polarization. The New York Times. 8 Sep. 2018. https://www.nytimes.com/2018/09/08/opinion/sunday/twitter-political-polarization.html

Bakshy, E.,Messing, S., Adamic, L. A. (2015) Exposure to ideologically diverse news and opinion on Facebook. Science. Vol. 348, Issue 6239. April 2015.

Balkin, J.M. (2019) The First Amendment in the Second Gilded Age. Buffalo Law Review, 2019 (forthcoming). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3253939, visited on 27 November 2018.

Bartlett, J., Reffin, J., Rumball, N., & Williamson, S. (2017) Anti-social media. Technical report, Demos.

BBC News (2018). Beyond Fake News: BBC launches huge new international anti-disinformation initiative. 9 Nov 2018. https://www.bbc.co.uk/mediacentre/latestnews/2018/beyond-fake-news, visited on 17 Nov 2018.

Beres, D. & Gilmer, M. (2018). A guide to 'deepfakes,' the internet's latest moral crisis. Mashable UK. 02 Feb 2018. https://mashable.com/2018/02/02/what-are-deepfakes/, visited 16 Nov 2018.

Bertrand, N. (2017). Trump retweeted a Twitter bot — then it got suspended. UK Business Insider. 7 Aug 2017. http://uk.businessinsider.com/trump-twitter-bot-nicole-protrump45-2017-8

Bickert, M. (2018). Oral evidence: Fake News – 8 February 2018. http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/digital-culture-media-and-sport-committee/fake-news/oral/78195.html

Booth R., Weaver M., Hern A., Smith S., Walker S. (2017), Russia used hundreds of fake accounts to tweet about Brexit, data shows, The Guardian, 14 November 2017, https://www.theguardian.com/world/2017/nov/14/how-400-russia-run-fake-accounts-posted-bogus-brexit-tweets, visited on 2 October 2018.

Bradshaw, S. & Howard, P.N. (2017) Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation. The Computational Propaganda Project. https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/89/2017/07/Troops-Trolls-and-Troublemakers.pdf, visited on 31 October 2018.

Bradshaw, S. & Howard, P.N. (2018) Challenging Truth and Trust: A Global Inventory of Organized Social Media manipulation. The Computational Propaganda Research Project. http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2018/07/ct2018.pdf, visited on 31 October 2018.

Brooker, K. (2018). Exclusive: Tim Berners-Lee tells us his radical new plan to upend the World Wide Web. Published on 29 Sep. 2018. https://www.fastcompany.com/90243936/exclusive-tim-berners-lee-tells-us-his-radical-new-plan-to-upend-the-world-wide-web, accessed on 27 Nov 2018.

Budak, C., Agrawal, D., & El Abbadi, A. (2011) Limiting the spread of misinformation in social networks. In Proceedings of the 20th international conference on World wide Web, ACM.

Buning, M.D.C. & al. (2018) A multi-dimensional approach to disinformation. Final Report of the independent High Level Group on Fake News and Online Disinformation. European Commission. https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation, visited on 31 October 2018.

Cadwalladr C. (2017), The great British Brexit robbery: how our democracy was hijacked, The Guardian, 7 May 2017, https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy , visited on 2 October 2018

Cadwalladr, C. (2017a). Revealed: Tory 'dark' ads targeted voters' Facebook feeds in Welsh marginal seat. The Observer. 27 May 2017. https://www.theguardian.com/politics/2017/may/27/conservatives-facebook-dark-ads-data-protection-election, visited on 2 December 2018.

Cadwalladr, C. (2017b). Revealed: Tory 'dark' ads targeted voters' Facebook feeds in Welsh marginal seat. The Guardian. 27 May 2017. https://www.theguardian.com/politics/2017/may/27/conservatives-facebook-dark-ads-data-protection-election, visited on 2 Dec 2018.

Cadwalladr, C. (2018). 'Plucky little panel' that found the truth about fake news, Facebook and Brexit. The Guardian. 28 Jul 2018. https://www.theguardian.com/politics/2018/jul/28/dcms-committee-report-finds-truth-fake-news-facebook-brexit, visited on 2 Dec 2018.

Campoy, A. (2018). More than 60% of Donald Trump's Twitter followers look suspiciously fake. Quartz. 12 October 2018. https://qz.com/1422395/how-many-of-donald-trumps-twitter-followers-are-fake, visited on 20 October 2018.

Chakraborty, A., Sarkar, R., Mrigen, A., and Ganguly, Niloy (2017). Tabloids in the Era of Social Media? Understanding the Production and Consumption of Clickbaits in Twitter. Available at SSRN: http://dx.doi.org/10.2139/ssrn.3034591.

Chan, M. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. Psychological science, 28(11), 1531-1546. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5673564/

Chavoshi, N., Hamooni, H., & Mueen, A. (2017). Temporal Patterns in Bot Activities. *WWW*. http://doi.org/10.1145/3041021.3051114

Colleoni, E., Rozza, A. & Arvidsson, A. (2014) Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data, Journal of Communication, Vol. 64, Issue 2. Pages 317–332, https://doi.org/10.1111/jcom.12084

Collins B. & Cox J. (2017), Jenna Abrams, Russia's Clown Troll Princess, Duped the Mainstream Media and the World, Daily Beast, 11 February 2017, https://www.thedailybeast.com/jenna-abrams-russias-clown-troll-princess-duped-the-mainstream-media-and-the-world , visited on 2 October 2018.

Confessore, N., Dance, J.X.A., Harris, R., & Hansen, M. (2018). The Follower Factory. New York Times. Published on 27 Jan 2018. https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html

Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1), 1-4.

Cook, J. , Ecker, U. and Lewandowsky, S. (2015). Misinformation and How to Correct It. In Emerging Trends in the Social and Behavioral Sciences (eds R. A. Scott and S. M. Kosslyn). doi:10.1002/9781118900772.etrds0222

Cresci, S., Di Pietro, R., Petrocchi, M.,Spognardi, A., Tesconi, M. (2016): DNA-inspired online behavioral modeling and its application to spambot detection. IEEE Intelligent Systems, 31(5).

Curtis, C. (2018). Deepfakes are being weaponized to silence women — but this woman is fighting back. The Next Web. 5 Oct. 2018. https://thenextweb.com/tech/2018/10/05/deepfakes-are-being-weaponized-to-silence-women-but-this-woman-is-fighting-back/, visited on 16 Nov 2018.

Davies, H. (2015). Ted Cruz using firm that harvested data on millions of unwitting Facebook users. https://www.theguardian.com/us-news/2015/dec/11/senator-ted-cruz-president-campaign-facebook-user-data.

Dance, G. X. J., Confessore, N., & LaForgia, M. (2018). Facebook Gave Device Makers Deep Access to Data on Users and Friends. The New York Times. 3 Jun 2018. https://www.nytimes.com/interactive/2018/06/03/technology/facebook-device-partners-users-friends-data.html, visited on 2 Dec 2018.

DCMS Report (2018). House of Commons Digital, Culture, Media and Sport Committee. Disinformation and 'fake news': Interim Report. HC 363. Published on 29 July 2018 by authority of the House of Commons. https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/363/363.pdf, visited on 31 October 2018.

DCMS HC 363 (2018). Oral evidence : Fake News – 8 February 2018 (George Washington University, Washington DC), HC 363. http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/digital-culture-media-and-sport-committee/fake-news/oral/78195.html, visited on 31 October 2018.

DCMS HC 1630 (2018). Disinformation and 'fake news': Interim Report: Government Response to the Committee's Fifth Report of Session 2017–19. HC 1630. Published on 23 October 2018. https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1630/1630.pdf, visited on 20 November 2018.

