OECD GOING DIGITAL TOOLKIT

POLICY NOTE

Disentangling untruths online: Creators, spreaders and how to stop them

OECD
BETTER POLICIES FOR BETTER LIVES

going digital

This Toolkit note was written by Molly Lesher, Hanna Pawelec and Arpitha Desai. It was reviewed by the Committee on Digital Economy Policy (CDEP), and it was declassified by the CDEP on 16 March 2022. The note was prepared for publication by the OECD Secretariat.

This Toolkit note is a contribution to the OECD Going Digital project, which aims to provide policy makers with the tools they need to help their economies and societies thrive in an increasingly digital and data-driven world.

For more information, visit *www.oecd.org/going-digital*.

#GoingDigital

Please cite this publication as:

Lesher, M., H. Pawelec and A. Desai (2022), "Disentangling untruths online: Creators, spreaders and how to stop them", *OECD Going Digital Toolkit Notes*, No. 23, OECD Publishing, Paris, *https://doi.org/10.1787/84b62df1-en*.

Note to Delegations:

This document is also available on O.N.E. under the reference code:

DSTI/CDEP(2021)19/FINAL.

Table of Contents

Figures

Disentangling untruths online: Creators, spreaders and how to stop them

While false rumours, inaccurate reporting, and conspiracy theories have existed for as long as there were people to create and spread them, the Internet has reshaped and amplified the ability to produce and perpetuate false and misleading content. Stopping the creators and spreaders of untruths online is essential to reducing political polarisation, building public trust in democratic institutions, improving public health, and more generally improving the well-being of people and society. This Going Digital Toolkit note discusses the importance of access to accurate information online and presents a novel typology of the different types of untruths that circulate on the Internet. It considers how untruths are spread online as well as the consequences, and it surveys the evidence base of false and misleading information online. It concludes by identifying approaches to fighting untruths online and mitigating their negative effects.

The global free flow of information was one of the main drivers of the early Internet movement. The pioneers of the Internet's architecture viewed an open, interconnected and decentralised Internet as a vehicle to bridge knowledge gaps worldwide and promote learning in disadvantaged communities, thus becoming a great information "leveller". Despite these idealistic beginnings, however, societies across the world are now confronted with the dystopian prospect that instead of being a facilitator of knowledge and information, the Internet has become a key conduit for spreading untruths with a speed and reach that is unprecedented (OECD, 2019[1]); (OECD, 2019[2]).

Untruths – particularly propaganda and disinformation – played an important role in the lead up to the Russian Federation's invasion of Ukraine, and they continue to circulate strongly in both countries (Barnes, 2022[3]). As the conflict unfolds, spreaders of untruths are employing co-ordinated efforts across multiple platforms, with social media such as Telegram and TikTok, mainstream media outlets, and propaganda-based websites all being used to sway the emotions and beliefs of ordinary citizens in the two countries and around the world (Frenkel, 2022[4]); (Scott, 2022[5]).

In the context of the COVID-19 pandemic, concerns about an "infodemic" – or information overload – have emerged. Some of this information is false and inaccurate, and it is being spread online (WHO et. al., 2020[6]); (OECD, 2020[7]). Such information is largely about the COVID-19 virus, its origins and effects, cures and remedies, and actions taken by the government or public health officials to manage the pandemic. More recently, untrue information relating to the safety of vaccines has caused vaccine hesitancy across the world, undermining the efforts of governments towards nation-wide inoculation (Jongh, Rofagha and Petrosova, 2021[8]); (OECD, 2021[9]).

Similarly, false information has prevented people across the globe from accessing accurate and truthful information about elections, which has in turn adversely affected democratic processes and institutions (Colomina, Margalef and Youngs, 2021[10]). Some countries have experienced foreign influence on elections in the form of disinformation campaigns that have increased voter fraud and suppression, and reduced trust in the legitimacy of elections (Taylor, 2019[11]). In other countries, "influence firms" have been found to illegally harvest personal data and profile users of online platforms for the purpose of delivering targeted political content to such users (Rosenberg, Confessore and Cadwalladr, 2018[12]).

Throughout history, false rumours, incorrect reporting, and conspiracy theories have existed. The harm caused by such untruths has varied from being innocuous to causing severe mental and physical harm (Ireton and Posetti, 2018[13]). Recent revelations, including document disclosures, have only intensified the impression that while the dissemination of falsehoods is not new, the Internet has reshaped and amplified the ability to create and

perpetuate content in ways that we are only just beginning to understand (Wall Street Journal, 2021[14]); (Politico, 2021[15]). Such inaccurate and misleading information can intensify social polarisation, erode public trust in democratic institutions, and harm people and society more broadly.

There is support from people across age groups and levels of education that action is needed to tackle inaccurate information online (Mitchell and Walker, 2021[16]); (Henry, 2021[17]). Technology firms themselves have also called for more oversight and regulation (Clegg, 2021[18]); (Schaake, 2021[19]). This Going Digital Toolkit note discusses the importance of access to accurate information online and disentangles the different types that circulate. It also considers how untruths are spread online as well as the consequences, and it surveys the evidence base of false and misleading information online. It concludes by identifying approaches to fighting untruths online and mitigating their negative effects.

Why is access to accurate information important?

The right to freedom of speech, thought and expression, coupled with a free and independent press, are indispensable for the healthy functioning of democratic societies. The concerns surrounding the quality and accuracy of information available through the press or on online platforms presents a challenge for the protection of fundamental human rights enjoyed by all individuals under international, regional, and national legal frameworks, including the International Covenant on Civil and Political Rights (ICCPR) and the Universal Declaration of Human Rights (UDHR). These rights include not only the right to freedom of thought, speech, and expression that is necessary for interacting within the public sphere, but also the right to health, the right to privacy, and the right to access and receive reliable information that allows for public participation in democratic processes, one of the cornerstones of the OECD's acquis on open government[1]. However, the propagation of untruths endangers these human rights, reduces trust in the media, and undermines democratic norms, national security and public order.

Article 21 of the UDHR grants citizens the right to choose their leaders in free, fair, and regular elections as well as the right to access accurate information about parties, candidates and other factors that may influence voting. The United Nations Human Rights Committee also imposes an obligation on Member States to ensure that "voters should be able to form opinions independently, free of violence or threat of violence, compulsion, inducement or manipulative interference of any kind" (UNCHR, 1996[20]). However, surveys

[1] The OECD Recommendation of the Council on Open Government defines open government as "a culture of governance that promotes the principles of transparency, integrity, accountability and stakeholder participation in support of democracy and inclusive growth" (OECD, 2017[88]).

suggest that political untruths negatively impact a country's politics, causing polarisation among communities, and also sow distrust in democratic institutions such as governments, parliaments, and courts as well as distrust of public figures, journalists and the media (CIGI-IPSOS, 2019[21]); (Green, 2020[22]).

In the context of the COVID-19 pandemic, which has significantly increased our reliance on technology and the Internet, health-related untruths have caused issues for public health systems worldwide (OECD, 2020[23]). Trust in the health information disseminated by entities such as the media, governmental bodies, and health professionals is essential, especially in pandemic times (Swire-Thompson and Lazer, 2020[24]). However, some users of online platforms – including elected representatives (Lerer, 2021[25]) – have taken to the Internet to spread misinformation and disinformation related to the global pandemic, thereby jeopardising our collective right to health.

Importantly, the right to freedom of thought, speech, and expression, which is the cornerstone of free democracies and protected under Article 19(2) of the ICCPR, is threatened by inaccurate and misleading information that interferes with people's ability to exercise socio-political and economic choices (OHCHR, 2001[26]). In the digital context, online platforms have become the arbiters of communication, where they balance free speech and, at the same time, require users to adhere to terms of service that can potentially limit speech (Heins, 2013[27]).

Recently, some examples of untruths have also unfairly interfered with the right to privacy and data protection of users of online platforms. Through content distribution techniques (e.g. "micro-targeting"), as well as algorithmic bias that delivers specific content to users based on their personal data, some individuals and entities have leveraged technology to spread falsehoods. Well-established systems of data collection contribute to this phenomenon and, in some cases, intrude on people's right to privacy and their right to form their ideas free from manipulation. The United Nations Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression highlights in its latest report on disinformation that sensitive personal data or information such as one's ethnicity or political affiliation could potentially be dangerous when untruths online fuel political violence (Khan, 2021[28]).

Beyond fundamental rights, there are other reasons that access to accurate information is important. Key issues in this respect include information about climate change, including its causes and impacts, non-COVID-19 health issues (such as the dangers of smoking tobacco), as well as conspiracy theories (e.g. related to cults or emotional events such as the origins of the 9/11 attacks) and hoaxes of different types.

