Artificial Intelligence and the Spread of Mis- and Disinformation

Annie Benzie and Reza Montasari

Abstract In a so-called post-truth era, research on the subject of the spread of mis- and disinformation is being widely explored across academic disciplines in order to further understand the phenomenon of how information is disseminated by not only humans but also the technology humans have created (Tandoc, Sociol Compass 13(9), 2019). As technology advances rapidly, it is more important than ever to reflect on the effects of the spread of both mis- and disinformation on individuals and wider society, as well as how the impacts can be mitigated to create a more secure online environment. This chapter aims to analyse the current literature surrounding the topic of artificial intelligence (AI) and the spread of mis- and disinformation, beginning with a look through the lens of the meaning of these terms, as well as the meaning of truth in a post-truth world. In particular, the use of software robots (bots) online is discussed to demonstrate the manipulation of information and common malicious intent beneath the surface of everyday technologies. Moreover, this chapter discusses why social media platforms are an ideal breeding ground for malicious technologies, the strategies employed by both human users and bots to further the spread of falsehoods within their own networks, and how human users further the reach of mis- and disinformation. It is hoped that the overview of both the threats caused by and the solutions achievable by AI technology and human users alike will further highlight the requirement for more progress in the area at a time when the spread of falsehoods online continues to be a source of deep concern for many. This chapter also calls into question the use of AI to combat issues arising from the use of advanced Machine Learning (ML) methods. Furthermore, this chapter offers a set of recommendations to help mitigate the risks, seeking to

A. Benzie (✉)
Hillary Rodham Clinton School of Law, Swansea University, Swansea, UK
e-mail: A.R.Benzie@Swansea.ac.uk

R. Montasari
School of Social Sciences, Department of Criminology, Sociology and Social Policy, Swansea University, Swansea, UK
e-mail: Reza.Montasari@Swansea.ac.uk
URL: http://www.swansea.ac.uk

R. Montasari (ed.), *Artificial Intelligence and National Security*,
https://doi.org/10.1007/978-3-031-06709-9_1

explore the role technology plays in a wider scenario in which ethical foundations of communities and democracies are increasingly being threatened.

Keywords Artificial intelligence · Misinformation · Disinformation · Social media · National security · Cyber threats · ML · Software robots

1 Introduction

The way that people communicate with each other has changed (Bakardjieva, 2005). Social media platforms such as Facebook and Twitter allow people to instantly access information on a wealth of subjects from users anywhere in the world, in real time. The World Wide Web which had its roots firmly in creating an environment which fostered freedoms (Benedek & Kettemann, 2020) and which connected people globally has replaced the millions of offline communities separated by distance, language, and culture, with a single global community. One where each of the billions of individuals who have access to an internet connection can contribute to the creation and dissemination of information quickly, and at low cost (Kreps, 2020). In reality, the abundance of information in a globalising world is not a new cause for concern. From James Madison in the nineteenth century to the work of Habermas and the many researchers who have followed since, there is one constant: a fear of both the controls and lack thereof, surrounding the ever-growing volume of information (Habermas, 1984). Offsetting the benefits of a more connected world are a string of worries as far-reaching platforms have become breeding grounds for mis- and disinformation (Tandoc Jr, 2019), which is disseminated by both humans and machines often with malicious underlying intentions. As such, the subject of how falsehoods are spread online has rightfully become a topic of intense interest for academics across an array of disciplines, such as computer science, psychology, and politics (Shu et al., 2020). This study was carried out with the aim of highlighting literature integral to this subject. The study adds to the existing body of research offering further insight into not only the manner in which mis- and disinformation is spread but also the challenges this presents. The study also focuses on ways in which the risks associated with widespread access to falsehoods may be mitigated. It is perhaps vital that we consider the challenges that modern technology has brought into the equation, in particular AI.

The remainder of this chapter is structured as follows. In Sect. 2, the meaning of information and truth in a post-truth world are explored. A particular focus is placed on the fine line between lies which refer to truth, and post-truth which seeks to delegitimise truth completely. The definitions of essential terms such as ‘misinformation’ and ‘disinformation’ are discussed, including exploration around the question of intent which distinguishes the two terms by definition. In Sect. 3, the role of AI in the spread of mis- and disinformation is considered, focusing on the types of disinformation and how fabricated content using techniques such as Generative Adversarial Networks (GANS), deepfakes, and multimodal content may be detected. This includes multifaceted approaches focusing on the content, organisation, emotion, and manipulation pertinent to the media content. In this section, the role of social media and AI bots are discussed, showing that although users may not actively consume content, they might disseminate said information, thus creating clusters within the network which align with and confine users to their personal beliefs. Section 4 further considers both the micro and macro impacts the spread of misinformation has on individuals and society as a whole. This includes a discussion on whether it is possible for disinformation to sway public opinion on such a scale as to conflict with democratic systems. Finally, ways in which the impact of widespread disinformation may be mitigated are considered, such as the introduction of a mass-collaboration model, increased education and focus on media literacy for users, as well as increasing the burden of responsibility of social media companies to find and employ effective solutions. Finally, after discussing recommendations for how to mitigate the risks of spreading falsehoods online, the synthesis of literature explored is concluded by providing final thoughts and considerations for the future, particularly in terms of research which would be of further benefit to the academic community.