De Vynck, G. & Wang, S. (2018). Russian Bots Retweeted Trump's Twitter 470,000 Times. Bloomberg. 26 Jan 2018. https://www.bloomberg.com/news/articles/2018-01-26/twitter-says-russian-linked-bots-retweeted-trump-470-000-times

Del Vicario, M., Bessi A., Zollo F., Petroni F., Scala A., Caldarelli G., Stanley, E. & Quattrociocchi, W. (2016). The spreading of misinformation online. Proceedings of the National Academy of Sciences, 113 (3), 554-559.

E. Denham, UK Information Commissioner (2018). Oral evidence: Disinformation and 'fake news', HC 363 . Questions 4274 – 4382. Digital, Culture, Media and Sport International Grand Committee. http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/digital-culture-media-and-sport-committee/disinformation-and-fake-news/oral/92924.html

Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Wong Sak Hoi, G. & Zubiaga, A. (2017) SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. In International Workshop on Semantic Evaluation, August 2017. http://www.derczynski.com/sheffield/papers/rumoureval-task.pdf

Duggan, M. (2015). Mobile Messaging and Social Media Users 2015. Pew Research. http://www.pewresearch.org/wp-content/uploads/sites/9/2015/08/Social-Media-Update-2015-FINAL2.pdf, visited on 31 October 2018.

Dungs, S., Aker, A., Fuhr, N. & Bontcheva, K. (2018). Can Rumour Stance Alone Predict Veracity? Proceedings of the 27th International Conference on Computational Linguistics (COLING), pages 3360–3370.

EC public consultation (2018). Synopsis report of the public consultation on fake news and online disinformation. Consultation results. 26 April 2018. https://ec.europa.eu/digital-single-market/en/news/synopsis-report-public-consultation-fake-news-and-online-disinformation, visited on 26 Nov 2018.

Edgett, J. S. (2017). United States Senate Committee on the Judiciary, Subcommittee on Crime and Terrorism, Testimony of Sean J. Edgett. Acting General Counsel, Twitter, Inc. 31 October 2017. https://www.judiciary.senate.gov/imo/media/doc/10-31-17%20Edgett%20Testimony.pdf, visited on 22 October 2018.

EDPS (2018) EDPS Opinion on online manipulation and personal data. European Data Protection Supervisor, Opinion 3/2018. https://edps.europa.eu/sites/edp/files/publication/18-03-19_online_manipulation_en.pdf, visited on 27 November 2018.

Ellis, E.G. (2018). People Can Put Your Face on Porn—and the Law Can't Help You. Wired. 26 Jan 2018. https://www.wired.com/story/face-swap-porn-legal-limbo/, visited on 16 Nov. 2018.

Eurobarometer 464 (2018) Final results of the Eurobarometer on fake news and online disinformation. https://ec.europa.eu/digital-single-market/en/news/final-results-eurobarometer-fake-news-and-online-disinformation

Facebook (2018). House Energy and Commerce Questions for the Record. http://docs.house.gov/meetings/IF/IF00/20180411/108090/HHRG-115-IF00-Wstate-ZuckerbergM-20180411.pdf

Faris, R., Roberts, H., Etling, B., Bourassa, N., Zuckerman, E. & Benkler, Y. (2017). Partisanship, Propaganda, and Disinformation: Online Media and the 2016 U.S. Presidential Election. Berkman Klein Center Research Publication 2017-6.

Flaxman, S., Goel, S., Rao, J.M. (2018) Filter Bubbles, Echo Chambers, and Online News Consumption, Public Opinion Quarterly, Volume 80, Issue S1, 1 January 2016, Pages 298–320, https://doi.org/10.1093/poq/nfw006

Frenkel, S., Confessore, N., Kang, C., Rosenberg, M. & Nicas, J. (2018). Delay, Deny and Deflect: How Facebook's Leaders Fought Through Crisis. The New York Times. Published on 14 Nov 2018. https://www.nytimes.com/2018/11/14/technology/facebook-data-russia-election-racism.html, visited on 27 Nov 2018.

Frier, S. (2018). Trump's Campaign Said It Was Better at Facebook. Facebook Agrees. Bloomberg, 3 April 2018, https://www.bloomberg.com/news/articles/2018-04-03/trump-s-campaign-said-it-was-better-at-facebook-facebook-agrees, visited on 31 October 2018.

Funke. D. (2018) Automated fact-checking has come a long way. But it still faces significant challenges. 4 April, 2018. https://www.poynter.org/news/automated-fact-checking-has-come-long-way-it-still-faces-significant-challenges, visited on 29 October 2018.
Garrett, R. K. (2013). Selective Exposure: New Methods and New Directions. Communication Methods and Measures, 7:247–256, 2013.

Giglietto, F., Iannelli, L., Rossi, L. & Valeriani, A. (2016). Fakes, News and the Election: A New Taxonomy for the Study of Misleading Information within the Hybrid Media System. Convegno AssoComPol 2016. SSRN, https://ssrn.com/abstract=2878774

Golbeck J. (2016) User Privacy Concerns with Common Data Used in Recommender Systems. In: Spiro E., Ahn YY. (eds) Social Informatics. SocInfo 2016. Lecture Notes in Computer Science, vol 10046. Springer, Cham

Golbeck, J. & Mauriello, M. L. (2016) User Perception of Facebook App Data Access: A Comparison of Methods and Privacy Concerns. Future Internet. doi:10.3390/fi8020009.

Golbeck, J., Robles, C., Edmondson, M. & Turner, K. (2011) Predicting Personality from Twitter, 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int. Conf. on Social Computing, pp. 149-156.

Golbeck, J., Robles, C. & Turner, K. (2011a). Predicting personality with social media. In CHI '11 Human Factors in Computing Systems (CHI EA '11). ACM. pages 253-262.

Gorrell, G., Roberts, I., Greenwood, M.A., Bakir, M., Iavarone, B. & Bontcheva, K. (2018) Quantifying Media Influence and Partisan Attention on Twitter During the UK EU Referendum. International Conference on Social Informatics (SocInfo), pp.274-290.

Gorrell, G., Greenwood, M. A., Roberts, I., Maynard, D., Bontcheva, K. (2018). Twits, Twats and Twaddle: Trends in Online Abuse towards UK Politicians. 12th Int. AAAI Conf. on Web and Social Media (ICWSM). https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17861/17060

Gorwa, R. & Guilbeault, D. (2018). Unpacking the Social Media Bot: A Typology to Guide Research and Policy. Policy & Internet, 10 August 2018. doi:10.1002/poi3.184

Gottfried, J. & Greco, E. (2018). Younger Americans are better than older Americans at telling factual news statements from opinions. Pew Research Center. 23 Oct. 2018. http://www.pewresearch.org/fact-tank/2018/10/23/younger-americans-are-better-than-older-americans-at-telling-factual-news-statements-from-opinions/, visited 17 Nov. 2018.

Graves, L. (2018) Understanding the Promise and Limits of Automated Fact-Checking. Reuters Institute report, http://www.digitalnewsreport.org/publications/2018/factsheet-understanding-promise-limits-automated-fact-checking/, visited on 31 October 2018.

Grimaldi, J. V. & Overberg, P. (2017) Millions of people post comments on Federal Regulations. Many are fake. The Wall Street Journal. https://www.wsj.com/articles/millions-of-people-post-comments-on-federal-regulations-many-are-fake-1513099188

Gyenes, N. & Mina, X. (2018) How Misinfodemics Spread Disease. The Atlantic. Published on 30 Aug. 2018. https://www.theatlantic.com/technology/archive/2018/08/how-misinfodemics-spread-disease/568921/, visited on 26 Nov 2018.