Disentangling the different types of untruths online

Given that there is no generally recognised typology of untrue content online, this Toolkit note surveys the literature to propose a coherent set of definitions to bring clarity to the international debate around untruths online. False, inaccurate, and misleading information often assumes different forms based on the context, source, intent and purpose. It is critical to distinguish between the various types of untrue information to help policymakers design well-targeted policies and facilitate measurement efforts to improve the evidence base in this important area.

- ***Disinformation*** refers to verifiably false or misleading information that is knowingly and intentionally created and shared for economic gain or to deliberately deceive, manipulate or inflict harm on a person, social group, organisation or country (EC, 2019[29]). Fake news[2], synthetic media, including deepfakes,[3] and hoaxes are forms of disinformation, among others.
- ***Misinformation*** refers to false or misleading information that is shared unknowingly and is not intended to deliberately deceive, manipulate or inflict harm on a person, social group, organisation or country (Ireton and Posetti, 2018[13]). Importantly, the spreader does not create or fabricate the initial misinformation content.
- ***Contextual deception*** refers to the use of true but not necessarily related information to frame an event, issue or individual (e.g. a headline that does not match the corresponding article), or the misrepresentation of facts to support one's narrative (e.g. to deliberately delete information that is essential context to understanding the original meaning). While the facts used are true (unlike disinformation) and unfabricated (unlike misinformation), the way in which they are used is disingenuous and with the intent to manipulate people or cause harm.
- ***Propaganda***[4] refers to the activity or content adopted and propagated by governments, private firms, non-profits, and individuals to manage collective attitudes, values, narratives, and opinions (EAVI, 2017[30]).

[2] ***Fake news*** refers to false information that is "purposefully crafted, sensational, emotionally charged, misleading or totally fabricated information that mimics the form of mainstream news" (Zimdars and McLeod, 2020[89]). Fake news can be wholly fabricated or a mix of fact and fiction.

[3] ***Deepfakes*** are synthetic media applications (e.g. videos or sound recordings) that alter a person's appearance or voice in an attempt to deceive viewers or listeners that what they are seeing or hearing is real (Somers, 2020[86]). Like fake news, deepfakes can be a mixture of real and unreal elements or completely fabricated.

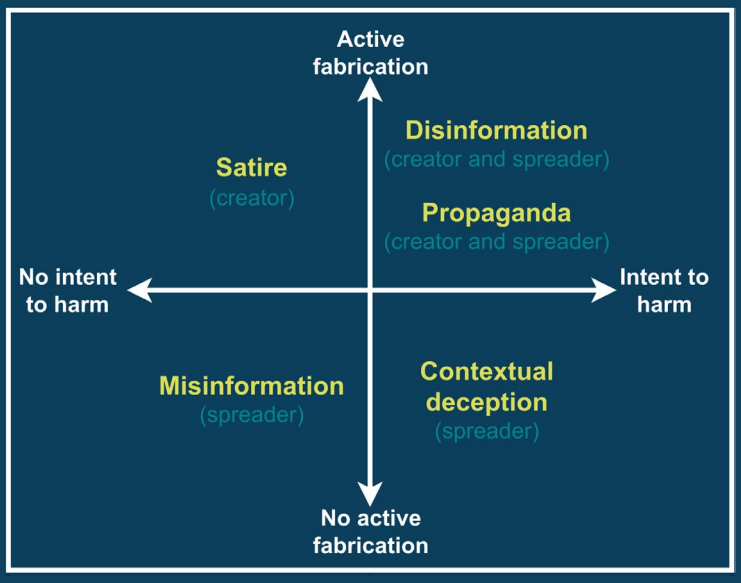
[4] Hate speech and terrorist and violent extremist content (TVEC) would also be considered propaganda in a broader sense, but it is not considered in the context of this note which deals exclusively with content that contains at least one untrue element or which is misleading.

While propaganda can contain both true and untrue elements, it is often used to appeal to an individual's or social group's sentiments and emotions rather than being informative (Neale, 1977[31]).

- ***Satire*** is defined as language, film or other works of art that use humour and exaggeration to critique people or ideas, often as a form of social or political commentary (OED, 2013[32]). Satire is an important form of social and political criticism, using humour and wit to draw attention to issues in society, and when satire is first published, the viewer often recognises the content as satire in part because of where and how they view it (e.g. directly from a satirical newspaper). However, as the content is shared and re-shared, this connection is sometimes lost intentionally (or not) by the spreader, leading new viewers to misunderstand the original meaning (Wardle, 2019[33]).

These definitions support a typology of false and misleading content that can be differentiated along two axes: 1) The intent (or not) of the information disseminator (spreader) to cause harm and 2) the degree of fabrication (if any) by the creator of the information content (e.g. altering photos, writing untrue article, creating synthetic videos) (Figure 1). This distinction is important insofar that some types of false and misleading content are not deliberately created with a view to deceive (e.g. satire) or they are not intentionally spread with a view to inflict harm (e.g. misinformation).

Figure 1. A typology of untruths online



Source: Authors.

From a policy perspective, it is important to differentiate the untrue content creators from the spreaders. There may be policies better suited to addressing

false content creators than spreaders, particularly those that disseminate falsehoods unknowingly (misinformation) or as part of societal or political commentary (satire). Before policy options can be devised, however, it is important to consider the context in and modalities by which false information diffuses online.

How are untruths spread online and what are the consequences?

The issues around inaccurate and misleading information online have emerged in tandem with the Internet's rise as a major news source. The share of people in the European Union (EU) who read online newspapers and news magazines nearly doubled in 10 years, with 65% of individuals aged 16 to 75 consuming news online in 2020 (Shearer, 2021[34]). Likewise, 86% of adult Americans access news on a digital device and it is the preferred medium for half of Americans (Shearer and Mitchell, 2021[35]). The Internet is also an important source of health advice. In 2020, 49% of men and 60% of women aged 16-75 sought health information online in the EU.

Mainstream and traditional media are one source of untruths, and once such content is disseminated, even if articles and posts are amended, the damage has already been done in many cases (Moschella, 2022[36]). Research shows that national figures, including politicians, are also a source of false and inaccurate information, and when such views are reported by mainstream news, some of this content circulates on social media without important context (Newman et al., 2021[37]). However, untruths online are primarily shared by individuals or organisations via online platforms such as social media platforms, private messaging services, and search engines.

A particular characteristic of the digital age is that false information can be more easily spread with digital technologies that were created for entirely different purposes (i.e. to increase user engagement, monitor user interaction, and deliver curated content without the aim or ambition of spreading inaccurate content) (Ávila, Ortiz Freuler and Fagan, 2018[38]). The digital technologies used to curate content are typically driven by algorithms or adopt AI-based approaches, making it sometimes difficult to track the source of misleading information, monitor its flow, and limit access to or block such content (Ávila, Ortiz Freuler and Fagan, 2018[38]). It also makes transparency about how these technologies work critical. Sophisticated disinformation attacks use bots, trolls, and cyborgs that are specifically aimed at the rapid dissemination of untruths (Paavola et al., 2016[39]).

Researchers from MIT Sloan showed that tweets containing false information were 70% more likely to be retweeted than the truth, and that false and misleading content reaches the first 1 500 people faster than true content

(Brown, 2020[40]). The authors also show that using bots is not necessarily required to spread untruths online; individuals are in fact more likely to share such content themselves. Another study from researchers at New York University and the Université Grenoble Alpes found that false information on Facebook attracts six times more engagement than factual posts (Edelson et al., 2021[41]).

The amplification of echo chambers and filter bubbles is a feature of the proliferation of untruths in the digital age. While such phenomena exist in an analogue world – for example, newspapers with a particular political inclination – it is easier and faster to spread information of any kind on the Internet. User-specific cookies log an individual's revealed preferences, and memberships in social networks and linkages to people or groups helps reinforce the type of content that is seen by individuals. When users consistently interact with or share specific content within their social networks that reinforces such beliefs, echo chambers that confirm existing biases emerge and grow (Karsten and West, 2016[42]). Recent document disclosures also reveal that targeting very small groups with falsehoods, in a tactic called "narrowcasting", has been successful in the spread of highly-viral untrue content (Wall Street Journal, 2021[14]).

Recent modelling work of the spread of misinformation over social media platforms suggests that filter bubbles can indeed help explain the spread of misinformation (Acemoglu, Ozdaglar and Siderius, 2021[43]). This research indicates that social media users are more likely to inspect articles that do not conform to their prior beliefs, and that a user will be more hesitant to share an article that does not conform to the views of those in his or her sharing network. According to this model, misinformation spreads when users decide to share an article without inspecting it, and users tend to share articles with others with similar beliefs (e.g. filter bubbles).