2 Misinformation Versus Disinformation

2.1 Information and the Meaning of Truth in a Post-Truth World

The conversation around information is one which has been ongoing for several decades. It is often argued that untrue information does not qualify as information (Floridi, 2004). In his ‘Outline of a Theory of Strongly Semantic Information’, Floridi presents the argument that in order to qualify as information, the sentence must be true. Under this theory, many statements which although may be considered important data to some, could not be considered information, if it is the case that there is insufficient evidence to suggest the statement to be concretely true (Floridi, 2004). Countering this theory, Fetzer (2004) suggests that even the mere existence of the terms ‘misinformation’ and ‘disinformation’ indicates that false information is not anomalous (Fetzer, 2004). Similarly, a conclusion drawn by Buckland argues that it is complex to definitively state that anything is not informative in some way; therefore, it must be defined as being information. With this in mind, a statement which is deemed as informative could therefore be considered as information, albeit false or misleading (Buckland, 1991). Outside of one’s own experiences, our perception of what is true is based on representatives of the truth that we deem reputable, for example, the news. According to Mingers (2001) and Carsten Stahl (2006) ‘Such meaning is only relevant if information can affect actions or perceptions’ (Carsten Stahl, 2006; Mingers, 2001). It is perhaps also key to consider the distinction between having information and being informed. While being informed requires that one has information, it is a much grander condition than having access to masses of information (Webster, 2014). In his 2006 work 'The Logic of Being Informed', Floridi considers that 'the informativeness of a message – raises issues of e.g., novelty, reliability of the source and background information' (Floridi, 2006).

As such, being informed is therefore dependent on the relative informativeness of the information. Furthermore, given that an overload of information, some of which will derive from unreliable sources, more information does not necessarily equate to being well-informed. If we understand that information does not equate to truth, then we can also establish that truth can be regarded as an 'ideal' type of information whereby the information provided in the statement is true (Carsten Stahl, 2006). This is echoed by Habermas (1984), who argues that ultimately rational speakers deem the truth to be the ultimate goal of communication (Habermas, 1984). However, there are specific terms used to represent different types of falsehoods. Pertinent to this chapter are, of course, the terms 'misinformation' and 'disinformation'. Another important consideration is the misalignment between the concepts of lies and post-truth. Bufacchi (2021) argues that while lies still predominantly refer to the truth, post-truth seeks to destroy the entire infrastructure on which the truth is based. 'Post-truthers are different: their aim is to delegitimize truth, since this is the best way to disarm the threat truth poses to them' (Bufacchi, 2021). The article also argues that although post-truth is a popular subject in the modern world, it is far from a new concept. As is outlined in the 1967 essay 'Truth and Politics' (Arendt, 1967), politics and truth are not necessarily a natural combination. Fifty-five years later, this concept has not changed. However, with platforms on social media being readily used, it is much easier for individuals to spread their ideologies and essentially employ modern technologies to act as a stage from which to shout. With new data being created and uploaded in real time, hunting for the truth could perhaps draw comparisons with hunting for a needle in a haystack, if the needle were to resemble every piece of hay in the stack. How does one distinguish truth from falsehoods? How can information be sorted into piles of information, and mis- or disinformation? Approaches to how this has been achieved using ML algorithms will be drawn upon in Sect. 3.

2.2 The Differences Between Misinformation and Disinformation

Misinformation may be defined as 'a claim that contradicts or distorts common understandings of verifiable facts' (Guess & Lyons, 2020). It has also more generally been defined as 'false or misleading information' (Lazer et al., 2018). While misinformation is false information which may be disseminated mistakenly, the OED defines disinformation as 'the dissemination of deliberately false information' (European Parliament, 2015). Guess and Lyons (2020) define the difference between the two in the intent (Guess & Lyons, 2020). While misinformation may be unintentionally false, disinformation is purposefully created to deceive or mislead (Ciampaglia et al., 2018). According to Habermas (2003), misinformation is not perceived to be an issue as it can be analysed and critiqued within discourse, and the speaker must explain their reasoning (Habermas, 2003). This is different to disinformation which can be considered more dangerous in the sense that it is deliberately disempowering, and the speaker therefore disregards the individual who may be alienated. This communicative approach can therefore be seen as strategic (Habermas, 2003). In fact, according to Vasu et al., disinformation is 'the most onerous given its impact on national security and social cohesion' (Vasu et al., 2018). In this article, the authors present the idea that falsehoods for entertainment may also be dangerous as some believe parody to be true. The issue pertaining to this is that extremist views may be masqueraded under the guise of something else using irony to make the idea seem mainstream. Essentially, this strategy allows individuals to push their ideas on a public stage, but gives them enough leeway to retreat if met with resistance (Vasu et al., 2018). According to his work 'Disinformation: The Use of False Information', disinformation bears similarities to lying, as we can state that lies are pieces of information 'that are false, that are known to be false, and that are asserted with the intention to mislead, deceive, or confuse' (Fetzer, 2004). The article draws upon a theory.

1. Offering information when not in the position to provide such knowledge.
2. The ignoring or bypassing of relevant information which may alter the opinion, or argument.
3. Attacking someone, such as an author, on 'misleading grounds' which are irrelevant to the author.
4. Creating 'a biased impression of a specific study by simply excluding the most significant features' (Fetzer, 2004).
5. Deliberately misleading someone by only using evidence which supports the preferred argument.

