CPSC 610: Topics in Computer Science and Law

Final Project

Regulating Falsehoods:

Lessons from Singapore's Protection from Online Falsehoods and Manipulation Act

Alden Tan

Professor Joan Feigenbaum

December 17, 2021

Table of Contents

Chapter 1: Background on Falsehoods

1.1. The rise of falsehoods

In recent years, the internet and social media have revolutionized the way we produce, communicate and distribute information. At the same time, the issue of falsehoods has gained prominence due to the ease and speed by which false information can be generated and spread. Social media platforms and other mobile and web applications provide intuitive and accessible editing and publishing technology that allows virtually anyone to create and communicate information. Furthermore, information is disseminated at much faster speeds with the prevalence of mobile phones and an accelerated news cycle. Information is often transmitted in real-time between peers rather than through traditional mainstream media, and this horizontal rather than top-down exchange of information has reduced the likelihood of any particular piece of information being checked or challenged for its veracity and accuracy.[1]

These trends have led to a rise in the prevalence of falsehoods in the digital sphere. During the 2020 Presidential Election, politicians like Donald Trump were able to spread falsehoods about election proceedings and outcomes, which led to serious consequences like the Capitol riot. During the COVID-19 pandemic, falsehoods about the effectiveness of vaccines and masks were spread across the world, which misled the public and hindered the ability of governments to respond effectively to the pandemic. Some falsehoods arose as a result of manipulation and interference by foreign actors. From 2014 to 2020, Russian operatives were found to have flooded the internet with false information in seven languages and across 300 social media platforms.[2] These posts

[1] Wardle, C. and Derakhshan, H., 2017. Information Disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe*.
[2] Brewster, T., 2021. *2,500 Posts, 300 Platforms, 6 Years: A Huge But Mysterious Pro-Russia Disinformation Campaign Is Exposed*. Forbes.

sought to spread pro-Russian propaganda around the world, such as by propagating fake tweets from US elected officials and conspiracy theories about COVID-19.[3]

This paper will study the regulation of falsehoods, with particular focus on Singapore's Protection from Online Falsehoods and Manipulation Act (POFMA), which is commonly regarded as one of the most comprehensive laws in the world that tackles falsehoods. This paper will examine POFMA as a case study on how to regulate falsehoods, evaluate its costs, benefits and challenges in the two years since it has come into effect, and provide policy recommendations to improve its effectiveness and legitimacy.

1.2. Definitions

Wardle and Derakhshan (2017) provide a framework to classify different types of information disorder, which is useful in defining what we mean by a falsehood in this paper. There are three types of information disorder: misinformation, disinformation and mal-information. As shown in Figure 1, these types differ based on whether they are false and harmful. Misinformation refers to information that is false but not harmful. Mal-information refers to information that is harmful but not false, such as genuine information designed to stay private that has been leaked into the public sphere. Disinformation refers to information that is both false and harmful. In this paper, the term "falsehoods" refers to disinformation.

[3] Brewster, T., 2021. *2,500 Posts, 300 Platforms, 6 Years: A Huge But Mysterious Pro-Russia Disinformation Campaign Is Exposed*. Forbes.

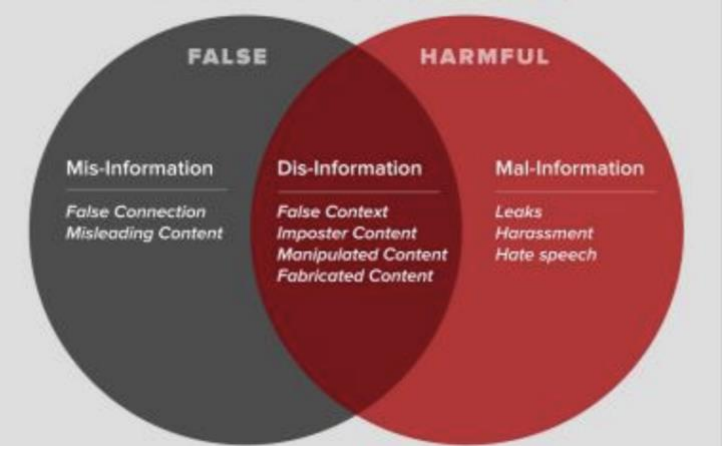

Figure 1: Types of information disorder (Wardle and Derakhshan, 2017)

We will avoid the term "fake news" because it is imprecise and inadequate to capture the complex phenomena of information disorder. The term "fake news" has been used to describe a number of different phenomena, including satire, parody, fabrication, advertising and propaganda, and it therefore lacks the desired precision for this paper.[4] Furthermore, the term has been co-opted by politicians around the world to describe information they disagree with, which further blurs the definition of "fake news".[5]

1.3. Digital technologies used to spread falsehoods

The phenomenon of falsehoods is not new, but modern technology has made the generation and dissemination of falsehoods easier, cheaper and more profitable.[6] Today, anyone is able to create and spread falsehoods that can potentially gain traction and virality, and this ability is not

[4] Tandoc, E., Lim, Z. and Ling, R., 2017. Defining "Fake News": A Typology of scholarly definitions. *Digital Journalism*, 6(2), pp.137-153.
[5] Wardle, C. and Derakhshan, H., 2017. Information Disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe*.
[6] Ibid.

just limited to well-resourced states and organizations. The key digital technologies that have enabled the rise of falsehoods are described below.

Social media platforms have been a primary medium through which falsehoods are spread in recent times. Anyone can create a social media account and make posts containing falsehoods that can potentially reach a wide audience. Due to lax or non-existent verification requirements, malicious actors can easily create inauthentic accounts that are then used to artificially amplify online falsehoods via basic social media functions like sharing, liking, re-tweeting, hyperlinking and hash-tagging. In addition, these actors can create accounts that impersonate a well-known government official, celebrity or organization, and use these accounts to make posts containing falsehoods, which may then achieve virality due to the popularity of the person or organization they are impersonating. For example, a troll Twitter account that impersonated the Tennessee Republican Party during the 2016 US Presidential Election had over 150,000 followers, which was much larger than the 13,800 followers of the Tennessee Republican Party's real Twitter account.[7] These fake social media accounts may be run by humans, known as "trolls", or by algorithms, known as "bots". Bots are automated social media accounts that act like real users and post content without human intervention. Human actors can also use bot technology to help them post faster and more frequently, and such human-machine collaboration is known as "cyborgs". Bots and cyborgs can be difficult to detect because they have short lifespans, and new bots can be created quickly. Cyborgs often display elements of genuine human interaction, which can make it even more difficult to detect such accounts.[8]

[7] "Sean Edgett's Answers to Questions for the Record", Senate Committee on the Judiciary, Subcommittee on Crime and Terrorism Hearing on Extremist Content and Russian Disinformation Online: Working to Find Solutions, October 31, 2017 (19 January 2018), pp 16 – 17.
[8] Thirteenth Parliament of Singapore, 2018. Report of the Select Committee on Deliberate Online Falsehoods – Causes, Consequences and Countermeasures.