Regardless of the veracity of specific content, online platforms curate the news feed of users based on their past engagement with similar posts and preferences to engage with particular topics, thereby creating and facilitating communication within echo chambers and reinforcing filter bubbles. In such situations, it is not technically feasible to monitor content in real time and hold users liable. As a result, some online platforms have adopted transparency and accountability measures in the form of community codes of conduct or content guidelines that restrict or prohibit specific forms of speech such as hate speech, obscene content, and misinformation and disinformation. Such transparency and accountability measures are important and have been implemented at a global level, but thought should be given as to whether national initiatives may also add value (e.g. the DIGI code of conduct that applies to Australia and online platforms, see Annex). Such initiatives may be particularly useful to ensure that culture and language are appropriately taken into account.

Though intermediary liability laws allow for take down of content when such content is reported by users or ordered by the government or courts, creators and spreaders of disinformation are often not held liable for their actions. More recently, several countries have either enacted (e.g., Singapore, Malaysia) or proposed disinformation laws (e.g., the United States, EU, and Korea) that call for imposing criminal sanctions on agents of disinformation in addition to take down of the false content from online platforms (Yadav et al., 2021[44]).

Though the constitutional contours of the right to freedom of speech and expression vary across jurisdictions, human rights advocates have raised concerns regarding such laws on the basis that they unreasonably restrict free speech and disregard the normative considerations for restricting free speech, namely the principles of 1) legality; 2) necessity; and 3) proportionality. However, some academics have also argued that criminalisation of untrue content could potentially curb the spread of false information online owing to their deterrent effect (Helm and Nasu, 2021[45]). However, it is sometimes difficult to come to an agreed upon definition of "truth", a question that has perplexed philosophers since the times of Aristotle (Blackburn, 2020[46]). Events, topics, and beliefs are subject to a range of individual and idiosyncratic factors that can lead to different interpretations, creating challenges particularly for technology-driven solutions that may be less nuanced for addressing untruths online. When such judgements are enshrined in law, freedom of expression may be negatively impacted.

AI and big data analytics can be leveraged to fight untruths online to help identify and remove false content online. In addition to the enhanced accuracy with which AI can detect false information or recognise disinformation tactics deployed through bots and deepfakes (Marcellino et al., 2020[47]), AI solutions are more cost-effective because they reduce the time and human resources required for detecting and removing false content. However, at the same time, the effective use of AI for countering untruths online depends on large volumes of data as well as supervised learning without which such tools run the risk of false positives and human biases (Woolley, 2020[48]).

There is a need to measure and evaluate the extent to which falsehoods are circulating online, as well as assess people's susceptibility to encountering and engaging with false content online. This can help identify the root causes of untruths and help people, firms and governments to develop measures to prevent the spread of untruths and ensure the protection of fundamental rights and access to accurate information on other important issues (e.g. climate change).

Surveying the evidence base of untruths online

While public discourse and policymakers have shown increasing interest in fighting untruths online, measuring this phenomenon has made less progress. This is in part because there is not a mutually agreed upon definition of the range of phenomena used to describe untruths online (Figure 1). Moreover, for the most part data on this topic is held privately and it is not systematically gathered either by national statistical organisations or by other entities. However, there is some data that can be pieced together to begin to show the extent of untruths online.

Individuals' perceptions of exposure to false and misleading content online

Social media users often share false content unintentionally because they believe it, which is why the problem of inaccurate and misleading information is so widespread. A CIGI–IPSOS poll (CIGI-Ipsos, 2019[49]) that surveyed 25 000 respondents in over 25 economies found that 86% of people around the world reported that they have been exposed to fake news, and 86% of them initially believed the false news at least once. As disinformation usually relies on highly emotional content that provokes shock or anger, those false social media posts draw more attention.

According to a Eurobarometer survey, 37% of EU respondents reported that they were exposed to fake news every day or almost every day, while four in five respondents indicated that they were exposed at least several times a month. Moreover, 85% perceived fake news as a problem in their country. In 11 emerging economies, research suggests that between 44% and 78% of social media and messaging app users indicated that they see occasionally or frequently articles or content that seems obviously false or untrue (Silver, 2019[50]). However, such surveys are inherently conceived with self-reporting bias and, in the context of "fake news", they measure more respondents awareness and perception rather than the real extent of the misinformation problem.

In a 2018 survey by Pew Research, more Americans reported that made-up news is a bigger problem than climate change or racism (Pew Research Center, 2019[51]). Most Americans indicated that they had come across inaccurate news, and over one-third said that their preferred news source reported "made-up information intended to mislead the public". At the end of April 2020, 63% of American adults indicated that they had seen some or a lot of news about COVID-19 that seemed entirely made-up (Pew Research Center, 2019[51]). A survey conducted in May 2020 among Japanese daily Internet users suggested that 72% of respondents reported that they saw or heard at least one piece of false or misleading content about COVID-19, and almost 36% had shared such

information with others (Japanese Ministry of Internal Affairs and Communications, 2020[52]).

In New Zealand, survey results of 2 301 people between February and March 2021 show that 75% of respondents considered misinformation as an "urgent and serious threat to New Zealand society", and they indicated that the Internet was an important vehicle for disseminating misinformation (Talbot and Nusiebah, 2021[53]). Over 80% indicated that misinformation is becoming more prevalent, and almost 60% reported exposure to misinformation in the past six months. The report also estimates that half of all Kiwis held at least one belief associated with misinformation, with as many as 19% of respondents holding three or more such beliefs.

Trends in untruths related to elections and democratic processes and institutions

Much media attention on inaccurate information online has focused on content related to elections and democratic processes or institutions. Researchers from the University of Oxford have been monitoring governments and political party actors engaging in manipulation of public opinion on social media annually between 2015 and 2020 (Bradshaw, Howard and Bailey, 2021[54]). In 2020, they found evidence of the use of social medial for political disinformation and propaganda in 81 countries, up from 70 in 2019. While almost all countries rely on accounts managed by humans, automated bots were also used in 57 countries. Since 2018, they also identified more than 65 "influence firms" providing political communication services (sometimes called "computational propaganda") to State actors , and the activities of these firms spread from 25 countries in 2019 to 48 countries in 2020 (Bradshaw, Howard and Bailey, 2021[54]). Given the extent and increasing reach of such content, it is perhaps unsurprising that in a 2018 Flash Eurobarometer Survey, 83% of the EU respondents agreed that fake news is a problem for democracy in general (EC, 2018[55]).

Some co-ordinated operations to manipulate public opinion have been taken down by Facebook and Twitter. Between 2017 and July 2021, Facebook (Facebook, 2021[56]) removed 180 networks engaged in influence operations or what the company calls "coordinated inauthentic behaviour". Overall, more than 60 thousand assets (e.g. Facebook accounts, pages and groups as well as Instagram accounts) were deleted. Around half of the networks were engaged in domestic interference, one-third in foreign interference, and the remaining groups did both. While identifying the country origin of false and inaccurate content can be challenging (i.e. spreaders may use a VPN to conceal their location), between 2017 and 2020 Facebook reported that the largest number of networks originated in the Russian Federation (27), followed by Iran (23), Myanmar (9), the United States, (9) and Ukraine (8). Foreign influence

operations targeted most often the United States (26), Ukraine (11) and the United Kingdom (11).

Researchers followed a sample of Internet users in the United States with their prior consent to understand their online behaviour and exposure to disinformation (Guess, Nyhan and Reifler, 2018[57]). They analysed web traffic of 2 525 Americans during the weeks preceding the 2016 United States presidential election and estimated that 27.4% of adult Americans visited an article from an unreliable news site in that period. However, the articles containing inaccurate or misleading content represented only 2.6% of all articles read on news websites focusing on national and global politics.

Trends in inaccurate content related to public health

The outbreak of the COVID-19 pandemic brought about a surge of false and misleading information worldwide. The Center for Countering Digital Hate (CCDH), a non-profit organisation, focuses on spreaders of disinformation in the context of the COVID-19 pandemic. Through their analysis of over 800 000 pieces of anti-vaccine content from Facebook and Twitter between February and March 2021, they identified 12 people responsible for 65% of all identified anti-vaccine posts (Center for Countering Digital Hate, 2021[58]). Following the publication of this report, Facebook closed some of the accounts linked to this group, which resulted in a loss of 5.8 million followers out of their total of 14.2 million followers (O'Sullivan, 2021[59]).

In total, the CCDH tracked 425 anti-vaccine accounts with 59.2 million followers across the platforms. While the report was criticised by some for overestimating the importance of these super-spreaders by not taking into account anti-vaccine accounts that were already removed by the company, another study confirmed that most interactions about COVID-19 were generated by a very small group of users. Indeed, researchers from Italy analysed 200 million interactions on Twitter related to the pandemic and concluded that 0.1% of users account for up to 45% of activities and 10% of the news that is shared (Sacco et al., 2021[60]).

