**Developing the methodology to detect “fake news” from fact-checked articles**

**MODULE DESCRIPTION AND GUIDELINES FOR THE INSTRUCTOR**

The module starts with an interactive introduction to the topic of fake news followed by 4 chapters on the following content:

1.Social Media: Response to Fake News

2.Fake News Detection Models

3.How to Recognise False Content - The 5Ws

4.The Future of Fake News: AI Generated Synthetic Media

Lastly there will be the possibility to reflect and summarise the outcome and time for concluding remarks.

Below each chapter will be explained briefly:

***Introduction***

Students will be introduced to the theme of the module mainly through a structured presentation of the definitions and ideas around the topic.

The introduction of social media has made news content easily available for consumption (Granik & Mesyura, [2017](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)). Even though the development of social media is a blessing to mankind. Conversely, it has also affected our lives negatively due to the untold suffering it has brought to us. Unlike the traditional media (newspaper, TV, and Radio) Social media has ushered in a new trend in news known as ‘fake news’ where malicious or misleading information is rapidly spread. The work of Klyuev ([2019](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)) posits that even though social media was created to enhance communication, it has almost replaced the mainstream media. A vast majority of people no longer watch television or listen to the radio, even if they listen to it, it will be done on social media.

The blossoming of false and unreliable news on the internet is a cultural, political, and technological phenomenon that’s hard to get your head around, let alone tackle. Conspiracy theories, misinformation, and disinformation run rampant on the internet, and it’s often difficult for people to tell what is true and what’s not.

In the next part of this chapter the definitions of the aspects that contribute to people believing fake news will be analysed.

Information Overload: When people are repeatedly exposed to the same stimuli, such as certain faces, they grow to like those stimuli more than those they have encountered less often. Such biases translate into an irresistible urge to pay attention to information that is going viral—if everybody else is talking about it, it must be important. In addition to showing us items that conform with our views, social media platforms such as Facebook, Twitter, YouTube and Instagram place popular content at the top of our screens and show us how many people have liked and shared something. Few of us realize that these cues do not provide independent assessments of quality. (M02/S01 slide 4)

Echo Chambers: Most of us do not believe we follow the herd. But our confirmation bias leads us to follow others who are like us, a dynamic that is sometimes referred to as homophily—a tendency for like-minded people to connect with one another. Social media amplifies homophily by allowing users to alter their social network structures through following, unfriending, and so on. The result is that people become segregated into large, dense and increasingly misinformed communities commonly described as echo chambers. (M02/S01 slide 5)

Rise of the Bots: Information quality is further impaired by social bots, which can exploit all our cognitive loopholes. Bots are easy to create. Social media platforms provide so-called application programming interfaces that make it fairly trivial for a single actor to set up and control thousands of bots. But amplifying a message, even with just a few early upvotes by bots on social media platforms such as Reddit, can have a [huge impact](https://doi.org/10.1145/2963104) on the subsequent popularity of a post. (M02/S01 slide 6)

The introduction of the link between fake news and social media leads to the next chapter where the social media response to fake news will be analysed. (M02/S01 slide 7)

***1.Social Media - Response to Fake News***

The first chapter starts with a short analysis of the timeline of events an individual goes through that leads to a better understanding of the question: Why does social media play an important role in the spread of fake news? (M02/S01 slide 8)

The understanding of the role of social media in the spread of fake news is very important for the following part of this chapter where we will look into the various social media response methods to fake news. The responses of the main social media are outlined below:

Google: Last year, Google.org provided Full Fact with $2 million and seven Googlers from the [Google.org Fellowship](https://www.google.org/our-approach/#fellowship), a pro-bono program that matches teams of Googlers with nonprofits for up to six months to work full-time on technical projects. The Fellows helped Full Fact build AI tools to help fact checkers detect 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. Over the past year, Full Fact boosted the amount of claims they could process by 1000x, detecting and clustering over 100,000 claims per day — that’s more than 36.5 million total claims per year!