Asking how the spread of misinformation threatens individuals bears similarities with questioning the harms of largescale deception. As an example, a study in 2020 reviewed the perceived credibility of various AI models in the face of shaping public perception on foreign policy (Kreps et al., 2020). The results demonstrated that people are unlikely to be able to differentiate between AI and human-generated text, thus proposing that there indeed exists a 'propensity for manipulation' in terms of users of the online space (Kreps et al., 2020). The article suggests that it is not the case that the content produced by AIs are able to outright change one's opinion or political beliefs, but rather they ignite a sense of confusion or distrust in the individual. In other words, it sows the seeds which may later grow (Weedon et al., 2017). Norris (1996) expresses the threat of people believing falsehoods as a way in which common reference points within society are broken down, thereby undermining systems (Norris, 1996).

3 AI, Bots, and the Spread of Mis- and Disinformation

3.1 Definitions of AI and ML

According to 'Aspects of AI', AI mimics activities which are deemed to be human-like, such as making rational decisions to solve problems. It is used in most sectors today and is very much embedded in our everyday lives (Neelam, 2022). AI was introduced in the 1950s by Alan Turing in his notable work 'Computing Machinery and Intelligence' which presented the idea of machines behaving with intelligence (Turing, 1950). Developing this theory more recently, Russel and Norvig conceptualised the idea of machines that can plan ahead and act on these plans independently (Russell & Norvig, 1995). According to El Naqa and Murphy, ML is a branch of AI which involves computational algorithms that are capable of learning from their environment and, therefore, imitating human behaviour (El Naqa & Murphy, 2015). In 1983, it was thought that ML had three research focuses: task orientated studies; cognitive simulation; and theoretical analysis. In his chapter 'Why Should Machines Learn?', Simon theorises that the main distinction between the learning process of humans and machines is that it generally takes far less time for a machine to learn how to perform a task, and how to do so to an extremely high standard (Simon, 1983). Consider the task of playing a game. We have seen that machines are not only able to learn the rules of a game immediately, but their ability to plan ahead means that they can improve their performance quickly to the point of becoming an expert in mere hours (Simon, 1983). However, ML is not restricted to gaming environments, and is commonly employed in a number of malicious settings, including in the fabrication of online content, as is discussed below.

3.2 Types of Disinformation and Detecting Fabricated Content

3.2.1 GAN Images

In recent years, it has been common to come across fabricated images which may be widely circulated on social media platforms. Generative Adversarial Networks (GANS) is a method in which fabricated images are machine-generated, often being realistic enough to manipulate users (Gragnaniello et al., 2021). GAN may be used to restore images and brings to the table some exciting features. However, enhanced photorealism has made the task of distinguishing false media content from authentic content a highly complex one (Gragnaniello et al., 2021). It is widely agreed that there is an urgent need for automatic tools which can detect synthetic media as often to the naked eye, false media is undetectable and often goes undiscovered (Shu et al., 2020). Over time, the presence of anomalies in images has diminished, including dissymmetry. However, studies have shown that invisible 'artificial fingerprints'

(Shu et al., 2020) can be recovered from a GAN-generated image which, as evidence, is much more difficult to eradicate completely. A number of studies focusing on GAN detection focus on training a classifier on a dataset of GAN-generated images produced by a pre-trained GAN model. The study, 'Detecting and Simulating Artifacts on GAN Fake Images' takes a slightly different approach by taking a closer look at the GAN generation pipeline, particularly at up-sampling layers which help to increase the resolution of the image (Zhang et al., 2019). This is before training the detection classifier with the frequency spectrum and not the raw RGB pixel. The results of this study demonstrate that this significantly increases the ability of the classifier to generalise (Zhang et al., 2019).

3.2.2 False Videos and Deepfakes

Technological advancements have made it possible to use generative adversarial networks to replace faces of a person in video content with the face of another person entirely. As a result, this creates content which may successfully manipulate viewers into believing that the content is authentic. Research into the impact of so-called deepfakes, as well as deepfake detection, has increased in recent years. George (2019) found that in cases where deepfakes are used to undermine politicians and political campaigns, this has the potential to have a resulting impact on the public sentiment towards the politician or political party in question (George, 2019). In terms of deepfake detection, interesting research strategies have been investigated. One such strategy is by Li et al. (2018), who sought to detect abnormalities in blinking under the assumption that this movement may appear stifled or may not appear at all in an artificial video (Li et al., 2018; Yuezun et al., 2018). This compares to studies such as that carried out by Güera & Delp (2018), who used convolutional LSTM to detect discrepancies between video frames in the facial features, for example, abnormalities in the lighting (Güera & Delp, 2018).

3.2.3 Multimodal Content

While most disinformation is based in one singular modality such as text or video, multimodal strategies are often used to make news articles more believable by having accompanying text alongside a fake video, for example. According to (Singh et al., 2020), it is essential to consider the authenticity of the content as a whole instead of each individual component, particularly given that news articles are becoming increasingly multimodal in structure. The detection of multimodal content has predominantly focused on intelligent text-processing strategies. The study of Singh et al. (2020) presents the need for a multimodal detection framework which has the capacity to review and identify abnormalities in different areas of the news item. They therefore consider a multifaceted approach to the detection of disinformation using a combination of text-based and visual features, and their relationship with the following four categories:

- Content: the topics that are identified in the item.
- Organisation: how these topics are presented to the reader or viewer.
- Emotion: this may include facial expressions present within images or video content.
- Manipulation: how the information may be distorted in favour of a particular viewpoint. The level of manipulation present in the text may be measured by excessive use of second-person pronouns (Horne & Adali, 2017), or the number of features within an image that have been tampered with.