Another study on misinformation analysed tweets with hashtags related to COVID-19 posted before March 2020 (Kouzy et al., 2020[61]). Out of the 673 tweets identified, about 25% contained misinformation and another 17% contained information that could not be verified. Another study showed that exposure to anti-vaccine misinformation decreased the share of people who indicated that they would definitely get a COVID-19 vaccine by 6.2 percentage points in the United Kingdom and by 6.4 percentage points in United States compared to the group that was exposed to factual information (Loomba et al., 2021[62]).

Before the COVID-19 pandemic, social media companies were often hesitant to moderate content posted on their websites, in part because they did not want

to limit free speech. But with the onset of the COVID-19 pandemic and the ensuing disinformation that followed, in March 2020 a group of seven platforms[5] proactively published a joint statement in which they committed to combat fraud and misinformation about the COVID-19 virus (Statt, 2020[63]). These platforms are also signatories of the EU Code of Practice on Disinformation and the Australian Code of Practice on Disinformation and Misinformation through which they commit to share progress reports.

Since the beginning of the pandemic and until the end of June 2021, 20 million pieces of content as well as 3 000 accounts, pages, and groups were removed from Facebook and Instagram for violating COVID-19 policies (Rosen, 2021[64]). In June 2021, deleted posts from the EU represented 11% of all deleted content, which is slightly more than in the previous months. Between March and December 2020, 110 000 pieces of content were removed in Australia which represents around 0.8% of all content taken down during that period (Facebook, 2021[65]). Since the start of the COVID-19 pandemic, Twitter reported suspending around 2 000 accounts, and it removed 50 000 pieces of content between January 2020 and August 2021 (Twitter, 2021[66]). To put the latter number in context, there are currently close to 10 000 tweets shared every second or around 25 billion a month, according to an estimate from Internet Live Stats (Internet Live Stats, 2021[67]) and in line with the figure shared by Twitter in 2013 (Twitter, 2013[68]).

Language plays an important role in content moderation

Fact-checking organisations have proliferated around the world and they emerged as one of the key actors during the COVID-19 pandemic. Social media platforms and researchers collaborate with fact-checking organisations to identify untruths online. Facebook marks the content that they find false or inaccurate with special warnings and downgrade it in the recommendation algorithms. So far, 190 million posts on COVID were labelled with a warning. For context, while Facebook does not share the global number of new posts, the platform had over 1.9 billion daily active users in June 2021.

In April 2020, an analysis of over 100 pieces of misinformation content that were flagged by fact-checkers to assess the effectiveness of Facebook's methods (AVAAZ, 2020[69]). They found that 68-70% of content in Italian and Spanish were not labelled with a warning, while the problem concerned only 29% of English-speaking content. Likewise, in a June 2021 report Twitter indicated that their machine-learning model that will be used to identify content violating COVID-19 policies will be trained on English-language content first (EC, 2021[70]). Other languages will only follow later.

[5] Facebook, Google, LinkedIn, Microsoft, Reddit, Twitter, and YouTube.

A moderation of two viral videos shows how non-English language content is treated differently than other languages. A 26-minute conspiracy video entitled "Plandemic the Hidden Agenda behind Covid-19" was published on various platforms on 4 May 2020. According to Digital Trends, it reached 1.8 million views just on Facebook before being quickly deleted across platforms (Gebel, 2021[71]). In November 2020, another conspiracy video, "Hold up", was released in French. It reached three million views in five days (Kayali, 2021[72]). As reported by Politico Europe and the EU DisinfoLab, six months later the video was still available on Facebook and YouTube. While it was initially removed, the video was later republished. Users who would look for "Plandemic documentary" were led to a COVID-19 information centre, while users who would look for "Hold up documentary" would find new uploads of the video, including one that reached over one million views on YouTube.

Going forward, the development of tools in multiple languages will be an important part of the fight against untruths online. For example, the Spanish-language Chequeabot tool (see Annex), which was developed by professional fact-checkers, has been helping to fight untruths in Argentina. Other approaches include co-operative arrangements between fact-checking entities and media partners (e.g. FactCheck Initiative Japan, see Annex) to support the identification of untruths online. Another innovative approach – crowdsourced fact-checking (e.g. the Birdwatch programme, see Annex) – could also support the moderation of non-English content in the future.

Approaches to fighting untruths online and mitigating their negative effects

Tackling untrue content online requires a multistakeholder approach where people, firms, and governments all play an active role in identifying and removing inaccurate content on the Internet, and all actors exercise judgment before sharing information online. It is likewise important to promote transparency, and to create an enabling environment for an independent, diverse, local and public service media to thrive, and to empower public communicators to help in the fight against untruths online (OECD, 2021[73]). Given the global reach of online platforms, a global approach is needed at least among "like minded" countries, although this will not be easy due to differences in culture, history and legal frameworks.

A better understanding of a range of complex and intertwined issues about untruths online is urgent to develop "best practice" policies to address this important problem, but in the meantime, concrete steps can be taken to begin the fight. This note argues that five broad steps can help fight untruths online:

1. Promote digital literacy initiatives,
2. Develop content moderation policies in a multistakeholder process and with independent oversight,
3. Integrate humans and technology in the fight against untruths online,
4. Increase transparency in spending on political advertisements online, and
5. Design a measurement agenda to improve the evidence base and inform more targeted policies to stop the creators and spreaders of untruths.

While none of these steps by themselves can effectively stop the spread of inaccurate and misleading content online, in combination they can go a long way toward protecting fundamental and other rights and to mitigating the negative effects from such content.

Create awareness about untruths online by promoting digital media literacy

People are generally not very good at identifying false and inaccurate information, with research showing that on average people are able to identify 47% of lies as deceptive and 61% of truths as non-deceptive (Bond and DePaulo, 2006[74]). As a result, an important way to stop untruths online is to promote digital media literacy among adults and children, including in schools (Khan, 2021[28]). It is also a very practical approach. Instead of trying to chase each piece of untrue content, which is impossible, digital literacy initiatives protect people by giving them the tools to distinguish false and misleading information and to disregard or ignore it.

Initiatives that have been implemented in countries worldwide by governments, schools, universities, online platforms, and non-profits help individuals to better assess and verify the accuracy of information online. Digital media literacy initiatives tend to focus on developing cognitive, critical, and technical skills that help discern fact from fiction, and enable meaningful participation in public interactions, discussions and debates (e.g. the Bad News and Go Viral! games, see Annex). Online platforms including Facebook, Google, and Twitter have launched extensive digital media literacy initiatives in partnership with international organisations, governments, and fact-checking organisations to raise awareness and educate people about how to spot potential false and misleading content.

While these initiatives provide resources to those accessing the Internet, scaling such initiatives to ensure widespread training across diverse demographics has been challenging. Such initiatives typically engage with only a tiny fraction of the population, such as politicians, journalists and school teachers that are concentrated in large cities (ERGA, 2020[75]). For example, research conducted

in the United Kingdom indicates that digital literacy inequalities correspond with other key elements of economic, social, and cultural inequality (Helsper, 2016[76]), with people of different ages possessing varied levels of digital media literacy. Efforts towards imparting digital media literacy would benefit from adopting a broader and more inclusive approach.

Develop and implement online platform content moderation policies in a multistakeholder process and with independent oversight

In light of the COVID-19 pandemic as well as concerns around election integrity in some countries, several online platforms have revised and expanded existing content moderation policies to include false and misleading content. While such activities move in the right direction, some online platforms implement content moderation policies that have been developed without public input and enforced with limited clarity (Kaye, 2018[77]). By not engaging in a multistakeholder process, these actions may raise the risk that such policies are not compliant with existing laws, including on free speech, notably if policies do not require independent oversight or transparency in the decisions leading up to the take down of problematic content.

Towards this end, the EU's forthcoming Digital Services Act calls on online platforms to be more transparent and accountable in their content moderation decisions through periodic reporting obligations. In its recent Joint Communication on Disinformation, the European Commission announced that online platforms that are required to comply with the Code of Practice on Misinformation would need to provide monthly reports on how they are dealing with misinformation.

While online platforms make information relating to takedown decisions public to enhance transparency, untruths may still go unchecked with problematic content remaining on platforms despite being declared untrue. Creating content moderation practices involving local stakeholders, including fact-checking organisations and researchers, and setting up independent audits of content moderation decisions could help make take-down decisions more consistent and further improve online content moderation at large.

For example, Facebook's Oversight Board, a governing body comprised of members from a variety of cultural and professional backgrounds, reviews content moderation decisions taken by Facebook. The Oversight Board's aim is to improve fairness and transparency around content and provide oversight and accountability (Oversight Board, 2021[78]). The Oversight Board recently rebuked Facebook for not being more forthcoming about how it exempts high-profile users from its rules (the "cross check" program), and said that it is drafting recommendations for how to overhaul the system (Schechner, 2021[79]). Oversight boards could potentially serve as a model for online

platforms, if the Board is truly independent, it has thorough access to information, and independent audits with well-trained auditors are carried out.