The AI-powered tools empower fact checkers to be more efficient, so that they can spend more time actually checking and debunking facts rather than identifying which facts to check. (M02/S01 slide 9-10)

Facebook: Facebook faces a credibility crisis as it is used as a vector for misinformation and manipulation of the public around the world. It has been linked to violence in many parts of the world, and so in February 2021, the social network held a briefing to explain how it is trying to limit misinformation on its platform. The briefing detailed the efforts made by the social media giant to curb the circulation of false information and fake news. Facebook emphasised that it uses a three-part strategy to address misinformation called “Remove, Reduce, and Inform”. Interestingly, the latest attempt by Facebook comes at a time when its credibility is at an all-time low. The US company has just been blocked in Myanmar for allowing misinformation and fake news that could lead to violence over the military coup. The platform is also being abused by individuals spreading fake information around the ongoing farmers' protest in India.

During the virtual briefing, [Facebook](https://gadgets.ndtv.com/tags/facebook) highlighted its partnership with over 80 third-party fact checkers that are certified by the nonpartisan International Fact-Checking Network (IFCN) and cover 60 languages around the world. These include prominent names such as AFP, BOOM, and Fact Crescendo that exist in a large part of Asia Pacific — alongside other countries worldwide.

Apart from the fact-checking partners, Facebook has its native similarity detection system that is touted to rate more content than what third-party fact checkers see.

Using inputs from fact checkers, Facebook is also claimed to reduce distribution of content that is deemed false, altered, or partly false on its platform and [Instagram](https://gadgets.ndtv.com/tags/instagram). It is also touted to reduce the dispense of pages and domains that spread fake news. Similarly, the platform claimed that it reduced the distribution of spammy and sensational content that could coincide with misinformation.

Aside from removing and reducing content spreading misinformation, Facebook said that it informed users about false content through labels across its platform and Instagram. (M02/S01 slide 11)

Twitter: Twitter has [focused on issuing suspensions and bans](https://www.bbc.co.uk/news/technology-56252545) to accounts which consistently spread harmful Covid-19 misinformation when they come to the company's attention.

It also began putting warning labels on such tweets in early 2020, announced [a collaboration with news organisations](https://www.bbc.co.uk/news/business-58065463) as part of an attempt to debunk false information, and started [a pilot scheme in January 2021 to allow a small number of people to submit "notes"](https://www.bbc.co.uk/news/technology-55806002) about misleading content. In addition in August 2021 Twitter announced that it will collaborate with two of the largest international news providers, Reuters and the Associated Press, to debunk disinformation on its messaging site. The news agencies will help Twitter give more context and background information on events which create a high volume of tweets. (M02/S01 slide 12)

Instagram: Similarly to Facebook, Instagram is partnering with 45 global third party fact-checking networks in order to debunk misinformation. As stated on Instagram’s blog on [combating misinformation](https://about.instagram.com/blog/announcements/combatting-misinformation-on-instagram), if the third party fact checking organization finds information posted is false, a few things happen including:

* The photo or video is removed from the explore page and hidden from hashtags used.
* Photo is labeled “False Information” and has an overlay that covers the picture.
* A warning is given to the account that posted the false information, that they have shared false information.
* A label is placed under the post giving the audience information on where to fact check. (M02/S01 slide 13)

Youtube: The platform has removed thousands of videos for violating its misinformation rules since the pandemic began, which has led to criticism for being too censorious.

In addition, beneath certain videos about health, the company will add information panels that will alert viewers if the source is “authoritative,” such as when the video comes from an accredited hospital. YouTube will also start displaying select videos more prominently on the site when people search for health terms, similar to how it now treats certain news topics.

For both the material in these panels and its ranking system, YouTube said it will rely on [a recent set of guidelines](https://nam.edu/programs/principles-for-defining-and-verifying-the-authority-of-online-providers-of-health-information/) for verifying online information from the National Academy of Medicine, a non-governmental organization. (M02/S01 slide 14)

Tiktok: To keep misleading information and deceptive content and accounts from the platform, TikTok launched three measures:

* Enhancing its misinformation policies: The updated [Community Guidelines](https://www.tiktok.com/community-guidelines?lang=en) prohibit misinformation that could cause harm to the TikTok community or the larger public, including content that misleads people about elections or other civic processes, content distributed by disinformation campaigns, and health misinformation. These updates were developed with industry experts, and the language reflects input from members of their [Content Advisory Council](https://www.tiktok.com/transparency).
* Expanding reporting options and fact-checking: In addition to working with their Content Advisory Council, which includes experts on deep fakes, free speech, inclusive AI, and more, they partnered with PolitiFact and Lead Stories to fact check potential misinformation related to the 2020 US election.
* Countering foreign interference: Ahead of the 2020 Presidential election they also worked with the US Countering Foreign Influence Task Force (CFITF) to help stop the threat and dangers of foreign influence on elections. (M02/S01 slide 15)