Targeted advertising is another tool that can be used to effectively amplify falsehoods. Online platforms like Google and Facebook offer cost-effective and simple-to-use targeted advertising options that anyone can use to communicate with specific users, based on targeting options provided by the advertising platform, such as user demographics, location, preferences and recent online activity.[9] Targeted advertising can enable malicious actors to target their falsehoods towards groups that are predisposed to believe them, which network theorists have shown helps the falsehoods spread further.[10] During the 2016 US Presidential Election, a Russian troll factory was able to spread Facebook advertisements to 126 million Americans with a cost of just US$100,000, including advertisements that were targeted at specific user profiles.[11] When falsehoods that are amplified by these methods gain popularity, they are then given a further boost by social media algorithms, which are designed to automatically promote popular posts. Therefore, social media algorithms can also contribute to the virality of falsehoods.

An increasingly popular type of falsehoods is known as "deepfakes", which are synthetic images, videos or audio recordings that are replaced with someone else's likeness. There exists user-friendly computer software that can transpose a picture of a person onto an existing video to create a fake video. One can also superimpose words and expressions onto the face or mouth of a person in a video such that it appears that they are saying those words. This has been done to politicians for the purpose of influencing elections.[12] Adobe's Project VoCo allows users to input a 10- to 20-minute clip of someone's voice into the application, which will then be able to dictate

[9] Meta. 2021. *Help your ads find the people who will love your business.*
[10] Buchanan, M., 2017. *Why Fake News Spreads So Fast on Facebook: Ad Technology has weaponised disinformation*.
[11] Cameron, D. and Conger, K., 2017. *Here Are 14 Russian Ads That Ran on Facebook During The 2016 Election*. Gizmodo.
[12] Khalaf, R., 2018. *If you thought fake news was a problem, wait for 'deepfakes'*. Financial Times.

any words given to the application in that person's voice.[13] This can clearly be used to generate audio deepfakes.

Chapter 2: The Protection from Online Falsehoods and Manipulation Act

2.1 Background

In response to the increasing trend of falsehoods and foreign manipulation observed in many countries, the Singapore Parliament convened a Select Committee on Deliberate Online Falsehoods in January 2018 to study and recommend how Singapore should respond to the problem of online falsehoods.[14] The Committee presented its report in September 2018. Using the Committee's findings and recommendations, POFMA was tabled to Parliament and subsequently passed into law in May 2019. POFMA came into effect in Singapore in October 2019.

POFMA, more commonly known as the "fake news law" in Singapore, seeks to prevent the electronic communication of "false statements of fact" that compromise the "public interest". It also seeks to counteract the effects of such communication and to safeguard against the use of online accounts for such communication. The POFMA legislation contains a total of 9 Parts, and this section will elaborate on the key features of POFMA.

2.2 Prohibition on falsehoods

Part 2 of POFMA criminalizes the communication of a statement in Singapore while knowing or having reason to believe that the statement is a false statement of fact, and that such

[13] Wardle, C. and Derakhshan, H., 2017. Information Disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe*.

[14] Thirteenth Parliament of Singapore, 2018. Report of the Select Committee on Deliberate Online Falsehoods – Causes, Consequences and Countermeasures.

communication is prejudicial to the public interest. In the legislation, communication of a false statement is prejudicial to the public interest when the communication is likely to harm the security, public health, safety, tranquility, finances or international relations of Singapore, influence the outcome of an election or referendum, incite feelings of hatred or ill-will between different groups, or diminish public confidence in the performance of any function of the government.[15]

Offenders are subject to a fine not exceeding S$50,000, or imprisonment not exceeding 5 years, or both.[16] If the offender is not an individual, such as in the case of an organization, the penalty is increased to a fine not exceeding S$500,000. If an inauthentic online account or bot is used for the communication for the purposes of accelerating the communication, the maximum penalties for both the fine and imprisonment are doubled. Importantly, any internet intermediary service or telecommunication service is excluded from this offence if the falsehood is communicated in the course of its provision of its services, because it is technically unreasonable to expect these intermediaries to verify the truth of statements made by their users in the context of the speed and volume of communication that exists in today's internet and media environment.

2.3 Directions to persons who communicate falsehoods

POFMA enables the executive branch of government, and in particular, the Ministers of any government Ministry, to issue Directions to persons or publishers, who are legally obliged to comply with such orders. These Directions are the primary tools the government has to tackle the problem of falsehoods. Part 3 of POFMA contains Directions that can be issued to a person who has communicated a falsehood.

[15] *Protection from Online Falsehoods and Manipulation Act 2019.*
[16] The exchange rate between Singapore Dollars (S$) and US Dollars is 1 Singapore Dollar = 0.73 US Dollars, as of December 12, 2021, according to Google Finance.

If a person has communicated a false statement of fact in Singapore, and any Minister is of the opinion that issuing a Direction is in the public interest, then the Minister may issue a Direction in relation to the false statement. The most common type of Direction that has been used is a Correction Direction, which requires that the person who communicated the falsehood puts up a notice saying that the subject statement contains a falsehood, and containing a link to a statement of facts provided by the government that clarifies why the subject statement was false. Usually, in the case of a social media post or web article, this notice must be placed at the top of the post or article, such that readers will be able to see the notice before reading the rest of the text.

For serious falsehoods, the Minister may issue a Stop Communication Direction, which requires the person to stop communicating the statement or any substantially similar statement. If a person fails to comply with a Direction, the Minister may require the internet service provider to disable access to the online location containing the falsehood.