Haddow, D.. (2016). Meme warfare: how the power of mass replication has poisoned the US election. https://www.theguardian.com/us-news/2016/nov/04/political-memes-2016-election-hillary-clinton-donald-trump, visited on 5 November 2018 .

Harford, T. (2018). How to burst your political filter bubble. 12 October 2018. http://timharford.com/2018/10/how-to-burst-your-political-filter-bubble/

Harvey, D. & Roth, Y. (2018). An update on our elections integrity work. Twitter. 1 October 2018. https://blog.twitter.com/official/en_us/topics/company/2018/an-update-on-our-elections-integrity-work.html, visited on 31 October 2018.

HLEG report (2018). A multi-dimensional approach to disinformation. Report of the independent High level Group on fake news and online disinformation. European Commission, March 2018. http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=50271, visited on 31 October 2018.

House of Commons (2018). Digital, Culture, Media and Sport Committee. Disinformation and 'fake news': Interim Report. HC 363. July 2018. https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/363/363.pdf, visited on 31 October 2018.

Howard, P.N. & Kollanyi, B (2016). Bots, #Strongerin, and #Brexit: Computational Propaganda During the UK-EU Referendum. SSRN, https://ssrn.com/abstract=2798311

Howard, P. N., Kollanyi, B., Bradshaw, S. & Neudert, L. M. (2018). Social Media, News and Political Information during the US Election: Was Polarizing Content Concentrated in Swing States?. arXiv preprint arXiv:1802.03573. https://arxiv.org/abs/1802.03573

Howard, P.N., Woolley, S. & Calo, R. (2018b) Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration, Journal of Information Technology & Politics, 15:2, 81-93, DOI: 10.1080/19331681.2018.1448735

Hui, J. (2018) How deep learning fakes videos (Deepfakes) and how to detect it? Medium. 28 Apr 2018. https://medium.com/@jonathan_hui/how-deep-learning-fakes-videos-deepfakes-and-how-to-detect-it-c0b50fbf7cb9 , visited 16 Nov 2018.

InVid Project, (2018), InVID Plugin surpassed 4000 installations – New release available, https://www.invid-project.eu/tools-and-services/invid-verification-plugin/, visited on 2 October 2018.

InVID Project, (Online 1), InVID Verification Plugin, https://www.invid-project.eu/tools-and-services/invid-verification-plugin/, visited on 2 October 2018.

Ireton C. & Posetti, J., eds. (2018) Journalism, 'fake news', & disinformation: handbook for journalism education and training, Unesco, https://en.unesco.org/fightfakenews, visited on 20 November 2018.

Jack, C. (2017) Lexicon of Lies: Terms for Problematic Information, Data & Society Research Institute.

Jeangène Vilmer, J.-B., Escorcia, A., Guillaume, M., & Herrera, J. (2018) Information Manipulation: A Challenge for Our Democracies, report by the Policy Planning Staff (CAPS) of the Ministry for Europe and Foreign Affairs and the Institute for Strategic Research (IRSEM) of the Ministry for the Armed Forces, Paris, August 2018.

Kang, C. & Frenkel, S. (2018). Facebook Says Cambridge Analytica Harvested Data of Up to 87 Million Users. The New York Times. https://www.nytimes.com/2018/04/04/technology/mark-zuckerberg-testify-congress.html

Kao, J. (2017) More than a million pro-repeal Net Neutrality comments were likely faked. Medium - Hackernoon. Published on 23 November 2017. https://hackernoon.com/more-than-a-million-pro-repeal-net-neutrality-comments-were-likely-faked-e9f0e3ed36a6, visited on 27 November 2018.

Khaldarova, I., & Pantti, M. (2016). Fake news: The narrative battle over the Ukrainian conflict. Journalism Practice, 10(7), 891-901.

Kelly, J. & François, C. (2018). This is what filter bubbles actually look like. MIT Technology Review. https://www.technologyreview.com/s/611807/this-is-what-filter-bubbles-actually-look-like/

Kirk A. (2017), EU referendum: The claims that won it for Brexit, fact checked, The Telegraph, 13 March 2017, https://www.telegraph.co.uk/politics/0/eu-referendum-claims-won-brexit-fact-checked/, visited on 2 October 2018

Kosinski, M., Stillwell, D., Graepel, T. (2013). Digital records of behavior expose personal traits. Proceedings of the National Academy of Sciences Apr 2013, 110 (15) 5802-5805.

Kumar, S., West, R., & Leskovec, J. (2016): Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes. In Proceedings of the 25th International Conference on World Wide Web.

Lacy, L. & Rosenstiel, T. (2015). Defining and measuring quality journalism. *Rutgers*.

Lapowski, I. (2018). Parkland Conspiracies Overwhelm the Internet's Broken Trending Tools. Wired. 21 Feb 2018. 6 Dec 2018.

Lazer, D., Baum, M.,Grinberg, N., Friedland, L., Joseph, K., Hobbs, W., & Mattsson, C. (2017). Combating Fake News: An Agenda for Research and Action. https://shorensteincenter.org/combating-fake-news-agenda-for-research/, visited 20 Nov. 2018.

Levien, R. & Aiken, A (1998) Attack resistant trust metrics for public key certification. In the 7th USENIX Security Symposium, San Antonio, Texas, January 1998.

Loewenstein, G. (1994) The psychology of curiosity: A review and reinterpretation. Psychological Bulletin 116, 1.

Lucas, E. (2018). Oral Evidence: Fake News HC 363. http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/digital-culture-media-and-sport-committee/disinformation-and-fake-news/oral/79824.html

Lucas, I. (2018a) We don't know who just spent £250k on pro-Brexit Facebook ads – that should worry us all. Nwq Stateman. 25 October 2018. , visited 4 December 2018.

Loewenstein, G. (1994) The psychology of curiosity: A review and reinterpretation. Psychological Bulletin 116, 1.

Lythgoe, L., & Dixon, H. (2016) EU-bashing stories are misleading voters – here are eight of the most toxic tales. The Guardian, 19 May 2016. https://www.theguardian.com/commentisfree/2016/may/19/inaccurate-pro-brexit-infacts-investigation-media-reports-eu-referendum, visited on 27 November 2018.

MacGuill, D. (2018) Did Facebook Flag the Declaration of Independence as Hate Speech? Snopes. https://www.snopes.com/fact-check/facebook-declaration-of-independence-hate-speech/

Mantzarlis, A. (2017) Repetition boosts lies — but could help fact-checkers, too. Poynter, 30 May 2017. https://www.poynter.org/news/repetition-boosts-lies-could-help-fact-checkers-too, visited on 20 November 2018.

Mantzarlis, A. (2017A) There are now 114 fact-checking initiatives in 47 countries. Poynter, 27 February 2017. https://www.poynter.org/news/there-are-now-114-fact-checking-initiatives-47-countries, visted on 26 November 2018.

Marconi, F. & Daldrup, T. (2018). How The Wall Street Journal is preparing its journalists to detect deepfakes. Nieman Lab. 15 Nov. 2018. http://www.niemanlab.org/2018/11/how-the-wall-street-journal-is-preparing-its-journalists-to-detect-deepfakes/, visited 16 Nov 2018.

Matsakis, L. (2018) Artificial Intelligence is Now Fighting Fake Porn. WIRED. 14 Feb 2018. https://www.wired.com/story/gfycat-artificial-intelligence-deepfakes/

McLaughlin, T.. (2018). Disinformation Is Spreading on WhatsApp in India—And It's Getting Dangerous. The Atlantic. https://www.theatlantic.com/international/archive/2018/09/fighting-whatsapp-disinformation-india-kerala-floods/569332/ September 5, 2018, visited on 20 October 2018.

Melford, C., Mousavizadeh, A., & Rogers, D. (2018). Global Disinformation Index—a step in the right direction. April 2018. https://medium.com/@cmelford/a-global-disinformation-index-a-step-in-the-right-direction-5165aee90198, visited on 4 November 2018.