Online Press:

Even though the main source of fake news is social media, online press is evolving in order to proactively tackle misinformation circulating online. A few online press outlets introduced the “fake news” column. This column can include articles debunking fake news or anything related to fake news. Examples of these kinds of online columns are “Fact Check”, “In Fact” and “Reality Check”. (M02/S01 slide 16)

In addition online media outlets like Reuters and BBC are taking public media literacy to the next step, introducing online courses and forums where the public can be informed and develop the methodologies to distinguish between fake and fact-checked news.

Reuters, partners with Facebook Journalism Project to help newsrooms around the world spot deepfakes and manipulated media. The course aims to combat these risks by teaching journalists about the various types of altered media and the ways in which newsrooms can be better equipped to handle misinformation in order to move forward to verify and publish genuine content from third-party sources. (M02/S01 slide 17)

The BBC's Trusted News Initiative is a partnership that includes organisations such as Facebook, Twitter, Reuters and The Washington Post. It is the only forum in the world of its kind designed to take on disinformation in real time. (M02/S01 slide 18)

**Activity: Discussion**

This activity aims to reflect on the chapter and create an opportunity for an open discussion. (M02/S01 slide 19) Example questions for discussion are:

* Were you aware of these measures?
* Did you came across to any of these measures? What was your reaction? Did it influenced your response to the information?
* Are you aware of any other measures?
* Do you believe that Social Media companies are doing enough about fake news?

By the end of this chapter students will be able to understand the important role of social media on the spread of fake news and the importance of the measures taken/to be taken by the social media industry in order to tackle fake news.

***2.Fake News Detection Models***

This chapter presents the different approaches in fake news detection used by the media. Even though these approaches are not applicable for individuals, they are important to be presented in order to provide a broader understanding in the process of fact-checking. This chapter aims to unveil the complexity of professional fact-checking, the advantages and disadvantages of each approach using statistics and research. (M02/S02 slide 4)

**Experts or professionals facts-checker approach**

Expert or professional fact-checkers are a small group of professionals in various disciplines who are capable of verifying the veracity of certain news items and decide whether such information is fake or authentic (Zhou & Zafarani, [2018](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)) posit that the strength of expert-based fact-checking techniques lies in the fact that they are small in number thus, easy to manage and have a high accuracy rate. A study by Pennycook and Rand ([2019](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)) explains that an expert-facts checker is a natural approach to verifying fake news which uses professional fact-checkers to determine which content is false, and then engaging in some combination of issuing corrections, tagging false content with warnings, and directly censoring false content .g., by demoting its placement in ranking algorithms so that it is less likely to be seen by users.

However, some Professional fact-checkers are not independent and work for an organization and often have a lot of limitations. The expert-fact checking technique is slow especially in a situation where they are given a large volume of information to verify due to their small number, also the fact that the process is manual. Most fact-checkers have often come under strong criticism for being biased and political. Some examples of prominent fact-checker sites includes; Snopes which deals with political and social issues, *Hoaxslayer* focusing on various field including health, religion, and economics, *Fullfact*, *TruthOrFiction*, *The Washington Post Fact Checker, PolitiFact, FactCheck* mostly focus on American politics. Due to the limitation of the expert-based fact-checkers, the crowdsourced technique is seen as a good alternative. (M02/S02 slide 5)

**Crowdsourced approach**

Crowdsourced or ‘wisdom of the crowds’ approach is based on the premise that no matter how smart someone is, the collective effort of individuals or groups supersedes any single individual intellectual capacity. Brabham ([1982](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)) sees crowdsourcing as, ‘an online, distributed problem-solving and production model that leverages the collective intelligence of online communities to serve specific organizational goals’. The weaknesses of expert-based fact-checkers have prompted many to seek the ‘wisdom of the crowds’ technique. Pennycooka and Rand ([2019](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)) used crowdsourced judgments of news sources on social media and discovered that the crowd is more effective than professional fact-checkers, in judging the news source quality laypeople got a similar rating with professional fact-checkers. The limitation of the ‘wisdom of the crowds’ approach is because the crowd is made up of laypeople of different fields and have little knowledge of some news site, consequently, news sites which they are unfamiliar with are marked as an untrusted site. (M02/S02 slide 6)