3.3 Software Robots

Software robots, or bots for short, may be created for various purposes depending on the intent of the creator. This may be simply to automate a process on a website, such as a chatbot which directs the user to relevant documentation to help resolve a query, or to create and interact with content online. Bots may require assistance to function by a human or be completely automated. Bots often do not work alone, and a group of bots working collaboratively is known as a botnet (Burkhardt, 2017). Bots are programmed algorithms which, in this case, assist with the dissemination of information by directing web content such as articles to the newsfeeds of social media users. This is a tactic which is used by the likes of political parties and news agencies which is extremely powerful. According to Burkhardt, bots are often programmed to identify the types of content the user tends to click on and push additional content which is similar to this. In this way, a user may only be able to see articles which are already aligned with their own views or published by similar sources (Burkhardt, 2017). This information may also be directed to those within the user's network, creating confined bubbles of information. Burkhardt argues that individuals within these echo chambers may feel as though their ideologies reflect those of the majority of the population, as they no longer see information that conflicts with their opinions (Burkhardt, 2017). The benefit of bots for parties wishing to disseminate information is that they work around the clock and work much faster than humans could. In the last few years, bots have increasingly been designed to spread not just targeted information, but misinformation. What is more, humans assist bots with their duties by engaging with the content by sharing across their own networks. In fact, one study showed that 55% of users spent less than 15 seconds on a page (Pedro Baptista & Gradim, 2021). Effectively, information is disseminated by bots and users who have often not read the content. Some users would not question this content, as they prefer to consume information that is within their own system of beliefs.

3.3.1 Types of Bots

According to Himelein-Wachowiak et al., content polluters are one type of bot which seeks to 'disseminate malware and unsolicited content' (Himelein-Wachowiak et al., 2021).

Spambots Giorgi et al. (2021) present their work which discusses the increased online presence of social spambots, a class of bot which can be defined as 'a computer algorithm that automatically produces content and interacts with humans on social media, trying to emulate and possibly alter their behavior' (Giorgi et al., 2021; Ferrara et al., 2016).

Social bots Social bots are programmed to engage with content online by interacting with posts from users as well as posting content themselves, acting in the same way as human users. Studies such as those by Shao et al. (2018) present that social bots are partly responsible for the infodemic, as bots target users who may be vulnerable to manipulation and thus share content shared by the bots to their own network (Shao et al., 2018). The main difference between spambots and social bots is that social bots act in a similar fashion to human users tend to act online, whereas spambots do not wish to hide their bot-like features.

Cyborgs Cyborgs are also common on social media platforms and have been defined as 'either bot-assisted humans or human-assisted bots' (Zhang & Ghorbani, 2020). For example, a user may set up automated posts and participate every so often by engaging with other users in their network. By demonstrating both manual and automated behaviour, lines may be increasingly blurred, which increases the difficulty for a user when it comes to differentiating a bot from a human user. A study by Chu et al. (Chu et al., 2012) aimed to classify Twitter accounts into human, bot, and cyborg categories with the ultimate goal of helping users to identify and avoid malicious users online and thus increase safety online. As such, the differences are divided into three categories: behaviour, content, and account properties, before data analysis is conducted on a dataset of over 40 million tweets (Chu et al., 2012). In terms of user behaviour, this included retrieving insight into whether tweets were posted periodically or sporadically, as posting according to a timetable is often a sign of automation. The study also expressed the need to check if tweets contained spam using text patterns and the account properties of the user (Chu et al., 2012).

3.4 Social Media and the Spread of Mis- and Disinformation

Many bots do not exist to participate in malicious activity, but instead actively help users navigate information online, such as chatbots which aid users to access information 24/7 (Chen et al., 2017). Despite this, as can perhaps be demonstrated by the above studies, research surrounding bots often depicts social media as an ideal location for bot detection. Chen et al. (2017) state that approximately 9–15% of active Twitter accounts are bots. Himelein-Wachowiak et al. (2021), refer to the burden of conflicting information being provided online surrounding the Covid-19 virus as an 'infodemic' (Himelein-Wachowiak et al., 2021). It presented that 33% of the US population reported seeing misleading information regarding the pandemic on several occasions on social media platforms. The article noted that within a subset of tweet data, 25% were deemed to contain misinformation; however, there was reportedly no difference between engagement levels of tweets containing misinformation or those containing accurate and reputable information (Himelein-Wachowiak et al., 2021). This emphasises that once information has been posted online, regardless of accuracy, there is the same level of interaction with the information by human users. At their core, social networking platforms represent networks of trust, which allow users to share information to their communities of friends and family. However, infiltrating this network of trust is malicious activity such as sharing hyperlinks to users under the guise of a safe URL, the link may contain malware, or direct the user to a phishing ploy to obtain account details (Himelein-Wachowiak et al., 2021).

There is little doubt that from increased use of social media platforms emerged a more connected global community which, in many ways, has had a significant impact on the awareness of world issues, such as climate change, and allowed individuals to collaborate and build movements to change the world in positive ways. However, on the other side of the coin, it can be dangerous when this information and corresponding movements are built upon a foundation of false information. It could perhaps be argued that this ever-growing global network has been exploited by users wishing to spread propaganda and misinformation. India, for example, which has 200 million monthly WhatsApp users, has experienced several instances of vigilante justice, whereby rumours of false crimes committed by innocent victims have gone viral on WhatsApp, resulting in their murders (de Freitas Melo et al., 2019). Section 4 of this chapter will discuss the impact of the spread of mis- and disinformation in further detail. A report by Kreps et al. (2020) presents that it is perhaps the structure and business models of social media platforms which seek to constantly analyse consumer behaviour that indirectly feeds into the problem (Kreps, 2020). The report argues that ultimately, we live in a world where netizens want instant gratification, which is provided through the personalisation of the content which appears online. This data is then harnessed by social media companies to tailor what users see to keep engagement on the platform high, regardless of the user's immediate interests and disinterests (Kreps, 2020). The report also argues that it is exactly this personalisation which has now been repurposed. No longer just a marketing strategy, political organisations personalise content for users to influence opinions and decision-making. Creating content to increase confusion and political divides can be labour-intensive at best for any human aiming to consistently churn it out, thus bots are employed to complete this task efficiently.