2.4 Directions to internet intermediaries

Part 4 of POFMA contains Directions that can be issued to internet intermediaries whose platforms contain falsehoods that are posted by its users. A Targeted Correction Direction can be issued to an internet intermediary, requiring it to communicate a correction notice (containing similar content as notices given under Part 3 Correction Directions) to all end-users in Singapore who have accessed or will access a particular falsehood. In effect, this is the same as a Part 3 Correction Direction, just that the notice is put up by the intermediary, instead of the person who communicated the falsehood. This usually occurs if the person who was issued a Part 3 Correction Direction has failed to comply with the Direction. A Disabling Direction may also be issued, requiring the intermediary to disable access by end-users in Singapore to the falsehood. The

Minister may also issue a General Correction Direction, which requires that the correction notice be communicated to all users of that service, rather than only to those who have seen the falsehood.

Part 6 of POFMA deals with inauthentic accounts and coordinated inauthentic behavior. Under this Part, the Minister may issue an Account Restriction Direction to an internet intermediary, requiring it to disallow a specified online account from communicating any statement in Singapore, if that account is an inauthentic online account or is controlled by a bot, has been used to convey a false statement of fact, and it is in the public interest to issue the Direction.

2.5 Declaration of online locations

The Minister may declare a website as a declared online location (DOL) if there are at least three falsehoods that were the subject of a Part 3 or 4 Direction hosted on that website in the prior six months. In effect, this serves as a blacklist of websites that are known to frequently perpetuate falsehoods. When a website has been declared a DOL, it has to publish a notice informing visitors that it is a DOL, and it becomes an offence to derive monetary benefit from operating the DOL. If paid content is hosted on the DOL, the Minister may require the internet service provider to disable access to the DOL. This provision prevents websites from profiting through spreading falsehoods, which is a trend observed in other countries. The nonprofit Global Disinformation Index studied 20,000 websites that had been found to publish falsehoods, and it found that advertisement technology companies spend about $235 million annually by running advertisements on such sites.[17] Advertising giants like Google have made it easy to monetize a website based on its traffic,

[17] Melford, C., 2019. *Tracking US$235 Million in Ads on Disinformation Domains – GDI*. Global Disinformation Index.

and one method that some websites use to generate traffic is by publishing sensational falsehoods.[18]

2.6 Appeals process

A person to whom a Direction has been issued may apply to the Minister to vary or cancel the Direction. If the application is unsuccessful, the person may then appeal to the courts to challenge the Direction. The courts are the final arbiter of truth, and may set aside a Direction if it is found that the person did not communicate the statement, the statement is not a statement of fact, the statement is in fact true, or it is technically impossible to comply with the Direction.

Chapter 3: Evaluation of POFMA

3.1 Overview of usage

Between October 2019 and October 2021, POFMA was used a total of 87 times, 43 of which was targeted at COVID-19 disinformation.[19] The breakdown of POFMA usage into the different types of Directions and Orders are shown in Figure 2. Correction Directions were by far the most common type of Direction issued.

[18] Funke, D., Benkelman, S. and Tardaguila, C., 2019. *Factually: How misinformation makes money*. American Press Institute.
[19] Singapore Internet Watch, 2021. *POFMA'ed Dataset.*

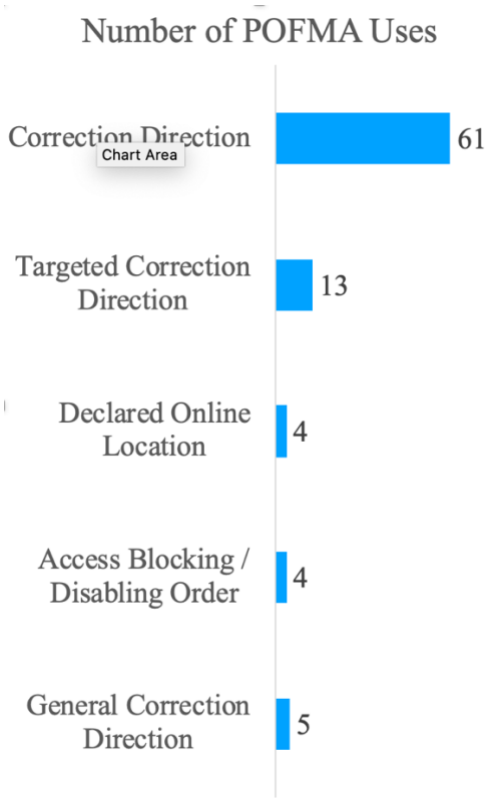

Figure 2: Types of Directions and Orders issued (Singapore Internet Watch, 2021)

The breakdown in terms of which platforms and websites hosted falsehoods that were subject to POFMA are shown in Figure 3. Notably, Facebook posts accounted for more than two-thirds of POFMA uses.

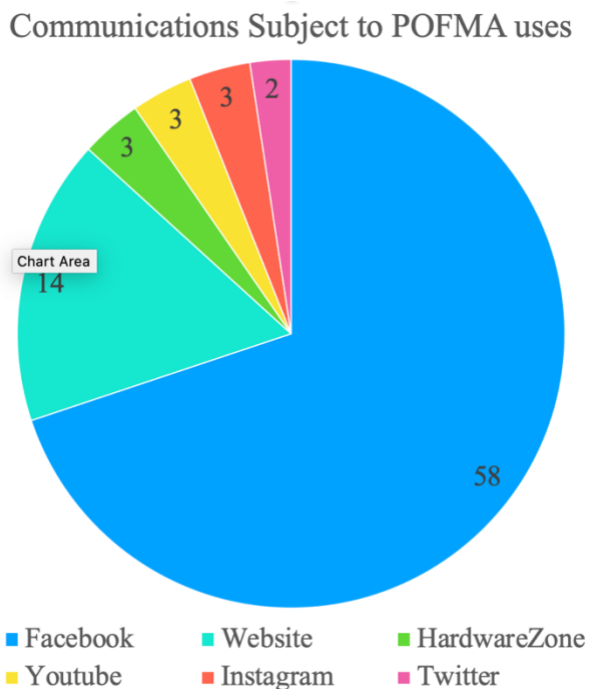

Figure 3: Breakdown according to platforms (Singapore Internet Watch, 2021)

3.2 Benefits of POFMA

Protecting the public interest by providing accurate information

A key requirement for a Minister to issue a Direction against a falsehood under POFMA is that issuing the Direction must be in the public interest. In other words, the falsehood which the Direction seeks to address must be harmful to the public interest. By requiring the publication of correction notices through Correction Directions, POFMA provides internet users in Singapore with more information to evaluate potentially misleading or untruthful content that they encounter online. This increases public awareness of the full circumstances and evidence surrounding an issue, and reduces the effectiveness of falsehoods in influencing public opinion or beliefs.