**Machine learning approach**

Early machine learning method in detecting fake news was proposed by Granik and Mesyura ([2017](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)), because it is assumed that fake news is created intentionally for the political and financial benefit, so they often have an opinionated and enticing headline, at such the extraction of the textual and linguistic feature is necessary for machine learning. Most Artificial intelligence tools for detecting and flagging fake news rely heavily on Click-Through Rates (CTR), the position of the stream page increases as the CTR increase and some fake news type such as clickbait articles usually have high CTR due to it enticing and appealing nature. Consequently, such an approach cannot be used to detect fake news types such as clickbait. Biyani et al. ([2016](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)) propose a machine learning model to detect fake news. Using Gradient Boosted Decision Trees (GBDT) their model achieves strong classification performance and saw that informality is a crucial factor of the ‘baity’ nature of web-pages. They formulate an automatic machine learning model for identifying and predicting an article whether it is a clickbait or not. Several features were used such as the URL, content, title etc. Using datasets from yahoo news aggregator, they collected 1349 (training set) clickbait and 2724 (testing set) non-clickbait web pages. They were able to identify spam and web pages by defining 8 types of clickbait such as Exaggeration, Teasing, Inflammatory, Graphic, Formatting, Bait-and-Switch, ambiguous, Wrong (Biyani et al., [2016](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)). By comparing clickbait articles, they assert that most clickbait has misleading information such as gossip and most appealing headlines aimed at enticing the reader to click on the link. The landing page is usually of low quality and thus, they contend that because news aggregator site i.e. yahoo news aims to serve its user with news article via its homepage, the proliferation of clickbait article which usually has low quality increases user's dissatisfaction rate and amplify their abandonment which is bad for business and hence detecting and removing clickbait site become inevitable. This approach is not without limitation, fake news is a broad issue with several types but this study focuses only on one type of fake news i.e. clickbait which has two ways of detecting it; firstly, it can be easily detected because the content is different from the headline and secondly, based on the fact that the content is of low quality. Some authentic news however has low quality content as well. (M02/S02 slide 7)

**Natural language processing technique (NLP)**

NLP work within automated deception detection technique which involves the application of lexical and semantic analysis, with the use of regression, clustering, as well as classification techniques such as binary classification of text where news are classified as real and not real, in a two-class problem, where it is difficult to detect, a third-class may be added such as partially real or partially fake. Linguistic features are a key factor to NLP including text content and style. Grammar and style detectors and syntactic analyser such as Stanford parser have been reported by Klyuev ([2019](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)) and it gives accurate results. The basic task is to identify some verbal and lexical cues which will point out linguistic differences when humans tell lies as opposed to when they tell the truth. For instance, deceivers produce more total words-count and sense-based words such as those that show lower cognitive complexity, the use of more *negative emotion words, extreme positive words*. (M02/S02 slide 8)

**Hybrid technique**

Hybrid detection techniques emerge as an alternative to several fake news detection methods, due to the complexity and ambiguous nature of fake news, the combination of other method is imperative. The failure of the single model in detecting fake news prompted scholars to find alternative measures to detect fake news accurately. Several Hybrid-based techniques do exist involving dynamic processes and a combination of some models. In this study, we discuss Hybrid Expert-crowdsource and Hybrid Machine-crowdsource detection method.

**Expert-crowdsource approach**

The hybrid expert-crowdsource approach is relatively a new method that emerges as a result of the weaknesses of the expert-based as well as crowdsourced-based fact-checker. This approach involves the combination of the two-manual fact-checking systems by applying human knowledge as opposed to automatic facts-checking involving the use of the machines. The key idea behind this approach is that where experts failed, the crowdsourced approach can complement and vice versa (Mahid et al., [2018](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)). The expert-based has often been accused of being politically biased, not independent and very slow in detecting fake news (Vedova et al., [1957](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)). While a study by Pennycook and Rand ([2019](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)) allude that the crowd is limited in many areas since they are composed of laypeople and at such, they will give the wrong prediction to content which they are unfamiliar with. Therefore, it is imperative that since the crowd is unbiased and acting independently, larger in number and thus can easily work on a large volume of information, the aggregation of the crowds’ decision can be sent to the expert which will yield better results since experts are familiar with many areas. (M02/S02 slide 9)

