How journalists’ knowledge and AI can help fight information disorder at scale

by Shaun Davies

Thesis submitted in fulfilment of the requirements for the degree of

Master of Arts (Research)

under the supervision of Professor Monica Attard and Professor Derek Wilding

University of Technology Sydney
Faculty of Arts and Social Sciences

April, 2025

Certificate of Original Authorship

I, Shaun Davies, declare that this thesis is submitted in fulfilment of the requirements for the award of Master of Arts (Research), in the Faculty of Arts and Social Sciences at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature:

Production Note:
Signature removed prior to publication.

Date: 20/4/2025

Table of Contents

Acknowledgement

This thesis would not have been possible without the generous support and guidance of many individuals. I extend my deepest gratitude to my principal supervisor, Professor Monica Attard, whose frank and fearless advice was instrumental in shaping this work. Her insights and unwavering support pushed me to refine my thinking and challenge my assumptions.

I am also indebted to my co-supervisor, Professor Derek Wilding, whose deep knowledge of legislation and literature enriched my understanding of the complex landscape surrounding this research. I would also like to acknowledge Dr Anne Kruger, formerly a co-supervisor, whose inspiring work in the mis- and disinformation space provided a strong foundation for my own explorations.

I am sincerely grateful to Dr Heather Ford and Dr Amelia Johns for their thoughtful reviews and constructive feedback during the stage assessments. Their insights helped me to strengthen my arguments and refine my approach.

My thanks also go to the members of the Centre for Media Transition who have offered invaluable advice and support. I particularly acknowledge Dr Michael Davis for his expertise on mis- and disinformation, Dr Sacha Molitorisz for his insightful contributions, and Dr Tim Koskie for his guidance on methodological approaches.

To my wife, Victoria, and my children, Bonnie and Cecile, thank you for your patience and understanding during those countless weekends when I disappeared. Your love and support kept me grounded throughout this process. I am also grateful to my dear friends, Hal Crawford and Andrew Hunter, for their wise counsel and unwavering encouragement.

Finally, I express my sincere appreciation to all the journalists and technologists who generously gave their time and shared their insights through interviews. Your contributions are the heart of this research.

This journey has been an invaluable learning experience, deepening my understanding of research and its rigour. It is my hope that the ideas presented in this thesis will make a small but meaningful contribution to addressing the complex challenges facing our information ecosystems.

Statement on Format of Thesis

This thesis adheres to the University of Technology Sydney’s guidelines for thesis format and style. The American Psychological Association (APA) 7th edition citation style is used throughout the document.

The thesis is written as a traditional monograph of six chapters comprising of an introduction and methodology chapter, a literature review, description of content moderation systems, two chapters of qualitative research findings based on the interviews with participants, and a final concluding chapter.

A bibliography and appendix are provided at the end of the thesis. The thesis is 46,051 words long including References and Appendices.

Abstract

Information disorder—including mis-, dis- and mal-information—threatens democratic processes, civic trust, and public well-being. Rapid developments in generative artificial intelligence (AI) worsen this landscape by enabling mass production of highly convincing but false content. Although scholars have examined technical and journalistic responses to misinformation, limited attention has been paid to how journalists and technologists might collaborate to achieve large-scale moderation that is both accurate and aligned with public interest values.

Methods

This study employed a structured qualitative approach, conducting 14 interviews with eight journalists and six technologists. Transcripts were coded using discourse analysis methods, supplemented by a thorough review of literature on misinformation, automated content moderation and journalistic verification practices. This methodology aimed to illuminate where and how journalistic input could enhance AI-driven content moderation systems.

Results

Participants agreed that misinformation and disinformation were urgent, multi-faceted issues. Journalists tended to focus on the societal impacts of information disorder, whereas technologists focused more on practical implementation. But both groups saw the need for rapid verification, nuanced cultural context and ethical oversight. Technologists also tended to have a more positive view of journalism than journalists themselves. Respondents identified a wide range of interventions that journalists could undertake—providing timely ground-truth data for AI training, flagging emergent misinformation trends before they go viral, and refining content policies with journalistic expertise in verification.

Discussion

Findings suggest that journalists can strengthen platform moderation by offering high-quality data, cross-cultural insight, and real-time fact-checking capabilities. A more radical proposal involves outsourcing some content moderation to journalistic organisations, who could provide embedded teams of journalists to work alongside content moderation teams at platforms in a combined advisor and watchdog role. The aim of this is instilling in platforms a stronger "public interest" ethos that results in them being a more responsible steward of the information ecosystems in which they have become dominant players.

Chapter 1: Introduction

The spread of false and misleading information has become a pressing global concern, with social media platforms struggling to curb viral rumours, distorted information and digital propaganda. Rapid advances in generative AI only intensify this crisis, creating content that can appear alarmingly credible. This thesis investigates whether closer collaboration between journalists and technologists can help build better, more accurate and more nuanced AI systems to find mis- and disinformation at scale. It also considers whether journalism's self-professed commitment to verifying information and acting in the public interest can provide a model for platform governance and regulation, and whether a deeper engagement could help secure new revenue streams for journalism while contributing meaningful value to digital platforms.

1.1 Tech platforms, journalism AI and information disorder

The internet's disruptive combination of communication, distribution and publishing technologies was initially greeted with a sense of optimism, even utopianism. As anyone inclined was empowered to broadcast their thoughts and opinions, it was assumed that free-flowing information would lead to a smarter, freer, and more equal world.

Early cyber libertarian John Perry Barlow (1996) went so far as to declare that governments—"weary giants of flesh and steel"—had no sovereignty in the world of the internet. "We are creating a world where anyone, anywhere may express his or her beliefs, no matter how singular, without fear of being coerced into silence or conformity," he wrote.

But the last decade has seen a sharp turn toward anxiety over how to govern a technology that touches every aspect of political and social life (Douek, 2021a; Gillespie, 2018; Napoli, 2019). Explosive growth in user-generated content and the rise of orchestrated disinformation campaigns (DiResta, 2018; Goldstein et al., 2023) have led to widespread concern and substantial research into problems in global information ecosystems. Governments have been taking notice of these phenomena, and several scholars have noted a "turn to regulation" that rethinks the idea of the internet as an inherently ungovernable space (Bossio et al., 2022; Flew & Wilding, 2020; Gorwa, 2024). False information floods social platforms faster than humans can refute it (Vosoughi et al., 2018; Zhou et al., 2023).

These issues are described using a variety of terms, including misinformation, disinformation, and mal-information. Information disorder (Wardle & Derakhshan, 2017) is a widely cited umbrella term that covers all three. For ease of reading, this thesis will interchangeably refer to both information disorder and misinformation (rather than mis- and disinformation), except in cases where disinformation is specifically meant.

Information disorder's spread has run in parallel to exponential growth in social media usage and digital platforms such as Facebook, YouTube and Twitter have been key distribution points for misinformation. Napoli contends that "social media platforms essentially fell into their role as significant news sources", which is "why they have proven to be particularly vulnerable to manipulation and exploitation by purveyors of disinformation" (2019, p. 4). Facing regulatory and reputational issues, platforms now employ tens of thousands of human moderators to manage issues with problematic content, including information disorder (McIntyre et al., 2022).

But the sheer scale of social media makes it impossible for humans to review every user post, photo and video. Platforms have attempted to solve this by deploying content moderation systems that utilise AI to identify, contextualise, downrank and even remove content in large quantities. A great deal of effort has been invested in content moderation systems, but the algorithms and policies that underpin them can be inaccurate and arbitrary, and many critics worry about private, US-based companies making decisions that impact the speech rights of billions of people around the world (Douek, 2021b; Gillespie, 2018; Napoli, 2019; Suzor, 2019).

Information disorder's growth coincided with a series of crises for traditional journalism. Trust in the profession is at an all-time low in many countries (Nielsen & Fletcher, 2024) and revenues have tumbled as audiences and advertising increasingly moved online, greatly benefiting Google, Meta and other tech companies (Posetti, 2018, pp. 57-72). Media companies and platforms are now battling over the value and necessity of news content, and governments around the world have introduced or are considering legislation to force platforms to strike commercial deals with news organisations (Bossio, 2024).

But these deals have proven unstable, with platforms claiming they overstate the value of the news content. In response to Canada's Online News Act, which obliged platforms to pay media outlets for their content, Meta opted to completely ban news from its platforms, stating that it did not believe users came to the platform to consume news content. At the end of 2024, Meta said it would not renew deals with Australian publishers, which were reached after the Australian government introduced its 2021 Media Bargaining Code, designed to force platforms to strike deals with media companies. Australia's government has since proposed a new "news media bargaining incentive" scheme that bolsters the code by threatening platforms with a tax levy if they do not strike (Bossio, 2024).

The positions of the platforms on issues relation to information disorder are shifting again as Donald Trump commences his second term as US President. Meta announced in January 2025 that it was ending its longstanding fact-checking program—widely interpreted as an attempt to neutralise criticism from President Trump and his allies. Founder and CEO Mark Zuckerberg stated that fact checkers were "biased," criticised the "legacy media" for drawing attention to misinformation on his platforms, and ordered a broad pullback in content moderation (Isaac & Schliefer, 2025).

Meta's global fact checking program is the largest of its kind in the world (Brashier et al., 2021; Vosoughi et al., 2018) and the most prominent example of platforms and newsrooms working together on misinformation problems. Zuckerberg's statements echoed the sentiments of Elon Musk, who drastically cut back content moderation at X and frequently attacks traditional media (Robison, 2024).

US Republican politicians are actively hostile towards content moderation and the study of misinformation, with incoming FCC chair Brendan Carr vowing to dismantle the "censorship regime" of social media platforms (Silva, 2024). These issues are global—in Australia, the federal Labor government failed in 2024 to pass its *Combatting Misinformation and Disinformation* bill due to concerns it would empower the government to censor legitimate speech (Butler, 2024).

Although these developments point toward an industry-wide platform pullback from fact-checking and stricter moderation, this thesis pushes against that trend. Building on Napoli's (2019) arguments for social media to embrace journalistic "public interest" principles such as independence, truth and accuracy in governance, the following chapters will argue that platform companies can better utilise the skills and practices of journalists, explore how such an integration might work and unpack the benefits of such an arrangement for both parties.

1.2 Research question, hypothesis and originality

My research question is: *How can the knowledge of journalists best be combined with scalable technological systems to combat information disorder?*

1.2.1 Originality

Research into information disorder has proliferated in the years since the 2016 US election. This substantial body of literature covers an array of topics and uses a variety of approaches: for example, catalogues of misinformation types and tactics (Sitek et al., 2020), detailed models outlining how disinformation is generated and propagated (Wardle & Derakhshan, 2017), psychological studies of why people believe in conspiracy theories (Schatto-Eckrodt et al., 2020), and practical ideas on how platforms can do a better job at protecting users from harms (Donovan, 2020). Data scientists and AI researchers have also considered at length the different types of AI techniques and technologies that can be applied to information disorder problems (Asr et al., 2024; Shu et al., 2017).

But while journalism researchers and data scientists have separately contributed to research on information disorder, scant attention has been paid to how journalists and technologists can work together. García-Marín et. al. (2022) explicitly call out this lack of co-operation in a study of the relationship between journalism and engineering in combatting misinformation. While young journalists are shown to have positive attitudes towards use of AI to detect disinformation, they say, most studies on this topic are found in IT journals and written by engineers and data scientists.

In response, the authors write: "In order to fight disinformation and improve journalism… (journalists) must have an increasing knowledge of engineering, and at the same time engineers must know more about communication. It is the symbiosis of both professions that can better fight disinformation" (p. 124).

A review of recent literature reveals that this gap persists. For example, while there are numerous studies that cover the use of crowdsourced data for training models to detect misinformation, these tend to cover the work of fact checkers and journalists only briefly, noting only that data from "experts" is expensive and not scalable for the large-scale data labelling exercises needed to train bespoke AI models (Allen et al., 2020; Pennycook & Rand, 2019; Roitero et al., 2023; Soprano et al., 2021). See Section 2.6 for more information.

This research will contribute to filling in the blanks and encouraging greater collaboration between the world of journalism and computer science to combat disinformation. One way it does this is through a direct comparison of the viewpoints of journalists and technologists regarding information disorder issues, something not seen elsewhere in the literature. This provides a more nuanced understanding of the points of similarity and difference between these two groups.

A further original contribution is the recommendation for platforms to partner more deeply with journalists on content moderation, an idea that has not appeared elsewhere in the literature. This proposal builds on the work of Napoli, who believes an ethos of public interest could improve platform governance (2019). It is the contention of this thesis that such partnerships can bring real value to both parties—the benefits and practicalities of this are discussed at length in Section 6.3.

1.2.2 Sub-questions

The research question is subsequently broken down into two sub-questions. Each has its own hypothesis.

Question 1: Are journalists and technologists sufficiently aligned on defining and understanding the problem of information disorder to enable meaningful collaboration?

Hypothesis: There may be deviance between journalists and technologists on issues such as definitions of misinformation, disinformation, and fake news; approaches to detecting and acting against information disorder; the contribution that AI can make to detecting and acting against information disorder; and the relative importance of user safety versus maximizing free speech. If there is a wide gap in understanding, it could hinder the ability of journalists to contribute to at-scale detection of misinformation in a meaningful way.

Question 2: How can journalists best contribute to making at-scale content moderation better, particularly at major social media platforms?

Hypothesis: Journalists and newsrooms are well-placed to assist in improving misinformation systems. They have unique practices relating to verification of information and will be particularly good as all-purpose experts in the development of moderation policy and enforcement of these policies. Journalists also have a strong public interest ethos that can be beneficial to platforms, where concerns about a healthy democracy have taken a backseat to profitability and rapid growth.

The research will also consider whether journalists and media companies could turn such a collaboration into an ongoing source of revenue, and what elements of journalistic governance and legislation could be adopted as jurisdictions worldwide create laws to better govern platforms in societies.

1.2.3 Structure of thesis

The remainder of **Chapter 1** will detail the data collection, methodology and ethical approach that the research has taken.

Chapter 2's literature review will consider a range of perspectives on information disorder, journalism and content moderation. It starts with definitions and moves on to the causes of information disorder. It also examines journalistic practices and journalism's response to misinformation, and considers academic research on content moderation, free speech and governance.

Chapter 3 takes a detailed look at how content moderation works. It considers how the rules of moderation systems are created and enforced, how AI systems are trained to detect misinformation and the role that human judges play in making decisions. It also considers how engineers establish "ground truth" for algorithms, the actions that content moderation systems take against violations of moderation policies and approaches to transparency.

Chapter 4 explores the perspectives of journalists and technologists on information disorder. This chapter presents insights gathered from interviews, focusing on whether these groups have a shared understanding of key terms like misinformation, disinformation, and propaganda. It offers an original analysis of the alignment and divergences in definitions and approaches between the two professions. This analysis contributes to understanding how journalism and technology might collaborate more effectively on content moderation challenges.

Chapter 5 shifts to analysing the role of journalists in contributing to platform moderation systems, particularly regarding their unique skill sets, such as real-time verification, contextual understanding, and ethical considerations. It evaluates how journalists' expertise can support and refine AI-driven approaches, discussing the socio-political insights journalists bring, which are essential for interpreting complex misinformation narratives and nuances that algorithms alone may overlook.

Chapter 6 concludes the analysis with recommendations. Drawing on findings from Chapters 4 and 5, it outlines steps for enhancing collaboration between journalists and technologists, including specific ways platforms could integrate journalists' skills to improve misinformation detection and moderation practices. It concludes by proposing a framework for collaboration between platforms and journalistic institutions, suggesting that newsrooms could provide training data, participate in decision-making to foster more effective content moderation systems, and potentially take an advisory and watchdog role in managing misinformation on a major platform.

1.3 Research plan and methodology

1.3.1 Rationale for qualitative approach

This research followed a structured qualitative methodology using interviews and data analysis to explore the intersection of journalism, AI and content moderation. Conducted in three stages, the research began with semi-structured interviews with experts, which was followed by rigorous thematic analysis, and culminated in a set of findings and actionable recommendations. Relevant literature was consulted throughout to ensure a robust theoretical framework for the study.

The aim of this study is to explore in-depth how journalists and technologists perceive the challenges of misinformation, as well as how journalists' skills can be integrated into AI-driven content moderation. A qualitative methodology is well-suited for this purpose—while quantitative surveys or large-scale experiments might show the frequency of certain opinions, qualitative interviews yield richer insights into the nuanced processes of verification, policymaking, and AI development. The rich detail that comes from the expert interviews in turn produces a rich set of opportunities and recommendations.

1.3.2 Ethics

My research proposal was approved by the UTS Ethics Secretariat on August 2, 2021. It was classified as low risk. For reference purposes, the approval number is UTS HREC REF NO. ETH21-6257.

The main risk identified was that a participant could disclose sensitive information during an interview that their employer wished to stay hidden from view. This could have put them in an uncomfortable situation or, in a more serious case, have ramifications for their employment.

I used four strategies to minimise risk:

1. All participant responses have been kept confidential.
2. My thesis uses generic terms—JN (e.g. J1) for journalists , TN (e.g. T3) for technologists—instead of actual names.
3. Explicit permission for participation was sought from subjects' employers.
4. I explicitly asked participants not to disclose any information they believed to be sensitive or commercial-in-confidence, or that could harm their employer's reputation, both in the participant information sheet and verbally before beginning the interview.

1.3.3 Recruitment

Participants were recruited from a cross-section of news media organisations, fact-checking institutions, academic bodies, and content moderation vendors that provide services to social media platforms. As noted above, the initial approach was made to organisations, rather than participants themselves. In some cases, the organisation was asked to identify suitable individuals based on expertise, while in others a candidate had been identified in advance. Individual participants were subsequently sent information packets, which facilitated informed consent and maintained confidentiality.

1.3.4 Participants

The study involved 14 participants, with eight from journalism backgrounds and six from technology. On the journalist side, the participant pool included diverse voices from major news organisations, fact-checking bodies, academic institutions, and moderation vendors, which added depth to the perspectives on misinformation challenges. Technologists came from moderation vendors, fact checking units, academic units and start-ups. Notably, only one engineer was sourced from a major technology firm, despite a concerted effort to involve multiple large platforms. Interviews took place between July 2022 and February 2024.

"Journalists" here covers three different types of editorial workers whose job involves misinformation detection. Four (J1, J6, J7, J8) come from fact checking organisations—of these, three were reporters before they transitioned to fact checking. The other has a hybrid background in both technology and journalism, but as their recent work has focused more strongly on editorial matters than engineering, they were included in the journalist group.

Two (J2 and J4) come from newswire services that provide fact checking services. J2 is a product manager with responsibilities for both human fact checkers and automated AI detection, while J4 has a leadership role in fact checking. Both started their careers in journalism and progressed through various newsroom positions. The remaining two journalists (J3 and J5) work in policy and investigations at third-party trust and safety platforms. These are companies that provide content moderation services, including misinformation detection, threat intelligence and "red team" testing to detect vulnerabilities in moderation systems, generally servicing major platforms. J5 was a journalist for many years before shifting into the investigations space. J3 does not have a background in journalism, but the investigative and documentary practices they employ in their work are journalistic in nature.

Two of the eight participants defined as "journalist" have non-traditional backgrounds, and many do not work in traditional newsrooms. It could be argued that a different term would better describe this group. But after considering a variety of alternatives (such as "editorial worker", "editor", and "investigator"), journalist is the most straightforward and readily understandable label. As such, the definition of journalist in this data and analysis concerns itself with the editorial practices used in day-to-day work, rather than educational background or work history, and should be interpreted with this understanding.

The six "technologists" also come from a variety of backgrounds. Two (T1 and T2) are in-house data scientists working on AI problems within fact-checking organisations. T1's job is to develop AI tools that fact checkers can use to find questionable claims to research and potentially debunk. T6 works as engineering lead at a company that provides content moderation services. They provide technical leadership on an at-scale misinformation detection system that where journalists contribute to data and insights.

The remaining two technologists are entrepreneurs. T5 founded a successful app for smartphones that identifies controversial trending news topic and presents users with views on this topic from across the political spectrum, with the aim of showing readers a broad range of perspectives. T3 is a data scientist and academic who has published widely on the topic of detecting misinformation with AI and started a misinformation detection platform based on this research.

Only one technologist (T4) comes from a major tech company. They work for a global product that requires misinformation detection across user comments and content from publishers, creators and other partners.

Participants came from a wide range of cultural and linguistic backgrounds, including Iran, Kenya, China, Israel, Australia, the US, the UK, and France. Variance on key issues can be seen based on participants' country of origin, so it is relevant to note this here.

It was easier to recruit journalists to participate in the research, leading to an imbalance between the two groups. The main cause of this was the reluctance of major platforms to participate. All the journalists interviewed work directly on information disorder issues, and many already work up-close with engineers and data scientists. This is valuable to this qualitative research, as they tend to have some knowledge of the technological problems that engineers and data scientists face, and put forward practical, informed ideas and solutions. But their responses should not be taken as representative of the entire journalism industry.

1.3.4 Interview design

Interviews were conducted in a semi-structured format and averaged one hour in length. Pre-determined questions focused on topics including misinformation detection, AI applications, verification processes, and free speech versus platform responsibilities. Tailored questions for each group ensured relevance: journalists provided insights on content policy, while technologists discussed AI design and implementation. This format allowed flexibility, enabling participants to elaborate on key points as necessary. A copy of the interview questions is included in the Appendix 2.

1.3.5 Data analysis

Transcriptions were imported into Nvivo 12, a qualitative research program, and the data was annotated via coding, a qualitative data analysis methodology that involves “attaching conceptual labels to data” (Urquhart, 2013, p. 2). Initial codes were refined and combined into larger groups to ensure similar concepts were nested together. This enabled an analysis of concepts, themes and definitions in the data, using techniques from discourse analysis, a research approach where “language material, such as talk or written texts, and sometimes other material altogether, is examined as evidence of phenomena beyond the individual person” (Taylor, 2013). The codes were discussed in detail with my academic supervisors to ensure that they were robust and representative.

1.3.6 Use of AI tools

Various AI tools were used in the creation of this thesis.

Otter.ai was used for recording and transcription of interviews (always with the explicit permission of participants). All transcriptions were checked manually and transcription errors corrected.

NotebookLM was used for data investigation and summarisation of sources. This tool was chosen because of its enabled deep investigation of large numbers of documents and includes extensive footnoting, which allows easy comparison of summarised information with primary source material to ensure that any transformed text is hallucination free. All references and claims were checked carefully against the source material to ensure factuality.

ChatGPT o1 is an advanced reasoning model from OpenAI. It was used to check and critique sections of the thesis – was anything missing, unclear, or open to objection? It was also used to create a draft synopsis, but this was heavily edited prior to publication.

1.3.7 Note on change in methodology

During the interviews, it became evident that participants had varied interpretations of the term "mainstream media," which was central to Section 4.3.6. After conducting the first six interviews, participants frequently sought clarification on this term, indicating that its meaning was not universally understood. In response, a decision was made to provide a standardised definition of "mainstream media" for the remaining eight interviews. This adaptive approach is consistent with established qualitative research practices, where iterative modifications are often necessary to enhance data collection instruments and ensure clarity (van Assche et al., 2023). Such adjustments are particularly common in exploratory studies, where participant feedback can reveal areas needing refinement to improve the accuracy and relevance of the data collected. Notably, the change did not cause significant variance between responses, as will be further analysed in Chapter 4.

1.4 Chapter summary

This chapter introduced the research context, highlighting the growing challenge of information disorder, exacerbated by digital platforms and generative AI. It established the central research question exploring how journalistic knowledge can be integrated with scalable technological systems to combat this issue more effectively. The chapter outlined the study's originality in bridging the gap between journalistic and technological perspectives on misinformation detection, detailed the specific sub-questions and hypotheses guiding the research, and provided a roadmap for the subsequent chapters. Finally, it detailed the structured qualitative methodology employed, including the rationale for the approach, participant recruitment and characteristics, ethical considerations, interview design, data analysis methods, and the use of AI tools in the research process.

Chapter 2: Literature Review

2.1 Definitions

Clearly and consistently defining terms such as misinformation and disinformation is widely acknowledged as a difficult task (Bernstein, 2021; Caplan et al., 2018; Gibbons & Carson, 2022; Tucker et al., 2018; Wardle & Derakhshan, 2017; Zeng & Brennen, 2023).

A lack of definitional rigor is consistently identified as a major challenge for the nascent field of misinformation research. Tucker et. al note in a review of scientific literature on social media and disinformation that “there is no real consensus across much of the academic literature on how to define many of the phenomena” commonly referred to in academic research, urging more definitional rigor and consistency around terms such as “hyperpartisan media”, “online propaganda” and “conspiracy theories” (2018, p. 55).

Gibbons and Carson (2022) have also examined the divergent understandings of misinformation and disinformation among key stakeholders in the Asia Pacific region. Through interviews with digital platform employees, civil society actors, academics, and journalists in Singapore and Indonesia, they found that existing definitions often developed in "information silos" and that an absence of uniform definitions can potentially lead to poor policy implementation or, conversely, government heavy-handedness.

2.1.1 Fake news

The evolution of the term “fake news” is illustrative of the difficulties inherent in defining these phenomena. The phrase came into common usage in 2016 and quickly became a politicised term (Caplan et al., 2018). While it was initially used to refer to websites creating false content for financial gain, populist commentators and politicians subsequently used the phrase to paint mainstream media sources as deceptive and illegitimate. Academics and mainstream journalists have also used the term to describe hoaxes, conspiracy theories, hyper-partisan content and state-sponsored disinformation. While the term is still sometimes used in everyday conversation and the media, it is generally accepted among both researchers and industry participants that the term is too politicised and contested to be useful in more formal study of these phenomena (Amakoh, 2020; Bounegru et al., 2018; Gibbons & Carson, 2022; Wardle & Derakhshan, 2017).

2.1.2 Misinformation and disinformation

The terms "misinformation" and "disinformation," frequently used in discussions about online falsehoods, are also politically charged. In the US, a Republican committee labeled research on these phenomena as a "political ruse" aimed at silencing dissenting voices (US Congress, 2023). This committee's actions, including subpoenas and public attacks on academics, drew criticism for its partisanship (Masnick, 2023), but it is worth noting here to show the level of disagreement over how misinformation and disinformation should be defined.