4 The Global Impact of the Spread of Disinformation

It could be argued that, on a micro-level, the effects of coming across mis- and disinformation are often ignored, partly due to the fact that we may not recognise false content when we see it, and therefore may not register it at all. However, it is perhaps interesting to consider the wider impact the spread of mis- and disinformation has had on, not only individuals and immediate networks, but on society and the global community. Interestingly, there exists some debate as to how significant the issue of misinformation actually is in terms of fake news content, with several authors presenting the viewpoint that despite the increased attention on the topic following the 2016 US presidential election and Brexit, fake news is not particularly prevalent online when compared to the magnitude of mainstream news available to the public online, particularly via social media platforms such as Twitter and Facebook (Watts et al., 2021). One such study, ‘Measuring the news and its impact on democracy’ (2021), estimates that the low impact of disinformation is due to the overall low engagement with online news more generally, finding that in the USA, individuals are five times more likely to consume news via television rather than online, while three quarters of Americans spend less than 30 seconds reading online news sources per day (Watts et al., 2021). The study argues that a broader overview must be taken into account when considering the impact of false information, which should be considered as a wider phenomenon which is not online and includes falsehoods posted online, but also incorrect information broadcast by news channels based on inaccurate information or lies disseminated by politicians for political motives.

Perhaps contrasting this view, it could be argued that it is not necessarily the rate of consumption of disinformation that poses the risk, but rather the mere existence of these falsehoods, combined with the complexity that comes with trying to distinguish between that which is true and that which is untrue. It reflects the instability of trust, which in itself possesses the ability to create fractures in systems and increase confusion and mistrust. It could also be argued that if misinformation is, in fact, overrepresented by recent news and research, this also may create a scenario in which individuals begin to question accurate reliable content, which may also become dangerous. For example, the term ‘fake news’ was used by former US President Trump and his supporters to describe accurate news stories which reflected badly on him, thus creating confusion as to the meaning of the term (Brown, 2019). In contrast, in a research conducted by Brown (2019), it is argued that misinformation prevents effective decision-making of citizens (Brown, 2019). ‘The fake news audience is small and comprises a subset of the Internet’s heaviest users, while the real news audience commands a majority of the total internet audience’ (Nelson & Taneja, 2018). Effectively, the study presents that misinformation and propaganda are equally destructive. This is given that many dangerous consequences of disseminated falsehoods, such as support for the Iraq war were based on misinformation in the form of misleading statements, not necessarily outright fabrications in order to sway public opinion from one mindset to another (Nelson & Taneja, 2018). Following this theory, the significance of the malicious intent separating misinformation from disinformation may be meaningless, although this is also widely contested.

It is also argued that although statistically the audience for false news stories is relatively limited, second-hand disinformation should also be considered, which occurs as individuals increasingly turn to social media platforms to engage in online news stories (Van Duyn & Collier, 2019). Misinformation carries the ability to tarnish the reputations of individuals, businesses, and alter public decisions as well as decrease the trust the public has in the media more generally (Tandoc et al., 2019; Diehl et al., 2018). According to Arthur C. Clarkes' Third Law, 'any sufficiently advanced technology is indistinguishable from magic' (Clarke, 1973). This goes to say that in a modern, technologically advanced world, it is more complex than ever to look at information and distinguish between truth and falsehoods. It could be argued that with the increase in both mis- and disinformation available online, social media platforms should be subject to the same levels of scepticism and awareness as other sources of information in everyday life, such as news agencies (Diehl et al., 2018). In turn, perceiving the online world in this way may allow government agencies to enforce further security measures to protect individuals online to the same degree as one would be protected in the offline world. Moreover, it is perhaps interesting to consider that the fundamental democratic principles which embodied the rise of the internet, and which allowed netizens to freely construct belief systems on a never-ending spectrum of subjects are now being threatened by the consequences of those very same online freedoms.

5 Mitigating the Impact of Mis- and Disinformation

Watts et al. (2021) argue that to improve the current situation and intercept and prevent the consequences of the spread of misinformation on democracy, a mass collaboration model is required, whereby insights could be reached quicker and with the combined skillsets and insights brought by a multidisciplinary pool of researchers (Watts et al., 2021). Equally, the author presents the idea of the power of a robust system for disseminating the insights and results to third party stakeholders who do not belong to the academic community. In this way, an open research agenda should be shared to ensure that data is readily available for researchers to use. This is one such way to solve the issue of the spread of misinformation on a much wider scale than is currently being explored. It is necessary to collaborate across multiple disciplines because, as it currently stands, research in the field of mis- and disinformation is carried out from a variety of perspectives and under different frameworks, making it difficult to obtain and contrast the research in a meaningful way. The importance of a collaborative effort to successfully tackle the spread of disinformation is shared by the EU Commission's 2018 report (updated in 2020) which sets out actions, including enhanced collaboration with both industry and social media platforms to tackle disinformation.