**Human-Machine approach.**

Most machine learning algorithms developed to automatically detect fake news has often failed. This is because all news does not have the same writing pattern and involves several topics with salient features. A study by Shabani and Sokhn ([2018](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)) found out that one of the limitations of automatic fake news detection is low accuracy, those machine algorithms developed to detect fake news through news contend are prone to low accuracy due to the fact that most language use in writing fake news bypass the detection process. While the wisdom of the crowd as seen already is a good approach but slow and time-consuming and lack expert knowledge because usually crowd are compose of laypeople (Pennycook & Rand, [2019](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)), the combination of machine learning algorithms and the collective effort of humans has proven to yield better fruits, especially in the area of detecting fake news automated by social bots.

One of the hybrid machine-crowdsource techniques was proposed by Shabani and Sokhn ([2018](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)), they propose a model that uses a hybrid machine-crowd approach to detect fake news and satire. Firstly, crowdsourcing was used to classify news from Satire and fake news and distinguish them which was difficult to detect by the machine. While the crowdsourcing task on fact-checking achieved an accuracy rate of 84%, they designed the hybrid fake news detection model, and they got an overall accuracy rate of 87%. This is a very good result with high accuracy as compared to previous studies. (M02/S02 slide 10)

**Graph-based method**

The homophily aspect of the social network ensures that connected users in a graph network have the same traits, and view similar content such as the same news article (Zhou & Zafarani, [2019](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)). On the other hand, malicious users will have similar behavioural patterns and hence articles from these malicious users can easily be recognized and flag as fake. Zoltan The graph network fake news detection model investigates news content from homogeneous and heterogeneous networks (Zhou & Zafarani, [2019](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)), a typical example is the truth tree or stance network: that is if one for or against, such a network is constructed and classify news articles that are fake from those that are authentic. A random walker traversing within the network beginning from authentic news will land in another authentic news and vice versa. The random work proposed by Jia et al. ([2017](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)) was effective in tracking fake accounts on social media, the work, although it is essential in truth detection. The work is aimed at combating Sybil attacks on social media which these Sybils or social bots as tools use in spreading fake news.

Coupled with that, Ren and Zhang ([2020](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)) propose a fake news detection model known as the Hierarchical Graph Attention Network (HGAT). Their method represents news articles as nodes and matches related nodes or vertices together and fake news is then identified from using classification tasks. Using a real-world dataset from PolitiFact, news articles were weighted as neighbours and the same neighbour nodes were aggregated in the schema node. The model employs the Heterogeneous Information Networks HIN-based fake news article and their result outperforms other states of the art fake news identification graph model. (M02/S02 slide 11)

**Recommendation system approach**

The recommender system has gained prominence recently with several companies relying on recommendation algorithms to target its audience such as Netflix (Gomez-Uribe & Hunt, [2015](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)), Amazon (Linden et al., [2003](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)), Youtube (Davidson et al., [2010](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)). These systems have proven to be useful in fighting fake news too on social media. In areas of academic recommendation, Hoang et al. ([2017](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)) examine academic event recommendations based on research similarity and the interaction between authors. The recommendation system approach tries to verify some news content deemed to be authentic and then recommend these news articles for consumption (Özgöbek et al., [2019](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)). Mizgajski and Morzy ([2019](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)) examine recommendation systems in the online news industry and posits that most recommendation systems use human emotions to recommend content. The collaborative filtering recommendation method (Garcin et al., [2012](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)) recommends news content based on ratings and comments from others. While the content-based method recommends news, items based on similar contents (Garcin et al., [2012](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)).