Zeng and Brennen (2023) argue that a universally accepted definition of misinformation may be unattainable, given the interdisciplinary nature of the field and the diverse conceptual approaches employed by scholars. Despite this, they stress the importance of striving for conceptual clarity and comprehensiveness in misinformation research. The authors highlight several key disagreements around defining mis- and disinformation, including whether misinformation can be considered "informative" at all, whether intent to deceive is a necessary component of the definition, and how to operationalise "falsity" in research.

They identify two different approaches in determining whether information is false: "falsity as property" and "falsity as process" (p. 7). The first approach views falsity as an inherent characteristic of misinformation, measurable against a "ground truth" established through evidence and expert consensus. However, this can be problematic, especially in domains like politics and science, where "truth" is often fluid, evolving and subject to debate. The second approach, "falsity as process", emphasises the social and contextual factors that influence how truth and falsehood are defined and perceived.

Wardle and Derakhshan (2017), in a widely cited work, offer a framework that moves beyond the binary of misinformation and disinformation, recognising the broader spectrum of problematic information online. “Information disorder” distinguishes between misinformation and disinformation based on the intention to cause harm. Misinformation is defined as information that is false but not created with the intention of causing harm, while disinformation is false information deliberately created to harm an individual, group, organisation or country.

They also introduce the concept of “mal-information”, which is information that is based on reality but used to inflict harm. This category highlights the potential for harm even when information is factually accurate, as it can be used out of context or with malicious intent to harm reputations or incite violence.

In addition, they propose a typology of seven types of problematic content, including satire, parody, fabricated content, misleading content, and imposter content. This typology highlights the diverse forms and intentions behind the spread of problematic information online. Information disorder has become a commonly cited term within media studies and information science and is used throughout this thesis.

Gibbons and Carson advance alternative working definitions for misinformation and disinformation based on the “dissensus” they see among key decision-makers in the Asia Pacific region. Their aim in doing so is the facilitate the types of holistic approaches needed to tackle the serious harms caused by online falsehoods.

For them, misinformation is the “spread of inaccurate or misleading content that may (or may not) cause emotional, physical, political, financial or intangible harm to individuals or institutions” (pp. 233-234). Disinformation is “the spread of misleading, inaccurate or deceptive content with decisive actions to mislead, deceive or otherwise cause emotional, physical, political, financial or intangible harms” (p. 234), inferring intent through concrete user behaviours such as using fake accounts or mass postings. Gibbons and Carson’s definition places emphasis the potential and actual harms caused by both misinformation and disinformation.

Legal Definitions

Definitions of misinformation and disinformation in government legislation also tend to put emphasis on the concept of “harm”. For instance, the Australian government’s unsuccessful misinformation bill (Parliament of the Commonwealth of Australia, 2024) defines misinformation as the dissemination of "verifiably false, misleading, or deceptive content that is reasonably likely to cause serious harm". Disinformation has the same characteristics, but with an additional element of intent to “deceive another person or to gain an economic or political advantage” (p. 44).

The EU's *Strengthened Code of Practice on Disinformation* (European Commission, 2022) also maintains a distinction based on intent, defining misinformation as "false or misleading content shared without harmful intent" and disinformation as "false or misleading content that is spread with an intention to deceive or secure economic or political gain". However, the EU framework also emphasises the potential for harm caused by both types of content, noting that misinformation, even if shared unintentionally, can still have harmful effects.

Intent, misinformation and disinformation

Altay et al. (2023) surveyed 150 experts and found that the most common definition of misinformation was "false and misleading information". However, they also found a lack of consensus on whether intentionality should be part of the definition of misinformation. Some experts believed misinformation must be spread unintentionally, while others thought it could be spread either intentionally or unintentionally. This difference of opinion appeared to be split along methodological lines, with researchers who use qualitative methods more likely to include intentionality in their definitions than those using quantitative methods.

Wardle and Derakhshan (2017), the EU disinformation code (European Commission, 2022), and the scuppered Australian legislation (Parliament of the Commonwealth of Australia, 2024) all use intent to draw a distinction between misinformation and disinformation.

By contrast, Zeng and Brennen (2023) argue that intent should not be a defining feature of misinformation. While acknowledging the common practice of distinguishing between misinformation and disinformation based on the presence or absence of intent to deceive, they propose an “intentionality-neutral” definition, contending that determining intent in real-world situations is often difficult or impossible. A strict requirement of intentionality could lead to the exclusion of important cases from research, they say, and there are very few instances where researchers can definitively establish a source's intention to deceive, except for some very limited cases like satirical websites (p. 6).

Instead, Zeng and Brennan say an intentionality-neutral definition of misinformation—encompassing both intentional and unintentional falsehoods—offers a more practical and inclusive approach, as both intentionally and unintentionally spread falsehoods can contribute to the broader phenomenon of misinformation and have detrimental societal consequences.

2.1.3 Propaganda

The literature on propaganda is rich, varied and has a long history. But like misinformation, the term’s definition is contested and evolving. The common definition of propaganda is pejorative—lies and manipulation that governments use to deceive and control populations (Tutui, 2017). But academic definitions of propaganda are not necessarily negative.

Bernays (1928) viewed propaganda as a necessary tool for shaping public opinion and maintaining order in modern society. He argued that in a complex world, informed decision-making is difficult for the average person. "Invisible governors" – influential figures and organisations – therefore use propaganda to simplify complex issues and guide the public towards desired actions and beliefs. This manipulation, Bernays claimed, was essential for the smooth functioning of democracy and the economy, although he conceded it could be used for both good and bad purposes.

Ellul (1957) notes the difficultly of distinguishing between information and propaganda, even for experts, saying: “Information is, by definition, a distortion of public opinion. But where and to what extent does the transition from information to propaganda take place?” Like Bernays, Ellul sees propaganda as a necessity in a complex society, but his definition emphasises technology and mass communication. He suggests that propaganda serves as a means of integrating individuals into a complex and technologically driven world.

As with disinformation, there are arguments over the role of intent in propaganda. Tutui (2017) contrasts Jowett and O'Donnell's view that propaganda is a systematic pursuit of selfish goals that is always insincere in intent (p. 115) and Stanley's view that propagandists may be sincere even as they exploit biases and social structures to distort reality (p. 117).

Tutui's own view is that there is no definition of propaganda that is free of problems, for three reasons: 1) usage of the term has changed over the years, making an all-encompassing definition difficult; 2) the pejorative view of propaganda causes many theorists to overestimate the role of the propagandist and underestimate the role of the receiver; and 3) the "social causes of the phenomena itself are more deeply rooted in the fabric of social reality than it is usually assumed" (p. 123).

The proliferation of internet and social media platforms has transformed the nature of propaganda, making it more pervasive and challenging to address. Guess et. al. (2020) note that propaganda and disinformation are "sometimes used interchangeably, with shifting and overlapping definitions". They define disinformation as a "subset of misinformation that is deliberately propagated" (p. 13) and involves the intentional spread of false or misleading information to deceive and manipulate audiences. Propaganda, they say, encompasses a broader range of persuasive communication tactics, including the use of truthful information to advance a particular agenda. Guess et. al. ultimately advance a very broad definition of propaganda as "any communications that are intended to persuade people to support one political group over another".

The scale of modern propaganda and its highly technological nature have also led to the evolution of new definitions. One key term that appears frequently in literature is computational propaganda (DiResta, 2018; Woolley, 2020), referring to the use of algorithms, automation, bots and human curation to spread misleading information over social networks.

DiResta (2018) defines computational propaganda as "a suite of tools or tactics used in modern disinformation campaigns that take place online" (p. 14). Its goals include manipulating public opinion, influencing elections, discrediting opponents and sowing discord. Bots are a key tactic of these campaigns, using computational scale to escalate attacks and message distribution at levels that exceed human capacity. The anonymity afforded by these technologies also provides anonymity and cover for their architects.

Paul and Matthews (2016) discuss another set of modern propaganda tactics that they dub the "firehose of falsehood", which they say originated in Russian state media but have since been adopted elsewhere. This is characterised by a high volume and rapid spread of messages across multiple channels without concern for consistency or truth. The objective is to overwhelm audiences with a barrage of information, making it difficult for them to discern accurate information from falsehoods.

Messages are disseminated rapidly and consistently, ensuring that audiences are constantly exposed to the disinformation, which complicates efforts to identify and counteract false narratives. The messages often contain blatant lies and contradictory statements, which serve to confuse and manipulate the audience. By flooding the information space with inconsistent content, the strategy aims to create uncertainty and doubt.

News outlets in democratic countries have also been accused of distributing both misinformation and propaganda, and again definitions are contested. Bauer et. al. (2022) argue that the ambiguity over whether Fox News should be considered legitimate news, a partisan outlet, or propaganda highlights broader definitional issues in journalism studies. They argue for development of a more nuanced, richer vocabulary to capture the realities of news in the 2020s, as partisan media's growing prominence means that definitions of news that emerged during the period of journalism at the end of the 20th Century are no longer sufficient.

2.2 Causes of information disorder

A consistent set of definitions for information disorder-related problems is crucial. But if we seek to find solutions to these problems, we must also understand why they have emerged.

A key question here concerns social media and content platforms, which have to a large extent displaced traditional media channels as people's primary conduits of information (Andi, 2021). Is the emergence of social media to blame for the issues relating to poor-quality and factually incorrect information being shared widely in global information ecosystems? Or are wider societal factors the cause?

Information disorder is not a new problem. Soon after its invention in around 1440 the printing press was being used to distribute questionable and dangerous claims and in the late 1800s US newspaper editors competed to outdo each other with outlandish stories during the era of "yellow journalism" (Hemanus, 2021). It was not until the 1920s that journalism began to adopt professional standards and ethical codes (Apps, 1990; Rauch, 2021, p. 122).

But Rauch argues that social media's impact on present-day information disorder is "one or two orders of magnitude higher" (p. 126) than previous technologies. His account summarises many of the key criticisms levelled at social media platforms in both academic literature and media commentary, and I will follow his line of reasoning here. [1]

In Rauch's view, social cohesion in democracies is underpinned by the activities of a "reality-based community", including journalists, scientists, intelligence agencies, academics, lawyers and other knowledge workers (p. 100-102). These professions share a common, if imperfect, commitment to ideals such as fallibilism, objectivity, accountability for mistakes, professionalism, viewpoint diversity and "no bullshitting" (p. 103-107).

Most major democratic countries have a relatively permissive attitude to public speech, and this ensures that a wide range of ideas are put forward into the public sphere. But only those that survive a rigorous process of checking and debate are published in key publications such as academic journals and major newspapers, to eventually become accepted knowledge. This system has a "positive epistemic valence" (p. 121)—something close to what's true becomes generally accepted.

Social media, Rauch says, has run this process in reverse to create a "negative epistemic valence". Platforms' focus on engagement and advertising revenue creates an information ecosystem that is "sensitive to popularity but indifferent to truth" (p. 125). In this system:

> "Instead of straining out error, they pass it along. In fact, instead of slowing the dissemination of false and misleading claims, they accelerate it. Instead of marginalising ad hominem attacks, they encourage them… Instead of validating claims, they share claims… Instead of identifying sources, they disguise them." (p. 124)

The all-up effect, in Rauch's view, is to plunge society into an "epistemic crisis".

[1] It should be noted Rauch's ideas are drawn from many researchers and that similar unified critiques of social media's design choices and business model have been made elsewhere. An example is the Netflix documentary *The Social Dilemma* (Orlowski, J. (2020). *The Social Dilemma* Netflix.)

There is some evidence to support this view. Vosoughi et al. (2018) analysed a very large dataset of news stories shared on Twitter and labelled them “True” or “False” based on judgements made by independent fact-checking organisations. Measuring the speed at which retweet “cascades” happened, they found false news spread faster than true news “by an order of magnitude”, with false political news spreading faster than any other category.

Like misleading or made-up news stories, propaganda techniques are nothing new. But two revelations from the 2016 US elections—that Russians used Facebook to target Americans with hyper-partisan content and hacked materials, and that analytics firm Cambridge Analytica helped candidate Donald Trump’s campaign micro-target political messages to voters via unethically harvested Facebook data—led many critics and researchers to conclude that social media’s advertising and content distribution systems are well suited to those who wish to manipulate a country’s voters from within or without (Frenkel & Kang, 2021).

Altay et al. (2023) surveyed 150 misinformation experts to understand the reasons behind misinformation spread and social media's role. While various factors were considered, experts overwhelmingly agreed (79%) that social media has worsened the problem. The experts in the study said that specific social media features contribute to partisanship. Echoing Rauch, they pointed out that platforms' incentive structures, like "like" and "share" buttons, tend to reward emotionally charged content regardless of accuracy. This can unintentionally boost misinformation, as emotionally driven content often spreads more rapidly.

Napoli (2019) argues that social media has significantly degraded the information environment, primarily by enabling the proliferation of false news and limiting the effectiveness of counterspeech to correct inaccuracies. He argues that in the early internet, people needed to actively "pull" information by searching the web. But social media algorithms “push” content to users in algorithmic news feeds and these algorithms are optimised for engagement and revenue, not factual accuracy. This shift is compounded by factors that make the environment more conducive to the spread of misinformation, including reduced resources for legitimate journalism, lower barriers to entry and undermined source credibility (pp. 92-102).

A further critique is that social media enables “firehose of falsehood” propaganda techniques designed to confuse and wear down citizens with a flood of incorrect and contradictory material. People then become unable or unwilling to distinguish truth from falsehood, and thus easier to manipulate (Paul & Matthews, 2016).

But other researchers believe that the effect of social media is overstated. Budak et al. (2024), in a review of evidence about the real-world harms of misinformation on social media, say that exposure to misinformation and extremist content is relatively infrequent and primarily concentrated among small, highly motivated groups. They argue that claims about algorithms playing a dominant role in exposing users to harmful content are exaggerated, with audience demand being a more significant factor, and question whether there is sufficient evidence to pin political polarisation on social media exposure.

The paper also highlights a significant geographic bias in research, which predominantly focuses on the USA and Western Europe. Budak et al. assert there is a critical need for more research in non-Western contexts, where the effects of misinformation may be more pronounced due to weaker media infrastructures and lower levels of media trust. They also call for better measurement of exposure among extremist groups, strategies to reduce demand for harmful content, increased transparency from social media platforms and expanded global research efforts.

Bernstein (2021) offers a similar counter-perspective on what he refers to as "big disinfo" – the journalists, researchers and lobby groups that strongly push a narrative of platform culpability for societal ills. While the article doesn't deny there are problems, it urges researchers and journalists to look beyond a "heroes and villains" narrative of technological determinism to a broader analysis that considers economic and social issues.

The article asks if social media is "creating new types of people or simply revealing long-obscured types of people to a segment of the public unused to seeing them" and questions the "supposedly objective stance" of journalists, think tanks, politicians and universities, noting that narratives of misinformation-led societal collapse could be read as attempts by gatekeeper institutions to reclaim cultural power from disruptive competitors.

Bernstein's argument provides a counterpoint to Rauch, who argues for "objective institutions" that filter out bad ideas and, through collective action, endorse something closer to "truth". The contrast points to an epistemic problem that's inherent in all discussions of misinformation – to what extent is it possible to know objective reality at all? [2]

This question becomes more important when considering the psychological underpinning of belief in conspiracy theories and misinformation. A key strategy for countering misinformation is debunking and fact-checking. But there is mixed evidence from psychological research to support the effectiveness of these practices at convincing people not to believe false information – it seems that some work as intended, others don't work at all, and some interventions even have the opposite of the intended effect.

In an example of what works, Bago et. al set out to test the impact of pre-existing beliefs and "motivated reasoning" on belief in false information, and whether giving people extra time to deliberate on a false headline would harden their opinions or cause them to reconsider. They found that "people made fewer mistakes in judging the veracity of headlines – and were less likely to believe false claims – when they deliberated, regardless of whether the headlines aligned with their ideology" (2020, p. 11).

Pennycook et. al. (2020) conducted an experiment that found priming users with subtle "accuracy nudges" nearly tripled their level of discernment when deciding on whether to share a news headline promoting COVID-19 misinformation. They recommended platforms adopt "simple and subtle reminders about the concept of accuracy [that] may be sufficient to improve people's sharing decisions regarding information about COVID-19" (p. 777).

In contrast, Nyhan et. al. conducted an experiment into whether aggressive fact checking "could correct the false belief that the (United States Obama-era) Affordable Care Act would create 'death panels'" (2013). It found that politically engaged supporters of former Republican vice-presidential candidate Sarah Palin became more likely to believe in death panels after being exposed to the fact check. Another US study on misinformation and climate change found that "corrections from Republicans speaking against their partisan interest are most likely to persuade respondents to acknowledge and agree with the scientific consensus on anthropogenic climate change" (Benegal & Scruggs, 2018).

[2] Establishing the nature of truth itself is obviously beyond the scope of this project, but I will make explicit my epistemic assumption – while I accept that no single person can know objective truth, I agree with Rauch that researchers ought to act as though it is possible to discover what is true. I cannot make this determination alone, but a collective effort by people committed to principles such as fallibility and intellectual honesty is able to get close.

People’s world views and mental models also influence their receptivity to misinformation and fact checks. Researcher Tommy Shane from anti-misinformation non-profit group First Draft (2020) makes a distinction between “facts” and “truth” – while these may seem to be the same at first glance, they can represent very different ways of knowing. Fact narratives (of which fact checks are a prime example) are granular, iterative, refer to authoritative sources, focus on falsification and verification, and are descriptive and technical. Truth narratives (as in, “we reveal the truth about X”), on the other hand, are skeptical of authority, use a causal logic that links together events in a way that explain the world with grand narratives, and rely on evidence from lived experience.

One of the challenges for fact checkers is the pace and volume at which misinformation is produced. Vosoughi et. al.’s (2018) investigation of the spread of true and false news stories on Twitter found the truth took six times as long to reach 1500 people as a lie, and that false political news was the most powerful type of lie, reaching 20,000 people three times faster than other types of false news reached 10,000 people.

2.3 The impact of AI on information disorder

Generative AI refers to powerful tools that allow users to create cohesive, creative and detailed text, images and videos via a simple natural language prompt. These models are created using a range of techniques that fall into the field of machine learning. Put simply, scientists developed methods to encode words and images into mathematical “representations”, and subsequently decode these back to words and images, while preserving as much information as possible. They then trained neural networks— complex webs of interconnected nodes, or "neurons," that mimic the human brain—on massive amounts of data, creating systems that were able to learn the relationships between billions of parameters and create human-like outputs in response to instructions in everyday language (Goodfellow et al., 2016, pp. 4-21).

But search engines, social media companies and other platforms are now encountering a very large amount of low-quality, auto-generated content that is commonly referred to as “AI slop”. A portion of this is misleading, malicious and harmful to users (Beres, 2024). While prominent AI companies like Open AI, Microsoft and Google put safety guardrails around generative AI tools such as ChatGPT to prevent misuse, these measures are susceptible to “jailbreaking” (Taylor, 2023), and open-source tools are being made available without safety limitations. The result is that anyone with time, dedication and modest capital can generate near-limitless amounts of text, image, audio and even video content through automated means.

In May 2023, Google researcher Geoffrey Hinton expressed concerns that a deluge of fake articles, images and videos would mean citizens “will not be able to know what’s true anymore” (Metz, 2023). OpenAI has published a paper with prominent disinformation researchers investigating possible ways to mitigate the harms that could be caused by use of generative AI in disinformation and influence campaigns, including creating AI models that are more “fact-sensitive”, stricter limits on the usage of these tools and media literacy campaigns – but ultimately concludes that “there is no silver bullet that will singularly dismantle the threat of language models in influence operations” (Goldstein et al., 2023, p. 4).

Generative AI tools have only been available to the general public for a few years, but some research on its implications for information disorder is beginning to emerge. One study used GPT3 to generate synthetic posts on X (here still referred to as “Tweets”, in line with the platform’s previous branding as Twitter), and found that:

> “Synthetic tweets containing reliable information are recognised as true better and faster than true organic tweets, while false synthetic tweets are recognised as false worse than

> false organic tweets. Moreover, GPT-3 does not perform better than humans in recognizing (sic) both information and disinformation. The results suggest that GPT-3 may be more efficient at conveying information because it can generate text that is easier to read and understand compared to text written by humans." (Spitale et al., 2023)

To combat this, the paper recommends a level of transparency and control around the datasets that are used to train models – if these contain sufficient unreliable information, models are likely to output convincing-sounding text based on inaccurate information.

2.4 What is journalism and who are journalists?

As the topic of this thesis concerns how journalists can help with the at-scale detection of information disorder, we need to examine some foundational concepts surrounding the field of journalism, starting with a deceptively simple question – what is journalism and who are journalists?

One commonly cited description of journalism is Kovach and Rosenstiel's *The Elements of Journalism* (2014). For Kovach and Rosenstiel, journalism's purpose is "providing people with the information they need to be free and self-governing". They divide the practices of journalism into 10 elements – reviewing each of these is beyond the scope of this essay, but two key elements are that journalism's first obligation is to the truth and that its essence is a discipline of verification (p. 9).

The idea of "truth" that Kovach and Rosenstiel have in mind is closely aligned with Rauch's. While philosophical arguments that truth is illusory or unobtainable may have philosophical merit, journalism concerns itself with the material world of human actions and politics, and as such needs to find a "practical or functional form of truth" that is subject to revision and emerges as a story develops through multiple retellings by a variety of sources (p. 57).

Kovach and Rosenstiel call the type of journalism that gets closest to practical truth "journalism of verification". This is contrasted with "journalism of assertion" (where news is published at speed, without adequate fact checking), "journalism of affirmation" (partisan media that interprets facts to fit a political narrative), and "journalism of aggregation" (where a technology platform disintermediates traditional news sources and usurps distribution of information) (pp. 64-65).

While journalists themselves are not objective, Kovach and Rosenstiel argue that the methodology of journalism can be objective, at least when practiced properly. This is achieved by not embellishing accounts or events, being transparent with audiences about what is known and relying on original reporting. There are also editorial techniques such as editing claims closely and with skepticism, using accuracy checklists, mistrusting non-primary sources (including accounts from officials) and exercising care with anonymous sources (pp. 129-136).

Some have contended that there is no such thing as objective journalism. Australian economics journalist Greg Jericho (2019) argues that in the face of misinformation and extremism, "neutral and detached" journalism leads to the promotion of incorrect ideas, which in his view largely come from the political right. On issues such as climate change and marriage equality, he says that journalists should become advocates who are "brave enough to stand up for truth and context and to be more concerned about being called out for ignorance than for bias".

In a response to this essay, former senior ABC journalist Alan Sunderland advances a view similar to Kovach and Rosenstiel's—rather than impartiality being the problem, it's a lazy approach to balance that leads to bad outcomes. He argues:

> “What does good, traditional journalism do? It seeks out the facts. All of the facts. It carefully weighs those facts to determine as far as possible where the truth lies. It listens to all views, works as hard as it can to identify and eliminate any prejudices or assumptions of its own. It then reports those facts, providing the context required to understand what is being reported. It understands the difference between facts (which are true and verifiable) and opinions, which are the many and varied views about those facts.” (2019)

2.4.1 Are journalistic practices unique and consistently applied?

The question of whether professional journalists possess unique practices that set them apart from other fields is complex and multifaceted. While journalism shares certain characteristics with other professions that prioritise truth-seeking and verification, such as law and academia, the literature suggests that journalistic practices are uniquely shaped by the field's evolving boundaries, its commitment to public service, and the constant negotiation between professional control and open participation in a rapidly changing media landscape. There is also variance between how journalists perceive themselves, and how audiences perceive the work of journalism.

Shapiro et al. (2013) found that while journalists agreed with the statement “the essence of journalism is a discipline of verification”, actual practices varied and compromises were common. An example of a compromise is writing a story based on single person’s version of events, rather than validating their claims via additional interviews, to meet a deadline. This is different from verification in the scientific method, where all data is held to a consistent standard. Shapiro et al. conclude that journalists’ commitment to verification could be viewed at least in part as a “strategic ritual” or boundary work – “something that legitimises a journalist’s role as being demonstrably different from other communicators” (pp. 668-669). Núñez-Mussa et al. (2024) similarly found that Chilean journalists interviewed for their study admitted verification practices often fell short of what was ideal due to practical challenges such as limited time and an over-reliance on official sources (p. 11).

A further question concerns what, exactly, makes someone a “journalist”. Digital publishing and distribution make it simple for anyone to practice journalism, even if they have not been explicitly trained in the craft. Citizen journalists routinely capture important news on camera phones and the crowdsourced investigation practised by Bellingcat shows that ordinary, motivated citizens are capable of investigations with higher standards of verification than many traditional journalism outlets (Higgins, 2021).

Anderson et. al. (2015) say that monolithic “industrial” journalism has been upended by three new entrants into the news ecosystem – individuals, crowds and machines.

> “Individuals are newly powerful because each of them has access to a button that reads ‘Publish’; material can now appear and spread, borne on nothing but the wings of newly dense social networks. Crowds are powerful because media have become social, providing a substrate not just for individual consumption but also for group conversation… And machines are newly powerful because the explosion of data and analytic methods opens whole new vistas of analysis." (p. 100).

But they argue professional and semi-professional journalists continue to have an advantage in areas such as originality, accountability and efficiency at finding certain types of information (for example, cultivation of sources). They have “moved higher up the editorial chain from the production of initial observations to a role that emphasises verification and interpretation” (p. 50).

Kovach and Rosenstiel agree that journalism can be practiced outside of traditional media companies. But they dispute Anderson et. al.’s assertion that journalists should move up the value chain and leave digging out facts and witnessing events to the crowd, as powerful interests will always hide crucial information from the public. Digging this out requires skills and access. “The fact that the White House now has a YouTube channel,” they say, “should not be mistaken for an administration being open or transparent". (p.30)

Shapiro (2024) points out the difficulty in legally defining "the press" in a world where anyone can publish information, leading to debates about accreditation and eligibility for state support. This ambiguity is echoed in Eldridge (2019) who examines the rise of "interloper media," actors operating outside traditional news organisations who nevertheless engage in journalistic activities. Lewis (2012) also highlights this blurring of boundaries, describing a tension between professional control, historically exercised by journalists, and the increasing participation of audiences in the news process facilitated by digital technologies.