Integrate humans and technology in the fight against untruths online

Existing approaches to reducing untruths online are often dependent on manual fact-checking, content moderation and takedown, and quick responses to attacks that involve human intervention and allow for a finer-grained assessment of the degree of the accuracy of content. For example, PolitiFact's "Truth-o-Meter" includes six ratings[6] to assess the degree of veracity. Collaborations between independent, domestic fact-checking entities and platforms can further help identify untruths (e.g. DIGI in Australia and FactCheck Initiative Japan, see Annex) and can also be useful to ensure that cultural and linguistical considerations are taken into account.

While human understanding is essential to interpreting specific content in the context of cultural sensitivities and belief or value systems, monitoring online content in real time is a mammoth task that may not be feasible entirely without technological assistance. Automation of certain content moderation functions and developing technologies that embed such functions "by design" could considerably enhance the efficacy of techniques used to prevent the spread of untruths online, although such approaches often provide less nuance on the degree of accuracy of content (i.e. content is usually identified as either "true" or "false").

Such approaches would also benefit from partnerships between local fact-checking entities and online platforms to ensure cultural and linguistical biases are addressed[7]. Advanced technologies such as automated fact checking (Dulhanty et al., 2019[80]) or natural language processing and data mining (Rahman, Chia and Gonzalez, 2021[81]) (Wang et al., 2018[82]) can be leveraged to detect producers of inaccurate information and prevent sophisticated disinformation attacks, although the spreaders of untruths have found ways to circumvent such approaches (e.g. through the use of images rather than words). In this regard, transparent use of digital technologies by online platforms to identify and remove untrue content can improve the dissemination of accurate information.

At the same time, the technical limitations of AI and other technologies (e.g. potential bias) point towards the need to adopt a hybrid approach that integrates both human intervention and technological tools in fighting untruths online. In such an approach, digital tools can help to monitor and detect

[6] 1) True, 2) mostly true, 3) half true, 4) mostly false, 5) false and 6) pants-on-fire.

[7] Algorithms trained mainly on American English have been shown to underperform on content using British English (Waterson and Milmo, 2021[87]).

inaccurate information online, and human expertise and value judgment can be used to determine the extent to which untruths are likely to harm the public. Approaches that marry technology-oriented solutions and human judgement may be best suited to ensure both efficient identification of problematic content and potential takedown after careful human deliberation, taking into account all relevant principles, rights and laws such as on free speech.

Increase transparency in spending on political advertisements online

With political parties and candidates spending large amounts of money on paid political advertising and content through different channels, online platforms have become a vehicle for disseminating untrue and misleading content. Such untruths often mislead voters and compromise electoral outcomes, undermining democratic elections and civic processes. In an effort to address this issue, online platforms now sometimes publish periodic transparency reports that disclose the identities of political advertisers as well as the amounts spent on such advertisements or content. However, such laws do not require disclosures by advertising agencies, consultancies or political organisations that spend money towards political advertisements and content on behalf of political parties and candidates.

To ensure enhanced transparency in online political advertising, including campaign spending, mandating political parties to disclose monies spent towards paid digital advertisements and content on a regular basis could potentially mitigate the harms caused by political disinformation and incentivise political parties and candidates to publish accurate and truthful information (Dunčikaitė, Žemgulytė and Valladares, 2021[83]). Another step in this direction that has been adopted by the EU is to restrict political advertisements and content to "issue-based advertising" that focuses on clear distinguishability from editorial content (EC, 2019[29]). Such issue-based advertising allows voters to make more informed decisions based on their own judgment and reasoning.

Design a measurement agenda to improve the evidence base

Without a solid evidence base, it is difficult to develop well-targeted policies for fighting untruths online. Ideally, indicators along the following dimensions would help shed important light on the scale, content and reach of untruths online:

- ***Who*** (age, gender, language, education and income levels) is spreading false information online?
- ***What*** types of false information online (e.g. health, elections, conspiracy theories) are most prevalent?

- ***Where*** does false information originate from and where is it disseminated?
- ***Why*** do people spread falsehoods online (i.e. what are their aims and ambitions, or is it accidental)?
- ***How*** and through which vehicles are untruths spread in the digital age?

While data on the dimensions noted above will be challenging to come by, not least because much of the data needed is proprietary and dispersed among private firms, it is nonetheless important that a co-ordinated measurement agenda be developed and implemented in a cross-country comparable manner and in partnership with the private sector.

Conclusion

Overall, while digital technologies are in and of themselves neutral, they can nonetheless be used intentionally and unintentionally to spread inaccurate and misleading information, thereby impacting individuals, social groups, and society in a variety of ways. While responses to untruths by individuals, governments, and firms vary across countries (see Annex), a multistakeholder approach is needed to reduce the spread of untruths online. Information producers, users of online platforms, and online platforms themselves all have an important role to play in stopping the creators and spreaders of untruths online and ensuring transparency and accountability.

Annex. A Selection of innovative approaches to fighting untruths online and mitigating their negative effects

Create awareness about untruths online by improving digital media literacy

Bad News and Go Viral! games

Responsible entity: University of Cambridge's Social Decision-Making Lab in collaboration with the UK Cabinet Office

Description: The Bad News and Go Viral! Games were developed in response to research from the University of Cambridge that found that educating people on the techniques used to spread false and misleading content on social media increases their ability to identify and disregard similar content in the future. Bad News was launched in 2018 and it has been played over one million times; Go Viral! is a shorter game that was launched in 2020. It focusses on falsehoods related to COVID-19 and it is easier to adapt for different languages and cultures; there are currently French and German versions.

Read more: *https://www.goviralgame.com*; *https://www.getbadnews.com*; *https://www.cam.ac.uk/stories/goviral*.

Be Internet Awesome initiative

Responsible entity: Google

Description: Google's media literacy initiative, Be Internet Awesome (BIT), aims to teach children how to spot untruths online and be safe, confident explorers of the online world. BIT is a multifaceted programme that includes an interactive, web-based game 'Interland', and an educational curriculum to teach children about digital safety.

Read more: *https://beinternetawesome.withgoogle.com/en_us*.

Check the Facts campaign

Responsible entity: Australian Associated Press (AAP) with the support of Facebook

Description: The AAP's "Check the Facts" digital literacy campaign aims to raise awareness about how to recognise reliable information based on techniques used in professional fact checking organisations. It includes videos and other resources to help spot untrue content by considering the source, whether the source is trustworthy, and the specific claims that are being made.

Read more: *https://www.aap.com.au/factcheck-resources/*.

Media and digital literacy programmes in Finnish schools

Responsible entity: National Audio-Visual Institute (NAVI) and the Finnish Ministry of Education and Culture, in collaboration with schools media professionals and fact checking organisations

Description: In response to fake news campaigns focussing on immigration, EU, Finland and the North Atlantic Treaty Organisation (NATO) members recognised the need to strengthen the population's resilience to digital untruths, and instituted a cross-sector approach to improve media literacy within the country, with a focus on children. Media and digital literacy skills are embedded across Finland's national school curriculum and implemented by the NAVI and the Finnish Ministry of Education and Culture in collaboration with media education professionals as well as a non-profit fact checking organisation, Faktabari, that provides fact checking and media literacy materials for schools.

Read more: *https://medialukutaitosuomessa.fi/mediaeducationpolicy.pdf*.

Section 51206.4 of the California Education Code

Responsible entity: State Government of California, United States

Description: In 2018, the state of California passed a bill requiring the State's Department of Education website to list resources and instructional materials on media literacy, including professional development programmes for teachers. Specifically, the bill aims to empower students to distinguish advertisements from news stories and make informed decisions online. The legislative intent behind this law was based on a 2016 Stanford University study that indicated that 80% of middle school students did not recognise an advertisement that was masquerading as a news story despite being labelled as "sponsored content" (Breakstone et al., 2019[84]).

Read more:
https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB830.

Develop and implement online platform content moderation policies in a multistakeholder process and with independent oversight

Birdwatch

Responsible entity: Twitter

Description: Launched in early 2021, Twitter's pilot programme, Birdwatch, aims to combat misinformation by adding fact-checking notes written by crowd-sourced volunteers. Initially restricted to a group of enrolled contributors, the flags on potentially misleading Tweets are now shown to Twitter test users in the United States. Notes must be approved by contributors who have shown to have diverging views in the past and can be further evaluated by users. A recent study by MIT researchers showed that ratings from a small, politically balanced group of regular people correlated with professional fact-checkers (Allen et al., 2021[85]).

Read more: *https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation*; *https://blog.twitter.com/en_us/topics/company/2022/building-a-better-birdwatch*; https://www.science.org/doi/10.1126/sciadv.abf4393.