MeniThings, (2017), Extreme Crosswind | Airliner spins 360, https://www.youtube.com/watch?v=AgvzhJpyn10 , 14 June 2017, visited on 17 October 2018.

Messing, S., & Westwood, S. J. (2014). Selective Exposure in the Age of Social Media: Endorsements Trump Partisan Source Affiliation When Selecting News Online. Communication Research, 41(8), 1042–1063. https://doi.org/10.1177/0093650212466406

Mitchell, A., Gottfried, J., Barthel, M. & Shearer, E. (2016). The Modern News Consumer. Pew Research. http://www.journalism.org/2016/07/07/pathways-to-news/

Monaco, N. (2018). What makes some people more susceptible to disinformation? NDI. 18 Apr 2018. https://www.demworks.org/what-makes-some-people-more-susceptible-disinformation

Moore M. & Ramsay G. (2017), UK media coverage of the 2016 EU Referendum campaign, King's College London, May 2017, https://www.kcl.ac.uk/sspp/policy-institute/CMCP/UK-media-coverage-of-the-2016-EU-Referendum-campaign.pdf , visited on 2 October 2018.

Morozov, E. (2018). After the Facebook scandal it's time to base the digital economy on public v private ownership of data. The Guardian. https://www.theguardian.com/technology/2018/mar/31/big-data-lie-exposed-simply-blaming-facebook-wont-fix-reclaim-private-information

Müller, K. & Schwarz, C. (2018). Fanning the Flames of Hate: Social Media and Hate Crime. Centre for Competitive Advantage in the Global Economy, University of Warwick. Working paper No.373.

NATO StratCom COE (2017). StratCom laughs. In search of an analytical framework

NATO Strategic Communications Centre of Excellence. ISBN: 978-9934-564-12-3.

NED (2018). Comparative Responses to the Global Disinformation Challenge. National Endowment for Democracy. October 4-5, 2018.

Nelson, R.A. (1996) A Chronology and Glossary of Propaganda in the United States, Westport, Connecticut: Greenwood Press, ISBN 0313292612.

Newman, N. (2011) Mainstream media and the distribution of news in the age of social discovery. Reuters Institute for the Study of Journalism, University of Oxford.

Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A. L., Nielsen, R. K. (2018). Reuters Institute Digital News Report 2018.

Newton, C. (2018) A partisan war over fact-checking is putting pressure on Facebook. Is it fake, or is it just clickbait?. The Verve, 12 September 2018. https://www.theverge.com/2018/9/12/17848478/thinkprogress-weekly-standard-facebook-fact-check-false, visited on 29 October 2018.

Ng, A. (2018). Facebook's real fake-news problem: It's the memes, stupid. Cnet. 15 Apr 2018. https://www.cnet.com/news/mark-zuckerberg-facebook-and-fake-news-its-the-memes-stupid/, visited 16 Nov 2018.

Nguyen, N.P., Yan, G., Thai, M.T., & Eidenbenz, S. (2012) Containment of misinformation spread in online social networks. In Proceedings of the 4th Annual ACM Web Science Conference.

Nyst, C. & Monaco, N. (2018) STATE-SPONSORED TROLLING: How Governments Are Deploying Disinformation as Part of Broader Digital Harassment Campaigns. Digital Intelligence Laboratory, Institute for the Future.

Quattrociocchi, W., Scala, A. & Sunstein, C.R. (2016). Echo Chambers on Facebook, SSRN, https://ssrn.com/abstract=2795110

Ofcom (2018) News consumption in the UK. Ofcom, 25 July 2018, https://www.ofcom.org.uk/research-and-data/tv-radio-and-on-demand/news-media/news-consumption, visited on 20 November 2018.

O'Sullivan D. (2017) A notorious Russian Twitter troll came back, and for a week Twitter did nothing, CNN, 19 November 2017, https://money.cnn.com/2017/11/17/media/new-jenna-abrams-account-twitter-russia/index.html, visited on 2 October 2018.

Pamment, J., Nothhaft, H., Agardh-Twetman, H., Fjällhed, A. (2018). Countering Information Influence Activities: The State of the Art. version 1.4. 1 July 2018. Department of Strategic Communication, Lund University.

Phillips, T. & Ball, J. (2017) Twitter Has Suspended Another 45 Suspected Propaganda Accounts After They Were Flagged By BuzzFeed News. BuzzFeed. 24 November 2017. https://www.buzzfeed.com/tomphillips/we-found-45-suspected-bot-accounts-sharing-pro-trump-pro, visited on 29 October 2018.

Phillips, T. (2017) This Is What The Twitter Abuse Of Politicians During The Election Really Looked Like. BuzzFeed. 23 July 2017. https://www.buzzfeed.com/tomphillips/twitter-abuse-of-mps-during-the-election-doubled-after-the

Pickles N. (2018), Letter from Nick Pickles, Head of Public Policy, Twitter UK, to Damian Collins MP, Chair, Digital, Culture, Media and Sport Select Committee House of Commons, 19 January 2018, https://www.parliament.uk/documents/commons-committees/culture-media-and-sport/180119%20Nick%20pickles%20Twitter%20to%20Chair.pdf, visited on 2 October 2018.

Posetti, J. & Matthews, A. (2018). A short guide to the history of 'fake news' and disinformation. A Learning Module for Journalists and Journalism Educators. International Center for Journalists.

Price, R. (2018) Russia has allegedly been spreading far-right propaganda on Facebook to try and influence the US midterms — here it is. Business Insider UK, 27 October 2018. http://uk.businessinsider.com/right-wing-propaganda-allegedly-posted-facebook-russia-2018-10, visited on 27 November 2018.

Quattrociocchi, W., A. Scala, and C. R. Sunstein. (2016). Echo Chambers on Facebook. Social Science Research Network. https://papers.ssrn.com/abstract=2795110

Qiu, X., Oliveira, D.F.M., Sahami Shirazi, A., Flammini, A. & Menczer, F. (2017). Limited individual attention and online virality of low-quality information. Nature Human Behaviour. Volume 1, Article number 0132. 26 June 2017.

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M.s 2010. Classifying latent user attributes in Twitter. In Proceedings of the 2nd international workshop on Search and mining user-generated contents (SMUC '10). DOI: https://doi.org/10.1145/1871985.1871993

Rapoza, K. (2018) Can fake news impact the stock market? Forbes, 26 February 2017, https://www.forbes.com/sites/kenrapoza/2017/02/26/can-fake-news-impact-the-stock-market/#40b380422fac, visited on 26 November 2018.

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In EMNLP, pp. 2931-2937. http://aclweb.org/anthology/D17-1317

Recasens, M., Danescu-Niculescu-Mizil, C., Jurafsky, D. (2013). Linguistic models for analyzing and detecting biased language. ACL, Vol. 1, pp. 1650-1659. http://www.aclweb.org/anthology/P13-1162

Recode (2017), Twitter's list of 2,752 Russian trolls, Recode, 2 Nov 2017, https://www.recode.net/2017/11/2/16598312/russia-twitter-trump-twitter-deactivated-handle-list, visited on 1 November 2018.

RES (2018) Social media bots influence stock market performance. Media briefing, March 2018. http://www.res.org.uk/details/mediabrief/10926361/SOCIAL-MEDIA-BOTS-INFLUENCE-STOCK-MARKET-PERFORMANCE.html, visited 26 November 2018.

Reynolds, M. (2018) How religious extremists gamed Facebook to target millions with far-right propaganda. Wired (UK), 2 February 2018. http://www.wired.co.uk/article/britain-first-facebook-jim-dowson-knights-templar, visited 17 October 2018.