This trend has continued and was reflected in the prominence of influencers and content creators in the 2024 US election. A Pew Research Center report (Stocking et al., 2024) that found 21% of Americans regularly get their news from news influencers on social media. The report found that male news influencers outnumber women by a ratio of two-to-one and are slightly more likely to be political right-of-centre. Strikingly, 77% have no background in news journalism.

But the literature also suggests there are core practices and values that, despite these blurred boundaries, remain central to journalism's claim to legitimacy. Eldridge (2019) notes that both traditional journalists and interlopers often share a commitment to reporting, gathering and verifying information, holding power to account, and serving the public. These shared values, even when practiced in diverse ways, contribute to a common understanding of what constitutes "journalism" and its distinct social role.

One way in which the journalistic practice of verification is seen as different from those of other professions is timeliness. While academics prioritise in-depth research and rigorous peer review, and lawyers focus on legal precedent and courtroom advocacy, journalists operate under tighter deadlines, often making decisions with incomplete information to inform the public about current events. Lewis (2012) describes this as a "professional logic of control", where journalists see their role as gatekeepers of information, responsible for filtering and disseminating news in a timely and accessible manner.

2.4.2 Audiences view journalistic practice differently

There is a difference in how journalists and the public perceive journalistic practices, especially when navigating the complexities of misinformation and the evolving media landscape. The Reuters Digital News Report (Nielsen & Fletcher, 2024) found that just 40% of respondents to a global survey trusted the news industry, with young people, people with less formal education and people from lower socio-economic groups trusting news the least. The report lists eight factors at play in whether audiences trust news, with transparency, high journalistic standards and fair representation of “people like me” being the three most important.

A similar result is found in Núñez-Mussa et al.’s (2024) analysis of the perspectives of journalists, editors, and various audience groups in Chile concerning misinformation. They found that journalists generally emphasise verification as a central tenet of their professional identity and a means of establishing credibility and trust. However, audiences often exhibit scepticism regarding the motivations underpinning journalistic work, viewing journalists as potentially influenced by political and commercial agendas. This perception contributes to distrust in the information presented by news organisations, particularly against the backdrop of rampant misinformation online, making it challenging to identify credible sources.

The paper recommends that journalists cultivate a better understanding of the diverse needs and expectations of their audiences. Rather than relying on social media engagement metrics or casual observations within their personal networks, the authors recommend incorporating qualitative research methods, such as focus groups, to gain more nuanced insights into how different audience segments interpret and evaluate journalistic practices. Additionally, they advocate for heightened transparency regarding journalistic processes, particularly those related to verification.

In Australia, Carson et. al. (2023) found that while trust in fact checkers is generally high among users who encounter fact checks on Facebook, trust levels were much lower among study participants who had a conservative ideology. Because of this, they recommend (among other things) that platforms consider showing a wide range of fact checking sources to users to ensure that their interventions are perceived as neutral.

Carlson (2016) uses the concept of "metajournalistic discourse" to describe public conversations about journalism that occur across various platforms, from traditional news outlets and opinion pieces to online discussions and academic debates. He contends that metajournalistic discourse plays a pivotal role in shaping both the meaning and boundaries of journalism, particularly during periods of significant change and uncertainty.

This discourse, Carlson says, is frequently characterised by tension and disagreement as journalists, non-journalists, media critics and academics engage in debates concerning the norms, practices, and legitimacy of journalism. He highlights how actors engage in "definition making," "boundary work," and "legitimation" through this discourse, actively shaping public perceptions of what constitutes "good" journalism.

The gap between journalistic and audience perceptions is important to this research. If technologists are sceptical of journalists' intentions and view them as compromised, they are unlikely to welcome additional journalistic input into their content moderation efforts for misinformation. This will be more deeply investigated in Chapter 4, where the views of journalists and technologists on journalistic practices will be reviewed and contrasted.

2.5 Journalism and information disorder

Information disorder problems present both a challenge and an opportunity for journalism. On one hand, getting a journalist to spread a false news story is often the ultimate aim of a misinformation campaign (Wardle & Derakhshan, 2017), and the news media as a whole has been regularly derided by some political actors as "fake news" (Caplan et al., 2018). On the other, the explosion of false information means that audiences need sources they can trust—this is an opportunity for journalists to re-establish credibility with their audiences and contribute to the public good (Ireton, 2018).

Kruger et. al. (2021) conducted a survey of 170 Australian journalists during the COVID-19 pandemic and found pervasive concern about misinformation, with just 14.1 % saying they had adequate training to deal with the problem and 44.3 % saying they had included mis- or disinformation in their reporting. The surveyed journalists said they wanted to upskill, but they believed newsroom leaders were unlikely to be supportive of giving them more time for fact checking or allowing them to train during work hours.

The following section considers four issues pertinent to information disorder and journalism – fact checking and UGC verification, transparency, strategic silence and reporting on unproven claims.

2.5.1 Fact checking and UGC verification

Fact checking and verification of user-generated content (UGC) are intertwined processes that draw upon journalistic expertise to assess the accuracy of material shared online. Traditional newsrooms, civil society groups, academic institutions and independent specialists all partake in this work, with notable examples including open-source intelligence agency Bellingcat, the British fact-checking unit Full Fact and The Washington Post's Fact Checker.

Techniques for verifying eyewitness-created social media content—often called open source intelligence (OSINT)—were pioneered by practitioners such as NPR reporter Andy Carvin, who relied on Twitter contacts to outpace established news outlets during the 2011 Arab Spring revolts (Hermida et al., 2014), and Eliot Higgins at Bellingcat, who assembled global networks of amateur investigators (Higgins, 2021).

Different tools and techniques apply to different content types, yet the fundamental elements of verification remain consistent: determining the content's provenance (who posted it), its source (the original author), its date (when it was created) and its location (where the event occurred) (Wardle, 2014). For instance, journalists frequently confirm the origins of images through reverse image searches (e.g. Google Reverse Image Search or TinEye) and employ mapping or geolocation software to verify a photo's date and place of origin (Barot, 2014).

Until recently, fact checking served as one of the most direct contact points between large social media platforms and journalism, typified by Meta's Third-Party Fact-Checking Program (3PFC) and TikTok's similar partnerships (Bélair-Gagnon et al., 2023). In these initiatives, outside fact-checking organisations flagged questionable content, rated its accuracy and, in Meta's case, applied labels to reduce the reach of falsities.

However, in January 2025, Meta abruptly ended 3PFC in the US, a move that threatens to decimate fact-checking teams reliant on Meta's significant financial support. In the wake of the announcement, PolitiFact said more than five per cent of its revenue came from Meta and Agence France-Presse (AFP), another major partner, indicated that Meta provided no warning before pulling the funding (Leingang, 2025).

The sudden loss of Meta's investment underscores the broader concerns identified by Graves and Anderson (2020) regarding the power platforms hold in shaping journalistic workflows. While fact-checkers gain visibility through these arrangements, they do so under constraints that often favour quick, digestible verdicts over nuanced reporting. Graves and Anderson caution that platforms, by defining the frameworks and resources for fact checking, effectively determine which claims merit scrutiny and how thoroughly they are examined.

Similarly, Meese (2023) writes about the media slipping into platform dependence, where a business or sector relies on platforms for its long-term survival, often leading to an alignment with platform goals and priorities. Meese argues that increasing reliance of the news media on platforms has significant implications for the sector's autonomy and its role within democratic societies.

Meta's abrupt withdrawal of 3PFC funding exemplifies this dynamic: a single corporate decision destabilises an entire subfield of journalism, raising the question of whether other media organisations or philanthropic sources will fill the funding gap.

Even so, fact-checking groups like PolitiFact and Lead Stories say they will keep operating regardless of the lost revenue. Their predicament illustrates how deeply reliant such initiatives have become on platform payments, provoking debate over whether truly independent fact-checking is possible in an ecosystem dominated by private companies. As Graves and Anderson (2020) note, meaningful transparency and accountability require more balanced power dynamics, where fact-checkers are not beholden to the very platforms whose policies they are meant to scrutinise. If fact checking is to remain a robust journalistic function, alternative funding models may need to be explored—particularly if Meta's recent exit signals a broader industry shift.

The technical aspects of how fact checkers and platforms interact will be considered at length in Chapter 3.

2.5.2 Transparency

Greatly increasing transparency around the mechanics of reporting is one of the ideas that is put forward as a way for journalists to re-establish trust with their audience. Ireton (2018) sees the volume and reach of misinformation and disinformation posing as news on social media as "a contagion that threatens further reputational damage to journalism", as "citizens struggle to discern what is true and what is false" (p. 33).

As a partial solution, Ireton urges transparency and renewed commitment to professional standards as a means of distinguishing journalism from other types of information, such as PR and propaganda. Journalism "needs transparency if the public is to trust that there is compliance with broad standards of verifiability and public interest" (p. 34). Wardle and Derakhshan similarly recommend that newsrooms "explain how the process of verification was undertaken" when debunking misinformation (p. 83).

A study of 579 US journalists found that a substantial number had increased their transparency practices in response to misinformation (Vu & Saldaña, 2021). This included limiting anonymity (such as exclusion of bylines on articles) and making clear how information was obtained. Journalists who saw their role as disseminators of information or watchdogs were most likely to have taken these steps.

2.5.3 Strategic silence

While reporting on misinformation has contributed greatly to public understanding, it can also inadvertently spread false claims more widely. Sometimes, "getting the mainstream media to amplify rumour and disinformation is the ultimate goal of those who seek to manipulate" (Wardle & Derakhshan, 2017, p. 13).

Donovan and Boyd (2021) advocate for journalists to use the tactic of "strategic silence"—refusing to cover certain topics to either reduce harms or prevent manipulation. This is an established media practice. For instance, media in many countries refuse to report on the details of suicide cases to prevent copycat incidents.

A real-world example of this was seen in the aftermath of the massacre of Muslim worshippers in Christchurch, when New Zealand media outlets collectively chose not to broadcast the testimony or views of the mass-shooter Brenton Tarrant, whose meme-laden manifesto was created to provoke the media into sharing his white supremacist ideas (Aigne Roy, 2020).

2.5.4 Reporting on unproven claims and scientific disagreement

Some major political flashpoints around information disorder involved decisions to ignore or underreport stories that later turned out to be either true or somewhat credible. One example of this is the New York Post's publication of materials from a laptop belonging to then-US presidential candidate Joe Biden's son Hunter during the 2020 election campaign (Morris & Fonrouge, 2020).

On the basis that the story may have been fabricated by Russian operatives, many mainstream news outlets ignored the story, and some social media platforms removed mentions of it. But the laptop was genuine and conservative news outlets and politicians strongly criticised the suppression of the story as political bias (Jenkins, 2021). Twitter CEO Jack Dorsey later apologised, saying that his company was wrong to block the story (Dorsey, 2020).

A second flashpoint were claims that COVID-19 may have been manufactured in a lab in Wuhan, China. Initially dismissed as a conspiracy theory by many in the scientific community, the theory was ignored by most mainstream news outlets and specifically banned from being mentioned in comments by some social media companies, only to gain credibility again in mid-2021.

In response to this, the now-defunct organisation First Draft put forward a series of recommendations for journalists covering the lab leak theory and similar incidents where facts have not been clearly established (Zhang, 2021). These include being transparent about the "ongoing and ever-changing nature of the investigations", identifying any complex political factors informing the debate, refraining from amplifying unproven information simply because it is being discussed by public figures or other media, using precise wording (particularly in headlines) and considering not writing about the topic at all (or, in other words, strategic silence).

O'Connor and Weatherall (2019) believe that journalists should trust scientific consensus and rarely go outside it. Their case is that the ethic of "fairness" leaves journalists vulnerable to misreporting scientific consensus in a way that slows progression to true beliefs and open to manipulation by propagandists – for example, the scientists who were paid by tobacco companies to produce studies that cast doubt on claims that cigarettes cause cancer. While not discounting that there are cases where a contrarian opinion turns out to be correct, they urge care be taken and alternative views not printed simply because they exist.[3]

2.6 Journalism and automated content moderation

It is common for disinformation literature from journalism researchers to mention at-scale content moderation. Likewise, it is common for literature from computer science researchers to mention the efforts of journalists. But there is only a small amount of literature that gives serious consideration to the ways in which journalism and technologists can work together on content moderation.

García-Marín et al. (2022) argue that there is a lack of collaboration between journalistic research into disinformation and technological research on how to detect disinformation using algorithms. They highlight the scarcity of publications on solutions to disinformation in communication journals. Instead, they say, the most impactful studies on using AI for automated fact-checking are found in journals focused on IT engineering, applied mathematics and IT systems management. Santos (2023) also notes that there is a gap between AI and journalistic practices.

This points to a divide between the disciplines, with communication studies largely overlooking the potential of technological solutions. García-Marín et al. note that most researchers studying AI-driven fact-checking are affiliated with IT engineering, computational sciences, machine learning, mathematics, and information management, and say the lack of engagement from the journalism

[3] It is worth noting that this will be a very controversial notion to many in the media industry, where skepticism and independence from authority are established norms.

scholars is a missed opportunity. They argue for a multidisciplinary approach to effectively combat disinformation, with input from both communication experts and those in technological fields.

A thorough search of available literature using several methods (Google Scholar, UTS Library and the AI research tool Consensus) bears out the observation that few papers consider how journalists and technologists can better collaborate. But while rare, there is some research that explicitly considers collaboration between journalists and technologists in the detection of misinformation at scale.

Vizoso et al. (2021) explore the strategies used by media and internet companies to combat the spread of misinformation, particularly deepfakes. They note that media companies and platforms share some similarities in their approaches to deepfakes. Both sectors increasingly recognise the importance of technology in identifying and mitigating the impact of misinformation. This has led to the development of sophisticated tools and algorithms for detecting manipulated content. Both sectors also engage with academic and research institutions to improve their methods and stay current with the latest developments in deepfake detection technology.

But the research also finds differences in how media companies and platforms address deepfakes. For example, media companies tend to focus on correcting misinformation rather than removing it. On the other hand, some platforms may opt to delete manipulated content to prevent further spread.

The paper sees increasing collaboration between media companies and platforms but provides just two examples of this. Firstly, Reuters partnered with Facebook to identify and debunk misinformation, particularly in the lead-up to the 2020 US election. Secondly, Facebook collaborates with fact-checking organisations to label false content and inform users.

Another paper that considers how journalists can contribute to at-scale content moderation is Schmitt et al.'s (2024) research into how explainability in tooling can aid collaborative human-AI disinformation detection. The researchers conducted a user study, and a crucial aspect of their methodology involved the inclusion of journalists as expert participants.

A total of 27 journalists participated, forming part of the 433 total participants, which also included 406 crowdworkers. This approach allowed for a comparison between expert (journalists) and non-expert (crowdworkers) users in terms of how they interacted with and perceived the AI system. Both journalists and crowdworkers engaged in a disinformation detection task using a "News Verification Dashboard", where they assessed news items sourced from accredited debunking websites (Snopes and Politifact). To test the influence of explainability features within the UI of this system, participants interacted with different versions, each providing varying levels of explanation.

Schmitt et al. found that journalists outperformed AI-only systems and AI-plus-crowdworker systems in correctly identifying misinformation. This conflicts with other studies showing crowdworkers achieving equivalent levels of performance to experts that will be discussed further in 3.5.1 Training data and ground truth.

A paper from Yameogo (2024) that proposes a system for analysing disinformation campaigns touches on the relationship between journalists and AI algorithms. Yameogo highlights the importance of expert feedback in this system, arguing that human intervention is crucial for refining the conceptual model and ensuring accurate and relevant analysis. Experts can provide contextual information, validate the results of the automated analysis, and suggest adjustments to the system based on their knowledge and experience.

While the paper doesn't explicitly name journalists as a desired expert type, the skills and knowledge required align strongly with journalistic practices. Journalists are explicitly mentioned as potential users of this type of system, which theoretically could help them better understand and respond to disinformation campaigns by providing semantically enriched data and insights into the connections between individual fake news items.

2.7 Critical perspectives on content moderation

In the early days of social media, the standard speech-rights position for Silicon Valley platform companies was First Amendment-influenced absolutism. Twitter's executives famously referred to the company as the "free speech wing of the free speech party" (Halliday, 2012), and Facebook CEO Mark Zuckerberg characterised his company as a "platform, not a publisher"—in other words, a neutral technology that was not obligated to play "gatekeeper" when it came to its users' speech (Marantz, 2019).

Russian interference in the 2016 presidential election raised critical questions about this laissez-faire approach to harmful speech (Napoli, 2019, p. 4; Suzor, 2019, pp. 94-95), a trend that intensified in the face of rampant misinformation about the COVID-19 pandemic (Douek, 2021a, pp. 800-804). Platforms were for a period expected to intervene more forcefully to control content on their services, and misinformation is a crucial category to be managed.

This focus on content moderation as a means of managing misinformation subsequently produced intense backlash from conservatives, who claimed that efforts to suppress claims about a rigged 2020 US election and Covid mandates were an attack on political speech and in violation of US First Amendment rights. The most visible manifestation of this was Elon Musk's purchase of Twitter, later renamed X. Musk sharply reduced the company's content moderation teams (Robison, 2024), leaked internal communications to sympathetic journalists in a project called the "Twitter Files" (Picchi, 2022), and during the 2024 US presidential election was repeatedly accused of using X to spread lies and distortions (Goldberg, 2024).

Platforms and academic institutions that study misinformation were the subject of an investigation and series of reports from the Judiciary Committee of the US Congress, led by Republican Jim Jordan. These reports were damning of both the platforms ("a consistent level of bias and patterns of capricious censorship in Silicon Valley") and of academics who study the phenomenon ("pseudo-experts").(Jordan et al., 2020; US Congress, 2023)

The committee, which consisted entirely of Republican lawmakers strongly supportive of Donald Trump, was in turn criticised as a witch hunt that misrepresented the work of researchers that shut down discussion of issues pertaining to misinformation while claiming to protect free speech (Masnick, 2023; Myers & Frenkel, 2023). Republican operatives also launched lawsuits against academic institutions and individual researchers (including some quoted in this thesis, such as Kate Starbird, Renee DiResta and Camilla Francois) (Myers & Frenkel, 2023).

Content moderation remains a controversial and fiercely contested field. Grimmelmann defines it as "the governance mechanisms that structure participation in a community to facilitate co-operation and prevent abuse" (2015, p. 47). All platforms moderate—if they did not, their products would be flooded with spam, pornography and other types of content that would make them unappealing to many users.

But as platforms such as Facebook, Twitter and YouTube have grown to enormous size, and as their content moderation efforts have expanded, a growing body of literature has expressed concern that a handful of mostly US-based technology companies has obtained the power to set rules for the speech of billions of people.

Gillespie says "the fantasy of a truly 'open' platform … is just that, a fantasy" (2018, p. 5). Platforms hide moderation from public view and largely disavow it. This has allowed them to maintain the illusion that they are neutral parties and stay in a "regulatory sweet spot", enjoying the safe harbour protections of Section 230 of the US Communications Decency Act, while avoiding onerous regulatory obligations. In fact, Gillespie says moderation is "in many ways *the* commodity platforms offer" (p. 13), as algorithms select content and display it to users in a fashion that will promote engagement with the product, thereby driving up revenue.

Klonick agrees that understanding how Section 230 "gives online intermediaries broad immunity from liability for user-generated content posted to their site" is important to understanding the platforms' approach to content moderation, but she also emphasises the tension that arises between these companies' professed commitment to free speech and complex business considerations relating to advertising revenue, user expectations and reputation. Platforms, she says, are "private self-regulating entities that are economically and normatively motivated to reflect the democratic culture and free-speech expectations of users" (2018, pp. 1602-1603).

2.7.1 Regulation and 'public interest'

Concern over platforms poses an obvious question – what, if anything, should be done to curb their power? Governments around the world have implemented or are considering laws to curb misinformation and other harmful forms of speech. Researchers have noted this turn to regulation (Flew & Wilding, 2020; Gorwa, 2024), but have also expressed concerns about the potential for bad laws that compel platforms to intervene even more aggressively in policing users (Douek, 2021b; Gibbons & Carson, 2022).

Napoli (2019) argues that the emergence of misinformation issues shows that social media should be considered through a media-centric lens, rather than a technology-centric lens, when considering regulation. Social platforms often argue that they are not media companies on the basis that they do not produce content, but Napoli says media companies also engage in distribution of content and display of content to users – functions that social media companies have become increasingly dominant in providing.

"Content creation / ownership has never served as a point of distinction in defining a media company from the perspective of those charged with regulating the media sector," he says. "A final irony is that some of the companies that made this argument [for example, YouTube] have subsequently vertically integrated into content creation." (p. 9)

Napoli also dismisses arguments that an abundance of engineers working at platforms proves they are technology companies, as engineers and technologists also led the way in the early days of television and radio. He points out that platforms are in the business of providing content to consumers and selling those audiences to advertisers, "a defining characteristic of the media sector" (p. 14).

This view contrasts with Klonick (2018), who argues that traditional categories like "broadcaster" or "editor" fall short of capturing the complexities of social media moderation. She proposes a "governance" framework as a more accurate model, highlighting the intricate sets of rules platforms have established to manage content. This framework, she argues, is essential for regulators to consider when crafting any new policies impacting online platforms. Klonick's analysis draws a parallel between content moderation and the legal system, pointing to the quasi-judicial role content moderators play (p.p. 1669-1670).

Napoli acknowledges that social media companies have made some efforts to address public interest concerns but criticises these efforts as insufficient due to platforms prioritising individual empowerment and profit over public service. Rather than calling for government regulation alone, he advocates for a broad range of media governance structures, involving stakeholders beyond government, such as the media, civil society groups, NGOs and users (pp. 163-198).

While he does not explicitly call for social media platforms to begin directly employing journalists, the concept of closer collaboration with the profession is implicit in many of Napoli's proposals for improving platform regulation. He advocates for a stronger public-interest orientation within social media, which would explicitly require the platforms to adopt journalistic principles such as independence, truth, and accuracy (pp. 171-172). To implement this, the platform companies would presumably need to collaborate closely with journalists and media organisations.

Napoli also calls for authoritative news content to be clearly distinguished from non-authoritative news on platforms (p. 175). And he says the news industry should consider the idea of a self-accreditation system for journalism so that platforms can use those authority signals to better rank content (p. 174). The need to define and prioritise 'trustworthy' sources also means that platforms would benefit from the knowledge and expertise of journalists. Napoli's argument has been taken up by scholars such as Meese (2023), who notes that "normative principles that supported journalism's important role within the wider system of liberal democracy" were challenged by the media's dependence on platforms for revenue and support. Meese says platforms' profit-maximising goals of platforms are not always aligned with the public interest and predicts "more specific (policy) reforms that focus on engagements between algorithmic systems and platforms as well as substantive structural interventions may be placed on the agenda soon".

The logic of these algorithms, often opaque to both news organisations and the public, can inadvertently or deliberately amplify certain types of content based on metrics that have little to do with journalistic standards of accuracy, context, or public value. This fundamental misalignment between the profit-driven algorithms of platforms and the public interest aims of journalism represents a significant and ongoing source of tension in their evolving relationship.

2.7.2 Free speech, transparency and human rights

In his book *Lawless: The Secret Rules that Govern our Digital Lives,* Nicolas Suzor argues that outsourcing regulation of the internet to intermediaries like social networks, search engines and internet infrastructure companies makes it a "lawless" space, because "so many of the decisions about what we can do and say online are made behind closed doors by private companies" (2019, p. 8).

Suzor argues for a rethink of "due process" in the age of massive online platforms via a "new constitutionalism – a new way of thinking about the power that technology companies wield and the discretion they exercise over our lives" (p.8). This would involve a "digital constitution" or bill of rights that builds "consensus about how the power of the network should be shared and limited, how those limits may be imposed, and by whom" (p. 113).

An international human rights framework provides a set of tools that could govern this consensus on minimum standards for technology companies (pp. 128-149). To achieve this, platforms would be required to have a principled approach to balancing competing interests – for instance, the rights of vulnerable groups to be protected from speech, and the rights of users to express themselves freely.

This would be “tailored to the independent norms of various networks and platforms, and the various needs and values of the communities of users they support” (p. 130).

Suzor proposes that all platforms should be required follow certain practices – increasing transparency with detailed and specific reporting, showing they are considering rights when rules are created, regularly reviewing rules to make sure they are in line with their commitments, and creating “scalable due process” that ensures users have adequate means to seek appeal and redress for decisions (pp. 142-149).

Suzor’s concern about transparency is widespread among content moderation researchers. Gillespie says the platforms’ rules, processes and outcomes are “shockingly opaque, and not by accident … content moderation should be much more transparent” (pp.198-199). Klonick says a lack of transparency makes it “hard to accurately assess the extent to which we should be concerned about speech regulation, censorship, and collateral censorship” (2018, p. 1665). And the Brookings Institute argues that “transparency is a necessary first step in creating a regulatory structure for social media companies” (MacCarthy, 2022).

Meese (2023) also advocates for transparency, but says the dynamic and evolving nature of platform algorithms means that opening “black box” systems to scrutiny is unlikely to yield good results. Instead, drawing on the work of Reider and Hoffman, he suggests that platform observability, which involves continuous and ongoing observation of platform behaviours and impacts offers a more effective way to understand the complex and unpredictable interactions between platforms and the information ecosystem.

Douek (2021a) argues for a content moderation framework that balances speech rights with other rights. She advocates a move away from what she calls a “posts-as-trumps” ethos – the laissez-faire, First Amendment-influenced approach to speech, which allows decision makers to sidestep “hard moral and political fights” (p. 775). Douek argues instead for a “proportionality” frame, where speech rights can be limited for legitimate purposes:

> “There are three main benefits: (a) it explicitly acknowledges interests other than the individual speech right, and thereby dignifies those interests and the importance of evaluating them in their particular context; (b) it is transparent about the value judgments inherent in constructing a system of freedom of expression; and (c) it encourages and rationalises remedial flexibility, rather than a binary “take-down/leave-up” paradigm of content moderation.” (p. 785)

This point is echoed by Napoli (2019, pp. 80-106), who says that with the advent of social media, the assumption that "more speech" will effectively counteract "bad" speech is no longer valid. The doctrine of counterspeech is undermined by the structure of the social media ecosystem, where individuals are more likely to encounter, trust, and share false information, often without being aware of the original source. This presents a fundamental challenge to the marketplace of ideas, and for democratic processes that rely on an informed citizenry.