DIGI Code of Practice

Responsible Entity: Digital Industry Group Inc. (DIGI)

Description: On 22 February 2021, DIGI launched a code of practice that commits a diverse set of technology companies to reducing the risk of harm from online disinformation and misinformation in Australia. The Australian Code of Practice on Disinformation and Misinformation has been adopted by Adobe, Apple, Facebook, Google, Microsoft, Redbubble, TikTok, and Twitter. All signatories commit to protect Australians by providing appropriate safeguards against harm from online disinformation and misinformation, and to adopting a range of scalable measures that reduce its spread and visibility. Participating companies also commit to releasing an annual transparency report about their efforts under the code, which will help improve understanding of online misinformation and disinformation in Australia over time. The Code was developed in response to the Australian Government policy announced in December 2019, where the digital industry was asked to develop a voluntary code of practice on disinformation, drawing on learnings from a similar code in the EU.

Read more: *https://digi.org.au/wp-content/uploads/2021/10/Australian-Code-of-Practice-on-Disinformation-and-Misinformation-FINAL-WORD-UPDATED-OCTOBER-11-2021.pdf*.

FIJ Fact-Checking Guidelines

Responsible entity: FactCheck Initiative Japan (FIJ)

Description: FIJ is a Tokyo-based non-profit organisation for the promotion of Japanese fact checking aimed at protecting society from untruths online. FIJ supports and co-operates with media partners that publish fact-checking articles according to FIJ Fact-Checking Guidelines that are based on the IFCN Code of Principles. In April 2020, FIJ worked with Yahoo! Japan – one of the largest sources of mainstream news in Japan – to create an English-version of their website to reach more people with COVID-19 related fact-checked information.

Read more: *https://en.fij.info/about/*.

International Fact-Checking Network (IFCN)

Responsible entity: The Poytner Institute

Description: In 2015, the Poynter Institute – a non-profit organisation that promotes freedom of expression, civil dialogue and truthful journalism – established the IFCN to bring together the international community of fact checkers to tackle false and misleading information. At the beginning of 2022, the IFCN had 108 active signatories to its Code of Principles from over 50 countries. The IFCN also provides training programmes and resources (including some in Spanish) to help develop the skills to identify untrue and misleading content. Many online platforms are collaborating with the IFCN and fact checkers, for example by applying fact-checked labels.

Read more: *https://www.poynter.org/ifcn/*.

Network Enforcement Act

Responsible entity: German government

Description: The *Netzwerkdurchsetzungsgesetz,* or Network Enforcement Act (the Act), went into force in 2017 and was subsequently amended in June 2021. Its aim is to address illegal content that meets the criteria of the Criminal Code (e.g. hate speech). It requires online platforms to remove content found to be offensive or "clearly illegal" within 24 hours after receiving a user complaint. If the illegality of the content is not obvious, the online platform has seven days to investigate and delete it. The Act requires transparency reporting, and fines for non-compliance can be assessed up to a ceiling of 50 million euros. The Federal Office of Justice has the power to issue fines for noncompliance and to supervise compliance with the Act.

Read more: *https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf;jsessionid=D7F76C9A3ED468A2F3BEECBE17F330BA.2_cid324?__blob=publicationFile&v=2*.

Oversight Board

Responsible entity: Facebook

Description: Facebook's Oversight Board is a governing body comprised of members from a variety of cultural and professional backgrounds that reviews content moderation decisions taken by Facebook and Instagram. The Board aims to promote free expression by making principled, independent decisions by issuing recommendations on the relevant content policies of the online platforms.

Read more: *https://oversightboard.com/*.

Twitter's guidance on content related to COVID-19, elections and other civic processes

Responsible entity: Twitter

Description: In March 2020, Twitter expanded its content moderation policy to address content that goes against guidance from global and local public health authorities on COVID-19 protocols. This includes sharing content that may mislead people about the nature of the COVID-19 virus' efficacy and/or the safety of preventative measures, treatments, or other precautions to mitigate or treat the disease. Twitter's content moderation policy also addresses untrue content related to official regulations, restrictions, or exemptions pertaining to health advisories, the prevalence of the virus, or the risk of infection or death associated with COVID-19. Content that could potentially endanger public health is either labelled as demonstrably false or

misleading, or removed from the platform. Similarly, Twitter also prohibits sharing content that could potentially compromise or interfere with elections and civic processes. Such content includes false or misleading information in relation to procedures or circumstances around participation in a civic process or content that seeks to confuse or manipulate voters using the platform.

Read more: *https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information*; *https://about.twitter.com/en/our-priorities/civic-integrity*.

TikTok's Community Guidelines

Responsible entity: TikTok/Bytedance

Description: In response to global concern about the propagation of disinformation and misinformation, TikTok updated its policies on misleading content to provide further clarity on what is and is not allowed on TikTok. TikTok has also added a policy that prohibits synthetic or manipulated content (for example, deepfakes) that misleads users by distorting the truth of events in a way that could lead to real world harm. Specifically, TikTok's Community Guidelines prohibit the sharing content that could cause harm to users or the greater public, including content that misleads people about elections or other civic processes, content distributed by disinformation campaigns, and untrue health information.

Read more: *https://www.tiktok.com/community-guidelines?lang=en*.

Integrate humans and technology to fight untruths online

Chequeabot

Responsible Entity: Chequeado

Description: In January 2018, the Argentine foundation Chequeado released Chequeabot, a bot tool that incorporates natural language processing and machine learning to identify claims made in the media and matches them with existing fact checks. Chequeado co-ordinated the development of Chequeabot with both the IFCN and Full Fact, underscoring the tightly linked global fact-checking community. Chequeabot is notable in that it is in Spanish and it was developed by professional fact-checkers. While it is used for global issues (e.g. the conflict in Ukraine), it is nonetheless strongly focussed on issues important in Argentina (i.e. inflation and IMF programmes are a strong focus).

Read more: *https://chequeado.com/tag/chequeabot/*.

Full Fact's AI-based fact checking tools

Responsible entity: Full Fact and Google

Description: In 2020, Google provided the non-profit Full Fact with 2 million USD and seven technical experts from the Google.org Fellowship to help Full Fact build AI-based tools to help fact checkers verify claims made by key politicians, then group them by topic and match them with similar claims from across press, social networks and even radio using speech-to-text technology. These tools helped Full Fact process 1 000 times more content, detecting and clustering over 100 000 claims per day. Importantly, the tools gave Full Fact's fact checkers more time to verifying facts rather than identifying which facts to check. Using a machine learning BERT-based model, the technology now works in four languages (English, French, Portuguese and Spanish).

Read more:
https://blog.google/documents/37/How_Google_Fights_Disinformation.pdf; https://blog.google/outreach-initiatives/google-org/fullfact-and-google-fight-misinformation/.

Increase transparency in spending on political advertisements online

Action Plan against Disinformation

Responsible entity: European Commission together with EU Member States

Description: As part of the Action Plan against Disinformation, the European Commission has recommended to Member States to focus on promoting the transparency of online political advertising, including campaign expenditure, and it invited all political parties to respect transparency recommendations. On 19 October 2020, the European Commission presented its 2021 work programme, which included as one of its priorities 'A New Push for European Democracy'. Under this priority, the Commission announced its intention to issue a proposal on greater transparency in paid political advertising.

Read more: *https://ec.europa.eu/info/sites/default/files/eu-communication-disinformation-euco-05122018_en.pdf; https://ec.europa.eu/info/strategy/priorities-2019-2024/new-push-european-democracy_en*.

References

Acemoglu, D., A. Ozdaglar and J. Siderius (2021), "Misinformation: Strategic sharing, homophily, and endogenous echo chambers", NBER Working Paper No. 28884, http://dx.doi.org/10.3386/w28884. [43]

Allen, J. et al. (2021), "Scaling up fact-checking using the wisdom of crowds", Science Advances, https://doi.org/10.1126/sciadv.abf4393 (accessed on 22 March 2022). [85]

AVAAZ (2020), "How Facebook can flatten the curve of the coronavirus infodemic", https://secure.avaaz.org/campaign/en/facebook_coronavirus_misinformation/ (accessed on 25 October 2021). [69]

Ávila, R., J. Ortiz Freuler and C. Fagan (2018), "The invisible curation of content", Web Foundation, http://webfoundation.org/docs/2018/04/WF_InvisibleCurationContent_Screen_AW.pdf. [38]

Barnes, J. (2022), "Russia Steps Up Propaganda War Amid Tensions With Ukraine", The New York Times, https://www.nytimes.com/2022/01/25/us/politics/russia-ukraine-propaganda-disinformation.html. [3]

Blackburn, S. (2020), "Truth", https://www.britannica.com/topic/truth-philosophy-and-logic. [46]

Bond, C. and B. DePaulo (2006), "Accuracy of deception judgments", Personality and Social Psychology Review, Vol. Vol. 10, No. 3, https://www.researchgate.net/publication/6927452_Accuracy_of_Deception_Judgments. [74]