Romm, T. (2018) Twitter will begin labeling political ads about issues such as immigration. The Washington Post. 30 August 2018. https://www.washingtonpost.com/technology/2018/08/30/twitter-will-begin-labeling-political-ads-about-issues-like-immigration/, visited on 31 October 2018.

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C.,Thies, J. & Nießner, M. (2018). FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces. https://arxiv.org/abs/1803.09179

RSF (2017) RSF joins coalition to launch new website tracking press freedom violations in United States, Reporters without Borders, 2 August 2017, https://rsf.org/en/news/rsf-joins-coalition-launch-new-website-tracking-press-freedom-violations-united-states, visited on 20 November 2018.

RSF (2018). ONLINE HARASSMENT OF JOURNALISTS - Attack of the trolls. Reporters without Borders. 25 July 2018. https://rsf.org/en/news/rsf-publishes-report-online-harassment-journalists

Russia Today (2017), Twitter reveals just 6 tweets posted from Russia to 'influence' Brexit vote... all from RT, Russia Today, 14 Dec 2017, https://www.rt.com/uk/413249-big-reveal-twitter-lists-just/, visited on 2 October 2018.

Salinas, S. (2018) The top trending video on YouTube was a false conspiracy that a survivor of the Florida school shooting was an actor. CNBC. 21 February 2018. https://www.cnbc.com/2018/02/21/fake-news-item-on-parkland-shooting-become-top-youtube-video.html, visited 20 October 2018.

Salinas, S. (2018A) Facebook has been talking up its third-party fact-checkers, but at least one says it's checking just one post per day. CNBC, 18 October 2018. https://www.cnbc.com/2018/10/18/facebook-talks-up-third-party-reviewers-but-one-isnt-reviewing-much.html, visited 27 November 2018.

Saslow, E. (2018). 'Nothing on this page is real': How lies become truth in online America. The Washington Post. 17 Nov 2018.
https://www.washingtonpost.com/national/nothing-on-this-page-is-real-how-lies-become-truth-in-online-america/2018/11/17/edd44cc8-e85a-11e8-bbdb-72fdbf9d4fed_story.html, visited 20 Nov 2018.

Shao, C., Ciampaglia, G.L., Flammini, A., & Menczer, F. (2016) Hoaxy: A Platform for Tracking Online Misinformation. In Third Workshop on Social News On the Web (WWW SNOW). Preprint arXiv:1603.01511.

Shao, C., Ciampaglia, G.L., Varol, O., Yang, K.-C., Flammini, A. & Menczer, F. (2018). The spread of low-credibility content by social bots. Nature Communications. Volume 9, Article number: 4787. https://www.nature.com/articles/s41467-018-06930-7#Sec13

Shore, J., Baek, J. & Dellarocas, C. (In print). Network structure and patterns of information diversity on Twitter. MIS Quarterly.

Silverman, C. (2017) 5 ways scammers exploit Facebook to feed you false information. BuzzFeed, 28 April 2017. https://www.buzzfeed.com/craigsilverman/how-facebook-is-getting-played, visited on 20 October 2018.

Silverman, C., Lytvynenko, J., and Pham, S. (2017) These Are 50 Of The Biggest Fake News Hits On Facebook In 2017. BuzzFeed, 28 December 2017. https://www.buzzfeednews.com/article/craigsilverman/these-are-50-of-the-biggest-fake-news-hits-on-facebook-in , visited on 20 October 2018.

Silverman, C. (2018) Twitter Allowed Hackers To Run An Ad On Its Platform That Pretended To Come From Twitter Itself. BuzzFeed News, 8 January 2018. https://www.buzzfeednews.com/article/craigsilverman/twitter-keeps-allowing-hackers-to-run-malicious-ads-that, visited on 31 October 2018.

Soltani, A. (2018) Oral evidence: Disinformation and 'fake news', HC 363. Questions 4274 – 4382. Published on 27 Nov 2018. http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/digital-culture-media-and-sport-committee/disinformation-and-fake-news/oral/92924.html

SRF. (2016) Die Propaganda-Analyse. https://swprs.org/srf-propaganda-analyse/

Staines, C. & Moy, W. (2018). Tackling misinformation in an open society. Full Fact report. 2 October 2018. https://fullfact.org/blog/2018/oct/tackling-misinformation-open-society/, 31 October 2018.

Starbird, K. (2017) Information Wars: A Window into the Alternative Media Ecosystem. In Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM-17).

Storyful team. (2017). Closed Networks and their Impact on the French Election. 21 Apr 2017. https://storyful.com/gamers-bots-and-memes-understanding-closed-networks-and-their-impact-on-the-french-election/, visited on 16 Nov 2018.

Subrahmanian, V.S., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., Zhu, L., Ferrara, E., Flammini, A., & Menczer, F. (2016) The DARPA Twitter Bot Challenge. Computer 49 (6): pp. 38–46. https://doi.org/10.1109/MC.2016.183

Swire, B., Berinsky, A.J., Lewandowsky, S., & Ecker, U.K.H. (2017) Processing political misinformation: comprehending the Trump phenomenon. Royal Society Open Science 4(3).

Swire, B., Ecker, U. K. H., & Lewandowsky, S. (2017b). The Role of Familiarity in Correcting Inaccurate Information. Journal of Experimental Psychology: Learning, Memory, and Cognition. May 2017.

Sydell, L. (2016) We Tracked Down A Fake-News Creator In The Suburbs. Here's What We Learned. NPR. 23 November 2016. http://www.cpr.org/news/npr-story/npr-finds-the-head-of-a-covert-fake-news-operation-in-the-suburbs, visited on 22 October 2018.

Tambuscio, M., Ruffo, G., Flammini, A. & Menczer, F. (2015) Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks. In Proceedings of the 24th International Conference on World Wide Web, ACM.

Tandoc Jr, E.C., Lim, Z. W., & Ling, R. (2017) Defining "Fake News": A Typology of Scholarly Definitions. Digital Journalism, 5 (7): 1-17.

The Guardian Pass Notes (2017) Jenna Abrams: the Trump-loving Twitter star who never really existed, The Guardian, 3 November 2017,

https://www.theguardian.com/technology/shortcuts/2017/nov/03/jenna-abrams-the-trump-loving-twitter-star-who-never-really-existed , visited on 2 October 2018.

Theocharis, Y.; Barberà, P.; Fazekas, Z.; Popa, S. A.; & Parnet, O. (2016) A bad workman blames his tweets: the consequences of citizens' uncivil twitter use when interacting with party candidates. Journal of Communication 66(6): 1007–1031.

Teyssou, D., Leung, J.M., Apostolidis, E., Apostolidis, K., Papadopoulos, S., Zampoglou, M., & Mezaris, V. (2017) The InVID plug-in: web video verification on the browser. In Proceedings of the First International Workshop on Multimedia Verification (pp. 23-30). ACM, October 2017.

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018) FEVER: a large-scale dataset for Fact Extraction and VERification. In NAACL 2018. https://arxiv.org/abs/1803.05355

Tunikova, O. (2018). 7 Memes that Drive Misinformation on Facebook. StopAd. 27 Apr 2018. https://stopad.io/blog/misinformation-memes-on-facebook, visited on 16 Nov 2018.

UNHRC (2018) A Human Rights Approach to Platform Content Regulation. The Special Rapporteur's 2018 report to the United Nations Human Rights Council. https://freedex.org/a-human-rights-approach-to-platform-content-regulation/, visited on 27 November 2018.

Vanderbiest, N. Online 1, A Russian influence on the French elections?, http://www.reputatiolab.com/2017/04/quelle-est-linfluence-russe-sur-la-campagne-presidentielle-francaise/, visited on 17 October 2018.