2.7.3 Mistakes and efficiency

A further point that Douek’s essay raises is that “content moderation is impossible” – the scale of speech to be judged requires the use of automated detection, making mistakes inevitable. Moving beyond an absolute approach to free speech to a proportional or probabilistic model allows for a realistic discussion of how content moderation works, crucially accepting the inevitability of errors (p. 791). Gillespie expresses a similar sentiment, saying that “while we sometimes decry the intrusion of platform moderation, at other times we decry its absence. We are partly to blame for having put platforms in this untenable situation, by asking way too much of them" (2018, p. 197).

Common (2020) argues against over-emphasis on efficiency metrics in content moderation, criticising it as a superficial approach that prioritises the rapid removal of vast quantities of content over thoughtfully addressing complex social and ethical issues inherent in regulating online expression. She sees platforms as fixating on easily quantifiable metrics, such as the speed and volume of content takedowns, instead of more critical issues like bias, transparency, and accountability.

Platforms tend to make moderation policy in response to controversies and criticisms, Common says, leading to superficial solutions. For example, platforms may increase the number of content moderators or implement new tools, but these reactive measures fail to address the underlying causes of harmful content and do not result in meaningful, long-term changes. (p. 135) One example of this approach was Facebook's response to the live streaming of violent events, such as murders and sexual assaults. Facebook pledged to hire 3,000 new content moderators to address the issue (p. 136), but Common suggests they should have considered whether live streaming itself was feasible given the difficulties in moderation this style of content.

Emphasis on the "efficiency narrative", Common argues, can lead to over-reliance on simplistic technological solutions. She says laws like Germany's Network Enforcement Act (NetzDG, now superseded by the EU's Digital Services Act) enshrined the efficiency narrative in legislation (pp. 136-137). Requiring social media platforms to remove illegal content within 24 hours under threat of substantial fines results in the over-removal of content as platforms err on the side of caution to avoid penalties.

2.8 Chapter summary

This chapter provided a review of the academic literature relevant to the study of information disorder, journalism, and content moderation. It began by examining the definitional challenges surrounding key terms like misinformation, disinformation, propaganda, and fake news, highlighting the lack of consensus and the impact of political context.

The review then explored the debated causes of information disorder, contrasting arguments that focus on the role of social media platforms and their algorithms with those emphasising broader societal factors and psychological vulnerabilities. The potential exacerbating effect of generative AI was also considered.

Subsequently, the chapter delved into the nature of journalism, discussing its core tenets like verification, the consistency of its practices, and the divergence between professional self-perception and public trust. It surveyed journalism's specific responses to information disorder, including fact-checking, transparency initiatives, and strategic silence.

Finally, the review addressed the literature on automated content moderation, noting the relative scarcity of research bridging journalistic and technological approaches, and concluded by examining critical perspectives on platform governance, regulation, free speech debates, and transparency.

Chapter 3: How at-scale content moderation works

Chapter 2 defined content moderation and touched on some of the debates that surround the field. But to determine how journalists can contribute to at-scale misinformation detection, it's necessary to understand how content moderation systems work. This chapter examines content moderation systems from many angles—the role of policies and rules, how automated detection works, the role of humans in these systems, and the different types of actions that platforms take against violating content.

3.1 The rules of the system

The bedrock of any content moderation system is its rules. There is no universally defined standard for how these rules should operate and different platforms may use varying terminology to describe the same concepts. Despite this lack of standardisation, all platforms possess at least some rules, and the development and enforcement of these rules follow some basic principles.

Some rules are uncontroversial, such as filtering out spam (unwanted mass email marketing), a fundamental task for any platform aiming to build a large user base (Klonick, 2018, p. 1637). Blocking spam is generally not considered a suppression of legitimate speech, even though many types of spam are not technically illegal. However, other types of moderation are hotly contested. While content moderation rules would ideally be made in an iterative and principled way, the reality is that most platforms have created rules reactively in response to specific controversies. Gillespie (2018, p. 66) aptly captures this phenomenon using the metaphor "every traffic light is a tombstone"—that is, rules in content moderation systems tend to emerge organically in response to specific controversies rather than through proactive planning.

To understand this reactive process, Common (2020) advances a three-stage model. The first stage, "creation," involves the initial drafting of rules. While this might appear to be a straightforward process of setting clear guidelines, Common notes the creation stage is often hampered by the vague and overly broad language used in platforms' terms and conditions. This lack of clarity creates uncertainty for users, as they struggle to understand how these rules will be applied in practice.

Common argues that the second stage, "enforcement", is the most crucial. Enforcement can involve flagging of content by users or algorithms, and review by human moderators. Ultimately, it results in a decision to retain, remove or take other actions against the content, the user or both. Common criticises the potential for bias in decision-making during the enforcement stage, where moderators' personal beliefs or cultural backgrounds can influence their judgment, leading to inconsistent application of the rules. She also highlights the problem of platforms prioritising efficiency over effectiveness, focusing on swift removal of content rather than tackling the root causes of harmful behaviour. This, she argues, results in superficial solutions that fail to create lasting change on the platform. Finally, Common notes that inconsistent enforcement, driven by factors like the popularity of content or its perceived newsworthiness, further exacerbates the problem of vague rules.

Common's model concludes with the "Response" stage, which considers the mechanisms available to users who wish to appeal moderation decisions or advocate for changes to the platform's rules. She argues that current internal appeal processes are often opaque and users are left with limited options to challenge decisions or influence policies. This lack of transparency and accountability, according to Common, undermines the legitimacy of platforms as regulators of online speech. She argues that platforms should move towards a "culture of justification," offering more detailed explanations for their rules and decisions, and providing clearer avenues for users to seek redress.

An internal policy team is typically responsible for creating rules. While in the early days of platforms such teams tended to evolve organically in response to the challenges thrown up by at-scale content generation, platforms' policy teams are now large and well-established, typically staffed by a mixture of lawyers, policy experts, content experts and others specialising in trust and safety issues. There are typically two sets of rules – those that are shown in public-facing documents such as Terms of Service and Community Guidelines, and more detailed internal rulebooks designed to allow human moderators to make consistent decisions. External-facing Community Guidelines tend to be written in plain, even casual language (Facebook, n.d.-b; LinkedIn, n.d.). They focus on broad themes such as how users should behave, and while they will usually explicitly state what categories of content are subject to moderation, they do not describe these categories in detail.

Internal guidelines, on the other hand, are much more detailed, although the level of detail depends to some extent on the size and maturity of the platform. Klonick draws on legal theory to make a distinction between two different approaches to internal guidelines—"standards" and "rules". Standards are less specific than rules, focusing on a platform's values and purpose – for instance, that users should not be hateful. This looser approach has the advantage of flexibility but leaves room for arbitrary enforcement that may reflect an individual judge's bias. Rules, on the other hand, are detailed and specific—for instance, that a user should not be denigrated based on their ethnicity. This provides more clarity and reduces arbitrary decisions, but prescriptive rules invariably contain "gaps and conflicts" that lead to unfair results. For example, banning all images showing women's nipples impacts breastfeeding mothers and breast cancer survivors who wish to tell their stories.

Klonick describes how the enormous growth of the major platforms pushed them from standards-based systems to rules-based systems for two reasons. Firstly, the sheer volume of content meant that flexible decision-making was not scalable. Secondly, platforms outsourced content moderation work to third-party companies, who employed large teams of moderators for low wages in locations such as the Philippines and India. Klonick interviewed YouTube and Facebook Trust & Safety employees, who described how standards-based systems were gradually abandoned in favour of rules-based systems that sought to ensure that any judge, regardless of cultural background or personal political tastes, would reach the same "objective" decision (pp. 1632-1634).

This shift, she argues, mirrors the development of legal systems, where vague standards are gradually codified into specific rules to ensure consistency and predictability. Content moderators, tasked with applying these rules to user-generated content, operate much like judges, interpreting and applying complex regulations to specific situations. This analogy is further reinforced by Klonick's analysis of leaked internal guidelines, termed "Abuse Standards", which she argues resemble legal precedents in their structure and application (pp. 1643-1646).

Common (2020) strongly critiques this position, arguing that while the processes may appear similar on the surface, the substantial difference in training, decision-making timeframes and accountability mechanisms undermines the validity of this analogy. Content moderators often receive only a few weeks of training, in contrast to the years of education required for legal professionals. Furthermore, content moderators are often pressured to make swift judgments, sometimes within seconds, due to the volume of content they must review. This stands in sharp contrast to the deliberative and meticulous processes expected in legal settings (pp. 133-134).

To address the issues of accountability and consistency in content moderation, Common (2020) makes a concrete proposal: the creation of a publicly available "body of precedents," akin to case law. This repository would document how specific rules are applied in practice, offering users a clearer understanding of platform decision-making processes. Such a system, she argues, would enhance procedural fairness by reducing uncertainty and enabling more informed participation by users in the content moderation system (p. 155). This "body of precedents" would serve as a valuable resource for users, providing insights into the interpretation and application of content moderation rules, and ultimately contributing to a more transparent and accountable system.

3.2 Specific rules for information disorder categories

Each platform has its own set of policies for information disorder issues, along with concepts and terminology to define these issues. These guidelines are often strikingly similar. LinkedIn forbids "false and misleading content" (LinkedIn, n.d.); TikTok bans "misinformation", which is defined as "content that is inaccurate or false" and that also "causes harm" to individuals, TikTok's user base or society as a whole (TikTok, 2020); and Snapchat forbids "false information that causes harm or is malicious" (Snapchat, n.d.). Facebook and YouTube defined explicit guidelines relating to COVID-19 misinformation, spelling out in detail the exact claims that users should not make on their platforms, including that COVID does not exist, that it is linked to the rollout of 5G mobile phone technology, and that alternative treatments are guaranteed to prevent or cure the disease (Facebook, n.d.; Twitter, n.d.; YouTube, n.d.).

But there are also some striking differences. For instance, Facebook uses the term "coordinated inauthentic behaviour" to describe "groups of pages or people [who] work together to mislead others about who they are or what they're doing" (Gleicher 2018). This definition is explicitly focused on intent—whether the content itself is false or against Facebook's community standards is immaterial. Facebook also forbids posting "false news", but provides scant detail on what precisely this means, noting it is a "challenging and sensitive issue" with implications for free speech.

In contrast, Google avoids focusing on the intent of agents and as such does not make a distinction between misinformation and disinformation, as this necessarily involves making a judgement about intent. Instead, the company focuses on "specific behaviours and types of content we seek to either prohibit, discourage or reward" (Google, 2021).

Platforms may also treat reputable publishers differently. Both Google and Bing rank news publishers based on authoritativeness. Bing makes its criteria clear in its rules for news publishers (Bing, n.d.) – authority is based factors such as whether the site clearly identifies ownership, its authorship labelling practices, the level of commercial content that the site produces, and whether it clearly labels opinion content as different from news content. Twitter's Blue Tick authentication was for many years a gold standard signal that you could trust a user's identity, but under new Twitter owner Elon Musk, its meaning has changed to mean that a person has subscribed to Twitter's paid service, leading many to question the its value as a mark of source authority (Konger, 2023).

3.3 How machines moderate content

The scale of content on major platforms means automated detection is required to classify and escalate content. Gorwa et al. (2020) provide a useful model for thinking about different types of automated moderation, dividing them into "classification" and "matching" systems.

Matching systems are the simpler of the two solutions. They identify copies of known bad content through a process called "hashing", which uses some data extracted from content to create a unique "hash" or string. Even if content is altered, there is sufficient identical data to enable a match. Major technology companies including Microsoft, Facebook and Google have collaborated to create important content moderation databases that rely on this kind of technology. The two best-known are GIFCT, which contains known copies of terrorist materials, and PhotoDNA, which contains known copies of Child Sexual Abuse Material (CSAM). Some individual platforms have also created hash-matching moderation systems, the most prominent of which is YouTube’s ContentID system for detecting unauthorised use of copyrighted material.

Classification systems, on the other hand, proactively assess new content and predict whether it violates platform rules. These algorithms are created using deep learning techniques. Historically, most content moderation algorithms have been trained via a process called supervised learning, where human judges who are trained in a company’s content moderation policies label very large datasets as positive or negative for a particular content trait.

The model extracts mathematical representations called features from this data. When presented with a new example, the model looks for similar features and makes a prediction whether it is positive or negative for the trait. For example, a nudity classifier will be trained using thousands or even millions of images of human-labelled images. The algorithm finds commonalities in the data for the positive and negative classes and when shown a new photograph makes a prediction about whether it contains nudity.

In recent years, generative AI algorithms have been introduced into content moderation technical stacks. Meta’s Oversight Board says that new generative models “present major potential improvements in the ability to automatically identify violations of specific policy lines” (The Oversight Board, 2024).

3.4 Automated misinformation detection

To date, there is no matching system database like GIFCT or PhotoDNA for misinformation. But classification systems are widely used in misinformation detection. Shu et al. (2017) divide classification systems that can detect misinformation into two categories: models focused on “news content features” and models focused on “social context features”.

News content features include the linguistic and visual elements of an article or social media post. For instance, a news article’s headline, body, author and source fields can be examined and grouped according to attributes such as writing style, total words, punctuation, adherence to grammar and characters-per-word. Images can also be grouped according to features such as clarity and the ratio of images to text. From these features, models can be trained to detect misinformation. For instance, statements and viewpoints in an article may be extracted from body copy, then checked against a database of fact checks to automatically detect dubious claims. While this may not definitively identify misinformation, questionable content can be escalated to humans for review.

Social Context Features focus instead on the engagement that surrounds news content. For example, the model could be trained on data associated with accounts that are known to share or interact with misinformation. This could include follower counts, account age and total number of posts. The model would then look for other accounts that share these features. Assuming the data is correlated with a higher chance of misinformation, they can be passed on to human judges for additional review and potential action.

A key challenge for automated misinformation detection is the lack of high-quality, real-world data for training AI models. Zeng et al. (2024) attempted to address this by generating synthetic data to bolster real-world data. Noting that synthetic data often fails to generalise to the real world, the researchers used data selection methods to both increase the similarity of synthetic examples to real-world examples in both semantics (content) and distribution. They found that this method improved detection for “out-of-context misinformation” (where incorrect text is paired with a genuine image) and for digitally manipulated content.

3.5 Roles for humans in at-scale misinformation detection

Human input is used at all levels of content moderation systems, from defining rules and providing "ground truth" for content moderation systems, through to making granular decisions on the truthfulness of individual claims and repercussions for offending users. The work of fact checkers and journalists comprises part of this, but low-cost tiers of judges from Business Process Outsourcing (BPO) firms and crowdwork platforms also have a role to play. This section will provide a brief description of the various human roles in these systems.

3.5.1 Humans provide ground truth and training data

Ground truth is a fundamental concept in machine learning. It refers to the reality that the algorithm is trying to model. Google's AI glossary defines it as: "The thing that really happened. For example, consider a binary classification model that predicts whether a student in their first year of university will graduate within six years. Ground truth for this model is whether or not that student actually graduated within six years." (Google, 2024a)

Ground truth is typically captured in what is called a "golden dataset", which is a set of manually curated data that has been labelled by experts who have domain knowledge (Google, 2024a). Golden datasets can be used both to train and to test algorithms, although the same dataset cannot be used both to train and to test a model, because if the data is contained within the model already, any subsequent test with the same data will not give a true indication of performance.

The work of journalists and fact-checkers already serves as ground truth in publicly available databases that are used to train AI models for detecting misinformation (Santos, 2023). Online repositories containing large volumes of fact-checked information began to emerge in 2017, directly addressing the initial challenge of limited data availability for training these models. Literature from data science on training misinformation detection models often acknowledges the use of these datasets for both training and evaluation (Asr et al., 2024; Asr & Taboada, 2019; García-Marín et al., 2022; Zeng et al., 2021)

Santos also notes that AI systems perform more effectively when they consider the contextual factors that human fact-checkers utilise. Human fact-checking involves assessing the historical context, individuals involved, locations, and other pertinent details related to the event in question. By incorporating these nuanced considerations into the AI model, its accuracy in detecting subtle forms of disinformation, such as sarcasm or irony, can be greatly enhanced.

3.5.2 Enforcement and response decisions

What Common (2020) calls the "creation" phase of content moderation, where policies are decided, codified and translated into rules for moderators to follow, is typically the responsibility of full-time employees at major platforms. The employees are usually very experienced and often have a background in law or public policy. But since about 2009, the work of enforcement and response has been typically passed on to Business Process Outsourcing (BPO) firms who employ cheaper labour in locations such as India, the Philippines and Latin America (Klonick, 2018, p. 1643).

The best-known arrangement is Accenture's extensive content moderation work for Facebook (Satariano & Isaac, 2021), but there are myriad other companies that take on this work. Moderators are trained on a platform's policies and provided tools for rating content. Content is typically reviewed first by an AI layer before escalation (as human review is typically costlier than an AI review). If a worker is unable to decide on a difficult case, there will typically be a channel for them to escalate to a higher tier of judge (such as a manager or platform employee) for a final adjudication.

For example, when content is flagged as potential misinformation on TikTok by users or detection algorithms, it is first routed to an internal team of moderators with “enhanced training”, who check the content against their claims database. If the moderators are unable to find a reference in the database, they subsequently pass it on to a fact checker for assessment. Content that is found to be false or unverified could be taken down, have its distribution reduced or be labelled as untrustworthy (TikTok, n.d.).

These outsourcing arrangements have been the subject of many critical reports, documenting the poor working conditions of underpaid workers who are asked to review thousands of horrific images at speed each day, with little support for their mental health (McIntyre et al., 2022; Roberts, 2021).

3.5.3 Fact checking

Platforms have used the work of fact checkers extensively and money from technology companies has been crucial to funding fact checking globally. However, this funding model can vanish overnight. Meta terminated its US fact-checking partnerships in January 2025 and may extend this globally (Isaac & Schliefer, 2025), which illustrates that short-term contract arrangements leave fact-checkers financially vulnerable and may compromise broader efforts to mitigate misinformation.

Fact checkers in Meta’s 3PFC program are provided with tools that allowed them to proactively identify false information on Meta platforms, but they are also paid to check the claims that Meta’s algorithms surface through a proprietary tool. The fact checkers review and rate the accuracy of stories – if found to be false, these stories may be labelled or have their distribution throttled, although they are not removed entirely unless they violate other Meta standards, such as coordinated inauthentic behaviour or hate speech (2021).

TikTok also pays fact checkers for their work, partnering with 19 organisations accredited by the International Fact Checking Network (IFCN), including AAP in Australia. The platform maintains a database of fact-checked claims (TikTok, n.d.).

Google funds fact-checking organisations through grants (Nyaricki, 2024), but does not have a pay-per-fact check system like Meta and TikTok. One crucial role it plays in the fact-checking ecosystem is its partnership with the Duke Reporters’ Lab on a standardised mark-up for fact checks called ClaimReview. This allows newsrooms and other accredited organisations to upload fact checks, either through a tool or via automated mark-up that search engines and social media platforms can read.

Some platforms treat ClaimReview fact checks as more authoritative than other sources and give them higher visibility. Google shows snippets from fact checks in response to search results concerning queries that are highly correlated with misinformation. Microsoft’s Bing search engine and Facebook also give higher prominence to ClaimReview factchecks (Graves & Anderson, 2020).

ClaimReview data is freely available to anyone through an API. Google has also created a tool called Fact Check Explorer that allows users to browse claims through a simple search interface, and Duke Reporters’ Lab has developed a similar Fact-Check Insights tool. (*The ClaimReview Project*)

Graves and Anderson (2020) call this “structured journalism”, which standardises and organises news content into machine-readable formats to make news stories more accessible and understandable for both human and machine audiences. The ClaimReview standard serves as an example of this concept, demonstrating how journalists can leverage technology to promote their work and ensure its continued relevance in the digital age. But as noted in the literature review, Graves and Anderson express concerns about the disciplining effect that this can have on journalistic practices, as data is homogenised and simplified to suit platform purposes.

Jiang et al. (2020) attempted to automate the process of creating ClaimReview entries when a new fact check was created, in response to the insight that less than half of fact checkers utilised the tool in 2019, primarily due to time constraints.

The key challenge for automating ClaimReview creation was identifying key “factors” such as “claim”, “claimant” and “verdict”, and extracting this data from long fact checks averaging more than 1000 words. While the research (which utilised a fine-tuned model of BERT, a Google model that was a precursor of its newer Gemini models) was promising, the variability of fact check structure meant that results for lesser-known fact checkers (who were underrepresented in training data) were significantly worse than for the top five fact checkers (who comprised 94% of training data).

3.5.4 Crowdworkers versus “experts”

Determining the truthfulness of claims is crucial for both AI model training and content moderation. This has led researchers to propose the use of crowdsourcing systems to cheaply source judgements at scale. In such a system, non-expert human judges are asked to judge the truthfulness of claims on platforms such as Mechanical Turk and Clickworker. Shabani and Sokhn (2018) advocate for a model that combines automated models with crowdsourcing, citing experimental evidence that this combination boosted accuracy in detecting false claims. They dismissed using journalists in such a system because while they were “well skilled” for the task, “employing experts becomes expensive and slow” (pp. 302-303).

However, research on crowdsourced truthfulness assessments has yielded inconsistent results. Barbera et al. (2024) investigated these inconsistencies, finding that crowdsourcing can be reliable for misinformation assessment under specific conditions: a high-quality platform, a well-designed task, and a carefully selected worker pool. They emphasise the importance of these factors to achieve higher agreement between crowd workers and experts.

However, Schmitt et al. (2024) found journalists to be significantly more accurate in identifying disinformation than both AI systems and crowdworkers. Crowdworkers were more prone to "blind trust" in AI, particularly when presented with incorrect AI labels, suggesting a lack of critical scrutiny. However, including free-text explanations improved crowdworker performance, bringing their accuracy closer to that of journalists.

Santos (2023) and Schmitt et al. (2024) highlight the potential for variability in crowdworker knowledge and biases. While experts provide higher quality data, their availability is limited. Therefore, robust quality control for crowdsourcing and strategies to maximise expert contributions are essential.

Demartini et al. (2020) propose a hybrid approach integrating AI, crowdsourcing and expert fact-checkers. This tiered system leverages the strengths of each: AI for efficiency, crowdsourcing for scalability, and experts for validation. This addresses limitations and promotes transparency, while mitigating bias and explainability concerns.

3.6 Treatment of ‘violating’ content

3.6.1 Content removal

The simplest way to deal with violations is to remove the content. But this kind of action is highly visible and potentially controversial. Content removal can come prior to publication (ex-ante moderation) or after publication (ex-post moderation). For high-severity content moderation problems like child sexual abuse material (CSAM) and terrorist content, takedowns are a standard action and may be mandated by law. The Australian government’s Abhorrent Violent Materials powers, established in 2019 after a notorious massacre of Muslim worshippers in New Zealand was livestreamed to Facebook, obligates platforms to remove content within 15 minutes of receiving a “blocking request”. Failure to comply can result in a fine of up to AUD$11.1 million or 10% of the annual turnover of the company (Attorney General's Department, 2019).

Douek notes that when it comes to misinformation, taking down content presents challenges. Platforms and fact checkers can make mistakes and suppress legitimate speech; authoritarian governments force platforms to take down content critical of them on the basis that it is misinformation; at-scale AI content moderation can be “blunt and stupid”; and removing misleading content does not address the underlying social conditions that may have produced it in the first place (Douek, 2021b). While takedowns can be used in critical misinformation scenarios, platforms typically prefer softer interventions.

3.6.2 Content Filtering

A less invasive action is to filter or downrank content so that it is visible to fewer users (for instance, only users who’ve explicitly searched for a term) or even just the user who posted it. But this technique is also controversial as it disguises the act of moderation – the person who posted the content may notice that it has less distribution than they expect, but it is difficult for them to interrogate the reason for this. The lack of transparency inherent in downranking content has led to accusations of deliberate “shadowbanning”, particularly of conservative accounts. For instance, in 2022 independent journalist Bari Weiss was provided access to Twitter’s internal message archive concerning content moderation, and found that Twitter engaged in a practice called “visibility filtering”, which reduced the reach of Tweets and blocked users from finding targeted accounts via search (Weiss, 2022).

3.6.3 Labelling / context

The most visible form of intervention against misinformation has been the addition of labels and contextual information to posts. The simplest of these labels will offer contextual information such as a link to an authoritative website that contradicts the labelled post. A more forceful intervention will attach a warning label to a user’s post, identifying the information as false and pointing out to a fact check. The most intrusive interventions will cover content and require a user to click to read the post.

Platforms continue to develop new ways of labelling content with context to help users make decisions about provenance and reliability. For instance, in March 2023 Google made a range of new tools available in the Asia Pacific region, including an expandable menu next to Google search results that shows additional context about the provenance of information, and panels on YouTube that give “topical context” on topics where misinformation is common (Liu, 2023). There are also standalone businesses that provide contextual labelling to platforms—NewsGuard has employed a team of journalists that give trust ratings to news websites in the US, Europe and Australia. Microsoft has made NewsGuard’s ratings available for free to all users of its Edge browser (NewsGuard, 2021).

An interesting approach to labelling is X's Community Notes feature (previously called Birdwatch). Instead of leaving annotation of misleading information up to professional fact checking organisations, Community Notes relies on ordinary Twitter users. To participate in the program, users sign up and are then able to make notes on posts. Notes that are identified as helpful by a wide range of people are applied to posts, with posts deemed helpful by users with opposing viewpoints given particular emphasis (*Community Notes Guide*, 2023). After ending its support of fact checking in the US in January 2025, Meta introduced its own version of Community Notes(Isaac & Schliefer, 2025).

3.6.4 User banning

The harshest and most controversial action a platform can take is to ban a user entirely. There have been a number of notable cases relating to misinformation, including the banning of former US president Donald Trump from most major social media platforms for claiming that the 2020 US election was rigged (Alba & Silver, 2021), the removal of accounts and content promoting the Q-Anon conspiracy theory (Porter, 2021), and the ejection of conspiracy theorist Alex Jones from all major platforms including YouTube (Coaston, 2018). Recently, many of these bans have been reversed, most notably at X under new owner Elon Musk, who has reinstated both Trump and Jones (Kleinman, 2022).