This includes a call for social media companies to actively: ‘Close down fake accounts active on their service, identify automated bots and label them accordingly, and collaborate with the national audio-visual regulators and independent fact-chequers and researchers to detect and flag disinformation campaigns’ (de Cock Buning, 2018). It has also been argued that increased education on the subject of misinformation and improved media literacy is required for internet users to reaffirm the importance of awareness of the issue and thus prepare users for not only the dangers of certain forms of misinformation, but also how they may be identified (de Cock Buning, 2018). Awareness for users may include methods in which individuals can mitigate misinformation through their actions when they see a post containing false information, such as not engaging with the content, reporting and blocking the user, and instead sharing official advice to counter the false content (Center for Countering Digital Hate, 2022a). Perhaps another way to increase awareness and to mitigate the issue is to apply further pressure on social media platforms to increase expenditure on security mechanisms to improve how they are able to tackle and detect disinformation online and thus protect users (Pourghomi et al., 2017).

Social media as an ideology is based on the premise of sharing information with multiple users within their network, with the ability to ‘go viral’, i.e. allow content to be easily and rapidly shared across networks, thus provoking large-scale engagement with the content (Berger & Milkman, 2013). It has frequently been reported that social media companies could do more to mitigate the risks associated with the spread of misinformation. For example, stricter disabling of the sharing mechanism, and increased budgeting and resources for user security. According to an internal memo from Facebook, which was disclosed by the New York Times (The New York Times, 2021), the company is aware that if they were to do nothing, the nature of the algorithms that largely contribute to the make-up of the platform promote the spread of misinformation. Facebook have previously addressed that there are indeed a select group of stakeholders that would require to take on more responsibility in terms of providing solutions (Weedon et al., 2017). In addition, Mark Zuckerberg later announced that Facebook have 40,000 members of staff who work in security of the platform (Facebook, 2022). It could be argued, however, that with 2.9 billion monthly users on the site (Dean, 2022), and therefore one staff member assigned to tackle the content of over 72,500 users, this is insufficient. Social media platforms have made concerted efforts to publish strategies on how to tackle the disinformation on topics such as climate change and COVID-19, where falsehoods are commonly shared online and have the potential to be damaging to public health. One method involves assigning a warning to a post if it is deemed to contain false information. However, according to a study carried out by the Center for Countering Digital Hate (CCDH), although Facebook began tagging posts with information labels and directing users to its Climate Science Information Center, 50.5% of the most popular posts in the sample, posted by far-reaching climate denial news websites, were not tagged (Center for Countering Digital Hate, 2022b).

It is argued that this may lead to the ‘implied truth effect’, which is described as ‘whereby false headlines that fail to get tagged are considered validated and thus are seen as more accurate’ (Pennycook et al., 2019). The study conducted by Pennycook et al. considered that the implied truth effect may be mitigated by removing any doubt as to whether a headline had been verified by labelling them as to their verification status. NATO researchers claimed that ‘Overall social media companies are experiencing significant challenges in countering the malicious use of their platforms’ (Choraś et al., 2021). As such, it could perhaps be argued that tighter restrictions placed on the information that can be shared would be welcome to mitigate sharing of falsehoods on the platform. However, this is complex given that the current system demonstrates a time lag between the content being flagged by a user and the checks that go on beneath the surface (Pourghomi et al., 2017). In addition, it must be acknowledged that the International Grand Committee (IGC) previously concluded that social media companies must not be held solely responsible for acting on the spread of falsehoods on respective platforms (Choraś et al., 2021). It is argued that although Facebook has taken a step in the right direction by implementing the likes of the right-click authenticate process, there are many failings which should be addressed (Pennycook et al., 2019). One way to do this could be to create an advanced ML model across social media platforms which would work to tackle the spread of misinformation on all platforms, instead of specifically one. By focusing efforts of the top companies on one overarching solution, this would be effective to protect users globally.

6 Conclusion

This chapter has presented that there is undoubtedly a complex relationship between falsehoods disseminated through AI methods such as bots, and humans who engage in content and disseminate it within their networks. As such, it appears that mis- and disinformation are interlinked with the online world, which alarmingly may then be spread further by humans offline. With this in mind, it is therefore evident that the issues posed by the spread of both mis- and disinformation are multifaceted and adapting rapidly in correlation with advancements in technology. Thus, this chapter cannot propose to solve these issues outright, but instead merely highlight recommendations to be considered by both academic communities and industry.

This chapter firstly considered the meaning of truth in a post-truth world, exploring various theories pertaining to what constitutes information in its ideal form, before delving into the definitions and variance in the terms ‘misinformation’ and ‘disinformation’ in an attempt to expel any discrepancies in their use. Following on from this, the role of AI in the dissemination of falsehoods was explored in Sect. 3, by discussing the methods of fabricating content and how GAN images, deepfakes, and false multimodal content are being identified and studied. The use and types of bots predominantly used online were explored, before contextualising their usage on social media platforms and how humans interacting with content disseminated by bots widens their reach and potential impact. Section 4 then considered the impacts of the spread of mis- and disinformation, including the debate as to the extent of its impact on democracy, and finally, Sect. 5 outlined how the impacts could potentially be reduced through a mass-collaboration framework which would allow research across multiple disciplines to be shared and approached on a wider platform. It is also considered that the burden of responsibility placed on social media companies could perhaps be increased, and that increased media literacy as well as further research into how AI itself could play its part in fighting the spread of falsehoods online could be highly beneficial. Despite the robust body of research which has already and is currently being conducted on the subject, it is crucial that further research is carried out and industry enforces a much stricter plan of action on how to tackle the spread of mis- and disinformation to reduce potential harm to users and thus, wider society.