Varol, O. Ferrara, E. Davis, C.A., Menczer, F. & Flammini, A. (2017) Online Human-Bot Interactions: Detection, Estimation, and Characterization. In Intl. AAAI Conf. on Web and Social Media (ICWSM).

Vlachos, A., & Riedel, S. (2015) Identification and Verification of Simple Claims about Statistical Properties. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 17-21 September 2015, Lisbon, Portugal. Association for Computational Linguistics, 2596-2601 .

Vosoughi, S., Roy, D., & Aral, S. The spread of true and false news online. 2018. Science: Vol. 359, Issue 6380, pp. 1146-1151.

Wang, A.H. (2010) Don't follow me: Spam detection in Twitter, in Proceedings of the International Conference on Security and Cryptography (SECRYPT 2010). https://ieeexplore.ieee.org/abstract/document/5741690

Wardle, C. & Derakhshan, H. (2017) Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making. Council of Europe report DGI(2017)09, October 2017, https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c, visited on 31 October 2018.

Wardle, C. (2017) Fake News. it's Complicated. First Draft, 16 February 2017. https://firstdraftnews.org/fake-news-complicated/, visited on 31 October 2018.

Waseem, Z. Davidson, T., Warmsley, D., Weber., I. (2017) Understanding Abuse: A Typology of Abusive Language Detection Subtasks. Proceedings of the First Workshop on Abusive Language Online, 78-85.

Weedon, J., Nuland, W. and Stamos, A. (2017) Information Operations and Facebook. Facebook. 27 April 2017, https://fbnewsroomus.files.wordpress.com/2017/04/facebook-and-information-operations-v1.pdf, visited on 31 October 2018.

Witness & First Draft (2018). Mal-uses of AI-generated Synthetic Media and Deepfakes: Pragmatic Solutions Discovery Convening. June 2018. https://witness.org/witness-leads-convening-on-proactive-solutions-to-mal-uses-of-deepfakes-and-other-ai-generated-synthetic-media/ and https://download1965.mediafire.com/tjdj74j339mg/q5juw7dc3a2w8p7/Deepfakes_Final.pdf, visited on 16 November 2018.

Woolley, S.C. & Howard, P.N. (2016) Automation, Algorithms, and Politics: Political Communication, Computational Propaganda, and Autonomous Agents. International Journal of Communication, 10 (October 2016) http://ijoc.org/index.php/ijoc/article/view/6298/1809

Wulczyn, E., Thain, N., & Dixon, L. (2017) Ex machina: Personal attacks seen at scale. In Proceedings of the 26th International Conference on World Wide Web, 1391–1399.

Zampoglou, M., Papadopoulos, S., & Kompatsiaris, Y. (2017) Large-scale evaluation of splicing localization algorithms for web images. Multimedia Tools and Applications 76:4: 4801-4834.

Zampoglou, Z. Papadopoulos, S., Kompatsiaris, Y., Bouwmeester, R., & Spangenberg, J. (2016) Web and Social Media Image Forensics for News Professionals. In Proceedings of the Tenth International AAAI Conference on Web and Social Media.

Zannettou, S., Caulfield, T., Blackburn, J., De Cristofaro, E., Sirivianos, M., Stringhini, G. & Suarez-Tangil, G. (2018). On the Origins of Memes by Means of Fringe Web Communities. arXiv:1805.12512. https://arxiv.org/abs/1805.12512v2.

Zhang, Ranganathan, *et al* (2018). A Structured Response to Misinformation: Defining and Annotating Credibility Indicators in News Articles. In WWW'2018.

Zollo F., Bessi A., Del Vicario M., Scala A., Caldarelli G., Shekhtman L., Havlin, S., & Quattrociocchi, W. (2017) Debunking in a world of tribes. PLoS ONE 12(7): e0181821. https://doi.org/10.1371/journal.pone.0181821

Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G., Tolmie, P. (2016) Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads. PLoS ONE 11(3): e0150989. https://doi.org/10.1371/journal.pone.0150989

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M. & Procter, R. (2018). Detection and Resolution of Rumours in Social Media: A Survey. ACM Computing Surveys, Vol. 51, No. 2, Article 32. Feb. 2018.

Zucconi, A. (2018). The Ethics of Deep Fakes. https://www.alanzucconi.com/2018/03/14/the-ethics-of-deepfakes/

Zuckerberg, D. (2018) How the Alt-Right Is Weaponizing the Classics. Medium. 15 October. https://medium.com/s/story/how-the-alt-right-is-weaponizing-the-classics-d4c1c8dfcb73

Zuckerberg, M. (2018A) A Blueprint for Content Governance and Enforcement. Facebook, 15 November 2018. https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/, visited on 27 November 2018.

9. ANNEX I: Survey Questions

Technological, social and legislative responses to fake news

9.1. Use of the information you provide

The information you provide by responding to this questionnaire will be used by the EU Disinfo Lab (http://disinfo.eu/) only for the study on "Technological responses to fake news", requested by Directorate-General for Parliamentary Research Services Directorate for Impact Assessment and European Added Value Scientific Foresight Unit (STOA) of the European Parliament.

The response provided, including your personal details provided in the next section, may be shared with STOA.

Your responses will be used anonymously in the publicly available final report of the study.

Your name may be used in the final project report only if you select so in the following questions.

The information you provide will not be used in any other way without your explicit consent.

Do you agree with the intended use of the information described above? *

Yes/No

Can we quote your name in the publicly available final report? *

Yes/No

Can we publish with due accreditation some of your responses in annexes of the publicly available report? *

Yes/No

9.2. Personal Details

First Name *

Last Name *

Email *

Organisation *

Your role within the organisation

9.3. Participation in initiatives related to fake news, misinformation or disinformation

Name of initiative/project/organisation or other activity you participate

Please describe or send a link of your initiative/project

Initiative sponsor

Initiative location

- North America
- South America
- Asia
- Africa
- Oceania
- Antartica

Initiative start date (if applicable)

Initiative end date (if applicable)

Brief overview of the aims and expected results (up to 200 words)

9.4. Problem addressed

What kind of mis/disinformation problem are you solving?

- fact checking
- content verification
- source trustworthiness verification
- clickbait detection
- rumour detection
- network-based disinformation spread analysis
- measuring impact on citizen beliefs/actions
- debunking methods
- legal/ethical aspects
- Other:

Are you analysing only a specific type of misinformation, e.g. information specific to elections?

What platforms are you addressing?

- Fake news sites
- Twitter
- Facebook
- Reddit
- WhatsApp
- Other:

What content in what modalities are you addressing?

- Text
- Images
- Video
- Audio
- Other:

Which of the following are you addressing within your initiative?

- short-term responses to the most pressing problems
- longer-term responses to increase societal resilience to disinformation
- a framework for ensuring that the effectiveness of these responses is continuously evaluated
- Other:

What are the current obstacles for your initiative to achieve its objectives?

Are you collaborating with other organisations/initiatives and how?

9.5. Technical Solutions used

What tech solutions are you currently using/developing for each of the mis/disinformation problems above?

If you are using third-party technology, who is the developer of that technology?

Pointers to further materials on the tech you are using/developing

What are their weaknesses?

How accurate is the technology you are using/developing?

What data/use cases have you tested/evaluated it on?

What technological limitations have you encountered?

What's preventing you from going further?

9.6. Legislation related to fake news, misinformation or disinformation

Which entities are the most relevant to deal with disinformation?

- The EU institutions
- National Governments
- Technology providers
- Civil society
- Academics
- NGOs
- Journalists
- Media agencies
- Other:

Are you aware of any national legislation related to fighting disinformation, and if yes, could you please provide us a pointer to a web page or give us its name?

Do you think that legislation is an effective way to fight fighting fake news?