Vo and Lee ([2018](https://www.tandfonline.com/doi/full/10.1080/24751839.2020.1847379)) proposed a fact-checking URL recommender system to combat fake news. The motivation of the study stemmed from the effort of the so-called ‘guardians’ in combating fake news on social media. Tweets from these guardians were collected which contained URLs ranging from different aspects such as politics and fauxtography. Their work is based on the premise that fake news is being spread through some URLs that can be traced and identified. In order word this method attempts to combat fake news by attacking the source. Comparing the tweets coming from certain URLs or URLs contained in tweets that are deemed to be an authentic website, these URLs can be flagged off the social media. (M02/S02 slide 12)

**Technology and fact-checking:**

* Official Media Bias Fact Check Icon:This extension shows an icon denoting the political bias for the current page. This extension will display a color-coded icon denoting the bias of the page you are currently viewing, according to Media Bias/Fact Check. You can click the icon to read more notes about the site, or visit MBFC for more details.
* News Guard: NewsGuard uses journalism to fight unreliable news. Their trained analysts, who are experienced journalists, research online news brands to help readers and viewers know which ones are trustworthy--and which ones aren't. The Green-Red ratings signal if a website is trying to get it right or instead has a hidden agenda or knowingly publishes falsehoods or propaganda, giving readers more context about their news online.
* Trusted News: Trusted News uses AI to assist newsreaders in evaluating the quality of the online content they read. In its first release, it scores the objectivity for a selected article, testing whether it is written from a neutral perspective as opposed to a personal one. For example, phrases like "in my opinion" or "I think" are used by authors to reflect their individual thoughts, beliefs, and attitudes. (M02/S02 slide 13-16)

By the end of this chapter, students will be able to implement the technology and tools for fact-checking in their everyday lives. The technology will develop their critical thinking about information and news online.

**Activity: Testing the technology**

A contemporary example of fake news (e.g., fake news about the 5g technology and Covid-19 pandemic) should be the starting point of this group activity. The students will then choose a fact-checking online tool such as (“Official Media Bias Fact Check Icon”, “News Guard” and “Trusted News”).

As a group the students will then browse the internet about their chosen fake news topic and test the tools for a range of articles and websites. (M02/S02 slide 17)

This activity aims to introduce students to the technological tools, to understand, evaluate and assess them with the following questions in mind:

* How do you feel about the tool?
* Was it helpful?
* What could be better about its performance?

In contrast with this chapter that focused on the technological methodologies for fake news detection, the next chapter will provide a different view on fake news detection focusing on the human level.

***3.How to recognise false content - The 5Ws***

The following chapter will introduce the students to the simple non-technological methodologies of fake news detection. The students will be presented with a 5-step thought process based on the 5Ws theory that can be easily implemented when searching for a topic online.

Almost a quarter of adults have shared a false news story, and we’re least likely to fact-check news and other things that come to us through people we know and trust on social networks (even though for many people these are their most common sources of news)?

Everyone – online platforms, news organizations, individuals and civil society – has an important part to play in keeping hoaxes and false news from spreading by making sure that something is reliable before we share it. This chapter focuses on what individuals can do to recognize false content online. (M02/S03 slide 4)

So when should we double-check before sharing something we find online? We don’t have to try to debunk everything that comes to our way, but we should ask ourselves these questions before we share, tag, retweet or forward:

* Could someone base an important decision (about their health, their career, travel, etc.) on this?
* Is it about a hot or controversial issue?
* Does this seem “too good to be true”? Is it something that we hope is real or would really like to believe?

In the next part of this chapter we will use the 5Ws method as a way to identify false content online.

**WHO? Authority**

Who is spreading it? Do they have a good track record for accuracy?

Adults are more likely to trust the content of a story if it was shared by someone they trust, so always track it back to the original poster. If the person who shared it with you didn’t give a source, do a search for it. But don’t assume that a newspaper, TV network or online news site is the original source.  
Once you’ve established the source, find out who they are and why they might be a valid source for what they’re spreading (M02/S03 slide 5) :

* If it claims to be from a newspaper or other news source, do a search to make sure it really exists. Check the Web address to make sure it’s the right one for the real website.
* If it’s a science or health story, is the author a real expert? Do a search on their name and see what comes up.
* If it’s a photo of something that’s happening in a particular place, do they really live there?
* Have they posted on this subject before? If not, be cautious.
* Do they post a lot of spreadable stuff? If so, be cautious.

You can also look at the source’s network. Are they connected mostly to:

* People and groups who all have similar opinions?
* Advertisers?
* People and groups who have no connection to the thing they’re spreading?
* Nobody?