References

Arendt, H. (1967). *Truth and politics*. The New Yorker.

Bakardjieva, M. (2005). *Internet society: The internet in everyday life*. SAGE Publications.

Benedek, W., & Kettemann, M. C. (2020). *Freedom of expression and the internet: Updated and revised* (2nd ed.). Council of Europe.

Berger, J., & Milkman, K. L. (2013). Emotion and virality: What makes online content go viral? *Insights, 5*, 18–23.

Brown, E. (2019). Propaganda, misinformation, and the epistemic value of democracy. *A Journal of Politics and Society, 30*, 194–218.

Buckland, M. K. (1991). Information as thing. *Journal of the American Society for Information Science, 42*, 351–360.

Bufacchi, V. (2021). Truth, lies and tweets: A consensus theory of post-truth. *Philosophy & Social Criticism, 47*, 347–361.

Burkhardt, J. M. (2017). *Combating fake news in the digital age. Library technology reports* (pp. 5–33). ALA TechSource.

Carsten Stahl, B. (2006). On the difference or equality of information, misinformation, and disinformation: A critical research perspective. *Informing Science Journal, 9*, 83–96.

Center for Countering Digital Hate. (2022a, July 3). *Home: Center for countering digital hate*. Retrieved from Center for Countering Digital Hate Website: https://www.counterhate.com/

Center for Countering Digital Hate. (2022b, February 23). *Facebook failing to flag harmful climate misinformation, new research finds*. Retrieved from Center for Countering Digital Hate Website: https://www.counterhate.com/post/facebook-failing-to-flag-harmful-climate-misinformation-new-research-finds

Chen, Z., Tanash, R. S., Stoll, R., & Subramanian, D. (2017). Hunting malicious bots on Twitter: An unsupervised approach. In *International conference on social informatics* (pp. 501–510). Springer.

Choraś, M., Demestichas, K., Giełczyk, A., Herrero, Á., Ksieniewicz, P., Remoundou, K., . . . Woźniak, M. (2021). Advanced ML techniques for fake news (online disinformation) detection: A systematic mapping study. *Applied Soft Computing, 101*, 1–22.

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing, 9*, 811–824.

Ciampaglia, G. L., Mantzarlis, A., Maus, G., & Menczer, F. (2018). Research challenges of digital misinformation: Toward a trustworthy web. *AI Magazine, 39*, 65–74.

Clarke, A. C. (1973). *Profiles of the future*. Harper & Row Publishers. Retrieved from NewScientist: https://www.newscientist.com/definition/clarkes-three-laws/

#:~:text=But%20perhaps%20the%20best%20known%20of%20Clarke's%20three,"Any%20sufficiently%20advanced%20technology%20is%20indistinguishable%20from%20magic"

de Cock Buning, M. (2018). *A multi-dimensional approach to disinformation: Report of the independent high level group on fake news and online disinformation*. Publications Office of the European Union.

de Freitas Melo, P., Coimbra, V., Garimella, K., Vaz de Melo, P. O., & Benevenuto, F. C. (2019). Can WhatsApp counter misinformation by limiting message forwarding? In *International conference on complex networks and their applications* (pp. 372–384). Springer.

Dean, B. (2022, January 5). *Facebook demographic statistics: How many people use Facebook in 2022?* Retrieved from Backlinko Website: https://backlinko.com/facebook-users

Diehl, T., Barnidge, M., & de Zuniga, G. (2018). Multi-platform news use and political participation across age groups: Toward a valid metric of platform diversity and its effects. *Journalism & Mass Communication Quarterly, 96*, 428–451.

El Naqa, I., & Murphy, M. J. (2015). What is ML? In I. El Naqa, M. J. Murphy, & R. Li (Eds.), *ML in radiation oncology* (pp. 3–11). Springer.

European Parliament. (2015, November). *Understanding propaganda and disinformation.* Retrieved from European Parliament Website: https://www.europarl.europa.eu/RegData/etudes/ATAG/2015/571332/EPRS_ATA(2015)571332_EN.pdf

Facebook. (2022, March 7). *Promoting safety and expression*. Retrieved from About Facebook Website: https://about.facebook.com/actions/promoting-safety-and-expression/

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. In *Communications of the ACM* (pp. 96–104). ACM.

Fetzer, J. H. (2004). Disinformation: The use of false information. *Minds and Machines, 14*, 231–240.

Floridi, L. (2004). Outline of a theory of strongly semantic information. *Minds and Machines, 14*, 197–221.

Floridi, L. (2006). The logic of being informed. *Logique et Analyse, 49*, 433–460.

George, S. (2019, June 13). *'Deepfakes' called new election threat, with no easy fix*. Retrieved from AP News: https://apnews.com/article/nancy-pelosi-elections-artificial-intelligence-politics-technology-4b8ec588bf5047a981bb6f7ac4acb5a7

Giorgi, S., Ungar, L., & Schwartz, H. A. (2021). Characterizing social spambots by their human traits. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 5148–5158.