3.7 Making decisions transparent

Another aspect of content moderation systems that needs to be considered is how results are made transparent to users and governments. As noted in Chapter 2, there is widespread concern among content moderation researchers about the opaque nature of these systems. Many researchers, politicians and civil society groups have criticised platforms for withholding data, obfuscating policy and ad-hoc enforcement of poorly explained rules. (Acker & Donovan, 2019; Douek, 2021b; Gillespie, 2018, p. 75)

Governments share this concern and are pushing to legislate transparency requirements for platforms. Most prominently, the EU's Digital Services Act (DSA) specifies a wide range of requirements that platforms need to adhere to when operating in EU member states. These include mandatory government takedown orders for content that is illegal in a particular country (such as images of Swastikas in Germany), a right for users to be informed of and appeal moderation decisions and provisions allowing researchers access to data and information about how content moderation and algorithms work (*Questions and Answers: Digital Services Act*, 2023).

Platforms have made some efforts to address transparency concerns. For instance, Facebook, Google and Reddit endorsed the Santa Clara Principles on Transparency and Accountability in Content Moderation, which sets an expectation that major platform companies will undertake a broad range of transparency efforts, including publishing a wide array of data, exposing algorithms to auditing, providing detailed notice to users whose content is removed and giving users an avenue of appeal. (*Santa Clara Open Consultation Report*, 2021)

In Australia, major platforms except for X voluntarily follow the Digital Industry Group's (DIGI) Code of Practice on Disinformation and Misinformation, publishing reports on their efforts to curb misinformation (DIGI, 2021). As noted in the literature review, the Australian government attempted to create new laws that would have given additional powers to the Australian Communications and Media Authority (ACMA) to police misinformation, but this legislation was abandoned in November 2024.

The DIGI code lays out seven objectives that the platforms may address, including having moderation safeguards in place to protect against harm from misinformation, disrupting monetisation incentives for bad actors, providing support to research and providing more transparency. Platforms are required to at minimum reduce the risks of harms that arise from mis- and disinformation and produce an annual transparency report, and may opt-in to other objectives in the code such as supporting researchers and disrupting financial incentives for mis- and disinformation if it applies to their business.

Transparency reports from major platforms submitted under the code so far tend to be very detailed on certain aspects of content moderation, but they also exclude some crucial information that would make clearer the ambiguities, trade-offs and inevitable mistakes that come with moderation. For instance, Facebook says it removed more than 14 million instances COVID-19 misinformation globally from March to December 2020, including 110,000 piece of content from Australia (Facebook, 2021).

Douek (2021a) makes the point that content moderation is inherently probabilistic, but public disclosures from platforms often obscure the reality that mistakes are unavoidable and expected. She proposes that an essential part of transparency should be reporting on the number of errors in moderation. The DIGI transparency reports to date contain no detail on AI error rates, which would establish how much content their systems are unfairly penalizing and how much misinformation is going undetected. There is also scant information on the interplay between automated and human review – how often are decisions made without a human in the loop?

3.8 Chapter summary

This chapter provided a detailed examination of how large-scale content moderation systems operate, particularly concerning information disorder. It outlined the foundational role of rules and policies, discussing their often reactive development, the distinction between external guidelines and internal rulebooks, and the shift from flexible standards to more rigid, scalable rules.

The chapter explored specific platform policies related to misinformation and disinformation, noting variations in terminology and approach (e.g., focus on intent vs. behaviour). It then explained the technical mechanisms of automated moderation, differentiating between matching systems (like hashing for child sexual abuse material) and classification systems (using AI/ML to predict violations based on content or social context features).

The crucial role of human input was detailed, covering the provision of “ground truth” data for AI training, the use of outsourced moderators for enforcement, the specific function of fact-checkers, and the merits of using crowdsourcing versus expert judgment. Finally, the chapter surveyed the range of actions taken against violating content—from removal and filtering to labelling and user bans—and discussed the ongoing challenges and demands for greater transparency in moderation processes and outcomes.

Chapter 4: How journalists and technologists see information disorder problems

4.1 Introduction

This chapter delves deeper into participant perceptions surrounding the complex phenomenon of information disorder, using data gathered in interviews with eight journalists and six technologists, as described in Chapter 1.

All participants were asked a series of structured questions at the beginning of each interview. This structured component consisted of two parts.

1. Participants were asked to provide a definition of four terms: misinformation, disinformation, propaganda, and fake news.
2. Participants were asked to indicate their level of support for seven controversial statements.

Both definitions and statements were derived from contentious points in the literature, and this chapter compares the data collected from the participants with views covered in the literature review. This analysis helps us answer the research question "how can the knowledge of journalists best be combined with scalable technological systems to combat information disorder?" in two important ways.

Firstly, for journalists and technologists to jointly work towards solving information disorder problems, it is necessary for them to have a shared understanding of the problem. There is some evidence in the literature for the existence of this shared understanding—for example, Shu et. al.'s "technologist" summary of machine learning techniques to identify misinformation (2017) cites Wardle and Derakhshan's "journalistic" information disorder paper (2017) to define the problem they are trying to solve.

But the views of journalists and technologists on key issues are not contrasted in any of the literature review material. Direct comparison of viewpoints in this data is an original contribution and even with a relatively small sample size provides a more nuanced understanding of the points of similarity and difference between these two groups.

Secondly, while the journalism industry views itself as a truth-seeking and producing industry that has a crucial role to play in democratic societies (Kovach & Rosenstiel, 2014), there are critics of the media who argue that journalists don't live up to their self-proclaimed standards (Kruger et al., 2021), or worse are unwitting propagandists who either knowingly or unknowingly participate in the manufacture of untrue narratives for political or commercial ends (Moschella, 2022; Zollmann, 2017).

Again, while there is a wealth of literature about the virtues and sins of journalism, there is not a lot of information on how technologists view journalists as partners in misinformation detection. Shabani and Sokhn (2018) do say that journalists are well-suited to fact-checking tasks but note that their involvement in content moderation systems faces scalability issues due to cost and speed. Graves and Anderson (2020) highlight how journalists' fact-checking processes can be integrated into content moderation systems through "structured journalism" but worry this risks oversimplifying complex news narratives in pursuit of algorithmic clarity.

Directly asking technologists about their views on the suitability of journalists for work in these systems provides an opportunity to generate new knowledge about the feasibility of deeper partnerships between journalists and technology companies to solve these problems.

It is also worth noting that all interviews took place prior to Meta's January 2025 decision to end its fact-checking program (Isaac & Schliefer, 2025). Had these changes occurred earlier, participants might have expressed deeper cynicism toward platform-led moderation.

4.2 Definitions of key information disorder terms

As noted in Chapter 2, definitional issues are widely acknowledged as a major challenge in the study of information disorder, a problem that is exacerbated by the multi-disciplinary nature of the issue, which has been addressed by fields as diverse as journalism, psychology, law, data science, engineering, and political policymaking. One large meta-study made a key recommendation that more rigour and specificity was required in defining foundational terms that are often used in the field (Tucker et al., 2018).

There were 14 individual responses to each of the four definitions (misinformation, disinformation, propaganda, fake news). These responses were annotated and closely compared to draw out commonalities and differences in perspectives at an individual and group level.

This paper makes use of the term information disorder, an umbrella definition advanced by Wardle and Derakhshan (2017). The term is not widely used outside of specialist and academic literature, so participants were not asked to define information disorder, but several did bring it up independently.

4.2.1 Misinformation

The definitions section of the literature view makes clear that there is broad is alignment among researchers that misinformation concerns inaccurate or misleading information disseminated in the public domain. As discussed above, one survey of 150 academic experts found broad consensus defined misinformation as "false and misleading information", although there was variance on the "importance of intentionality and what exactly constitutes misinformation" (Altay et al., 2023).

Australia's abandoned legislation on misinformation echoes the language of "false, misleading, or deceptive content" but adds a further dimension that it is "likely to cause serious harm" (Parliament of the Commonwealth of Australia, 2024). The EU's *Strengthened Code of Practice on Disinformation* uses a very similar definition of "false or misleading content shared without harmful intent" (European Commission, 2022).

Data gathered from participants reflects the discussion in the literature review. When asked to define misinformation, all participants gave an answer that boiled down to "false and misleading information".

There were subtle but meaningful distinctions in focus and emphasis between the journalist and technologist groups. While all concurred on the basic premise that misinformation involved the spread of false or inaccurate information, journalists in this sample were somewhat more likely to elaborate on the socio-psychological aspects that lead to the spread of misinformation within communities, emphasising the role social dynamics play in its spread and the consequences of this spread.

J1, for instance, noted that the unintentional sharing of false information was "driven by socio-psychological factors". They said: "Online, people perform their identities, they want to feel connected to their 'tribe'. The pandemic helped to [highlight] a nuanced point about the definition of misinformation in that people were often unwittingly sharing information that they didn't realise was false but did so as they thought they were protecting their loved ones."

J6's definition for misinformation focused on harmful outcomes in the real world. They said: "(It's) anything that's not true and that is broadly harmful to people… if it's not harming anyone then it's not worth getting excited about". This echoed government legislation in Australia and Europe, where an emphasis on harm was a critical component of misinformation definitions (European Commission, 2022; Parliament of the Commonwealth of Australia, 2024).

Tech participants tended towards plainer and more concise definitions. T4, for instance, referred to misinformation as a misleading "claim … that becomes viral, spread frequently in a domain", while T1 simply said "misinformation is information that is false and spread by people who believe it is true". But many journalists were also concise – J5's initial response was: "False information – I'll keep it as simple as that."

As in the literature, participants were split on whether intent mattered in the definition of misinformation. Some believed that false information should only be called misinformation when it was shared by someone who genuinely believed it. Others said that any false information spread online could be called misinformation.

4.2.2 Disinformation

Disinformation is defined by participants as intentionally false information spread to deceive. All participants said that disinformation was characterised by the deliberate, intentional dissemination of misleading or false information. T5 defined disinformation as "information that is deliberately, deliberatively inaccurate", while T3 defined it as "intentionally spread false information".

However, the debate from the literature around intentionality was also reflected in the participants' responses. Many researchers view misinformation and disinformation as distinct categories, with misinformation unintentional and disinformation intentional. Examples of this include Wardle and Derakhshan's information disorder framework (2017), Australia's abandoned misinformation legislation and the EU's code of practice (European Commission, 2022; Parliament of the Commonwealth of Australia, 2024).

There are other definitions that either see both intentional and unintentional spreading of false information as misinformation (and so disinformation is a form of misinformation) (Altay et al., 2023), or believe that intentionality is difficult and often impossible to ascertain, so not useful in a definition of misinformation or disinformation (Zeng & Brennen, 2023).

Six participants (five journalists and one technologist) offered a comment on the hierarchical relationship between misinformation and disinformation. J5 and J8 explicitly placed both terms into the rubric of information disorder (Wardle & Derakhshan, 2017). "Misinformation, if you look at it as a bubble, will probably be bigger and disinformation would sit inside it, so you could argue that this is a subset of [misinformation], but I would sort of think of them all as underneath the information disorder rubric," J5 said.

J3 argued that while the two terms were connected, disinformation was not a subset of misinformation, based on the centrality of intent to understanding disinformation, saying: "Users who are part of a network that seeks to spread disinformation use misinformation narratives, and then those misinformation narratives get spread by other users who are not necessarily part of a disinformation campaign, and then it becomes misinformation."

J4, on the other hand, said that "disinformation is really a subclass of misinformation… it's so confusing the way that people use the two words". This confusion was echoed by T4, who admitted that their demarcation between mis- and disinformation was not totally clear. While noting that motive mattered for disinformation, they said: "I usually refer to all of it as misinformation, to simplify, but I know there are more fine-grained definitions and categories." Interestingly, T4 was an AI generalist who came to work on misinformation problems relatively recently and so was not as familiar with the academic debates surrounding definitions as some other participants.

The variety of views expressed here points to an opportunity to drive further clarity on the precise separation between misinformation and disinformation. The fact that information disorder was mentioned independently and without prompting by two participants shows the impact that Wardle and Derakhshan's framework has had in the field.

Another aspect of Wardle and Derakhshan's framework that showed up in the data was the importance of analysing the motivations of agents who spread disinformation. There was broad consensus that an agent spreading disinformation stood to gain something. A notable 79% (11 out of 14) of participants said those who spread disinformation aimed to gain politically or financially, or to sow discord and harm among target groups. T2 said that "disinformation is the intentional spreading of false narratives against at-risk groups with the intention of creating harm," while J1 said "disinformation is content that is intentionally false and designed to cause harm".

Journalists were more likely to include an assessment of the societal and political implications of disinformation in their definitions (5 out of 7, 71%), perhaps reflecting the broader engagement with public discourse that their profession demands. J7 described the act of sharing disinformation as "anarchistic" and viewed it as a "threat to the democratic institutions that regulate us". J5 said that agents of disinformation could be inspired by "religious views, US political views, financial motivation … and sometimes people who just want to stir some shit".

But these differences were subtle. Overall, there was little variance between journalists and technologists in their base-level view of disinformation.

4.2.3 Propaganda

Propaganda is widely discussed in literature on misinformation. But defining propaganda is tricky and definitions have evolved over time (Tutui, 2017). This ambiguity is reflected in the participants' responses, where varying perspectives emerged on the nature of propaganda, the role of intent, and the actors involved.

All participants recognised the persuasive intent behind propaganda, echoing the views of Guess et. al. (2020) who define propaganda as "any communications that are intended to persuade people to support one political group over another". Participants consistently defined propaganda as information deliberately crafted and disseminated to influence public opinion or advance a specific agenda. J2's response stating that "propaganda is typically a state-sponsored activity for hard political ends" exemplifies this understanding.

Participants also acknowledged that propaganda could utilise truthful information, aligning with Guess et. al.'s (2020) assertion that propaganda can involve "the use of truthful information to advance a particular agenda." T1, for instance, described propaganda as potentially involving information that is "exaggerated, maybe half a story, but it's misleading in order to present something political." J6 explicitly framed up propaganda as "something that might actually be true… [but] you're not being fair". These perspectives echo Stanley's view, cited in Tutui (2017), that propaganda is not necessarily insincere, even though it may exploit biases and social structures to distort reality.

Interestingly, the participants' views on the source and nature of propaganda diverged. While most (10 out of 14, 71%) associated propaganda exclusively with political actors, a minority (4 out of 14, 29%) held a broader view, encompassing any effort to use information for strategic advantage. This mirrors broader points of view expressed by Ellul (1957) and Bernays (1928).

From the politics-only side, T1 said propaganda was "state sponsored information, with the intent to persuade and opinion one way or another", and J3 said it was "the intentional dissemination of information to influence public opinion by state actors or … pro-national [actors]."

T6, on the other hand, presented a broader view of propaganda as “a particular perspective on a narrative or topic to paint one group as the benefactors and the others as the victims or detractors.” J8 offered the broadest definition of all participants as “basically any content deliberately spread to make one side in a conflict look good … I mean, sometimes advertising is propaganda, like ‘the camera on our phone is 10 times better than the other guy's camera’."

Another interesting observation came from, J1 who brought up the concept of "ampliganda" in the context of propaganda (DiResta, 2018). J1 described ampliganda as "the distribution of information tailored to influence public opinion" but uniquely facilitated by online personalities and influencers, rather than solely by state actors or official media outlets. This trend represents a rebalancing of power towards individual creators over mainstream media organisations, where anyone with a significant online following can shape public discourse, blurring the lines between personal opinion and orchestrated propaganda campaigns.

4.2.4 Fake News

As noted by Caplan et al., the phrase fake news came to prominence in 2016 after a Buzzfeed article on explicitly fake, viral election stories about the US election, but later became politicised. US President Donald Trump, his supporters and sympathetic media used the term to critique mainstream media, while scholars and researchers used it to refer to false stories that mimicked the aesthetics and signifiers of news coverage (Caplan et al., 2018). Scholars have since moved away from using the term in serious research, arguing that it should be replaced by more specific and descriptive terms that are less tied to political debate (Wardle & Derakhshan, 2017).

The research data reflects this sentiment. Almost all respondents (13 out of 14, 93%) criticised fake news as a term that has been co-opted by political figures to dismiss or discredit information that contradicts their viewpoints or interests. Several respondents explicitly noted US President Donald Trump’s use of the term as the turning point for its politicisation. (T1: “Donald Trump used the phrase to dismiss everyone who disagreed with him, so it ceased to mean in anything a few years ago now.”)

Most respondents (12 out of 14, 86%) also said use of the term fake news was imprecise and subjective. Several noted that they try not to use the term, particularly for any serious discourse. (T5: "Such a bastardised term at this point, but sadly used to just characterise anything that doesn't fit with your view”; J2: “It's become a trope used by so many different people that it has multiple definitions. It's a term I would try to avoid at all costs.”)

Four respondents (29%, three journalists and one technologist) mentioned a more specific definition of fake news as a website set up to mimic a mainstream news organisation to give their fabricated news stories a veneer of credibility, whether for propagandistic or commercial purposes. J5 provided a short history of the term: “[Fake news] has been around for quite a while. It was most recently popularised by (Buzzfeed reporter) Craig Silverman, who did fantastic reporting into how websites were being set up with completely false stories and spread to gullible audiences for further spread. The most famous of these being Pope Francis endorses Trump.”

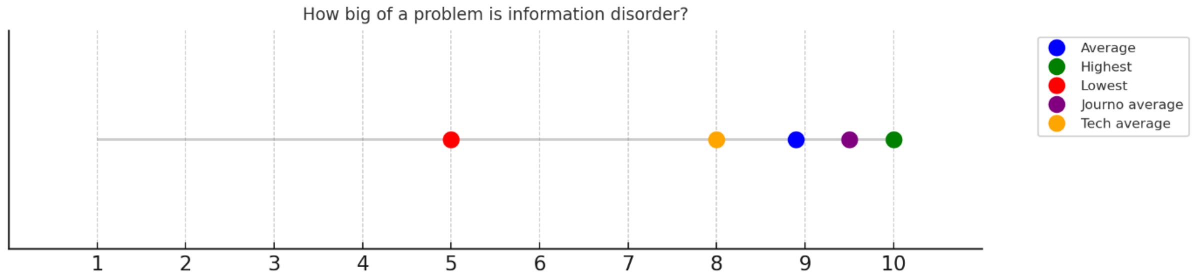
Some respondents also framed fake news within the broader landscape of media integrity and political manipulation, with four (all journalists) highlighting its role in undermining public trust in journalism. J4 said: “I don’t think Donald Trump was the first to do it, but it’s a term used to discredit the media… we try to avoid (the term) in our copy.”

4.3 Statements on key information disorder issues

Beyond definitions, it's also important to understand how the two groups frame up the problem of information disorder. To what extent do journalists and technologists agree about the size of the information disorder issue, the causes of these issues, and the trade-offs inherent in questions around algorithmic transparency and free speech versus safety?

Seven statements were derived from the literature on misinformation, focusing on issues where there was disagreement, ambiguity or conflict. Respondents were asked to rate each of the seven statements between 1 (strongly disagree) and 10 (strongly agree) and provided a short explanation for the rating. These ratings were recorded and used to produce average scores for journalists, technologists and both groups combined. The ratings for each statement are shown in charts. Beyond this quantitative measure, the explanations from each participant are used to determine the level of alignment between the two groups on these issues.

4.3.1 Statement 1: Misinformation and disinformation are a major problem for society.



The literature review suggests that most researchers view information disorder—encompassing misinformation, disinformation, and other forms of harmful information—as a significant societal issue (Núñez-Mussa et al., 2024; Rauch, 2021; Wardle & Derakhshan, 2017). This belief is driven by studies showing the rapid spread of false information on social media, its impact on public trust, and the role of platforms in amplifying misleading content. However, some scholars argue that the problem is overstated. Critics suggest that the focus on misinformation may be part of a broader narrative driven by traditional gatekeepers trying to reclaim authority, arguing that social media might simply reveal long-existing beliefs rather than create new societal problems (Bernstein, 2021).

Most journalists and technologists in the study believed that mis- and disinformation was indeed a major problem for society, with an average rating of 8.9 across all participants. While journalists showed higher agreement (9.5) with the problem statement than technologists (8.0), this variation does not show a fundamental disagreement on the importance of addressing misinformation.

Technologists expressed concern over misinformation's impact on decision-making and behaviour. T1 said that "people make decisions based on information they have, and if that information is wrong, it leads to poor decisions. Misinformation is therefore a significant issue." T2 further underscored the behavioural consequences: "What people read influences behaviour. Misinformation can lead to harmful behaviours, making it a serious problem."

Journalists presented the issue with a more pronounced sense of urgency, with 75% of them rating their agreement with the statement at the maximum 10. For example, J6 said: "If you can manage to manipulate information, you can manipulate democracy. Misinformation is a direct threat to the foundation of our society." Technologists also expressed concern, but their views tended to be expressed in more sober language.

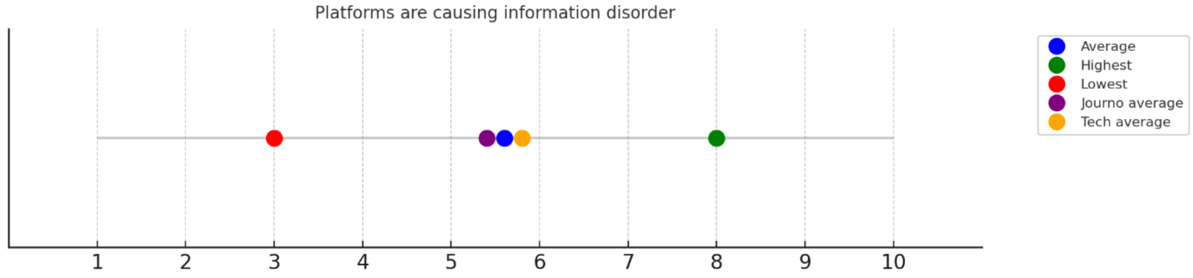
Despite these differences in emphasis, both groups believed that information disorder in its various forms was a major threat to society and expressed a need for effective countermeasures. The slight variation in perceived severity did not indicate a deep-seated divide between the two groups.

The lower average severity rating among technologists is in large part due to the response of T5, whose view significantly diverged from the consensus. T5 rated the problem of misinformation and disinformation as a 5, and they were skeptical about the impact of misinformation. They said:

> "I think the scope of that is overstated. (Misinformation) doesn't reach as many people as we believe it does … I don't think it's a high percentage of what you consume in a day. So, I think it's overstated. That said, it is a problem because the public's ability to sometimes differentiate between accurate information and inaccurate seems to be going down. Our willingness to engage critical thinking skills seems to be going down so that worries me but the scale of the problem is not yet very high."

This echoes the Bernstein's critique (2021) that the impact of misinformation from social media is overstated in attempts to explain societal ills and disorder in information ecosystems. In this view, there have always been people who believe in wild or unproven ideas (for example, UFO enthusiasts, 9-11 truthers and JFK conspiracy theorists). The internet has just made them more visible, and the deeper issue is political polarisation and lack of trust in institutions that has taken hold in western democracies.

4.3.2 Statement 2: Social media platforms are causing "information disorder" and disrupting democracies



Once again, this question was derived from conflicting views expressed in the literature. Many scholars argue that platforms amplify false information due to their design and engagement-driven algorithms, significantly contributing to these issues (Altay et al., 2023; Gillespie, 2018; Rauch, 2021). However, other researchers believe these problems are part of a broader set of societal symptoms. Budak et al. (2024) suggest that misinformation exposure is often overstated and primarily affects small, highly motivated groups, arguing that audience demand, rather than algorithms, plays a more significant role in harmful content exposure.

Both tech experts and journalists were reluctant to solely blame platforms for information disorder problems. Tech experts rated the statement at 5.8 and journalists at 5.4 on average.

Journalists showed slightly more understanding toward social media companies than technologists and reflected on the balancing act platforms face in managing misinformation while upholding free speech. J1 noted that while platforms have played a role in making the information ecosystem worse, they could be considered "collateral damage" as their objective is to connect people, rather than confuse or lie to them. "An exception is the newer platforms that are established for political gain, sometimes disguised under the grander notion of free speech, such as Trump's Truth Social," they said.

Looking closely at the perspectives offered by technologists and journalists, it’s clear the inclusion of "cause" in the question influenced rating, leading the majority to gravitate towards the middle of the scale in the 5-6 range. Most participants openly acknowledge the issue's complexity. Also worth noting is the lack of very extreme ratings – no participants rated this statement as strongly disagree (1 or 2) or strongly agree (9 or 10).

J8 colourfully reflected on the role of misinformation in the French revolution: “Rumors spread through Paris that got to people on the other side of town, literally, and these were based on false information. All those heads chopped off and they didn't have Facebook … or Twitter, so you don't really need [social media] to get a lot of mischief and problems.”

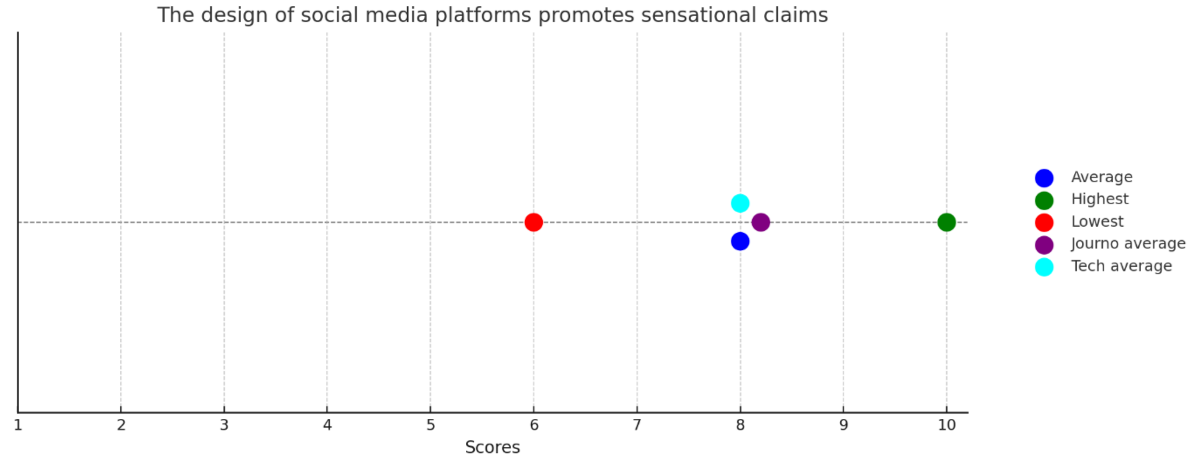
The two lowest ratings (both 3) came from participants who grew up and worked in countries and regions where democracy is not a given and the government is intolerant of free speech. Their responses show that the answer to this question may be different depending on where you live in the world. For those who live in restrictive environments, the freedom that social media offers may appear more likely to usher in democracy rather than disrupt it.