Bradshaw, S., P. Howard and H. Bailey (2021), "Industrialized disinformation: 2020 global inventory of organized social media manipulation", Oxford Institute, https://demtech.oii.ox.ac.uk/research/posts/industrialized-disinformation/. [54]

Breakstone, J. et al. (2019), "Students' civic online reasoning: A national portrait", https://stacks.stanford.edu/file/gf151tb4868/Civic%20Online%20Reasoning%20National%20Portrait.pdf. [84]

Brown, S. (2020), "MIT Sloan Research About Social Media, Misinformation, and Elections", MIT, https://mitsloan.mit.edu/ideas-made-to-matter/mit-sloan-research-about-social-media-misinformation-and-elections (accessed on September 15 2021). [40]

Center for Countering Digital Hate (2021), "The disinformation dozen: Why platforms must act on twelve leading online anti-vaxxers", https://www.counterhate.com/disinformationdozen. [58]

CIGI-IPSOS (2019), "CIGI IPSOS Global Survey", Vol. 3, https://www.cigionline.org/sites/default/files/documents/2019%20CIGI-Ipsos%20Global%20Survey%20-%20Part%203%20Social%20Media%2C%20Fake%20News%20%26%20Algorithms.pdf (accessed on 5 August 2021). [21]

CIGI-Ipsos (2019), Fake News: A Global Epidemic, https://www.ipsos.com/en-us/news-polls/cigi-fake-news-global-epidemic (accessed on 15 September 2021). [49]

Clegg, N. (2021), "Facebook's Nick Clegg calls for bipartisan approach to break the deadlock on internet regulation", https://www.cnbc.com/2021/05/24/facebooks-nick-clegg-a-bipartisan-approach-to-break-the-deadlock-on-internet-regulation.html (accessed on 29 October 2021). [18]

Colomina, C., H. Margalef and R. Youngs (2021), "The impact of disinformation on democratic processes and human rights in the world", https://www.europarl.europa.eu/RegData/etudes/STUD/2021/653635/EXPO_STU(2021)653635_EN.pdf. [10]

Dulhanty, C. et al. (2019), "Taking a stance on fake news: Towards automatic assessment via deep bidirectional transformer language models for stance detection", https://arxiv.org/pdf/1911.11951.pdf (accessed on 12 August 2021). [80]

Dunčikaitė, I., D. Žemgulytė and J. Valladares (2021), "Paying for Views: Solving transparency and accountability risks in online political advertising", https://images.transparencycdn.org/images/2021_Report_PayingForViews-OnlinePoliticalAdvertising_English.pdf. [83]

EAVI (2017), "Beyond Fake News", https://eavi.eu/beyond-fake-news-10-types-misleading-info/. [30]

EC (2021), "Reports on June Actions - Fighting COVID-19 Disinformation Monitoring Programme", https://digital-strategy.ec.europa.eu/en/library/reports-june-actions-fighting-covid-19-disinformation-monitoring-programme. [70]

EC (2019), Code of Practice on Disinformation, European Commission, https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation. [29]

EC (2018), "Flash Eurobarometer 464 (Fake news and disinformation online)", https://data.europa.eu/data/datasets/s2183_464_eng?locale=en. [55]

Edelson, L. et al. (2021), "Understanding engagement with U.S. (mis)information news sources on Facebook", IMC '21: Proceedings of the 21st ACM Internet Measurement Conference, https://dl.acm.org/doi/10.1145/3487552.3487859. [41]

ERGA (2020), "Improving Media Literacy Campaigns on Disinformation", https://erga-online.eu/wp-content/uploads/2021/01/ERGA-SG2-Report-2020-Improving-Media-Literacy-campaigns-on-disinformation.pdf (accessed on 23 August 2021). [75]

Facebook (2021), "Coordinated inauthentic behavior", https://about.fb.com/news/tag/coordinated-inauthentic-behavior/ (accessed on 15 September 2021). [56]

Facebook (2021), "Facebook response to the Australian disinformation and misinformation industry code. May 2021", https://australia.fb.com/wp-content/uploads/sites/69/2021/05/Facebook-commitments-under-disinfo-and-misinfo-code-final-1.pdf (accessed on 23 September 2021). [65]

Frenkel, S. (2022), "TikTok Is Gripped by the Violence and Misinformation of Ukraine War", The New York Times, https://www.nytimes.com/2022/03/05/technology/tiktok-ukraine-misinformation.html. [4]

Gebel, M. (2021), "Facebook takes down viral 'Plandemic' coronavirus conspiracy video", Digital Trends, https://www.digitaltrends.com/news/facebook-will-take-down-viral-plandemic-coronavirus-conspiracy-video/ (accessed on 24 September 2021). [71]

Green, T. (2020), "Americans are confident in tech companies to prevent misuse of their platforms in the 2020 election", https://www.pewresearch.org/fact-tank/2020/09/09/few-americans-are-confident-in-tech-companies-to-prevent-misuse-of-their-platforms-in-the-2020-election/ (accessed on 5 August 2021). [22]

Guess, A., B. Nyhan and J. Reifler (2018), "Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign", European Research Council, Vol. 9/3, https://about.fb.com/wp-content/uploads/2018/01/fake-news-2016.pdf. [57]

Heins, M. (2013), "The brave new world of social media censorship", Harvard Law Review Forum, Vol. 127, pp. 325-330, https://harvardlawreview.org/2014/06/the-brave-new-world-of-social-media-censorship/. [27]

Helm, R. and H. Nasu (2021), "Regulatory responses to fake news and freedom of expression: Normative and empirical evaluation", Human Rights Law Review, Vol. 21/2, pp. 302-28, https://doi.org/10.1093/hrlr/ngaa060. [45]

Helsper, E. (2016), "The social relativity of digital exclusion: applying relative deprivation theory to digital inequalities", Communication Theory, Vol. 27/3, pp. 223-242, https://onlinelibrary.wiley.com/journal/14682885. [76]

Henry, D. (2021), "Fake news: Majority of Kiwis think spreading misinformation should be illegal", The New Zealand Herald, https://www.nzherald.co.nz/nz/fake-news-majority-of-kiwis-think-spreading-misinformation-should-be-illegal/WC3Q45YGOB2IXQI6QGNFS4TEOU/ (accessed on 23 September 2021). [17]

Internet Live Stats (2021), "Twitter usage statistics", https://www.internetlivestats.com/twitter-statistics/ (accessed on 26 October 2021). [67]

Ireton, C. and J. Posetti (2018), Journalism, fake news & disinformation: Handbook for journalism education and training, UNESCO, https://unesdoc.unesco.org/ark:/48223/pf0000265552. [13]

Japanese Ministry of Internal Affairs and Communications (2020), "Information distribution survey on novel coronavirus infectious diseases", https://www.soumu.go.jp/main_content/000693295.pdf (accessed on 8 February 2022). [52]

Jongh, D., B. Rofagha and L. Petrosova (2021), "Countering online vaccine misinformation in the EU/EEA", https://www.ecdc.europa.eu/sites/default/files/documents/Countering-online-vaccine-misinformation-in-the-EU-EEA.pdf. [8]

Karsten, J. and D. West (2016), "Inside the social media echo chamber", Brookings Institution, https://www.brookings.edu/blog/techtank/2016/12/09/inside-the-social-media-echo-chamber/. [42]

Kayali, L. (2021), "A French coronavirus conspiracy video stayed on YouTube and Facebook for months", Politico Europe, https://www.politico.eu/article/french-viral-covid-19-conspiracy-documentary-stayed-months-on-youtube-facebook/ (accessed on 23 September 2021). [72]

Kaye, D. (2018), Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, U.N. Human Rights Council, https://digitallibrary.un.org/record/1631686/usage?ln=en. [77]

Khan, I. (2021), "Disinformation and freedom of opinon and expression", https://undocs.org/A/HRC/47/25 (accessed on 12 August 2021). [28]

Kouzy, R. et al. (2020), "Coronavirus goes viral: Quantifying the COVID-19 misinformation epidemic on Twitter", Cureus, Vol. 12/3, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7152572/. [61]

Lerer, L. (2021), "How Republican vaccine opposition got to this point", New York Times, https://www.nytimes.com/2021/07/17/us/politics/coronavirus-vaccines-republicans.html. [25]

Loomba, S. et al. (2021), "Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA", Nature Human Behaviour, Vol. 5/3, pp. 337–348, https://www.nature.com/articles/s41562-021-01056-1. [62]

Marcellino, W. et al. (2020), "Human-machine detection of online-based malign information", RAND Corporation, https://doi.org/10.7249/RRA519-1. [47]

Mitchell, A. and M. Walker (2021), "More Americans now say government should take steps to restrict false information online than in 2018", Pew Research Center, https://www.pewresearch.org/fact-tank/2021/08/18/more-americans-now-say-government-should-take-steps-to-restrict-false-information-online-than-in-2018/. [16]