**WHAT? Accuracy**

What kinds of false content should I watch out for? There are many different kinds of false information being spread around. (M02/S03 slide 6)

Before you share something, make sure it isn’t one of these:

* Hoaxes and false news: These are spread on purpose to mislead people. Sometimes these are motivated by malicious or mischievous intent; sometimes they are motivated for ideological or political purposes; other times they’re done for financial gain.
* Scams: Sometimes the purpose of a fake story is to separate you from your money, to get you to give up your personal information, or to get you to click on a link that will download malware onto your computer.
* Ads: Some things that are spread around are obviously ads, but others are disguised as “real” content.
* Legitimate satire and parody content are fun and entertaining, but should not be confused with actual news!

**WHEN? Currency**

When did it start spreading?

A lot of things get spread more than once, like some of the photos of flooding that go around every time there’s a big storm. You can do a reverse image search to see if a photo has been posted before, or search a description of the photo. (M02/S03 slide 7)

* How long has the original poster’s account been active? If an account appears new or recently started posting with no prior history, be cautious.

For Websites look at:

* When was the site created?
* Is there a last updated date?
* Are there "dead" links?

**WHERE? Repetition**

Where else can I find out if something is real? (M02/S03 slide 8)

* Do a search for the subject with the words “hoax”, “fake”, “viral” or “scam”
* For pictures, you can do a reverse search for images at TinEye (www.tineye.com). This will tell you where else the picture has appeared, and also show you similar pictures (which is a good way to find out if a picture has been photoshopped). TinEye can also be installed as a browser plugin, so you can check a picture just by right-clicking it.
* Check out hoax-busting sites like Snopes.

Also you should ask yourself questions such as:

* Have you heard of the publisher before?
* Does the publisher take responsibility for the content?
* Is it a *peer-reviewed* or *refereed* source?
* Was it self-published?

**WHY? Purpose and Point of View**

Why is it being spread around?

Is it trying to scare you? To make you laugh? To make you angry? Does it use emotionally-loaded words or images to get a rise out of you?

Be especially wary of stories you want to believe. Some fake news sources target you with stories they think you'll hope are true to get you to click and spread them. (M02/S03 slide 9-10)

***4.The Future of Fake News: AI Generated Synthetic Media***

The last chapter focuses on an advanced level of fake news which is AI Generated Synthetic Media. The definitions and the general terms around the topics will be presented in order to provide students with the general understanding on the topic.

For the introduction to the topic three video examples will be presented:

* Nancy Pelosi: A video circulating on social media shows House Speaker Nancy Pelosi speaking in a slurred and awkward manner. One popular post boasts 91,000 shares on Facebook and bears a caption reading: “This is unbelievable, she is blown out of her mind, I bet this gets taken down!” The video, however, has been manipulated to make Pelosi appear drunk and incoherent. (M02/S04 slide 4)
* Mark Zuckerberg: “Mark Zuckerberg reveals the truth behind Facebook: predicting the future behaviours of its users and manipulating them.” However, Zuckerberg never pronounced these words. It’s a deepfake created by artists Bill Posters & Daniel Howe. (M02/S04 slide 5)
* Nixon: Burgund and co-creator Francesca Panetta created [an art installation in 2019 that combined actual footage of Nixon’s resignation speech](http://news.mit.edu/2020/mit-tackles-misinformation-in-event-of-moon-disaster-0720), and the text of an in-memoriam draft speech that had been written by Nixon speechwriter Bill Safire in case of a failed moon landing. The result is a deepfake video that, despite the creators’ attempts to be transparent about the fabrication, still tricked some viewers into thinking it was an unaired version of the speech. (M02/S04 slide 6)

Advancements in artificial intelligence (AI) and cloud computing technologies, GPU VMs (Graphic Processing Units Virtual Machines) and Platform Services, have led to rapid development in the sophistication of audio, video, and image manipulation techniques. Access to commodity cloud computing, public research AI algorithms, and abundant data with availability of diverse and vast media have created a perfect storm to democratize the creation of synthetic media. This AI-Generated synthetic media is referred to as deepfakes. The distribution of synthetic media is democratized at scale by social platform.

Deepfakes are synthetic media (fake) generated using the Artificial Intelligence technique of deep (deep) learning.