T4, who grew up in Iran before moving to the US, noted that in countries where the traditional media is controlled by governments, social media can act as a counter to disinformation and an aid to democracy: “Even without social media, these issues would exist, perhaps through different channels. However, social media can also aid democracy by giving a voice to everyone, not just those with specific agendas and funding … In countries like Iran, for example, social media is a major driver of democracy. The government controls most news and is a main source of disinformation. Social media is essential for fighting that censorship.”

Journalist J6, who worked for 10 years as a journalist and fact checker in Kenya, also disagreed with the premise, saying “people have always found ways to spread misinformation and platforms just change the medium”. J6 said they were more suspicious of the mainstream media in Africa, as many outlets were willing to repeat the government’s talking points during an election. Like T4, they saw social media as a legitimate outlet for critical opposition voices.

At the other end of the spectrum, T5 rated the statement as an 8, saying “the incentive systems of many of these platforms is to amplify non-representative viewpoints of society… It's painting an inaccurate picture of many societies and communities. And this, unfortunately, is sometimes used as evidence to support some of these positions, whether by politicians or organisations.”

4.3.3 Statement 3: The design of social media platforms means they are likely to promote sensational claims over sober claims, regardless of the truth of those claims.



All respondents agreed to some extent that the design of social media platforms inherently favours sensational content and can be indifferent to truthfulness. The average rating for all respondents was 8.0, and both groups also averaged 8.0. The range of responses was also relatively limited, with the highest rating at 10 and the lowest rating at 6. Statement 3 saw the strongest alignment among participants of all the statements. This aligned with a point of view that surfaced repeatedly in the literature review (Gillespie, 2018; Rauch, 2021; Douek, 2021).

The most common view expressed by the participants was that sensational claims often eclipse sober, factual content because such claims are more likely to engage users and make money for the platforms. But there was disagreement over whether the platforms cynically design their recommendation algorithms to favour sensational content or product leaders had made naïve but honest mistakes rooted in incorrect assumptions.

T4, for example, said the platforms had not intentionally created algorithms that search for sensationalism, but this was a byproduct of focusing on engagement and growth metrics, and that any harm caused was accidental. “(The algorithms) are designed to promote engagement and views. It's human nature to be more engaged with something shocking or interesting. A boring, yet factual news item might not attract as much attention, depending on the audience. If the audience is informed and cares about reality and facts, they won't engage with sensational but untrue content. However, for those who seek entertainment, social media will promote content that's written in an interesting or shocking way, even if it's not factual.”

T5, a tech entrepreneur, also expressed a level of sympathy for big tech companies: “A platform made up of people being angry all the time won't pay off in the long run. Eventually, people get burned out and leave, so I don't think (Meta founder) Mark Zuckerberg is really trying to build a system where you are mad all the time … I think it's hard for them to change this monster they've created.”

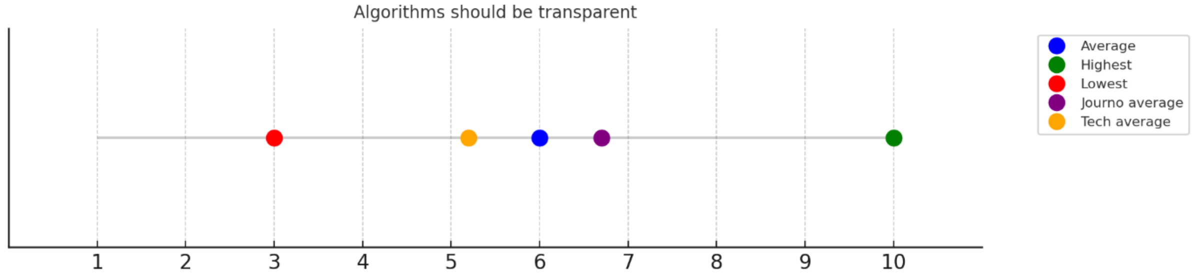
Journalists also highlighted the interplay between sensationalism and revenue, but they were less likely to view this as an accident, instead remaining neutral or expressing suspicion that the platforms were happy to let bad content proliferate for profits, provided they weren’t suffering unfavourable public criticism. J6 said: "I think (the platforms) are designed to make money, which sensationalism does,” underscoring the perception that sensational content's dominance is a strategic decision tied to revenue generation.

Some participants believed the platforms had improved in recent years. T6, for instance, rated their agreement at 7, but said their view had been “changing quite a lot, where if we were to have this conversation maybe four years ago, I think there would have been [a higher rating] … it has been moving in a better direction where there are real attempts at improvement”. As noted in the introduction to this chapter, the interviews took place before Meta ended its fact checking programs and wound back its moderation efforts. Had this interview occurred in 2025 we can assume T6’s rating would have been higher than 7.

J7 also noted the improvements that platforms had made, but was far less sympathetic, saying efforts to improve the quality of information in the platforms were “nowhere near as effective” as needed. “I'd say they facilitated [these problems] by not having proper rules and regulations, not screening their content, and aggressively being oblivious to the hurt and the damage that has been caused in pursuit of the dollar,” they said.

J2 gave a unique perspective based on their time working in the advertising industry, saying that an advertising-based business model did not automatically mean a social media system would need to maximise engagement to make money. “I don't necessarily agree that the advertising model is the fault… I don't think the type of engagement that social media tries to sell, the type of targeting and engagement-based targeting, is a prerequisite for selling advertising on those platforms … could you have a different form of advertising with brand advertising with less of a less of a sense of that engagement work? Quite possibly.”

4.3.4 Statement 4: Tech companies should be forced to make their algorithms and data transparent so that researchers can interrogate them.



A common theme in the literature is that platforms should be forced to make their algorithms and data more transparent. Acker & Donovan (2019) describe the lack of access to accurate data and metadata as a “thumb trap” that prevents researchers, journalists, policy-makers and civil society groups from understanding the impact of technology on information disorder and society at large. Douek (2021a) advocates for a level of mandatory transparency on moderation (particularly error rates), but also notes some costs and problems associated with more transparency, such as burdensome reporting costs that may discourage the development of new products in the social media space.

Among participants, there was a notable divergence in viewpoints. The overall agreement rating was a 6, with journalists averaging 6.7 and technologists averaging 5.2. There was also a very wide range of views expressed, with the lowest rating at 3 and the highest at 10.

Journalists, on average, rated the issue as more critical, and some advocated for maximum transparency. J7 rated the issue a “super 10”, saying they were “totally” for transparency. But while positively inclined towards transparency, most journalists did acknowledge that transparency presented risks for users around privacy, proprietary information and whether algorithms could be reliably understood and interpreted.

For example, J2 (rating: 5) said: “The idea that there's a single algorithm is sort of naïve. We shouldn't ask companies to reveal all the secrets of their special sauce that drives their ability to differentiate themselves from other people. But there are certain elements regarding the tuning of AI models, and the data that it's been trained on … that should absolutely, in my view, be regulated to be transparent to the public and to the users.”

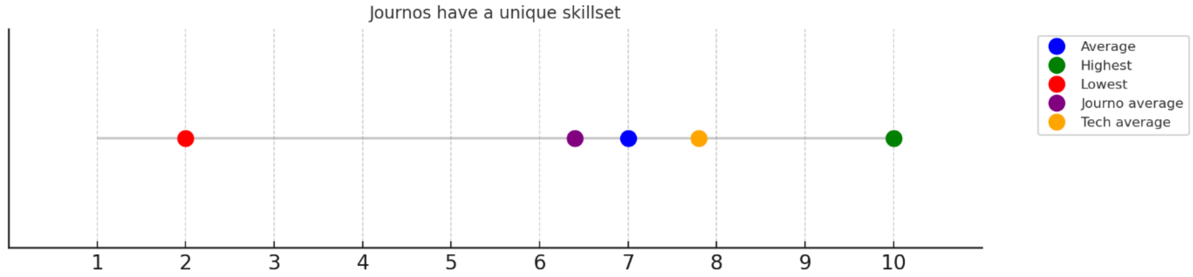
Most technologists expressed concern and urged caution in taking the right approach to achieve transparency. While all agreed that some level of transparency was a good thing for society, they raised concerns about the utility of suddenly making complex algorithms “transparent” to the public, as they could be impossible to interpret, even for experts or in some case people who played a role in building them. This echoes Meese’s (2023) point that “platform observability” (i.e., continuous and ongoing observation of platform behaviours and impacts) is likely to provide better transparency than demanding access to complex “black box” algorithms.

T6 said: “Sharing the algorithm, having just the bare code available, might not necessarily be a great indicator of how it actually operates on a daily basis.” To illustrate, they pointed to an internal document from Facebook (Franceschi-Bicchierai, 2022) that showed the scale of data that was being used within the company’s advertising algorithms. In response to regulations from Europe’s General Data Protection Regulation (GDPR), Facebook estimated it would take an individual engineer at least 600 years to do the necessary work to ensure compliance. “From a research point of view, an algorithm without data, I don't think massively moves the needle on trying to understand how they're working.”

T4 expressed a similar sentiment: “I'm more toward disagreement. Tech companies, after ensuring privacy and data protection, should share some data for research purposes. However, making algorithms and detailed data public poses risks. Bad actors could exploit the algorithm or find loopholes, harming other users. There should be a mechanism to increase visibility for trustworthy researchers, but we should not make everything public.”

Both groups acknowledged the importance of transparency, albeit with different emphases. T1 said "transparency is crucial for building trust with the public, yet we must balance this with the need to protect privacy and intellectual property,” reflecting the tension between the desire for transparency to enable accountability and the need to safeguard sensitive information. This tension reflects broader debates in society about the balance between openness and privacy, innovation, and regulation. The broad range of perspectives shared by the participants highlight the complexity of navigating these issues.

4.3.5 Statement 5: Trained journalists possess skills and practices unique to their profession that allow them to find out verifiably true things.



As we saw in Chapter 2, Kovach & Rosenstiel (2014) outline 10 “elements” of journalism. These include working as an independent monitor of power, maintaining independence, and maintaining balance. But the most important practice is “a discipline of verification”. While journalists continue to hold the view that verification is a key skill that sets their industry apart, public trust in news institutions and journalists is at an all-time low and audiences are skeptical of these claims about authority and fact finding (Nielsen & Fletcher, 2024; Núñez-Mussa et al., 2024). Research has also found that the verification practices of journalists can be inconsistent, ill-documented and prone to compromise (Shapiro et al., 2013).

Professional journalists are also seeing increased competition from social media influencers and creators who report on the news and are often seen as more trustworthy than “mainstream media” (Stocking et al., 2024). And other professionals such as lawyers, police officers, academics and scientists also have a professional practice of verification (Rauch, 2021).

If technologists believe that professional journalists cannot be trusted to verify facts, or that other professionals are better able to achieve this, it is possible that they may not wish to work closely with journalists to establish ground truth for algorithms that detect misinformation. To that end, this statement was primarily included to gauge how technologists feel about the skills and practices of journalists, and to compare this to journalists' own self-image.

This statement provoked a wide range of responses, with journalists offering a markedly more negative assessment of the skills and practices of their industry. Technologists, with an average rating of 7.8, tended to hold journalists in high regard, perhaps due to an external appreciation of the nuanced skill set required for effective journalism.

Journalists' ratings averaged 6.4. The lower ratings typically related to a detailed understanding of the commercial pressures that journalists face to produce large amounts of engaging content, echoing Shapiro et al.'s (2013) observations that compromises and shortcuts were common during journalistic verification work. J7 said: "You can't make that rule about all journalists. Some journalists are very good at what they do (but) there's far fewer of them around than they used to be. I think the pressures of the profession where you need to turn around stuff quickly, often from press releases, without checking it out, means that journalists have less time to be accurate."

The lowest rating of 2 came from J2, the fact checker from Kenya. "(My country) just had an election, and the President was sworn in two days ago," they said. "I strongly felt that mainstream journalists were acting as a conduit for the spread of misinformation around the election … they didn't seem to have fact checking skills in their day-to-day reporting. So they were essentially mouthpieces of … politicians and other political actors."

This reflects the observations of Núñez-Mussa et al. (2024), who discuss how journalists in Chile struggle with verification practices due to limited resources and heavy reliance on official sources, which are often untrustworthy but difficult to independently verify. This reflects broader challenges faced by journalists in less democratic or highly corrupt environments, where external pressures and constraints undermine journalistic integrity.

Whether journalists had "unique" skills was another point of contention. Some respondents pointed out that there were other professions where investigative and fact-finding skills are paramount, or that non-professionals could take on a journalistic mindset. T4 pointed to "detectives" as another class of specialists who engage in evidence gathering and analysis. J2 pointed to jurisprudence, medicine, science, and forensics as "fields where the rigorous analysis of factual information in order to convey and communicate a message is absolutely the lifeblood of what they do".

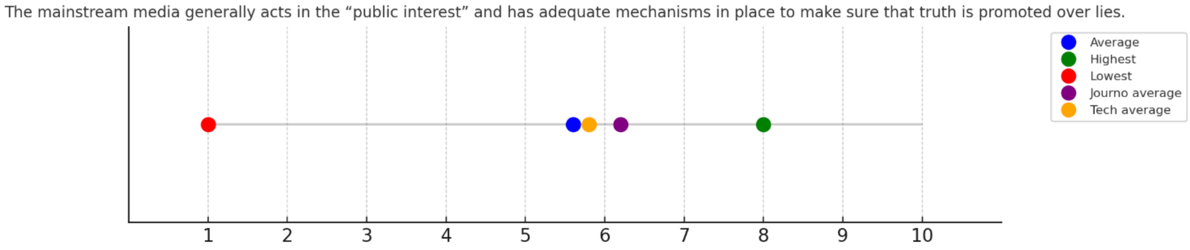
But one potentially unique aspect of journalistic practice in the literature concerns the speed at which journalists work to verify claims sets them apart from other professions (Lewis, 2012). T1 echoed this observation: "I'm very impressed with the journalists I've worked with … they really can dig into things quickly understand things very rapidly and figure out what's true and what's false."

J8, who came to fact checking via technology rather than a traditional newsroom, described journalism as "an attitude". "A lot of it is be willing to put in the work. And I agree that for most of the general public, who has time for that? If you see something on social media, will you pick up the phone and call the spokesperson? (But) it's by no means an impossible skill that requires years of training." Anderson et al. (2015) also highlight that the rise of "post-industrial journalism" has brought in individuals, crowds, and machines, blurring the lines of who qualifies as a journalist and how verification is practiced.

T5 made a similar statement, saying "you can get some of these skills without the formal journalism training," pointing to OSINT organization Bellingcat as an example. "They are not professional journalists in most cases, but some of their research and picking up (of information) is good. I think it would make most journalists say, ‘I want to be like that’.”

To summarise, journalistic skills were well respected by technologists, particularly their versatility, speed and ability to work with ambiguity. But they also recognised there were similar competencies in other professions, and that the level of commitment to ideals such as verification varied between journalists, publications, and media cultures within countries and regions.

4.3.6 Statement 6: The mainstream media (by which I mean newspapers, television and websites with large audiences who produce original reporting and claim to adhere to journalistic codes of practice) generally acts in the “public interest” and has adequate mechanisms in place to make sure that truth is promoted over lies.



Scepticism towards the practices and motives of mainstream media is very common in journalism studies literature and more generally in society. Kruger et al. (2021), discusses how journalists struggle with balancing truth and commercial pressures, often resulting in compromises that affect public trust in media, and the most Reuters Digital News Report showed very low levels of trust for journalism worldwide (Nielsen & Fletcher, 2024).

If mainstream newsrooms are regarded as untrustworthy entities, or even vectors of misinformation themselves, they are unlikely to be welcomed as partners in the fight against information disorder.

Participants expressed widespread scepticism towards the statement, with technologists averaging 5.8, and journalists averaging 5.5, for an all-up average of 5.6. Both groups engaged critically with the concept of "mainstream media” and its implications for public discourse.

As noted in the methodology section, the question was initially presented without the definition for mainstream media as “newspapers, television and websites with large audiences who produce original reporting and claim to adhere to journalistic codes of practice”. Participants responded negatively to this ambiguity, questioning its broad application and suggesting a need for more precise definitions. J8, for instance, challenged the term's usefulness, arguing that it was “too broad to describe large, often dominating or monopolistic media organisations, which may have vastly different missions and goals". T6 referred to it as “a wonderfully loaded phrase” and J8 said it was a term used by “conspiracy theorists.”

In response to this feedback, an adjustment was made midway through the interviews to provide a standardised definition of "mainstream media" for eight of the 14 interviews. This change was implemented to reduce ambiguity and ensure that responses were focused on the intended concept rather than participants' individual interpretations. Participants who requested clarification in earlier interviews were also provided with a similar definition to the one provided to later participants. The rationale for the change is discussed in more detail in Section 1.3.7.

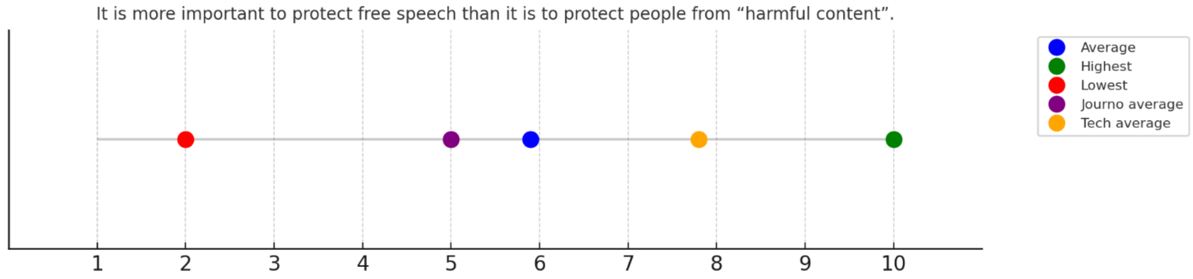
To determine whether this change had a major impact on the responses, the results of respondents who were provided a definition of “mainstream media” were compared with the results of those who were not. Data collected before and after the adjustment revealed that the impact on participants' responses appeared to be minimal. Those who were not provided a definition of "mainstream media" gave an average rating of 5.66, while those who received the definition rated it 5.63, resulting in an overall average of 5.64. While it is impossible to know if any participant might have responded vastly differently without the additional definition, with the lack of variance between the groups it seems reasonable to analyse all responses in a single section.

A shared concern among respondents revolved around the media's balancing act between truthfulness and external influences, whether political or economic. T1 criticised what they saw as the media's prioritisation of engagement over truth, noting that it was "too often trumped by political and business considerations". This sentiment mirrored Rauch's (2021) discussion of how the media landscape often favours sensationalism due to economic incentives, compromising the public interest mission. J4 highlighted the amplification of partisanship in the media, reflecting broader concerns about polarization as documented in Vu and Saldaña (2021), who found that newsrooms are increasingly pressured to cater to partisan audiences, further complicating their commitment to objective truth.

J2 said it was difficult to make "blanket statements" as "the mainstream media in Russia is a very different animal to the mainstream media in Western Europe," reflecting similar findings by Núñez-Mussa et al. (2024).

Overall, while there is acknowledgment of the value of mainstream media, the literature and participant feedback highlight significant challenges, including commercial pressures, partisanship, and varying standards across political contexts, which undermine the ideal of serving the public interest.

4.3.7 Statement 7: It is more important to protect free speech than it is to protect people from "harmful content".



Conflicting viewpoints on the need to maximise free speech on one hand, and protect users from harmful content on the other, is at the heart of global debates surrounding misinformation and content moderation. There is a sharp disconnect between conceptions of free speech in different jurisdictions. One the one hand, the most-used social media platforms (except TikTok) were founded in the USA, where freedom of expression is often viewed as a "trump" that takes precedence over other rights. But Europe, which is the most active jurisdiction regulating social media, tends to have a more proportional view of rights, where speech is one right weighed off against many others, and this approach is supported by many researchers (Douek, 2021a; Gillespie, 2018; Suzor, 2019).

Several respondents commented that the speech issue, which has broken along partisan lines, was difficult or even impossible to resolve. A very wide range of views was expressed.

Technologists on average rated the importance of protecting free speech as higher than journalists. The average rating from technologists was 7.3 and most technologists clearly favoured the protection of free speech (ratings of 7 and above), with only one giving an “on-the-fence” rating of 5.

Tech entrepreneur J5 was strongly aligned with this ethos, rating the statement with the maximum score of 10. “Free speech enables everything,” they said. “It's like the bedrock of a functioning democracy. I admired the (American Civil Liberties Union) for fighting for the right of Nazis to have their propaganda out. It's distasteful but that's the world we need to be comfortable with. Words do not hurt people as much as they think.”

This aligns with observations in academic literature that the tech industry tends towards strongly pro-free speech perspectives informed by America’s First Amendment (Douek, 2021a). J5’s statements are reminiscent of Elon Musk’s public statements regarding his decisions to drastically wind back content moderation at Twitter (now X) upon buying the company: “By ‘free speech’, I simply mean that which matches the law I am against censorship that goes far beyond the law. If people want less free speech, they will ask the government to pass laws to that effect. Therefore, going beyond the law is contrary to the will of the people.” (Musk, 2022)

This point of view is criticised in the literature, particularly by Napoli (2019) who says the tendency of social media platforms to show individually curated feeds and not clearly identify authoritative information reduces the effectiveness of counterspeech as a cure for the distribution of incorrect or harmful information.

Other technologists in the study advocated for free speech but put some weight on the need to protect users from harms. T4, for example, said their experience growing up in Iran had influenced their perspective towards the need for free speech. “In countries where censorship is strong, freedom of speech is crucial as it can uncover a lot of truths. However, defining 'harmful content' is complex… While freedom of speech has great benefits, the human nature of not doing due diligence can also cause harm. So, I value freedom of speech highly but acknowledge the potential harm from misinformation.”

Journalists were more likely than technologists to give moderate ratings and place greater emphasis on the need to protect users from harmful content, although there was a range of views. J5, who rated the statement a 4, said that while they had always valued freedom of speech due to their journalistic heritage, their perspective shifted as they worked on misinformation problems and realised that many related harms were extremely serious.

They said: “I lean more towards the removal of harmful content because we're talking about real lives and deaths. If we're talking about antivaccine narratives, for example, we just saw a report released saying 20 million people were saved by these vaccines. The number who died as a result of misinformation is hard to quantify, but it's not insignificant. So if sometimes a post is taken down, and it shouldn’t have been taken down, perhaps we cause less harm in the world by over-indexing on (harm reduction).”

T2 diverged from the consensus “free speech” view among technologists, giving the question a neutral rating (5). They said: “free speech should not equal free reach”. In their view, free speech was generally preferable, but there should be mechanisms for making sure that information that is given wide distribution has been carefully thought through, and not “tweeted off in 200 characters while sitting on the toilet”. They concluded that “the economic incentive for the companies that provide infinite reach needs to change”. T2’s view has parallels with Rauch’s concept of an information ecosystem with “positive epistemic valence”. In such a system, free speech enables the expression of a very wide range of views, but to be distributed widely and enter the realm of accepted knowledge, these views need to pass through various layers with a commitment to truth-seeking control – for example, a claim that is tested in court or subject to rigorous investigative reporting. (2021)

J1 rated their agreement with the statement as a 2. While at first glance it may appear that J1 believes protecting people from harms is far more important than protecting free speech, the rating is out step with their comment that "in an ideal world both should be held as equally important", which suggests the low rating was more a reaction to the phrasing of the question, rather than a radically strong commitment to harm prevention.

4.4 Chapter summary

The data in this section provides direct, comparable evidence on how journalists and technologists define and view the problems surrounding information disorder. While both groups recognise the critical importance of combating misinformation, there are noticeable differences in their perspectives. Journalists tend to focus on the socio-political impacts and the ethical responsibility of reporting, emphasising the need for narrative complexity and human oversight in content moderation systems. Technologists, on the other hand, often prioritise scalability, speed, and the efficiency of automated systems.

Despite these differences, the foundation for collaboration exists: both groups see the severity of the issue and acknowledge the need for each other's expertise to create more robust solutions. This gap between how they conceptualise the problem—journalists focusing on context and technologists focusing on systematisation—can be viewed as an opportunity for dialogue and collaboration, rather than a barrier. As Graves and Anderson (2020) point out, the integration of journalistic insights into more structured content moderation systems could provide significant benefits, but it requires careful balance to avoid oversimplifying the complexities journalists bring to the table.

The second contribution of the data is in determining whether technologists view journalists as desirable partners in content moderation systems. The answer from the participants is largely positive, albeit with some qualifications. Technologists in the study were notably more optimistic about the role journalists could play than the journalists themselves. They valued journalistic skills in verification, ethical reporting, and identifying misinformation. However, they also recognised that these contributions depend on journalists being given the right resources, training, and editorial freedom to uphold the standards of truth and verification. Interestingly, while some technologists expressed scepticism about the current state of the media—particularly in regions where journalism is either underfunded or compromised by government influence—many believed that journalism, when practised well, was an essential ally in fighting misinformation.

This optimism among technologists contrasts with the journalists' own concerns about the erosion of their professional standards due to budget cuts and pressure to produce click-worthy content. This aligns with concerns noted in the literature, such as Shapiro et al.'s observation that commercial pressures and the lack of a clear accreditation system means that there is often a large gap between journalists' view of verification activities and their actual day-to-day reporting, with many taking shortcuts or simply rewriting the copy of others (2013). The gap between these views—where technologists see potential and journalists see systemic problems—will be explored further in Chapter 6, which will analyse how these differences can be bridged to create stronger collaborations in content moderation efforts.

Chapter Five: How journalists help with AI detection for misinformation

5.1 Introduction

This chapter explores specific ways in which journalists can contribute to AI systems to mitigate information disorder problems. Participants had a variety of viewpoints on how journalists' skills could be of use in these technological systems. They saw AI as both a challenge and a tool in detecting misinformation and were generally positive about the ways in which a collaborative approach that combined journalistic integrity with AI's scale could protect against information disorder.