Moschella, D. (2022), "It's not just Facebook—"Old media" spreads misinformation, too", https://itif.org/publications/2022/01/10/its-not-just-facebook-old-media-spreads-misinformation-too. [36]

Neale, S. (1977), "Propaganda", Screen, Vol. 18/3, pp. 9-40, https://doi.org/10.1093/screen/18.3.9. [31]

Newman, N. et al. (2021), Reuters Institute Digital News Report 2021, https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2021-06/Digital_News_Report_2021_FINAL.pdf. [37]

OECD (2021), "Enhancing public trust in COVID-19 vaccination: The role of governments", https://read.oecd-ilibrary.org/view/?ref=1094_1094290-a0n03doefx&title=Enhancing-public-trust-in-COVID-19-vaccination-The-role-of-governments. [9]

OECD (2021), OECD Report on Public Communication: The Global Context and the Way Forward, OECD Publishing, Paris, https://doi.org/10.1787/22f8031c-en. [73]

OECD (2020), "Combatting COVID-19 disinformation on online platforms", https://read.oecd-ilibrary.org/view/?ref=135_135214-mpe7q0bj4d&title=Combatting-COVID-19-disinformation-on-online-platforms. [23]

OECD (2020), "Transparency, communication and trust: The role of public communication in responding to the wave of disinformation about the new Coronavirus", https://www.oecd.org/coronavirus/policy-responses/transparency-communication-and-trust-the-role-of-public-communication-in-responding-to-the-wave-of-disinformation-about-the-new-coronavirus-bef7ad6e/. [7]

OECD (2019), An Introduction to Online Platforms and Their Role in the Digital Transformation, OECD Publishing, Paris, https://dx.doi.org/10.1787/53e5f593-en. [2]

OECD (2019), Going Digital: Shaping Policies, Improving Lives, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264312012-en. [1]

OECD (2017), Recommendation of the Council on Open Government, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0438. [88]

OED (2013), Oxford English Dictionnary, Oxford, https://www.oed.com/viewdictionaryentry/Entry/171207. [32]

OHCHR (2001), "Monitoring human rights tn the context of elections", U.N. Human Rights Office of the High Commissioner, https://www.ohchr.org/Documents/Publications/Chapter23-MHRM.pdf. [26]

O'Sullivan, D. (2021), "White House turns up heat on Big Tech's Covid 'disinformation dozen'", CNN Business, https://edition.cnn.com/2021/07/16/tech/misinformation-covid-facebook-twitter-white-house/index.html (accessed on 29 October 2021). [59]

Oversight Board (2021), Oversight Board, https://oversightboard.com/ (accessed on 12 August 2021). [78]

Paavola, T. et al. (2016), "Understanding the trolling Phenomenon: The automated detection of bots and cyborgs in the social media", Journal of Information Warfare, Vol. 15/4, pp. 100-111, https://www.jstor.org/stable/26487554. [39]

Pew Research Center (2019), "Many Americans say made-up news is a critical problem that needs to be fixed", https://www.pewresearch.org/journalism/2019/06/05/many-americans-say-made-up-news-is-a-critical-problem-that-needs-to-be-fixed/ (accessed on 26 October 2021). [51]

Politico (2021), The facebook papers, https://www.politico.com/tag/the-facebook-papers, accessed 1 November 2021 (accessed on 1 November 2021). [15]

Rahman, M., A. Chia and W. Gonzalez (2021), "Using NLP to fight misinformation and detect fake news", https://omdena.com/blog/fighting-misinformation/ (accessed on 12 August 2021). [81]

Rosenberg, M., N. Confessore and C. Cadwalladr (2018), "How Trump consultants exploited the Facebook data of millions", New York Times, https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html (accessed on 5 August 2021). [12]

Rosen, G. (2021), "Community Standards Enforcement Report, Second Quarter 2021", Facebook, https://about.fb.com/news/2021/08/community-standards-enforcement-report-q2-2021/ (accessed on 22 September 2021). [64]

Sacco, L. et al. (2021), "Emergence of knowledge communities and information centralization during the COVID-19 pandemic", Social Science & Medicine, Vol. 285, https://www.sciencedirect.com/science/article/pii/S0277953621005475. [60]

Schaake, M. (2021), "Big Tech calls for 'regulation' but is fuzzy on the details", https://www.ft.com/content/a0a7f8de-f365-4e4e-a755-284df91c6e3a (accessed on 29 October 2021). [19]

Schechner, S. (2021), "Facebook is rebuked by Oversight Board over transparency on treatment of prominent users", Wall Street Journal, https://www.wsj.com/articles/facebooks-oversight-board-says-company-wasnt-fully-forthcoming-on-treatment-of-high-profile-users-11634817601. [79]

Scott, M. (2022), "As war in Ukraine evolves, so do disinformation tactics", Politico, https://www.politico.eu/article/ukraine-russia-disinformation-propaganda/. [5]

Shearer, E. (2021), "News use across social media platforms in 2020", Pew Research Center, https://www.pewresearch.org/journalism/2021/01/12/news-use-across-social-media-platforms-in-2020/. [34]

Shearer, E. and A. Mitchell (2021), "More than eight-in-ten Americans get news from digital devices", Pew Research Center, https://www.pewresearch.org/fact-tank/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices/. [35]

Silver, L. (2019), "Misinformation and fears about its impact are pervasive in 11 emerging economies", Pew Research Center, https://www.pewresearch.org/fact-tank/2019/05/13/misinformation-and-fears-about-its-impact-are-pervasive-in-11-emerging-economies/ft_19-05-13_misinformation_exposuretoincorrectinformationwidespread/. [50]

Somers, M. (2020), "Deepfakes, explained", MIT Sloan School, MIT Sloan School, https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained (accessed on 29 October 2021). [86]

Statt, N. (2020), "Major tech platforms say they're 'jointly combating fraud and misinformation' about COVID-19", The Verge, https://www.theverge.com/2020/3/16/21182726/coronavirus-covid-19-facebook-google-twitter-youtube-joint-effort-misinformation-fraud. [63]

Swire-Thompson, B. and D. Lazer (2020), "Public health and online misinformation: Challenges and recommendations", Annual Review of Public Health, Vol. 41/1, pp. 433-451, https://doi.org/10.1146/annurev-publhealth-040119-094127. [24]

Talbot, H. and A. Nusiebah (2021), "The edge of the infodemic: Challenging misinformation in Aotearoa", https://www.classificationoffice.govt.nz/assets/PDFs/Classification-Office-Edge-of-the-Infodemic-Report.pdf. [53]

Taylor, M. (2019), "Combating disinformation and foreign interference in democracies: Lessons from Europe", https://www.brookings.edu/blog/techtank/2019/07/31/combating-disinformation-and-foreign-interference-in-democracies-lessons-from-europe/. [11]

Twitter (2021), "COVID-19 misinformation", https://transparency.twitter.com/en/reports/covid19.html#item2:2020-jul-dec (accessed on 22 September 2021). [66]

Twitter (2013), "New tweets per second record, and how!", https://blog.twitter.com/engineering/en_us/a/2013/new-tweets-per-second-record-and-how (accessed on 21 September 2021). [68]

UNCHR (1996), "The right to participate in public affairs, voting rights and the right to equal access to public service", U.N. Committee on Human Rights, https://www.equalrightstrust.org/ertdocumentbank/general%20comment%2025.pdf. [20]

Wall Street Journal (2021), The facebook files, https://www.wsj.com/articles/the-facebook-files-11631713039 (accessed on 1 November 2021). [14]

Wang, L. et al. (2018), "Five shades of untruth: Finer-grained classification of fake news", http://dx.doi.org/10.1109/ASONAM.2018.8508256. [82]

Wardle, C. (2019), "First Draft's essential guide to understanding information disorder", http://creativecommons.org/licenses/by-nc-nd/4.0/ (accessed on 9 August 2021). [33]

Waterson, J. and D. Milmo (2021), "Facebook whistleblower Frances Haugen calls for urgent external regulation", The Guardian, *https://www.theguardian.com/technology/2021/oct/25/facebook-whistleblower-frances-haugen-calls-for-urgent-external-regulation* (accessed on 8 February 2022). [87]

WHO et. al. (2020), "Managing the COVID-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation", *https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation*. [6]

Woolley, S. (2020), The Reality Game: How the Next Wave of Technology Will Break the Truth, Public Affairs, New York. [48]

Yadav, K. et al. (2021), "Countries have more than 100 laws on the books to combat misinformation. How well do they work?", *https://thebulletin.org/premium/2021-05/countries-have-more-than-100-laws-on-the-books-to-combat-misinformation-how-well-do-they-work/* (accessed on 5 August 2021). [44]

Zimdars, M. and K. McLeod (2020), Fake news: Understanding media and misinformation in the digital age, Cambridge, Massachusetts ; London, England : The MIT Press. [89]