Almost half of the POFMA use cases were against COVID-19 disinformation. Examples of falsehoods that were subject to Correction Directions include claims suggesting that vaccines

were ineffective and harmful, that certain drugs like ivermectin were safe and effective for COVID-19 treatment despite minimal evidence, and that there were COVID-19 cases at particular transportation nodes (which generated significant public anxiety early on in the pandemic where there were few COVID-19 cases).[20] Clearly, falsehoods that misrepresent or lie about the effectiveness of vaccines or drugs can fuel confusion and result in people making choices with misleading or wrong information. These choices affect the public health of a country, and are especially crucial during a pandemic when the lives and livelihoods of people are at stake. Falsehoods that seek to generate public anxiety and disquiet can also adversely affect the stability and tranquility of a country. In these cases, Correction Directions were issued to present the facts to readers, so that readers may make an informed decision after contemplating the full range of information available. For example, in the case of COVID-19 vaccines, the correction notice included worldwide data on the percentage by which vaccines reduced COVID-19 cases, as well as data on adverse reactions to the vaccines.[21] The enforcement against COVID-19 disinformation via POFMA could be one of the possible reasons why Singapore has one of the highest COVID-19 vaccination rates in the world, where 96% of the eligible population has been fully vaccinated.[22]

Breaking the virality of falsehoods

POFMA enables the executive branch of government to react quickly to any falsehood that may be gaining traction, and require the author or intermediary to publish a correction notice within a short time frame, usually within 24 hours after the Direction has been issued. By allowing the

[20] Singapore Internet Watch, 2021. *POFMA'ed Dataset.*
[21] Government of Singapore, 2021. *Corrections and Clarifications regarding content about COVID-19 Vaccines in a blog post by Cheah Kit Sun.*
[22] Reuters, 2021. *Singapore close to vaccinating all eligible people against COVID-19.*

Minister to make a preliminary assessment of the falsity and public harm of a statement, rather than involving a separate review board or the courts, POFMA enables falsehoods to be quickly clarified and addressed, thus mitigating the harm that falsehoods can cause. Thereafter, aggrieved authors or publishers may then appeal to the courts for final adjudication.

Balancing deterrence and freedom of speech

POFMA exerts a deterrence effect against the publication of falsehoods because of the criminality it ascribes to the communication of falsehoods. Although most people who are issued POFMA Directions have not been charged with the offence of communicating a falsehood, the possibility of a fine or imprisonment for knowingly communicating a falsehood can serve as a deterrent to those who may consider misrepresenting the facts on a particular issue. It is important to note that POFMA only criminalizes the communication of falsehoods if the person knows the statements are false, so a person who has no knowledge of the falsity of a statement cannot be charged under POFMA, though they can still be issued a Direction. At the same time, POFMA attempts to balance regulation with freedom of speech, by allowing the offending content to remain in the public domain in most cases. When Correction Directions are issued, a correction notice has to be published, but the original statement can still remain published, and viewers are left to come to their own conclusions.

3.3 Costs of POFMA

Lack of sociological legitimacy

While the power given by POFMA to the executive branch of government enables falsehoods to be quickly addressed, the perception that the Ministers are the arbiters of truth

compromises the sociological legitimacy of the legislation. Sociological legitimacy refers to the public's willingness to respect and obey a piece of legislation.[23] Firstly, the Ministers may not have the full facts and circumstances surrounding an issue, and so a decision on the falsity and public harm of a statement made by the Ministers alone can be perceived as unfair or incomplete. Secondly, because Ministers are political office-holders, there may be a perception that POFMA is used for political purposes, such as to quell dissent by opposing voices. Although the government has repeatedly assured the public that POFMA will only be used against statements of facts, which would exclude critical opinions of the government, the process through which POFMA works may nevertheless compromise its legitimacy in the public's mind. Furthermore, there were 17 Directions that were issued to political figures from opposition parties, which may contribute to this perception of politicization, where one party has the power to unilaterally label content as false.[24]

In October 2021, the Court of Appeal ruled that when an appeal against a Direction is brought before the courts, the person who allegedly made the false statement holds the legal burden of proving that the statement is true, rather than the Minister who issued the Direction.[25] While it may be fair to expect that the person making a statement has a duty to substantiate their claims, it is also reasonable to argue that the government, with more resources and data, may be in a better position to present evidence, especially since the Minister must have considered evidence of the falsity of the statement when making the decision to issue the Direction. Therefore, sharing the

[23] Frost, A., 2019. *Academic highlight: Fallon on "Law and Legitimacy in the Supreme Court".*
[24] Singapore Internet Watch, 2021. *POFMA'ed Dataset*.
[25] Lam, L., 2021. *Falsehoods, freedom of speech and burden of proof: Key findings from Apex Court's landmark POFMA judgment.* Channel News Asia.

burden of proof between both parties may be a more reasonable approach that would strengthen the sociological legitimacy of POFMA.

Chilling effect on speech

While POFMA deters the communication of falsehoods, the flip side of the argument is that it can exert a chilling effect on speech in Singapore. Public discourse often involves communicating statements even when one does not have the full facts on an issue, and it is such communication that enables the exchange of ideas and encourages discussion on important issues of social and national importance. However, the possibility of being charged for communicating a falsehood may discourage people from speaking up about controversial and important issues where the facts may not be in the public domain. Furthermore, there is little information on how falsehoods are being collated and identified by the Minister, and so there may be a sense that the government is carrying out surveillance on social media platforms for the purposes of identifying falsehoods, which may further exert a chilling effect on speech.

Privacy concerns

POFMA creates legal obligations for internet intermediaries that give rise to privacy concerns. When a Targeted Correction Direction is issued to an internet intermediary, the intermediary is required to communicate the correction notice to users in Singapore who have accessed the content. This creates a legal obligation for intermediaries, such as social media companies and messaging services, to keep records of what users view, so that they may identify

all users who have looked at infringing material before it was labeled as a falsehood.[26] Although many intermediaries may already implement some form of tracking of user behavior, POFMA goes one step further by legally requiring that intermediaries are able to obtain and use records of user activity for the purposes of communicating correction notices whenever a Direction is issued.