Participants identified a range of challenges involved in detecting information disorder at scale. These challenges have been grouped into six categories:

1. The dynamic and evolving nature of information disorder
2. Linguistic and cross-cultural challenges
3. Data quality, quantity and potential for bias
4. Structure of fact-check data
5. Definitions, policies and ground truth
6. The right model for platform regulation

Each of these challenges is presented as a separate section. The issues will be introduced and explained, and the participants' views on the ways that journalists can use their skills to improve these systems will be discussed, as will relevant information from the literature.

5.2 Dynamic and evolving nature of information disorder

Machine learning is well-suited to problems where features (i.e. the attributes that the AI detects for) are relatively consistent and easy to define. For example, while there may be many ways that a dog can be represented in a picture, it is not difficult to describe what a dog is, entirely new breeds of dogs appear rarely, and even new breeds have four legs and a wet nose.

In contrast, misinformation is constantly evolving, new narratives emerge rapidly, and the truth of claims is often contested and sometimes never settled at all. This presents a major challenge for at-scale automated detection. Shabani and Sokhn (2018), for example, note that the linguistic similarities between false and real news mean algorithms that detect for misinformation make a lot of mistakes.

But the dynamic and evolving nature of misinformation (which we will refer to as "novelty") is not only a problem for machines. Making correct decisions can also be challenging for human judges. Common argues that human moderators rely on heuristics when asked to make decisions about content that is unfamiliar to them or ambiguous, which increases the risk of inconsistent identification due to biases or ingrained assumptions (2020).

Common sees decisions about hate speech and extremist content as particularly prone to this kind of error, but decisions about misinformation are arguably even more exposed to subjective interpretation, as demonstrated by debate over whether misinformation is a legitimate field of study or a type of pseudo-expertise (US Congress, 2023).

Five out of six technologists interviewed mentioned the issue of novelty. T1 said that finding new types of misinformation was one of the hardest problems for at-scale detection: "Most NLP [Natural Language Processing] is based on finding synonyms and paraphrasing things. If it's a whole new type of misinformation or a new set of lies [it's harder to determine] what's common there." While other signals such as virality could be indicative of misinformation, most viral content is "perfectly true", T1 said, which makes it a noisy and less useful signal where "you're always going to be playing catch-up".

T4 said that as new claims of misinformation could emerge at any time, the AI needed a continuously updated database to stay relevant. "Once a classifier is trained and stored, its knowledge base is fixed to that point in time… having a complete set of facts is practically impossible. Even complex solutions like GPT can't be used for misinformation detection as they tend to create facts and hallucinate, mixing truth and misinformation."

T6, who was interviewed in 2021, referenced Google's LLM BERT (a predecessor to the current Gemini family of models), which was released shortly before the COVID-19 pandemic. "(With) COVID, a huge amount of vocabulary that is now common simply just didn't exist … or only (existed) in medical research papers. So (BERT) wasn't aware of these words, and when it comes across them, it doesn't quite have a way to understand its relationship to other parts of its vocabulary," they said.

"Anti-vaccine misinformation (was) using carrot emojis instead of a syringe to identify posts related to (vaccine misinformation), sort of a nudge and a wink to those in the know… that kind of gives you a sense of the sort of moving target that these algorithms are trying to chase," J2 said. They also believed that many current AI detection systems were based on "naïve" assumptions and could not deal with nuanced and evolving tactics. "For all their levels of scientific and engineering brilliance… these models are kind of dumb and don't really understand the ways that mis- and disinformation actually happens", such as subtle distortion or manipulation to make factually true information appear sinister or threatening.

5.2.2 How journalists can help

Identifying, verifying and publishing new information is a core activity in journalism (Kovach & Rosenstiel, 2014) and most participants mentioned novelty as a key area where journalists could help at-scale misinformation detection systems. These contributions are not limited to traditional reporting but extend to improving technological solutions for a more effective response to misinformation.

Analysis of the data reveals five different ways that journalists can contribute to solving the problem of novelty.

i. Quickly identify emerging misinformation

The interviews reveal that rapid identification and assessment of "ground truth" for misinformation is one way that journalists, fact checkers and analysts are already contributing to at-scale misinformation systems. Ground truth was defined in Section 3.5.1 and refers to the highest quality and most accurate data available to train and test a model.

Fact checkers already undertake such rapid assessments to get ahead of newly emerging misinformation narratives. The now-defunct misinformation research group First Draft dubbed the activity "prebunking" and defined it as "the process of debunking lies, tactics or sources before they strike" (Garcia & Shane, 2021). But such information may not be making its way into algorithms and closer co-operation between platforms on the one hand, and fact checkers and media organisations on the other, provides an opportunity to close detection gaps more rapidly.

Participants said the ability of trained journalists to swiftly verify facts and discern the credibility of information could allow AI systems to respond faster to misinformation threats. This ability to quickly identify and verify information is crucial, as research has shown that false news can spread significantly faster and more broadly than true news on social media platforms. Vosoughi et al. (2018) found that on Twitter, false news stories reached 1,500 people six times faster than true news stories, and false political news was particularly potent, reaching 20,000 people three times faster than other types of false news reached 10,000 people.

J3, who runs the investigations team for a commercial trust and safety platform, said their team's key activity was "finding viral misinformation and trends that are emerging". J3's team creates bespoke reports that document misinformation narratives emerging from known influencers and channels on platforms such as Rumble and Telegram. These are sent to clients, mostly major online platforms, who use the data to train their AI systems and human moderators to improve detection.

The success of this team is measured in terms of how far in advance of mainstream media they identify the trend. J3 said: "No platform is a silo. If (misinformation) emerges on one platform, it's going to emerge on your platform." This work is still a largely manual process for J3's team, as the newness of the information means that AI and keyword detection methods are less useful, as "you don't know what you're looking for".

J5, head of the editorial operations for a misinformation detection platform that has since been acquired by a major technology platform, described their team's work in similar terms. "We would like to think that a competitive advantage is that because we're very nimble and agile, and small and flexible, and because misinformation is evolving so quickly, we can be on top of it in close to real time."

ii. Prioritising important information

Some participants said that journalists brought contextual knowledge and nuance regarding politics, history, and major news events. Not all misinformation is harmful – stories about Bigfoot or the Loch Ness Monster are unlikely to result in real life injury or death – and journalists were seen as having the expertise to help technologists focus on the most important problems.

This point was highlighted by J3, who said: "I think the more years we're dealing with (misinformation), the more we realise that … removing basic misinformation from years ago is not necessarily going to help anyone. It's more about what's happening now, what's happening today. We need to figure out what's important, and what's not important."

T1 likewise said that platforms should be making use of journalists to say whether a new claim is going to be important or can be ignored. "The way AI normally works is through training data—you get a bunch of examples and find similar examples in the future. But what do we mean by similar?" For example, they said, take a quote from the CEO of Pfizer—to an AI, all quotes from this person may appear similar. But a journalist can more quickly understand the context in which the quote is appearing—is it a benign business publication, or is it a wellness influencer who is using the quote out-of-context to support an anti-vaccination narrative?

T4 also said that data science teams working on misinformation detection systems needed help to focus on the most important information. "My goal would be to automate as much as possible and utilise the journalists' expertise where it's most needed, ensuring we cover the most critical topics effectively," they said. "We would focus on misinformation that becomes viral enough to warrant attention, as we can't address every single piece. I would automate the process to identify these examples and then ask the journalists to verify the solidity of the information. Their input would help to increase the coverage of my AI solution."

iii. Tracking the evolution of misinformation tactics and language

Journalists' deep understanding of communication strategies and their ability to document changes in the presentation of misinformation can also be helpful for AI systems to adapt to evolving threats.

J5 said that the broad themes of misinformation often did not change greatly over time, but the specific language and tactics being used evolved rapidly. For example, anti-vaccine narratives have not changed greatly in the last 200 years, but the targets of a misinformation campaign and the specific keywords being used are often entirely novel. “One way to put it is we need to understand the evolution of a (misinformation) movement, so then the question becomes, how do you define the movement?”

Journalists' ability to track these changes is particularly valuable given the "firehose of falsehood" propaganda techniques identified by Paul and Matthews (2016), where a high volume of misleading messages is spread rapidly across multiple channels. By understanding how these tactics evolve, journalists can help AI systems detect and counter them more effectively.

iv. Online verification skills

Another aspect of establishing ground truth on emerging misinformation narratives is the techniques of Open Source Intelligence (OSINT). OSINT describes a range of online practices relating to verification of information that is shared online. Examples of these practices include using maps to establish whether an image matches its purported location, or examining metadata to determine when an image was created, and by whom (Wardle, 2014).

J3 said that a combination of journalism and OSINT skills was the ideal combination for their trust and safety team when they assessed new examples of possible misinformation. “You need the OSINT skills to help you tap into and find information where you wouldn’t necessarily know exactly how to validate it,” they said. J4, who works in fact checking for a major newswire service, said there was a growing need for OSINT, and that more journalists should invest in upskilling in this area.

5.3 Linguistic and cross-cultural challenges

A frequent critique of misinformation detection and fact checking in the literature concerns its focus on wealthy Western democracies, particularly the US (Budak et al., 2024; Gillespie, 2018; Zeng & Brennen, 2023). For example, Budak et al. recommend funding additional research in the Global South on misinformation, particularly in authoritarian countries where exposure to untrue narratives may be higher (p. 50).

Responses from participants echoed this concern. T1 said that AI systems they had worked with were less performant in terms of their accuracy in languages outside English, and in countries outside America. “I think most of the tooling only works in English, because it's been developed in England or in the US… yet most of world doesn’t speak English.” They said their organisation was working to make such tools available in other languages “as a kind of moral, social good”.

J5 noted that covering multiple regions had thrown up difficulties for their misinformation detection platform. They cited an example where their automated systems failed to accurately translate the nuances of the Turkish language, resulting in missed opportunities for detection. “I’d love to be having a conversation that’s truly global, that’s not just about Donald Trump. It’s about Africa and Asia and all the different parts of the world that don’t get as much attention, but are critical,” they said.

T4 noted that the development of highly performant LLMs to some degree solved the issue of adequate translation, at least in “high resource” languages where the model had sufficient examples to train on. But even with relatively seamless understanding between languages, detecting misinformation required up-to-date facts and knowledge of local issues and customs. "AI struggles with language and cultural nuances, affecting its detection accuracy across different regions," they said.

5.3.1 How journalists can help

There are journalists and fact-checking operations all around the world. Some platforms have already put these networks to work in detecting misinformation, most notably Meta and TikTok with their third-party fact-checking programs (Amakoh, 2020; Vizoso et al., 2021). During the interviews, Meta's program was repeatedly mentioned as both an essential source of funding and distribution for fact checks. In fact, without funding from Meta, participants felt that fact checkers globally may struggle to stay solvent. This will be covered in more detail in Section 5.4.3.

Participants' suggestions on how journalists can help with linguistic and cross-cultural challenges can be grouped into three categories.

i. Nuance and cultural sensitivity

The potential for journalists to help improve automated decision-making via their cultural and language skills was raised by many participants. For example, when asked about the best way that journalists can help with at-scale misinformation detection, J1 said employing journalists in non-English languages as it was "necessary to involve native speakers who understand the nuances of their language and the cultural context that comes with it". This aligns with the view of Zeng and Brennen (2023) who argue for misinformation interventions to consider the social and contextual factors that influence how truth and falsehood are defined and perceived.

T6 said their misinformation detection platform employed a team of 20 international experts, including journalists, who worked on ground truth for their algorithms. "(Even) if you remove the language aspect from it, there can be really interesting cultural quirks that if you haven't had a high level of exposure to that particular culture, you wouldn't pick up on those signals, which can be very valuable." J3, who also works for a misinformation detection service, has a team of analysts (including journalists) working across four global regions, with a particular focus on elections.

ii. Test and advise on AI tooling for non-English languages

Another opportunity identified was in the creation of AI tooling to help fact checkers detect for misinformation. Most of these tools are currently in English and need to be adjusted for use in other languages and cultures. J6 noted that they had worked on adapting tools to help detect misinformation in Africa. While initially these models did not generalise well to non-English contexts, as they were fed more accurate data in the native language, J6 said they saw a notable improvement in the quantity and quality of posts being flagged.

T1 said they were actively working on a project to extend their organisation's misinformation detection software to non-English languages. They hoped to create a model that would be able to generalise well into other languages when identifying posts but said local experts on the ground would need to "do their own fact checking" to make a final judgement on individual examples. This could create incremental improvement, where better AI detection allowed more culturally sensitive detection, thus helping fact checkers find the most critical information. These checks could then by fed back into the algorithm, improving it further.

iii. Provide reliable information in restrictive regions

T4 said independent-minded journalists with domain knowledge could help build more equitable AI solutions, emphasising their ability to uncover trustworthy information and establish a fair yardstick for ground truth in regions lacking freedom of speech or robust media infrastructures. "With advancements in LLMs and their ability to work multilingually, we can approach problems independent of language. However, for (detecting) misinformation, we still need factual information," they said.

One challenge with this is identifying and employing such independent journalists to work on moderation systems. Directly contracting with journalists in countries with strict media censorship laws could land those journalists in legal trouble. Collaborating with third parties who are critical of restrictive regimes and operate outside their borders is another option, but such groups may be themselves unreliable or politically motivated in their presentation of facts.

5.4 Data quality and quantity

Both journalists and technologists highlighted concerns about the quality of data used to train AI systems. J5, who worked directly on data annotation for a misinformation detection platform, worried that huge datasets scraped from the internet would inevitably reflect biases that are already present in the world. “When you think about a problem as tricky as misinformation, that question of bias and perspective becomes even more worrisome,” they said.

T5, whose product presents users with opinions on news from across the political spectrum, said that when training algorithms it was important not to encode your own biases into the ranking system. "(Many) of these machine learning models are based on scraping millions of data sets from the Internet, resulting in poor quality datasets which can reinforce biases," they said. Similarly, T2 said that while it was not difficult to “train AI if you have millions of dollars and all the compute on factfinding of history”, that biases would be present in the data because “history is written by the winners”.

To get data in sufficient quantities to train models, AI companies and data scientists tend to either harvest it from publicly available sources or use crowd-sourcing platforms such as Amazon’s Mechanical Turk, where low-cost workers annotate data at scale with simple labels. Journalists, on the other hand, produce small amounts of detailed, high-quality data at high cost.

J6 said an average fact checker could complete one or two fact checks in a day, depending on complexity. Making use of such small amounts of data machine-learning systems requires creative solutions. And in addition to high-quality data being scarce and expensive to produce, T3 said it was often qualitative and context-dependent, which created challenges for data scientists whose work is ideally based on data that is objective and quantitative.

“I think not just in disinformation, but in any machine learning endeavour, data is by far the most challenging part of the whole process,” T6 said. “Everybody wants to do the models, but nobody wants to do the data work.”

5.4.1 How journalists can help

i. Provide high-quality and unbiased “ground truth” data

As noted in Section 3.5.1, “ground truth” is high quality data that is used to train and test machine learning models (Google, 2024a). It represents the reality of what the model is trying to discover and influences the weighting of the parameters in the model that ultimately determine its outputs. Ideally, this data is empirical and measurable, such as weather observations or images of tumours known to be cancerous.

Ground truth labelling often comes from Subject Matter Experts (SMEs) (Santos, 2023) and journalists are already used as a source of ground truth for misinformation (Meta, 2021; TikTok, n.d.). But journalists and fact checkers are expensive to use (Shabani & Sokhn, 2018). As noted in the literature review, some researchers also claim that crowdsourced fact checking, where non-experts are used to judge whether content is misinformation and a decision is made on the aggregate of their scores, can be as effective as using journalists (Barbera et al., 2024), although this claim is contested.

That journalists can and should be more involved with the provision of ground truth data is a key statement that this research aims to test. The technologists interviewed saw journalists as strong candidates for identifying ground truth in misinformation detection systems and as noted in the previous section they tended to hold journalists' skills in higher regard than journalists themselves. The perception among the technologists was that journalists were information generalists, skilled in verification, who could quickly analyse a problem and come to a principled conclusion about it.

One area where journalists can help with ground truth is identifying the stylistic traits of misinformation for models that detect such linguistic markers. T3, who has worked directly on such models, said that "because journalists work in the information domain, they may know some common strategies that fake news writers may leverage".

"For example, if you think of a short TikTok video, young teenagers are attracted to (misinformation) videos for some reason. Maybe there's some psychological reason. Maybe they try to combine some very controversial, very emotional cues in the video or twist something. What are these emotional and controversial cues? Can we quantify that? Journalists can help with that."

T1 said that when building an AI detection tool for fact checkers, they regretted not being more engaged with journalists for labelling during the development process. "We did use them for things like determine what types of (false) claims were worth categorising, lists of claim types, how to describe claim types… In hindsight, we didn't do enough. We should have done more checking with the fact checkers themselves to make sure the tool solved the problems they had."

ii. Inputter and reviewer roles for data

J5, who works directly on ground truth labelling for misinformation detection, said that journalists and other SMEs tended to produce "smaller datasets that are high quality and have a very clearly defined reason for everything that's included". The challenge is how to "make use of less, but higher-quality data". One way that journalists can work in a system like this, J5 says, is in dual roles as "inputters" and "reviewers".

"I would suggest that there is an inputter of data and there is a reviewer of data, both of whom might be journalistic in their background, separate to the technology or the technologist who builds on top of this data. The job of the inputter is to look at the guidelines, the formulations, the policies, the definitions, and collect data as they see fit. It's then the job of the reviewer to check that the data is in fact aligned with everything that's agreed in the definitional documentation.

"So, the role of the journalist is both collector and inspector, if you like. We need the reviewer to challenge the inputter to make sure that not everything just goes through because you know we don't have time to check things or whatever it might be so. I think where journalists need perhaps some outside help is from our friends in policy and governance and to help us realise where the red lines are, where it makes sense to collect data." They noted that there were similarities here to the traditional newsroom roles of reporter and sub-editor.

iii. Weak labels and synthetic data generation

Scarcity of high-quality data is a problem for the detection of misinformation, and researchers have begun to experiment with augmenting real-world datasets with synthetic datasets that semantically and distributionally resemble false information seen in the real world (Zeng et al., 2024).

T3, a computer scientist, also saw detecting new types of misinformation as a major challenge for the field. They saw a path to overcome this issue by annotating data via “weak supervision”, where noisy or qualitative data from journalists or other SMEs could be used to create an initial model that works at a better-than-random level. The “weak labels” may not be perfect, T3 said, but “as long as the weak label contains some knowledge, then (the model) will be able to learn and it will be helpful to the detection”.

Taking this one step further, T3 said the model could be further augmented by the generation of synthetic data that was like the original examples identified by the journalists. This could be done using an LLM, which would be tasked with generating variations on what had already been identified. Journalists could then be shown the synthetic data and asked to judge whether it was sufficiently realistic to use in the training data for the model.

T3 said: “If the human experts can identify a common problem from this generated text, then when we design our model, we can think of some kind of metrics to optimise and avoid these issues …. I think human experts are very important in assessing the quality of the generality of data, and it can help us to better generate the synthetic data to help the detection.”

iv. Identify untrustworthy data sources

In addition to assessing “news content features” such as writing style or article length, misinformation detection systems can also use “social context features” to predict if a new article is likely to contain mis- or disinformation (Shu et al., 2017). One such social context feature is the authority of the source that produced the article. Several participants noted that journalists were well-sorted for this type of ground-truth labelling.

This is a key activity for trust and safety platforms that sell misinformation services to platforms. J2 said that a core service they provided to platforms was “a database of potential actors” who engage in the distribution of misinformation. Finding these bad actors was a manual process, they said, involving human researchers searching through platforms such as Telegram, Rumble and X to find key influencers who were repeatedly spreading misinformation and driving new narratives.

Similarly, J5 said the first step for their team when assessing a new topic was to understand “who is the super spreader, who's the influencer for this particular area that we're focused on”.

> “Whether it's anti-vaccine or whether it's white supremacy, who are the leading voices? We then do deeper dives into other people who are maybe not quite so well known, but who are super influential in the movement itself. They're probably the ones who actually shape the evolution and the changing of (narratives) as much as anyone. Often the super-spreaders jumping on stuff they've seen elsewhere.”

Even T5, who due to concerns around media bias did not use journalists or human labellers directly in developing their algorithmically driven “all sides” news app, relied on labels from the website Media Bias / Factcheck to assess the credibility of sites.

5.5 Structuring fact check data

One solution to the scarcity of data challenge would be to combine work from many journalistic and fact-checking organisations into a single industry-wide dataset. This is a solution that has precedents, particularly with Child Sexual Abuse Material (CSAM) and Terrorist and Violent Extremist (TVEC) content, where standardised hash databases have allowed mass suppression of illegal and disturbing content (Gorwa et al., 2020).

Most fact checkers, including those interviewed for this paper, already make use of ClaimReview markup. This standardised system, developed with assistance from Google, provides a flexible schema of elements (like the title, the body, and the rating of the fact check) that fact checkers can funnel their content into. Search engines and platforms can map to this standardised schema to readily identify and present fact checks without needing to do a bespoke integration for each individual fact checker (Bélair-Gagnon et al., 2023).

But each fact checking organisation has its own way of representing its claims and even with ClaimReview, this presents challenges for training models. Verdicts can be presented as a true-false binary, on a numbered scale, or as text-only. Some are entirely unique ("3 Pinocchios", "Pants on Fire") (Jiang et al., 2020). ClaimReview's schema does provide some flexibility for bespoke data structures and ratings. At minimum, fact checkers need to enter a text rating stating their verdict on the fact check – this could be a single word ("False") or a sentence ("We believe this is false because…). But they can also to give the fact check a numerical score (e.g. 3), and define the range of scores possible in their scale (e.g. worst=0, best=5) (Google, n.d.).

If a platform attempted to create a shared database of up-to-date training data based off ClaimReview, it would somehow need to account for the various scales and methodologies used by the fact checkers. This would involve finding a way to standardise the data for the myriad rating methodologies and scales used by fact checkers around the world.

Jiang et al. (2020) note that while ClaimReview offers a standardised schema for fact-checks, the flexibility it allows in representing verdicts poses challenges for training models. Variability in the verdict structure makes it difficult for models to learn consistent patterns and accurately extract the necessary information. When they used an LLM (BERT) to automatically extract verdicts to help fact checkers, they found that while their system worked for well-known fact checkers with many examples, results were worse for less prominent fact checkers.

Beyond the technical structure of the data, there is also the editorial structure of the data. Fact-check information is surfaced by AI systems to users who may be disengaged and time-poor. J8 noted that fact checks needed to compete with misinformation narratives whose shocking nature tended to make them highly engaging for users. They also said that a poorly presented fact check could inadvertently spread a lie further, as users may only notice the false claim, and not the debunk.

5.5.1 How journalists can help

i. Open standard for fact check verdicts

Several participants mentioned the possibility that fact checkers and journalists could collaborate to create a structure for fact-check verdicts that was more suited to use in algorithms than the wide variety of formats in use today. J2, for example, said a "common open standard" for verdicts would be "a very useful thing". T3 said they would like to see "common ground among the different sites" on whether binary or multi-class rating should be used for each fact check help solve this classification problem.

But this presents a challenge—getting such a disparate group of organisations with competing aims to adopt a universal standard. Referring to the automated hash detection database for terrorist content run by the Global Internet Forum to Counter Terrorism (GIFCT), which allows platforms to automatically detect for known terrorism material, J5 said: “Could we do that (create an industry-wide database) for misinformation? And I think the challenge then becomes how can we agree on anything?” This kind of standardisation effort has some precedent in other areas of platform-journalist collaboration.

5.6 Definitions, policies and truth

Who, if anyone, should decide what is true and what is not?

This question is the biggest obstacle to effective content moderation of misinformation, and the difficulty in answering it played out in the failure of the Australian government’s Combatting Misinformation and Disinformation Bill ("Combatting Misinformation and Disinformation (Cth).", 2024). The bill proposed giving the Australian Communications and Media Authority (ACMA) various powers relating to platform transparency and oversight of a self-regulatory code. Most controversially, it gave the ACMA powers to create and enforce a regulatory standard of its own if it was not satisfied with the co-regulatory code, or in case of an emergency. But it stopped short of allowing ACMA to make decisions about individual pieces of content and set a high bar for whether something was “harmful” enough to warrant attention.

The bill was fiercely criticised when it was first released. Conservative-leaning media fiercely and repeatedly attacked the proposed legislation as “Orwellian” (Markson, 2023). But a wide range of groups criticised elements of the code. Submissions from TikTok (2024) and Google (2024b) focused on concerns about regulatory burden, a lack of definitional clarity about misinformation, concerns that ACMA’s powers were too broad, and disproportionate penalties. In contrast, the Centre for Media Transition’s submission (2024) said that the bill’s scope was too limited as it should consider content-level decisions, but recommended that any such decisions should be sent to an independent board rather than having ACMA act as an “arbiter of truth”. The bill was subsequently abandoned in November 2024.

Other jurisdictions—notably Europe, which is integrating its previously voluntary Code of Practice on Disinformation into the Digital Services Act effective July 1, 2025 (Jahangir, 2025)—have had more success in setting up a regulatory regime. But even without regulation, most platforms will opt to continue acting against misinformation. As Suzor notes, a lack of rules leaves it to the platforms themselves to regulate speech on the internet, making it a “lawless” space where companies operate behind closed doors (2019, p. 8).

One participant (T5) said that the platforms were in a “no win” situation where they would be criticised no matter what they did. They were doubtful that “censorship” of posts was the right approach for platforms to take, although they did advocate for some posts to be aggressively downranked and said that as companies, platforms had the right to takedown any content they wished.

5.6.1 How journalists can help

Discovering the truth, or something resembling the truth, is core to journalism’s self-perception, however imperfectly it is pursued. Kovach and Rosenstiel say that “the essence of journalism is verification” (2014, p. 9). Because journalism concerns itself with the material world, journalists should strive for “a practical or functional form of telling the truth”, which evolves as stories are reported and more facts emerge.

Platforms may be able to use some of these journalistic concepts to find a version of “practical truth” for content moderation, and even employ journalists with the intent of having them advise on their overarching policies and practices regarding misinformation.

i. Collaborate on the development of policies and guidelines.

There was consensus among participants on the need for robust policies to govern content moderation approaches to misinformation, underscoring a collective acknowledgment of the complexities inherent in defining and combating misinformation. J1 said: "Robust guidelines that help measure what constitutes misinformation need to be in place so that journalists, data reviewers and content moderators can work against an agreed set of rules and lessen the impact and the arbitrary nature of human preferences and biases… Labellers should be part of the discussion and decision-making process as well as ongoing refinement of the definitions and rules."