Gragnaniello, D., Cozzolino, D., Marra, F., Poggi, G., & Verdoliva, L. (2021). Are GAN generated images easy to detect? A critical analysis of the state-of-the-art. In *2021 IEEE international conference on multimedia and expo (ICME)* (pp. 1–6). IEEE.

Güera, D., & Delp, E. (2018). Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)* (pp. 1–6). IEEE.

Guess, A., & Lyons, B. (2020). Misinformation, disinformation and online propaganda. In N. Persily & J. A. Tucker (Eds.), *Social media and democracy* (pp. 10–33). Cambridge University Press.

Habermas, J. (1984). *The theory of communicative action: Reason and the rationalization of society. Reason and the rationalization of society*. Wiley.

Habermas, J. (2003). *Truth and justification*. Polity Press.

Himelein-Wachowiak, M., Giorgi, S., Devoto, A., Rahman, M., Ungar, L., Schwartz, H. A., ... Curtis, B. (2021). Bots and misinformation spread on social media: Implications for COVID-19. *Journal of Medical Internet Research, 23*(5), e26933.

Horne, B. D., & Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *arXiv:1703.09398*.

Kreps, S. (2020). *The role of technology in online misinformation*. Brookings.

Kreps, S., McCain, R. M., & Brundage, M. (2020). All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science, 9*, 104–117.

Lazer, D., Baum, M., Benkler, Y., Berinsky, A., Greenhill, K., & Menczer, F. (2018). The science of fake news: Addressing fake news requires a multidisciplinary effort. *Science, 359*, 1094–1096.

Li, Y., Chang, M.-C., & Lyu, S. (2018). In ictu oculi: Exposing AI created fake videos by detecting eye blinking. In *2018 IEEE international workshop on information forensics and security (WIFS)*. IEEE.

Mingers, J. (2001). Embodying information systems: The contribution of phenomenology. *Information and Organization, 11*, 103–128.

Neelam, M. (2022). Aspects of AI. In J. Karthikeyan, T. Su Hie, & N. Yu Jin (Eds.), *Learning outcomes of classroom research* (pp. 250–256). L'Ordine Nuovo.

Nelson, J. L., & Taneja, H. (2018). The small, disloyal fake news audience: The role of audience availability in fake news consumption. *New Media & Society, 20*, 3720–3737.

Norris, A. (1996). Arendt, Kant, and the politics of common sense. *Polity, 29*, 165–191.

Pedro Baptista, J., & Gradim, A. (2021). "Brave new world" of fake news: How it works. *Journal of the European Institute for Communication and Culture, 28*, 426–442.

Pennycook, G., Bear, A., Collins, E., & Gertler Rand, D. (2019). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science, 66*, 4944–4957.

Pourghomi, P., Safieddine, F., Masri, W., & Dordevic, M. (2017). How to stop spread of misinformation on social media: Facebook plans vs. right-click authenticate approach. In *2017 international conference on engineering & MIS (ICEMIS)* (pp. 1–8). IEEE.

Russell, S., & Norvig, P. (1995). *AI: A modern approach*. Prentice-Hall.

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C. F., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications, 9*(1), 4787.

Shu, K., Bhattacharjee, A., Alatawi, F., Nazer, T. H., Ding, K., Karami, M., & Liu, H. (2020). Combating disinformation in a social media age. *WIREs Data Mining and Knowledge Discovery, 10*, 1–23.

Simon, H. A. (1983). Why should machines learn? In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *ML: An AI approach* (pp. 25–37). Elsevier Inc.

Singh, V. K., Ghosh, I., & Sonagara, D. (2020). Detecting fake news stories via multimodal analysis. *Journal of the Association for Information Science and Technology, 72*, 3–17.

Tandoc, E. C., Jr. (2019). The facts of fake news: A research review. *Sociology Compass, 13*(2), 1–10.

Tandoc, E. C., Lim, D., & Ling, R. (2019). Diffusion of disinformation: How social media users respond to fake news and why. *Journalism, 21*, 381–398.

The New York Times. (2021, October 25). *Facebook wrestles with the features it used to define social networking*. Retrieved from The New York Times Website: https://www.nytimes.com/2021/10/25/technology/facebook-like-share-buttons.html

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*, 433–460.

Van Duyn, E., & Collier, J. (2019). Priming and fake news: The effects of elite discourse on evaluations of news media. *Mass Communication and Society, 22*, 29–48.

Vasu, N., Ang, B., Teo, T.-A., Jayakumar, S., Faizal, M., & Ahuja, J. (2018). *Fake news: National security in the post-trust era*. S. Rajaratnam School of International Studies.

Watts, D. J., Rothschild, D. M., & Mobius, M. (2021). Measuring the news and its impact on democracy. In *Advancing the science and practice of science communication: Misinformation about science in the public sphere* (pp. 1–6). PNAS.

Webster, F. (2014). *Theories of the information society*. Routledge.

Weedon, J., Nuland, W., & Stamos, A. (2017, April 27). *Information operations and Facebook*. Retrieved from About Facebook Website: https://about.fb.com/br/wp-content/uploads/sites/3/2017/09/facebook-and-information-operations-v1.pdf

Yuezun, L., Chang, M.-C., & Lyu, S. (2018). Exposing AI generated fake face videos by detecting eye blinking. In *IEEE international workshop on information forensics and security (WIFS)* (pp. 1–7). IEEE.

Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management, 57*, 2–26.

Zhang, X., Karaman, S., & Shih-Fu, C. (2019). Detecting and simulating artifacts in GAN fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)* (pp. 1–6). IEEE.