3.4 Challenges of regulating online falsehoods

Personhood and identity in the digital sphere

In issuing a Direction against a falsehood, the Minister must be able to identify a person to whom the falsehood can be attributed. However, this can be challenging and sometimes impossible in the online sphere. When making posts on social media, blogs or other websites, users often use pseudonyms as their username and do not reveal their true identity. Even if there is a seemingly genuine name, the account could be impersonating someone, and there is no clear way of establishing the true identity of the person behind an online account. This occurs because social media platforms often have lax or nonexistent user verification policies, resulting in very low barriers to entry and exit. While it may be possible to track the perpetrator down based on the Internet Protocol (IP) address from which the offending post is made, the existence of easy-to-use and cheap or free Virtual Private Networks (VPN) enables perpetrators to mask their identity behind the digital veil of anonymity. Therefore, even when a Minister identifies a falsehood, there is a possibility that they are unable to attribute the falsehood to a person, and cannot issue a Correction Direction or hold the person accountable. Furthermore, even if access to an inauthentic

[26] Daskal, J., 2019. *This 'Fake News' Law Threatens Free Speech. But It Doesn't Stop There.* New York Times.

account is blocked via an Account Restriction Direction, the perpetrator can easily set up new accounts and continue perpetuating falsehoods.

Enforceability in a "cat-and-mouse" game

In a similar vein, even if a falsehood can be attributed to a person, and Directions are issued to that person, they may nevertheless set up new accounts or go to other platforms to continue communicating falsehoods. This becomes a problem when the perpetrator is outside the country and cannot be held responsible for breaching the criminal offence of knowingly communicating a falsehood. A case in point would be Mr Alex Tan, a Singaporean based in Australia who posted multiple falsehoods on his Facebook page "States Times Review".[27] He was issued Correction Directions for these falsehoods but failed to comply.[28] Furthermore, his page was designated as a Declared Online Location, but he failed to comply with the requirements as well.[29] As a result, Facebook was issued a disabling order to restrict access to Mr Tan's page in Singapore. However, Mr Tan simply set up a new page "Singapore States Times" that continued to perpetuate falsehoods. This process repeated itself two more times, with Mr Tan creating a total of four Facebook pages.[30] The falsehoods he made include the claim that Singapore had run out of face masks, that Malaysia had rejected Singapore's request for a cross-border travel agreement amidst COVID-19, and that the government had arrested someone in relation to whistleblowing.[31] Although Mr Tan would be guilty of an offence under POFMA, the fact that he operates overseas means that it is difficult to hold him accountable and prevent him from continuing his falsehoods. This example clearly

[27] Ang, M., 2019. *States Times Review founder Alex Tan refuses to comply with POFMA order, claims he's now an Australian.* Mothership.
[28] Ibid.
[29] Singapore Internet Watch, 2021. *POFMA'ed Dataset.*
[30] Ibid.
[31] Ibid.

highlights a key challenge of regulating online falsehoods, where regulators can be stuck in a cat-and-mouse game with overseas perpetrators who can simply use new online locations to continue spreading falsehoods in Singapore.

Granularity of online locations

Another challenge that POFMA faces is in defining the granularity of online locations when declaring a DOL (which acts like a blacklist of recalcitrant websites) or restricting access to a website. In the legislation, a declaration of an online location must contain the Universal Resource Locator (URL), domain name, and any other unique identifier. However, it is not clear what the granularity of such a declaration should be. In the context of a falsehood made in a Facebook post, it would be reasonable to declare as a DOL the homepage URL of the Facebook page or account responsible for the post, as well as the URLs to any posts made by the page or account. Clearly, it would be too wide a scope to declare the URL "facebook.com" as the DOL, and it would be too narrow to declare only the URL of the post containing the falsehood as the DOL. While it may be clear what the approach should be for Facebook, it is not so clear for a forum like Reddit or Hardwarezone (a Singapore-based forum owned by Singapore Press Holdings), or a personal blog or website. For example, in the case of Reddit, if there have been multiple falsehoods made by a particular user or group of users on a certain thread, should the entire thread be declared a DOL? Doing so may seem too harsh if there are many other innocent users on the thread. In the case of a personal website, there may be multiple webpages that users can navigate to using a navigation bar. If there are multiple falsehoods made across a subset of the webpages, should the URLs of all the webpages be declared as DOL? Again, it might seem too harsh to completely take down all the pages of a website, yet restricting only some parts of the

website would be ineffective in preventing the person from spreading further falsehoods. Indeed, deciding on the granularity at which action should be taken is a key challenge in the implementation and enforcement of POFMA.

Administrative effort

The administrative effort required to trawl the web in search of falsehoods is a significant challenge in the enforcement of POFMA. The speed and volume at which online communication takes place makes it immensely challenging for the government to enforce against falsehoods at scale. Of course, there may not be a need to comprehensively identify all falsehoods, because the ones that should warrant regulatory action are precisely those that gain sufficient traction such that they would be easily identified in the first place. However, the effort required to track those who have shared or reposted a falsehood, and subsequently issue them with Directions, can also represent an uphill challenge because of the scale and volume that social media operates in. Furthermore, the government will have to investigate the falsehood and prepare the facts surrounding the issue, and also monitor if Directions have been complied with over the specified period of time. The significant administrative effort required in these tasks presents an opportunity for automation that will be suggested under this paper's policy recommendations.

Legal culpability in the context of autonomous language models

Rapid improvements in natural language processing and language models have given rise to the development of autonomous language models that are able to produce coherent and seemingly genuine text output without human intervention. An example is GPT-3, a deep learning

neural network trained by OpenAI with over 175 billion model parameters.[32] The model is able to generate large amounts of realistic human text when given just a small text input, and it has performed well in a variety of context such as news articles, poetry, blog posts and dialogue.[33] However, such autonomous language models can be prone to generating falsehoods, regardless of the intention of the human user. The University of Oxford and OpenAI conducted a joint study and found that when language models were asked a series of questions, the best-performing model was truthful on only 58% of questions, falling short of human performance at 94%.[34]

Autonomous models thus pose a challenge for the regulation of online falsehoods, because it is unclear who should bear legal responsibility when it as an artificial intelligence (AI) artifact that created the falsehood. While it may be reasonable to hold accountable the person who ran the model, that person may not have intended for a falsehood to be generated in the first place. Furthermore, it is conceivable for a model to be trained and then "set loose" by a human, such as in the case of Tay, which was Microsoft's Twitter chatbot that learned racist and sexually-charged messages when it was allowed to learn through its dialogue with other users on Twitter.[35] While an autonomous bot posting falsehoods may not gain traction initially, it is possible for a bot, under the guise of an authentic user profile, to gain credibility over time and eventually cause harm through falsehoods. The question of legal culpability in such scenarios would be complicated because of the autonomous or semi-autonomous nature of today's modern language models.