What kind of policy actions could be taken and what kind of effects you expect they might have.

10. ANNEX II: EU initiatives roadmap

Name	URL	Approach	Stakeholders	Focus	Brief Description
CoInform EU Project	http://coinform.eu/	Tech	EU Project	Research and Technology Development	A Horizon 2020 EU-funded project to promote the interaction between researchers, journalists, private sector, non-profit sector and citizens against misinformation.
Council of Europe	https://www.coe.int/en/web/portal/home	Legal	Government	Regulation	The Council of Europe advocates freedom of expression and of the media, freedom of assembly, equality, and the protection of minorities. It has campaigned on a wide range of issues including child protection, online hate speech, and the rights of the Roma, Europe's largest minority. The Council of Europe helps member states to fight corruption and terrorism and to undertake necessary judicial reforms. Its group of constitutional experts, known as the Venice Commission, offers legal advice to countries throughout the world.
COMRADES EU Project	https://www.comrades-project.eu/	Tech	EU Project	Research and Technology Development	A H2020 EU project developing open-source, community resilience platform, designed by communities, for communities, to help them reconnect, respond to, and recover from crisis situations. It is developing technology for detecting crisis-related misinformation.
Dante EU Project	http://www.h2020-dante.eu/	Tech	EU Project	Research and Technology Development	DANTE supports law enforcement authorities and counter terrorisms in their everyday work. The proposed system will reduce the required investigation time by utilizing automatic processes for detecting relevant terrorist-related data in surface/deep/ dark web.
EU Stratcomm Taskforce	euvsdisinfo.eu	Social	Government	Fact-checking	This initiative was established to deliver a response to pro-Kremlin disinformation. The 'EU versus Disinformation' campaign is run by the European External Action Service East Stratcom Task Force.
European Audiovisual Observatory	https://www.obs.coe.int/en/home	Legal	Government	Governmental Solution	Providing essential market and legal information on the audiovisual industry, a unique information network in Europe and a varied range of information products and services.

Fandango EU Project	https://fandango-project.eu/	Tech	EU Project	Research and Technology Development	The aim of FANDANGO is to verify different typologies of news data, media sources, social media, open data, so as to detect fake news and provide a more efficient and verified communication for all European citizens in three specific domains Climate, Immigration and European Context
InVid EU Project	https://www.invid-project.eu/	Tech	EU Project	Research and Technology Development	InVID has developed services to detect, authenticate and check the reliability and accuracy of newsworthy video files and video content spread via social media.
Joint Research Centre (JRC)	https://ec.europa.eu/info/departments/joint-research-centre_en	Tech	Academic	Research	JRC scientists carry out research, in order to provide independent scientific advice and support for EU policy makers
Journalism Trust Initiative	https://rsf.org/en/news/rsf-and-its-partners-unveil-journalism-trust-initiative-combat-disinformation	Social	Non-profit	Journalistic Standards	RSF is now one of the world's leading NGOs in the defense and promotion of freedom of information
The High Level Expert Group on Fake News and Online Disinformation	https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation	Legal	Government	Regulation	The European Commission set up the High Level Expert Group (HLEG) to advise on policy initiatives to counter fake news and the spread of disinformation online.
PHEME EU Project	https://www.pheme.eu/	Tech	EU Project	Research and Technology Development	The PHEME project develops machine learning algorithms to identify, track, and verify the veracity of online rumours.
Public Data Lab	http://publicdatalab.org	Social	Academic	Research and Technology Development	A pan-European, multi-disciplinary collaboration that explores the use of digital methods to study false viral news, political memes, trolling practices and their social life online and much more.
REVEAL EU Project	https://revealproject.eu/	Tech	EU Project	Research and Technology Development	The REVEAL project developed technology for image verification and high-level analysis of social media, including reputation, influence and credibility of information.
SoBigData EU Project	http://sobigdata.eu/index	Tech	EU Project	Research	The SoBigData project is developing a research infrastructure (RI) for large-scale social media analysis. Analysis of online disinformation and social debates during elections is one of the target applications.

11. ANNEX III: Initiatives in Member States roadmap

Name	Approach	Country	Stakeholder	Focus	Brief Description
Mimikama	Social	AT	Civil society	Media literacy	Austrian association providing Media literacy and fact-checked news on a web portal
Belgian Overlegplatform	Social/Legal	BE	Civil society	Regulation	Following the report of a Belgian expert group, the government created a platform to gather civil society actors fighting misinformation. It will collect useful resources and provide kickstart funding for innovative projects in Belgium, focusing on media literacy, academic research and improving journalism.
Lie Detectors	Social	BE	Civil Society	Media Literacy	Lie Detectors delivers media literacy training for schoolchildren in Europe aged 10-15. The goal is help children understand propaganda, distorted facts online, and news media, as well as make informed choices and resist peer pressure as they assemble their worldview
Aspen Institute: Commission on Trust, Media and Democracy	Legal	BE	Non-profit	Research	The Aspen Institute has earned a reputation for gathering diverse, nonpartisan thought leaders, creatives, scholars and members of the public.
Hercule Platform	Tech	BG	Company	Technology Solution	Hercule is an AI and Big Data platform for finding and fact - cheking breaking news
Human and Social Studies Platform	Social	BG	Academic	Research	Social and political science and journalism research on propaganda narratives.
Center for the Study of Democracy	Social	BG	Academic	Research	Research on Russian propaganda narratives.
Manipulatori	Social	CZ	Company	Fact-checking	Manipulatori.cz is a Czech publishing platform, with focus on political marketing, public relations and communication. It supports fact-checking and projects leading to the improvement of public discourse.
Fact Czech	Social	CZ	Non-profit	Fact-checking	Fact-checking and verification training resources and courses for Czech students in journalism; guides for combatting misinformation and propaganda.
Faktograf.hr	Social	HR	Civil Society	Fact-checking	Fact-checking initiative in Croatia.

Correctiv	Social	DE	Journalist	Fact-checking	Using collaborative tools and open source technology to provide fact-checking
German "Network Enforcement Act"	Legal	DE	Government	Governmental Solution	The Act set up a procedure for online platforms to remove “obviously illegal” posts within 24 hours or risk fines of up to €50 million. The Act makes the platforms accountable for the monitoring and removal of content. In particular, the handling of complaints shall be monitored via monthly checks by the social network’s management.
Faktenfinder	Social	DE	Journalist	Fact-checking	Initiative from ARD to gather fact-checking articles on one common platform
Wafana	Social	DE	Journalist	Media literacy	Initiative aiming at training german-speaking journalists to spot misinformation online
Crowdalyser	Tech	DE	Civil society	Technology Solution	Initiative developed at the MediaLab Bayern bringing software engineers and journalists to visualise narrative evolution and social media manipulation through sockpuppet accounts.
CaptainFact.io	Social	FR	Non-profit	Technology Solution	Collective and real-time fact checking.
Decodex	Tech	FR	Journalist	Fact-checking	Decodex is an initiative from Le Monde and Les Decodeurs to create an index of reliability and truthiness for online information. Based on journalism criterions, the Decodex ranks website from green to red as trustworthy sources.
Les Décodeurs	Social	FR	Journalist	Fact-checking	A fact-checking initiative from Le Monde which verifies news and statements from politicians.
French "Anti Fake News" law	Legal	FR	Government	Governmental Solution	The French parliament is discussing a proposal that would enable law enforcement from the judge on fake news spread during electoral campaigns.
Visibrain	Tech	FR	Company	Fact-checking	Visibrain is a French company specialized in the monitoring and analysis of data circulating on the internet. They've worked with bloggers on disinformation and misinformation mapping.
CNRS - IRISA project	Tech	FR	Academic	Research	This project aims to combine a multi-modal research (source detection, similar context and similar images) to automatically detect misinformation content.
Climate Feedback	Social	FR	Academic	Fact-checking	Scientists fact-checking climate news
FightHoax	Tech	GR	Company	Technology Solution	FightHoax uses bulk requests to give contextual analysis to a news piece and track similarities with content posted elsewhere.