Innovation and research in GAN techniques combined with the growing availability of computing have led to improvements in the quality of synthetic data at a surprising pace. New tools, many of which are publicly available, can manipulate media in increasingly believable ways, such as creating a copy of a public person’s voice or superimposing one person’s face on another person’s body. GAN and deepfakes have evolved from research and academic topics to practical applications for businesses to innovate, entertain, and engage in social engagement.

Cheapfakes are simple manipulation through easy conventional editing techniques like speeding, slowing, and cutting, as well as nontechnical manipulations like restaging or recontextualizing existing media. Cheapfakes or Shallowfakes are defined as manipulated media created by using more straightforward image and video editing techniques to spread mis/disinformation or to change the narrative of a story. (M02/S04 slide 7)

Types of Deepfakes:

Deepfakes have become synonymous with face swapping and lip-syncing. There are many other types of AI-based manipulation of audio, video, and images that can be qualified as deepfakes. (M02/S04 slide 8-9)

* Face-swapping: Face swapping is when one person’s face is replaced or reconstructed by another person’s face or key features from another face. Face swapping or manipulation with filters is a common feature of almost all the social media, video chatting apps.
* Puppeteering: Puppeteering is rendering manipulated full-body actions and behavior using AI. It is a technique to create a 3D model of the target face and body in a video to act and say as the puppeteer. It is also known as full body deepfakes.
* Lip-sync: Lip Synching is a technique to render mouth movements and facial expressions to make the target say things with their voice and the right tone and pitch. AI algorithms can take an existing video of a person talking and alter the lip movements in the video to match new audio.
* Voice Cloning: Voice Coning is a deep-learning algorithm that takes in the voice recordings of an individual to generate synthetic voice that is overly like the original voice. It is a technique to create custom voice font of an individual and then use the font to generate speech.
* Image Synthesis: Image Generation or Image Synthesis is a technique to use computer vision technology, deep Learning and Generative Adversarial Networks (GANs) to synthesize new images. It can produce a computer-generated image of a person or any object that is not real.
* Text Generation: Text Generation is a method to automatically generate text, write stories, prose, and poem, create abstracts of long documents, and synthesize using AI techniques for text and deep learning.

Malicious use

As with any new technology, evil actors will take advantage of the innovation and use it for their benefit. GAN and Deepfakes have become more than research topics or engineering toys. Starting as an innovative research concept, now they can be used as a communication weapon. Deepfakes are becoming easy to create and even easier to distribute in policy and legislative vacuum. (M02/S04 slide 10)

A few examples of malicious use are:

* Fabricate media
* Damage reputations
* Fabricate evidence
* Defraud the public
* Undermine trust in democratic institutions
* Thread to individuals
* Thread to society
* Thread to democracy
* Thread to businesses

Positive Use

Technology is very empowering and a great enabler. Technology can give people a voice, purpose, and ability to make an impact at scale and with speed. New ideas and capabilities for empowerment have emerged because of the advancements in data science and Artificial intelligence. AI-Generated Synthetic media has many positive use cases. (M02/S04 slide 11) A few examples of positive use are:

* Education (i.e. historical figures can come to life and create more impact and engagement)
* Art (i.e. “Mona Lisa” was brought to life using Deepfake technology)
* Autonomy & Expression (i.e. Deepfake can be used to anonymise voice and faces to protect their privacy)
* Reach and Message Amplification (i.e. Localise video and audio content to broaden the reach of a public message)
* Public Safety & Digital reconstruction (i.e. Reconstructing the crime scene)
* Innovation (i.e. Deepfake news presenter, Fashion trends on virtual customers)

**Activity: Detect Political Fake News**

This activity is based on the MIT research project about detecting fake news. Students will need to visit this [link](https://detectfakes.media.mit.edu/) in order to proceed with the activity. Students will watch a variety of media snippets including transcripts, audio files, and videos. The aim is to share how confident they are that the individual really said what they show. Half of the media snippets that we present are statements that the individual actually said.

This activity aims to challenge the students to test their knowledge on detecting fake news with real-life examples that would be possible to encounter online. (M02/S04 slide 12)

***Summary***

Open time to reflect and summarise what has been learned. Some inspiration for possible questions for reflection:

* Are the new technologies for the detection of fake news beneficial for you?
* In your opinion, what are the important steps in tackling fake news?
* Critical thinking: What can actually help prevent the spread of fake news?