[32] Schmelzer, R., 2021. *GPT-3.* SearchEnterpriseAI.
[33] Ibid.
[34] Wiggers, K., 2021. *Falsehoods more likely with large language models.* Venture Beat.
[35] Schwartz, O., 2019. *In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation.* IEEE Spectrum.

Chapter 4: Policy Recommendations

4.1 Recommendations to improve transparency and sociological legitimacy

Having reviewed the benefits and costs of POFMA, as well as the challenges of regulating online falsehoods, we now turn our attention to policy recommendations to improve the legitimacy and effectiveness of both POFMA and the regulation of falsehoods in general.

Independent review board

Rather than having the Ministers be the initial arbiters of truth, I propose the formation of an independent review board that is responsible for fact-checking statements and issuing Directions under POFMA. The board should comprise experts from different fields that will enable it to effectively carry out its fact-checking duties on a wide variety of topics. The board would receive complaints of false statements from the government and members of the public, after which it would review the alleged falsehood and determine whether any Direction should be issued under POFMA. This recommendation introduces more transparency into the POFMA decision-making process and prevents the perception of politicization by the incumbent government.[36]

Legal requirement to explain decisions

The POFMA legislation should include a legal requirement for the government to explain its decision when issuing a Direction. This explanation should include why the statement was deemed as false, and how it is prejudicial to the public interest. While the government currently already explains its decisions when issuing Directions under POFMA, there should be a legal

[36] Mahmud, A., 2020. *In Focus: Has POFMA been effective? A look at the fake news law, 1 year since it kicked in.* Channel News Asia.

requirement encoded in law to ensure that this practice is carried out under all circumstances to preserve public trust and confidence in POFMA.

Sharing burden of proof during appeals process

When an appeal against a Direction is considered by the courts, both the defendant who allegedly made the falsehood and the government should share the burden of proving whether the statement is true or false. Currently, the burden of proof lies with the defendant, but the government usually has more data and information that would shed light on the issue. In the process of deciding whether to issue the Direction, the government would already have considered evidence of whether the statement is false. Therefore, the government should share the burden of proof by virtue of its greater domain knowledge and resources to obtain the relevant information.[37] However, the government should not bear the full burden of proof because this may open the way for abuse, where a person can make a false allegation and then go to the courts to extract information from the government.[38]

4.2 Recommendations to enhance effectiveness of the law

Strategic silence

In circumstances where a falsehood has not gained public traction, and is unlikely to gain significant public attention when left unresolved, I propose that the government adopt a policy of strategic silence where the falsehood should not be addressed via POFMA. When the threshold

[37] Tham Y., 2020. *Judge wrong in placing burden of proof on Government in Pofma cases: AGC.* The Straits Times.
[38] Ibid.

for legal intervention is set too low, POFMA may not only suppress legitimate expression, but it may also backfire by giving media attention to the falsehood, thereby increasing its reach.[39]

Nurturing a fact-checking coalition

Fact-checking and combating falsehoods should be a whole-of-society endeavor, rather than one unilaterally carried out by the government. Therefore, the government should nurture a coalition of fact-checking organizations, news outlets and industry partners to investigate and debunk falsehoods. This not only reduces the administrative burden on the government in regulating falsehoods, but it also serves to improve the sociological legitimacy of the fact-checking process and better engage all members of society in the collective fight against falsehoods.

4.3 Technical recommendations

Crowdsourcing and distributed moderation

In section 4.1, I proposed that members of the public be allowed to submit complaints on falsehoods to an independent review board which will decide on whether a Direction should be issued under POFMA. However, this may generate a large volume of statements that the board has to process. Therefore, I also propose a crowdsourcing and distributed moderation mechanism which will help to narrow down the volume of statements and allow the board to only focus on those which are most likely to be falsehoods that harm the public interest.

Drawing inspiration from Reddit's moderator system, I suggest that moderators be recruited to aid the board in its work. Any citizen can become a moderator. Moderators will review

[39] George, C., 2019. *Meeting the challenge of hate propaganda*. Written Representation to the Select Committee on Deliberate Online Falsehoods.

the falsehoods that have been submitted by the public, and assign a rank from 1 to 5 on each of the following measures: the falsity of the statement, the public harm brought about by the statement, and the extent to which the statement is a statement of fact. The rankings will be aggregated into a score (with a higher score indicating a higher likelihood that the statement is a harmful and false statement of fact), and the board will only review statements that have received a certain score and above. Furthermore, after the board has made its decision, moderators will receive "experience points" if their ranking matches the board's decision (for example, if the board decided to issue a Direction, and the moderator correctly recommended a high rank for the measures); otherwise, they will lose "experience points". Moderators who have more "experience points" are subsequently given more weight when the rankings from all moderators are aggregated.

This system brings a few benefits. Firstly, it reduces the workload of the board by filtering out statements that are unlikely to be false or harmful to the public interest. Secondly, moderators are incentivized to make an accurate determination of the falsity and public harm of the statements they review because a determination that matches with the board's eventual decision would give them more "experience points". Thirdly, moderators who have proven their credibility and gained more "experience points" are given more weight in determining which statements get reviewed by the board. However, it is important to ensure that the group of moderators appropriately reflects the diversity of society, and that no moderator is given too much weight in deciding what statements are surfaced to the board. Therefore, the marginal increase in the weight given to a moderator should be decreasing in the number of "experience points". Put another way, as a moderator accumulates a higher number of "experience points", the weight assigned to their ranking will increase by a lesser amount. Furthermore, some statements with low relevance scores

will be chosen at random to be surfaced to the committee, so that any systematic bias of issue selection due to any possible skewed composition of moderators can be prevented.

It is also possible to group moderators into different areas of expertise, and assign more weight to moderators who have more experience or knowledge in the subject domain which the falsehood falls into. A study by Bhuiyan et al. found that crowdsourced credibility assessment performance differed depending on rater demographics and the scope of tasks the crowd was assigned to rate.[40] Therefore, moderators may have differential performance in their credibility assessment of statements in different topics, which motivates the grouping of moderators into different areas of expertise.