EllinicaHoax	Social	GR	Company	Fact-checking	Resource of false information debunked by journalists and bloggers.
Truly Media	Tech	GR/DE	Civil Society	Fact-checking	Collaborative content verification and fact-checking tool for newsroom journalists
Transparent Referendum Initiative	Social	IE	Civil Society	Fact-checking	Volunteers providing fact-checking for the Irish referendum, to allow for an open, truthful and respectful debate.
Elezioni2018	Tech	IT	Academic	Research	A project on political news coverage on the 2018 Italian general election by the Italian mainstream, digital and alternative media. It analyzes the level of engagement on Facebook and Twitter and measure its polarization as the fraction of overlapping active social media audiences across different media sources.
Laboratory of Data Science and Complexity at the University of Venezia	Tech	IT	Academic	Research	Develop data-driven computational models of complex socio-cognitive systems. We work to develop innovative mathematical models and computational tools to better understand, anticipate and control massive social phenomena with a complex systems approach. The focus is on information and misinformation spreading and their effect on opinions. Recent findings showed the pivotal role of confirmation bias on informational cascades online as well as the inefficacy of debunking in contrasting misinformation spreading.
Polizia Postale	Legal	IT	Government	Governmental Solution	The Postal Police (a division from national police) can be contacted regarding published misinformation. Then , a fact-check is operated by the police and published on the Postal Police website.
4Facts	Social	LV	Company	Technology Solution	Aimed at opening up and popularising fact-checking so that misinformation is identified faster and consequently its distribution is minimized.
Fake news monitoring	Legal	LV	Government	Governmental Solution	The Latvian government is considering to set up an institution monitoring fake news in Latvia.
Debunk	Social	LT	Civil Society	Fact-checking	Civil society collaboration between media organisations, strategic communication experts and "elves" to debunk false stories in Lithuania.
Drog	Social	NL	Civil Society	Media literacy	An online game teaching its users to create fake news pieces. By doing so, it gives citizens understanding of the disinformation mechanisms and strengthens their awareness of the issue.

The Hague Center for Strategic Studies (HCSS)	Social	NL	Civil Society	Research	Research institute focused on Security with records on cybersecurity and disinformation studies.
Olimpiada Cyfrowa	Social	PL	Government	Media literacy	Olimpiada Cyfrowa is a project launched in 2002 by the Modern Poland Foundation (Fundacja Nowoczesna Polska) and funded by the Ministry of Education and Sport and Ministry of Culture and National Heritage. The project targets secondary school pupils to raise their awareness about media skills and literacy, including critical analysis of information, media ethics, the language of media, and internet security. The project also motivates teachers to discuss security issues, the internet, media, and digital education in social media.
Trudat	Social	PL	Journalism	Fact-checking	Trudat is an initiative from Polish media Natemat to fact-check information in Poland. They're also asking the general audience what information should be checked by journalists.
TrustServista	Tech	RO	Company	Technology Solution	TrustServista uses advanced Artificial Intelligence algorithms in order to provide media professionals, analysts and content distributors with in-depth content analytics and verification capabilities
Factba.se	Social	SE	Company	Technology Solution	Factbase is an initiative from Factsquared, using AI and data collection to create a repository of publid declarations and statrements from politicians and official organisations. This repository is made accessible for everyone to fact-check or verify information.
MSB	Social	SE	Government	Governmental Solution	The Swedish Civil Contingency agency (MSB) has raised awareness around disinformation during crisis time through the diffusion of leaflets. They also monitored disinformation within the agency.
The Beacon Project	Social	SK, CZ, HU, PL	Civil Society	Research	The Beacon Project's ICT data collection tool, >versus<, is being developed to aid local media monitors and researchers to obtain qualitative and quantitative data from a wider range of sources. This database will be built by developing a collaborative standardized methodology to provide quality data on the scale and effectiveness of disinformation for the first time. In addition, this data will be further supported by new national polling initiatives.

Security Committee of the Spanish Congress	Legal	SP	Government	Governmental Solution	Non-binding resolution from the Security Committee of the Spanish Congress to strengthen the media in order to fight disinformation
Open Evidence	Tech	SP	Company	Research	Open Evidence, in consortium with RAND Europe, has been commissioned by the European Commission – DG CONNECT, to conduct the study on“Media Literacy and Online Empowerment issues raised by Algorithm-Driven Media Services”
Observador	Social	PT	Journalist	Fact-checking	Fact-checking website
Breakthrough	Social	UK	Company	Research	Full-service communications business by real-time audience engagement. Did some research around disinformation narratives
Britain National Security Communications Unit	Legal	UK	Government	Governmental Solution	The unit is tasked with "combating disinformation by state actors and others". It will also help to "deter" the actions of those creating fake news.
Digital Culture, Media, and Sport Committee on Disinformation and “Fake News”	Legal	UK	Government	Governmental Solution	A UK Parliamentary committee investigating the causes and effects of online disinformation. The transcripts of all oral witness sessions and all written evidence submitted to the committee are publicly available, as well as the reports published by the committee itself
Computational Propaganda project	Tech	UK	Academic	Research	Investigating the use of algorithms, automation and computational propaganda in public life
Factmata	Tech	UK	Company	Technology Solution	Factmata is a company building algorithms for detecting hyperpartisan content, junk websites and hate speech. One of its main project is to develop a news platform that could help journalists and editor to verify information more quickly through the use of algorithms.
FactCheckNI	Social	UK	Civil Society	Fact-checking	Organisation providing fact-checking, tools and media literacy activities in Northern Ireland

FullFact	Social	UK	Non-profit	Journalistic Standards	Full fact is a UK charity fact-checking information online and providing a toolkit to fight disinformation. It coooperates on a regular basis with FirstDraft and Facebook. FullFact also provides a toolkit to spot misinformation.
Freedom House report	Legal	UK	Non-profit	Research	Monitor the freedom of the press in the world by creating a world map on the basis of numerous indices.
Global Council to Build Trust in Media and Fight Misinformation	Social	UK	Non-profit	Research	Trust in media and the fight against misinformation is a global problem that requires a global solution. This is addressed through: global trust and misinformation repository,global trust and misinformation network, global trust and misinformation innovation and a voice for the industry
Institute for Strategic Dialogue (ISD)	Tech	UK	Company	Research	ISD has been running election monitoring (Germany, Sweden) on social media to document narratives and foreign influence allegations.
LSE Truth, Trust, and Technology Commission	Social	UK	Academic	Research	The Commission deals with the crisis in public information.It searches to identify trends, policy and strategy opportunities which provide ideas for solving the pressing challenges caused by online misinformation.
Serelay	Tech	UK	Company	Technology Solution	Platform for verification of photos and videos.

This study maps and analyses current and future threats from online misinformation, alongside currently adopted socio-technical and legal approaches. The challenges of evaluating their effectiveness and practical adoption are also discussed. Drawing on and complementing existing literature, the study summarises and analyses the findings of relevant journalistic and scientific studies and policy reports in relation to detecting, containing and countering online disinformation and propaganda campaigns. It traces recent developments and trends and identifies significant new or emerging challenges. It also addresses potential policy implications for the EU of current socio-technical solutions.

QA-04-19-194-EN-N

This is a publication of the European Science-Media Hub
EPRS | European Parliamentary Research Service

ISBN 978-92-846-3945-8 | doi: 10.2861/368879 | QA-04-19-194-EN-N