Automated detection of potential falsehoods

Other than using the wisdom of crowds, natural language processing can also be used to identify potential falsehoods based on the words, grammar and context of a particular statement. The idea is similar to spam detection, where natural language processing has enabled models that perform well in identifying spam and phishing messages. For example, Google's spam detection model is able to detect spam and phishing messages with 99.9% accuracy.[41] However, the task of predicting credibility and falsity is significantly more challenging because it requires knowledge of the true state of affairs in the real world, often contemporaneously, which cannot simply be inferred from large amounts of data. Nevertheless, there may still be features like bias, syntax, and common words that are more prevalent in falsehoods than other statements. Indeed, a study by Fairbanks et al. found that it is possible to detect bias in articles using natural language processing,

[40] Bhuiyan, M., Zhang, A., Sehat, C. and Mitra, T., 2020. Investigating Differences in Crowdsourced News Credibility Assessment. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2), pp.1-26.
[41] Lardinois, F., 2017. *Google says its machine learning tech now blocks 99.9% of Gmail spam and phishing messages.* TechCrunch+.

which is useful in identifying particularly one-sided, emotionally charged or incendiary statements that may be more likely to contain falsehoods.[42] Another study by Zhang et al. found that indicators like tone, the presence of a "clickbait" title, the presence of citations, and the amount and type of advertisements present in a web article are useful in predicting the credibility of the article.[43] Therefore, it is possible to train models that are capable of predicting, with a reasonable accuracy, the likelihood that a particular statement or article contains falsehoods. Such a model can be used to scan different websites and platforms for potential falsehoods, which will greatly reduce the administrative workload required to enforce POFMA.

Automated detection of statements similar to already-identified falsehoods

One major challenge mentioned in section 3.4 was that even after a Direction has been issued regarding a particular falsehood, that falsehood may have already been reproduced or shared widely by other people, whether unwittingly or intentionally. Furthermore, the person who posted the falsehood may set up new social media accounts or new websites to continue perpetuating similar falsehoods. These problems pose a significant challenge and burden to the enforcement of POFMA.

Therefore, similar to the previous recommendation, I propose the use of natural language processing to automatically identify statements that are substantively similar to falsehoods that have already been identified and issued with Directions. There are numerous sentence similarity methods that make use of semantic matching between words in two sentences to compute a

[42] Fairbanks, J., Fitch, N., Knauf, N. and Briscoe, E., 2018. *Credibility Assessment in the News: Do we need to read?*

[43] Zhang, A., Ranganathan, A., Metz, S., Appling, S., Sehat, C., Gilmore, N., Adams, N., Vincent, E., Lee, J., Robbins, M., Bice, E., Hawke, S., and Karger, D. *A Structured Response to Misinformation: Defining and Annotating Credibility Indicators in News Articles*. The Web Conference, April 2018.

similarity score between two input sentences.[44] Statements with high similarity scores when compared with established falsehoods can then be highlighted for review by the independent review board.

Government and social media collaboration via PolicyKit

PolicyKit is a software infrastructure proposed by Zhang et al. that allows online users to author a range of governance procedures that can automatically be implemented on their platforms.[45] There are two categories of policies that can be written and enforced using PolicyKit: platform policies and constitution policies. Constitution policies define the rules of how platform policies are approved (e.g. a policy must be accepted by two-thirds of users before it is approved). Any action done by a user on the platform will be checked against all existing platform policies, and will successfully execute only after all platform policies have been complied with.

There is potential of implementing PolicyKit for regulating online falsehoods. The government can work together with social media companies to implement PolicyKit on social media platforms, and write the constitution policies that will define the rules of policymaking on the platform. The constitutional policies will also limit the platform policies to the purpose of regulating falsehoods. Thereafter, the government can introduce platform policies that automatically regulate the communication of certain falsehoods. For example, after a falsehood has been issued a Direction under POFMA, the government can introduce a platform policy that the same falsehood and substantially similar statements cannot be posted on the platform, or that such statements must be posted with a specified correction notice. Additionally, the government

[44] Wang, Z., Mi, H., and Ittycheriah A., 2017. *Sentence Similarity Learning by Lexical Decomposition and Composition.*
[45] Zhang, A., Hugh, G., and Bernstein, M., 2020. *PolicyKit: Building Governance in Online Communities.* Association for Computing Machinery.

can add conditions that must be fulfilled before the policy is triggered. For example, a post containing an established falsehood must reach a certain viewer count before it will be taken down. This eliminates the need for the government to laboriously search the platform for similar statements. However, to prevent the government from acting as the sole arbiter of truth, the platform policy must be approved by a certain proportion of users on the platform first, as defined by the constitution policy. This recommendation attempts to strike a balance between government regulation and freedom of speech, which can continuously be tuned via platform policy parameters like view count, as well as adjustments to the constitution policies.

Requiring more robust identity verification on social media

Many of the challenges outlined in this paper stem from the fact that social media platforms today have lax or nonexistent identity verification policies, and malicious actors are able to easily create new accounts on social media while abandoning old ones to evade accountability. Therefore, I propose that the government requires social media companies to implement robust identity verification tools during user registration. In particular, social media companies must require that a user provides their legal name during the account registration process, and submit proof of identity in order for the account to be registered successfully. In Singapore, where each citizen has a digital QR code in the SingPass application system that proves their identity, social media companies should be required to verify a user's identity using such a QR code. In essence, the idea is that each social media account in the virtual space should be tied to the identity of a person in the physical space, so that when the need to identify the person behind an online account arises, there is a way to do so. Therefore, when a social media post is determined to contain a falsehood

harming the public interest, it is ensured that the person behind the post can be identified and issued with a Direction or otherwise held accountable.

Chapter 5: Conclusion

This paper studies the regulation of online falsehoods, using Singapore's POFMA as a case study on how such regulation can be implemented. We analyze the costs and benefits of POFMA that have arisen during the two years since it was passed into law, and find that while POFMA has been effective in breaking the virality of falsehoods and providing important, factual information in issues concerning the public interest, there is still room for improving its sociological legitimacy and transparency. We also identify the key challenges of regulating falsehoods in the context of the modern digital age, covering issues of online personhood and identity, enforceability, granularity and legal culpability. Finally, this paper provides policy recommendations on both improving POFMA and on bolstering the regulation of online falsehoods in general. We cover qualitative recommendations that target the legitimacy and effectiveness of POFMA, as well as technical recommendations that make use of crowdsourcing, natural language processing and automation to improve and strengthen the detection and prevention of online falsehoods.

References

Ang, M., 2019. States Times Review founder Alex Tan refuses to comply with POFMA order, claims he's now an Australian. (2021). Available at: <https://mothership.sg/2019/11/states-times-review-alex-tan-mha-pofma/>

Bhuiyan, M., Zhang, A., Sehat, C. and Mitra, T., 2020. Investigating Differences in Crowdsourced News Credibility Assessment. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2), pp.1-26.

Brewster, T., 2021. *2,500 Posts, 300 Platforms, 6 Years: A Huge But Mysterious Pro-Russia Disinformation Campaign Is Exposed*. Forbes. Available at: <https://www.forbes.com/sites/thomasbrewster/2020/06/16/2500-posts-300-platforms-six-years-a-huge-but-mysterious-pro-russia-disinformation-campaign-is-exposed/?sh=1334589e185c>

Buchanan, M., 2017. *Why Fake News Spreads So Fast on Facebook: Ad Technology has weaponised disinformation*. Available at: <https://www.bloomberg.com/opinion/articles/2017-08-31/why-fake-news-spreads-so-fast-on-facebook>

Cameron, D. and Conger, K., 2017. *Here Are 14 Russian Ads That Ran on Facebook During The 2016 Election*. Gizmodo. Available at: <https://gizmodo.com/here-are-14-russian-ads-that-ran-on-facebook-during-the-1820052443>

Daskal, J., 2019. *This 'Fake News' Law Threatens Free Speech. But It Doesn't Stop There.* Available at: <https://www.nytimes.com/2019/05/30/opinion/hate-speech-law-singapore.html>

Fairbanks, J., Fitch, N., Knauf, N. and Briscoe, E., 2018. *Credibility Assessment in the News: Do we need to read?* Available at: <https://snap.stanford.edu/mis2/files/MIS2_paper_17.pdf>

Frost, A., 2019. *Academic highlight: Fallon on "Law and Legitimacy in the Supreme Court".* Available at: <https://www.scotusblog.com/2019/06/academic-highlight-fallon-on-law-and-legitimacy-in-the-supreme-court/>

Funke, D., Benkelman, S. and Tardaguila, C., 2019. *Factually: How misinformation makes money*. American Press Institute. Available at: <https://www.americanpressinstitute.org/fact-checking-project/factually-newsletter/factually-how-misinformation-makes-money/>

George, C., 2019. *Meeting the challenge of hate propaganda*. Written Representation to the Select Committee on Deliberate Online Falsehoods.

Government of Singapore, 2021. *Corrections and Clarifications regarding content about COVID-19 Vaccines in a blog post by Cheah Kit Sun.* Available at: <https://www.gov.sg/article/factually291121>

Khalaf, R., 2018. *If you thought fake news was a problem, wait for 'deepfakes'*. Financial Times. Available at: <https://www.ft.com/content/8e63b372-8f19-11e8-b639-7680cedcc421>

Lam, L., 2021. *Falsehoods, freedom of speech and burden of proof: Key findings from Apex Court's landmark POFMA judgment.* Channel News Asia. Available at: <https://www.channelnewsasia.com/singapore/falsehoods-freedom-speech-and-burden-proof-key-findings-apex-courts-landmark-pofma-judgment-2230541>

Lardinois, F., 2017. *Google says its machine learning tech now blocks 99.9% of Gmail spam and phishing messages.* TechCrunch+. Available at: <https://techcrunch.com/2017/05/31/google-says-its-machine-learning-tech-now-blocks-99-9-of-gmail-spam-and-phishing-messages/>

Mahmud, A., 2020. *In Focus: Has POFMA been effective? A look at the fake news law, 1 year since it kicked in.* Channel News Asia. Available at: <https://www.channelnewsasia.com/singapore/singapore-pofma-fake-news-law-1-year-kicked-in-688816>

Melford, C., 2019. *Tracking US$235 Million in Ads on Disinformation Domains – GDI.* Global Disinformation Index. Available at: <https://disinformationindex.org/2019/08/tracking-us235-million-in-ads-on-disinformation-domains/>

Meta. 2021. *Help your ads find the people who will love your business.* Available at: <https://www.facebook.com/business/ads/ad-targeting>

Protection from Online Falsehoods and Manipulation Act 2019.

Reuters, 2021. *Singapore close to vaccinating all eligible people against COVID-19.* Available at: <https://www.reuters.com/business/healthcare-pharmaceuticals/singapore-close-vaccinating-all-eligible-people-against-covid-19-2021-12-01/>

Schmelzer, R., 2021. *GPT-3.* SearchEnterpriseAI. Available at: <https://searchenterpriseai.techtarget.com/definition/GPT-3>

Schwartz, O., 2019. *In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation.* IEEE Spectrum. Available at: <https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation>

"Sean Edgett's Answers to Questions for the Record", Senate Committee on the Judiciary, Subcommittee on Crime and Terrorism Hearing on Extremist Content and Russian Disinformation Online: Working to Find Solutions, October 31, 2017 (19 January 2018), pp 16 – 17.

Singapore Internet Watch, 2021. *POFMA'ed Dataset.* Available at: <https://pofmaed.com/>

Tandoc, E., Lim, Z. and Ling, R., 2017. Defining "Fake News": A Typology of scholarly definitions. *Digital Journalism*, 6(2), pp.137-153.

Tham Y., 2020. *Judge wrong in placing burden of proof on Government in Pofma cases: AGC.* The Straits Times. Available at: <https://www.straitstimes.com/politics/judge-wrong-in-placing-burden-of-proof-on-government-in-pofma-cases-agc>

Thirteenth Parliament of Singapore, 2018. Report of the Select Committee on Deliberate Online Falsehoods – Causes, Consequences and Countermeasures.

Wang, Z., Mi, H., and Ittycheriah A., 2017. *Sentence Similarity Learning by Lexical Decomposition and Composition.* Available at: <https://arxiv.org/abs/1602.07019>

Wardle, C. and Derakhshan, H., 2017. Information Disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe*. Available at: <https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c>

Wiggers, K., 2021. *Falsehoods more likely with large language models.* Venture Beat. Available at: <https://venturebeat.com/2021/09/20/falsehoods-more-likely-with-large-language-models/>

Zhang, A., Hugh, G., and Bernstein, M., 2020. *PolicyKit: Building Governance in Online Communities.* Association for Computing Machinery.

Zhang, A., Ranganathan, A., Metz, S., Appling, S., Sehat, C., Gilmore, N., Adams, N., Vincent, E., Lee, J., Robbins, M., Bice, E., Hawke, S., and Karger, D. *A Structured Response to Misinformation: Defining and Annotating Credibility Indicators in News Articles*. The Web Conference, April 2018.