This sentiment was echoed by T3, who said that journalists were experts in issues such as “bias, credibility and sentiment”, and conceded that “maybe there are some signals we don't know as a computer scientist." Meanwhile, T6 emphasised the foundational nature of policy guidance and saw journalists as good candidates to provide it: "Particularly in mis- and disinformation, you need to start with really some guidance…some sort of a policy orientation."

This discussion calls to mind Common's (2020) argument for a publicly available "body of precedents" in content moderation, which would document how specific rules are applied in practice. While Common envisions the platforms’ own policy teams being responsible for maintaining this body, journalistic organisations could also contribute to the development of such a resource by providing examples of how they apply their own editorial standards in challenging cases. A further role for journalistic organisations in policing such a resource will be discussed in Chapter 6.

ii. Introduce fact-finding and verification frameworks to engineers and product managers working on misinformation issues at platforms.

Engineers, data scientists and product managers who work on misinformation issues come from a technical background and may not have sufficient experience working with news and information to create a sufficiently robust framework for understanding the content they are dealing with. J2, who works in fact checking and misinformation detection at a major newswire, saw an opportunity for journalists to contribute frameworks for “identifying and dealing with information”.

"We need the data scientists and engineers building AI systems to understand the frameworks of manipulation,” they said. “Then they can design systems that can identify mis- and disinformation. Good journalists can absolutely bring their specialism to bear in the design of AI models and frameworks.”

J4, who also works in fact checking at a different newswire, said that platforms had in the past collaborated with them on guideline development. But they said the level of interest in consulting directly with journalists on information disorder problems varied depending on whether the platforms were under scrutiny. “In the past, they saw it as a priority, but now I don’t think they do so much. It seems to be dependent on the amount of pressure on the platform.”

iii. Contribute to high-level principles that guide platform governance

Napoli advocates for misinformation issues to be considered through a media lens and argues that the concept of “public interest”, which is prominent in journalism, should be brought into the world of platforms and personalised feeds (2019, pp. 9-16). There is some evidence that platform companies have moved in this direction – Meta’s Oversight Board includes three prominent journalists who contribute to decisions and recommendations that impact the company’s direction.

Some respondents expressed support for this idea. J2 felt that engagement with journalistic institutions could be beneficial for platforms: "It would be sensible to consult with journalists and journalistic organisations known for their rigor and approach… Not everyone involved in journalism has the best of intent and I think that's why that's why journalistic organisations and institutions are very important."

J4 was positively disposed towards the idea but did not believe it was realistic. "What if platforms saw themselves as having responsibility to provide accurate information? That's very much a journalistic way of seeing things… I feel like (the platforms) are quite keen to show that intent when there's a really big event such as COVID."

J6 also said it would be good for platforms to consult journalists on governance but was also doubtful of their willingness to do so: "I don't think most of them have public interest as their top priority. I think their first interest is generating as much revenue as possible. Public interest is there but maybe somewhere down the list."

But if platforms are not willing to engage, another option is for journalists and fact checkers to proactively take up issues with the platforms and apply public pressure. J6 noted that fact checkers banded together to publish an open letter to YouTube in 2022, which said that disinformation was rampant on the platform, especially in non-English countries and the global south. The letter demanded "meaningful transparency" from YouTube on disinformation, provision of context and labels on debunked content, taking action against repeat offenders and extending support in languages other than English.

J6, who was a signatory to the letter, said that this had led to "very promising discussions" with YouTube, and expressed hope that the platform and fact checkers would soon announce a program to solve the problem publicly.

5.8 Chapter summary

This chapter has explored how journalists can contribute to AI-driven systems designed to mitigate the spread of misinformation at scale. The challenges are substantial: the dynamic nature of misinformation, linguistic and cross-cultural complexities, limitations of data quality, and the structure of fact-check data all point to solutions beyond the purely technological. The statements of the participants indicate that both technologists and journalists believe that the news media's emphasis on verification, contextualisation, and a commitment to the public interest can complement to these technological efforts.

A recurring theme was the importance of novelty. Participants highlighted the limitations of AI in identifying and responding to new narratives, echoing concerns about the rapid spread of false information online (Vosoughi et al., 2018). The skills of journalists, particularly in rapidly identifying and verifying new information and prioritising its importance, were consistently identified as valuable in this context. Their ability to discern credible sources, use OSINT techniques, and determine whether claims are sufficiently newsworthy to bother investigating, add valuable insight that can make end-to-end moderation systems more accurate and efficient.

Journalists are also good candidates to assist with linguistic and cross-cultural challenges. Participants underscored the limitations of current AI systems in non-English languages and outside Western contexts, a concern also voiced in the broader literature (Budak et al., 2024; Gillespie, 2018; Zeng & Brennen, 2023). Participants believed that journalists with deep knowledge of specific regions and cultures can help ensure content moderation systems are both accurate and equitable, particularly in regions with limited press freedom.

The provision of high-quality, unbiased "ground truth" data is another area where journalists can contribute. While crowdsourcing can provide data at scale, misinformation demands the careful, contextualised analysis that journalists are trained to provide.

Also highlighted was the potential for greater standardisation in the structure of fact-check data. The prospect of a "common open standard" for fact-check verdicts, as suggested by some participants, could enhance the usability of this data for algorithmic systems. Finally, the challenges faced by platforms in developing and enforcing consistent policies underscore the need for a principled approach grounded in a commitment to public interest. Insights from journalists regarding verification frameworks and policy development could prove invaluable.

Chapter 6: Conclusion

This final chapter synthesises the insights from Chapters 4 and 5 and provides a set of recommendations, starting with straightforward ideas before moving on to more radical solutions. Consideration will also be given to gaps in feasibility, non-Western contexts, and funding mechanisms. The end goal is to forge closer alignment among journalists, technologists and platforms, grounded in a renewed commitment to the public interest.

6.1 Analysis

6.1.1 Shared definitions, different emphases

The interviews in Chapter 4 show that while journalists and technologists broadly agree on definitions of misinformation (unintentionally false or misleading information) and disinformation (intentionally deceptive content), some lingering ambiguity remains. Nonetheless, these definitional nuances do not appear to be a fundamental obstacle to productive collaboration.

The most polarised term was fake news, which nearly all participants viewed as politicised and unhelpful. Meanwhile, propaganda was defined by all as deliberately persuasive communication—though some treated it as government-driven, while others viewed it more generally as *any* effort to strongly shape public opinion. These definitional overlaps and variations underscore that while there is enough common ground to collaborate, journalists and technologists sometimes bring different emphases: as noted in Chapter 4, journalists often underscore societal, political, and cultural factors, whereas technologists more readily spotlight system design and engineering challenges.

6.1.2 Attitudes toward platforms and "information disorder"

Participants agreed that misinformation represented a serious threat, though some technologists were more measured about the scale of real-world harms. Journalists expressed higher urgency, often citing democracy, public well-being and civic trust as at risk.

Neither group placed total blame for "information disorder" on platforms alone. However, journalists tended to see platform design and business models as exacerbating sensationalism and fuelling harmful content. Some technologists emphasised that these harms may be accidental byproducts of revenue-driven optimisations and pointed to end-user responsibilities. Both sides recognised that platform designs nudge user behaviour in ways that can amplify misinformation.

On free speech, technologists generally leaned more libertarian, while journalists weighed speech against potential harms—particularly around hate speech, vaccine denialism, or orchestrated political campaigns. This was a surprising result, given journalism's generally strident stance on free speech issues. On algorithmic transparency, journalists advocated for insights into content-ranking decisions. Technologists cautioned that "full code release" is fraught with privacy, security, and intellectual property issues and that fully explaining large-scale AI models is non-trivial. Most participants saw a middle ground in targeted transparency, such as publishing error rates, aggregate statistics, and moderation "case law."

6.1.3 Journalism's role in the evolving information disorder threat

Chapter 5 shows that misinformation is a moving target, requiring human context and editorial judgment. Journalists are well-positioned to:

- Rapidly detect emergent narratives
- Prioritise harmful vs. harmless rumours
- Provide cultural or linguistic nuance

- Conduct forensic investigations
- Establish “ground truth” in contested domains
- Rank the credibility of individual sources

These activities can help improve machine-learning pipelines. Although fact checks yield relatively small amounts of carefully vetted data, their utility in algorithms remains considerable, whether to ensure quality control with data from cheaper, but less reliable, sources, or as a seed for high-quality synthetic data. Moreover, journalism’s public-interest ethos provides a counterweight to purely commercial or risk-avoidance pressures.

6.2 Recommendations

6.2.1 Cultivating Common Ground

While the participant interviews do not reveal a large gap in understanding between journalists and technologists, definitional issues are an ongoing issue for the study of misinformation, disinformation and propaganda. This lack of coherence leaves the field open to critiques, including complaints that any attempt to moderate speech is “censorship” (Markson, 2023), and that those who study the lies, distortions and misrepresentations routinely broadcast through digital platforms are “pseudo-experts” (Jordan et al., 2020).

To ensure both journalists and technologists working on misinformation have a shared understanding and lexicon, the following ideas can be considered:

i. Create basic definitions toolkits

Professional bodies or research institutes could publish succinct guides (e.g. clarifying misinformation, disinformation and propaganda, and clearly calling out where definitions diverge or overlap) for use by both journalists and technologists. These can be part of standard onboarding for newsroom staff, trust and safety teams, and fact checkers.

ii. Encourage basic AI literacy in newsrooms and editorial literacy among technologists

Despite working in the field of misinformation, some of the journalists interviewed showed scant knowledge of either content moderation or AI. Journalists working on information disorder issues could benefit from workshops or short courses on how content-classification models work, how to interpret algorithmic error rates, and how to spot AI bias.

Likewise, engineers and product managers who are working on issues relating to verification and truthfulness of information could benefit understanding the techniques and ethos of journalism, which has long been on the front line of these tricky issues. This could involve contracting with media trainers to provide courses in source verification, open-source intelligence (OSINT) and journalistic ethics—clarifying how journalists approach truth claims in real-world events.

One objection to this is that journalists often do not live up to their own ideals and are themselves biased and liable to produce inaccurate news. But this statement is overly general—there are many newsrooms that continue to live up to high standards, and this commitment is tangible through practices such as issuing corrections, being public about decision-making processes, and having clear and standardised processes for editing and fact-checking of copy. Provided that technology companies partner with newsrooms that retain a commitment to verification and truthfulness, there is much value to be found here.

iii. Establish joint workshops and pilot projects

A further idea is to co-develop regular “misinfo labs” or hackathons where journalists and technologists build prototypes. This would have several benefits:

- Journalists and technologists who participate in such hackathons will gain an enhanced understanding of each other’s viewpoints, blind spots and unique knowledge.
- Journalists can learn in detail about AI concepts and practices relating to testing, safety, and data.
- Technologists will get exposure to journalistic ways of thinking about verification and how to work through situations where “ground truth” is not easily or immediately identifiable.
- There may be successful prototypes and these interactions would lay the groundwork for more extensive collaboration on larger problems.

6.2.2 Strengthening fact checking and data-sharing ecosystems

i. Unify fact-check data structures

Currently, fact-check verdicts are highly heterogeneous. Industry players—wire services, fact-checking organisations—could:

- Standardise a set of core verdict types (e.g., “False,” “Mostly False,” “Unproven,” “True,” “Context Needed”).
- Use a consistent numeric or categorical scale to enable cross-organisation aggregation or “translation” between idiosyncratic verdict types.

That said, the dramatic improvements in function of LLMs may render this work unnecessary. While Jiang et al. (2020) were only partially successful in using Google’s BERT to automatically generate standardised verdicts from varied fact check formats, it would be worth repeating such an experiment with frontier models such as GPT4o or Gemini 2.5.

ii. Use journalists’ expertise for “weak labels” and synthetic data

As participant T3 noted, journalists could be used to seed small, high-quality datasets that data scientists augment with synthetic text or images. Journalistic review can also help ensure that the expanded synthetic data truly mirrors real misinformation narratives. This related to a more general financial opportunity for newsrooms to supply high-quality data for the training of LLMs, which will be discussed in section 6.3.

6.2.3 Truly independent oversight board

Determining the best legislation for platforms is a complex and evolving problem with substantial deviation between countries.

Meta’s Oversight Board provides a real-world example of a multi-stakeholder advisory with a charter that emphasises the public interest. It has had mixed success and been the subject of criticism, especially after responding positively to Meta’s pullback from moderation in January 2025 (Stokel-Walker, 2025; West & White, 2024).

But conceptually, the establishment of a similar multi-stakeholder bodies that include senior journalists and fact checkers, alongside representatives from civil society groups and academia, has utility both to facilitate establishment of stronger public interest governance in platforms, and as a third-party independent check on particularly difficult moderation decisions. Considering the positions that Meta and X have taken to roll back content moderation and criticise media, it is likely that such a group would need to be imbued with government legitimacy and regulatory powers to be effective.

6.3 Outsource more moderation to journalists

The platforms' disintermediation of the news media from both their audiences and their advertisers has been a major concern for legislators worldwide, but efforts to make platforms pay for news have had mixed results. As noted in Section 1.1, Meta opted to pull all news from its platforms in Canada after the government's Online News Act demanded they pay for content, and Australia is considering additions to its News Media Bargaining Code after Meta said it would not renew deals with publishers (Bossio, 2024).

One reason these revenue-sharing deals have foundered is the wide gap between what news publishers believe their content is worth and how much platforms are willing to pay. While the news industry continues to claim it is being underpaid, Meta's actions suggest that the company does not value news content enough to sustain partnerships. From a purely financial point of view, the platforms may be sincere in asserting that news is simply not core to their strategy: it is controversial, expensive to moderate, and less appealing to advertisers than soft or entertaining content.

But profit is not the only lens for evaluating the importance of news. Napoli argues that social media platforms have placed profits above the public interest, portraying their algorithms as neutral while ignoring real-world harms (2019). In his view, platforms must be compelled to take their public role more seriously, particularly as they have become the primary news source for many audiences. One way to do this, Napoli says, is to adopt some of journalism's public interest values, including independence, truth and accuracy (pp. 171-172).

While forcing platforms to pay for news is one approach, a more enduring solution would bring substantial value to both sides of any arrangement and protect public interest—not merely patch over revenue disputes. What if platforms "outsourced" moderation not just to business process outsourcing (BPO) firms, but to reputable newsrooms?

Rather than only pay for low-value, low-cost review against platform-established rules, platforms would fund news organisations to lend their editorial expertise and public-service outlook to moderation systems. This shift has implications far beyond revenue: it would acknowledge that large platforms, by virtue of their scale and influence, share responsibility for the health of the information ecosystem.

6.3.1 Embedded Journalist Teams (EJTs)

A core component of this approach would be teams of journalists embedded within platforms' content moderation teams, yet independent from their management structure, acting both as expert advisors and watchdog. These Embedded Journalist Teams (EJTs) could be entirely funded by platforms, or co-funded by platforms and government grants. The EJTs—drawn from large or mid-sized news organisations—would have several functions.

1. Act as a strike team to detect and catalogue emerging disinformation narratives that have a high potential for public harm.
2. Provide very high-quality data on disinformation and the credibility of specific sources for training and augmenting models.
3. Provide third-party adjudication on the most difficult content moderation decisions that the platforms face and provide advice on how to respond to specific emerging disinformation threats.

4. Audit platforms' performance on moderating and report publicly on this to increase transparency.

The watchdog function of an EJT aligns with the investigative tradition of journalism, injecting oversight and transparency into the moderation process. Both sides of this arrangement would have skin in the game—when moderation inevitably goes wrong, the EJTs will share in the responsibility for mistakes. One strength of high-quality journalism companies is owning up to mistakes in a clear, comprehensible and public fashion, and bringing this deeply held ethos into content moderation governance would help platforms move away from a mindset of risk management and profit protection, and towards an honest embrace of accountability and public interest.

6.3.2 Benefits

i. Credible rulings

Journalists trained in investigative methods, and news organisation with a track record of high-quality verification and investigation, can deliver consistent, authoritative decisions on questionable content.

ii. Sustainable funding

Platforms' moderation budgets could underwrite high-quality journalism, in contrast to the purely transactional "licensing deals" that have yielded limited public benefit. Platforms would receive more value for money and be more likely to continue paying for the service.

iii. Public interest lens

The watchdog function of an EJT aligns with the investigative tradition of journalism, injecting oversight and transparency into the moderation process. One strength of high-quality journalism companies is owning up to mistakes in a clear, comprehensible and public fashion, and bringing this deeply held ethos into content moderation governance would help platforms move away from a mindset of risk management and profit protection, and towards an honest embrace of accountability and public interest.

iv. High-quality data for AI

Many major platforms have AI systems that require up-to-date and accurate information to provide utility to users and avoid incorrect information. A close partnership with a major newsroom would provide a constant source of very high-quality data that could be fed into models, particularly those dealing with real-time search results. This would have an advantage over current deals in that the data produced would be highly suited to the specific issues that platform faces.

Collaborating with EJTs would help platforms evolve beyond what Common (2020) terms the "Efficiency Narrative"—an ethos that encourages rapid takedowns that may overlook societal and ethical nuances. It furthers Napoli's argument that platforms need strong legislative nudges or incentives to accept their public-interest obligations. Journalists, with their expertise in contextual understanding and ethical scrutiny, naturally fit that role. By funding or co-funding newsroom watchdogs, platforms could harness this expertise on an ongoing basis, rather than relying on patchwork, after-the-fact interventions.

6.3.3 Objections

i. Objectivity and bias

Some news outlets lean partisan. Robust codes of practice, editorial independence structures, and transparent oversight would be essential to combat perceived or real biases. Carson et. al. (2023) found that while trust in fact-checking in Australia was generally high, ideologically right-wing participants had lower levels of trust in fact-check verdicts. They recommend that platforms consider deploying different fact-checking groups to reach different ideological groups. This is an important point - handing over platform moderation to a publisher who was not well-trusted by one side of politics could lead to distrust in decisions and accusations of political bias and free-speech suppression, even if that publisher was generally doing a good job.

ii. Liability Issues

If journalists moderate borderline or defamatory content, legal frameworks must clarify how lawsuits or free-speech disputes are handled, including the mechanisms for appeals. This issue would need to be defined in each jurisdiction where journalists operate on behalf of the platforms. A particularly thorny issue will be indemnification for the financial costs of any major error.

iii. Coverage Gaps

Mis- and disinformation are global problems that extend across cultural, linguistic and legislative borders. How would embedded journalist teams achieve sufficient coverage to combat the well-documented Western centrism in content moderation, while also maintaining a level of consistency and internal coherence regarding enforcement decisions? Some options are discussed in 6.3.3.

iv. Institutional capture

Graves and Anderson (2020) and Meese (2023) both note that relationships with tech companies have pushed journalists to adjust their practices to suit the needs of platforms, which can undermine the public interest ethos of journalism. Further integration between newsrooms and content moderation teams could increasingly push journalists towards adapting their practice to benefit platforms. This is why a somewhat independent watchdog function is essential in a EJT model. Large and credible organisations will also have more scope to resist this institutional capture.

6.3.4 Operationalising Newsroom Content Moderation in Platforms

Given the objections above, how could a program using Embedded Journalist Teams work in practice? The list below identifies some practical starting points for making this achievable.

i. Choose partners with a great reputation

From a platform's perspective, any journalistic partner needs to meet some key criteria. Firstly, they need to be highly trustworthy, with a reputation for factual, unbiased reporting and correcting mistakes on the record. Broad geographical and linguistic coverage is also necessary, preferably reaching into countries where trustworthy information may be hard to come by.

Two potential paths for achieving this would be:

1. **Single Global Partner**
   Collaborate with a major wire service (e.g. AP, Reuters, AFP, Deutsche Welle) that has broad language coverage, a strong reputation and an established public-interest ethos. The partner can use their existing scale to increasing coverage in areas that are key for the platform. Reputationally, wire services have held up strongly compared to more partisan news services (Orth & Bialik, 2024).

2. **Coalition of News Outlets**
   Convene multiple reputable organisations committed to public-interest journalism—allowing for regional specialisation and greater cultural nuance. A set of criteria should apply to each candidate newsroom, including their ownership structure, evidence of independence from commercial or political interests, and a track record of impactful and preferably prize-winning journalism. This approach ties in with Carson et. al.’s (2023)

recommendation to use a range of fact-checking sources to reach different ideological groups

Platforms can also choose to combine these two strategies, potentially using the large global partner to manage the partnerships with regional, local and politically diverse news organisations on their behalf.

ii. Joint governance charters

Once partners are chosen, it is essential to establish formal contracts outlining the responsibilities of both platform and newsroom. These contracts must define scope (topic areas, languages), timelines, indemnification and the precise roles of the platform moderation team and the EJT. This minimises legal and editorial ambiguities. But for such a partnership to be impactful, the platform team should be prepared to hand over a meaningful level of scrutiny to the EJT, with mechanisms and requirements to act on the advice that is given.

iii. Tiered rollout and pilot programs

Begin with small-scale implementations (e.g. start with countries where a wire agency has strong representation, or some tricky smaller markets where there is an excellent and reputable local partner), then expand. Evaluate success through metrics like error rates, user satisfaction and median takedown times. Ensure robust governance to protect platform companies, newsrooms and individual journalists from legal and government threats.

iv. Scale to global operations

If the pilot programs prove successful, the model can be gradually scaled up to global levels. One major challenge with this is the enormous variety of legislative and information environments that can be found globally. Digital platforms need to create a level of consistency for their moderation enforcement and must abide by their own publicly statement ethical guidelines. If EJTs are used globally, the activities and decisions of those teams also need to negotiate this complexity.

If the platform chooses to partner with a global journalistic organisation, the simplest path here is close alignment between the partner and the platform's global legal team. Each market launch of an EJT would require sign-off from the appropriate legal representative at the platform company and be governed by a specific set of rules and requirements relating to the specifics of the jurisdiction.

If multiple newsrooms are involved globally, governance becomes trickier. In this case, a governing body with representatives from the platform and each newsroom partner (or, if this is too unwieldy, a regional representative with specialist knowledge) could investigate and approve the rules and requirements for each new EJT jurisdiction.

6.4 Chapter summary

The global information ecosystem stands at a crossroads. Generative AI has accelerated the distribution of untruthful content, leaving platforms under pressure to moderate at scale and journalists grappling with diminished business models. Yet the research and interviews in this thesis suggest that collaboration between platforms and newsrooms could be highly advantageous to multiple parties.

- **Platforms** can gain from journalists' deep expertise in verification, contextual analysis, and investigative methods. This collaboration can also help allay government fears about public interest and disinformation. It also gets them access to pools of high-quality journalistic data for AI training that can be targeted to their specific needs.

- **Newsrooms** secure new revenue streams and alliances and help bring their public-interest ethos and verification skills more deeply and meaningfully into the digital sphere.
- **Governments** can push platforms towards a stronger public interest ethos that values good societal outcomes over profits while providing newsrooms with a more sustainable revenue stream than requiring platforms to pay for news content.

Misinformation and disinformation will continue to be a highly contested and politicised field for the foreseeable future. The sudden dismantling of Meta's fact-checking program in early 2025 (Isaac & Schliefer, 2025) highlights the fragility of progress made over the past decade. Far from cementing stronger ties between social media platforms and reputable fact-checking organisations, some platforms appear to have reversed course—further underscoring the urgency of building systemic and sustainable collaborations.

Technological advances continue to provide bad actors with new methods to produce, distribute and reinforce politically and financially motivated lies, so the ability to detect and act against this content (particularly disinformation) continues to be an essential activity for platforms and a major concern for journalists and governments. Bridging the divide between technology and journalism will help all parties through more effective moderation, greater confidence in online information, and a healthier digital public sphere.

Appendices

Appendix 1 – Participant details

Code	Class	Subclass	Workplace type	Country
J1	Journalist	Fact checker	Fact check org	Australia
J2	Journalist	Product manager	Wire service	UK
J3	Journalist	Investigator	Trust & Safety Platform	Israel
J4	Journalist	Fact checker	Wire service	France
J5	Journalist	Policy	Trust & Safety Platform	Ireland
J6	Journalist	Fact checker	Fact check org	US / Kenya
J7	Journalist	Fact checker	Fact check org	Australia
J8	Journalist	Fact checker and product lead	Fact check org	Belgium
T1	Tech	Data scientist	Fact check org	UK
T2	Tech	Data scientist	Fact check org	US
T3	Tech	Data scientist / Entrepreneur	University / Startup	US
T4	Tech	Data scientist	Platform	US / Iran
T5	Tech	Entrepreneur	Startup	US
T6	Tech	Engineering lead	Trust & Safety Platform	Ireland

Appendix 2 – Interview Questionnaire

Introduction for all interviews

Script: Firstly, I want to reiterate that I do not want you to share any information with me that you believe could be commercially sensitive or that should be kept confidential. The purpose of this research is to collaborate with industry on finding great solutions to combatting misinformation. I want all interview subjects to have confidence that they can do so safely.

I may ask you questions that are beyond your scope of knowledge. In that case, answer as best you can. It's perfectly OK to tell me, "I know nothing about this".

My research question is: How can the skills and practices of journalists be put to use in AI systems to combat information disorder?

This is further broken down into seven related sub-questions. Some of these relate to governance, and others to the practical aspects of detecting and taking action against misinformation and disinformation. Note that these are not the questions I am asking you to answer directly – this is context for you to better understand the purpose of my research.

- Practical concerns:
  - Do journalists and platform practitioners have shared definitions and similar mental models regarding mis- and disinformation?
  - How can journalists best contribute to expert labelling / annotation that is crucial to improving machine learning systems?
  - Can some concepts or practices from journalism be imported into AI systems? (For example, "strategic silence" on some topics, or verification practices?)
  - Could platforms in some circumstances embrace the journalistic notion of "practical truth" in their AI systems? If so, what's the best mechanism to determine that "practical truth" has been reached and adjudicate between competing claims?
- Governance concerns:
  - Would platforms benefit from involving senior journalists in discussions about key policy issues relating to news distribution and information integrity?
  - Could platforms use media self-regulation as a model for their own AI self-regulation—setting minimum moderation standards for any accredited platform and adopting practices such as transparency about mistakes?
  - How can the concept of "public interest", so prominent in journalism (at least, journalism at its best) be adopted into AI systems that deal with news selection and distribution?

Common questions

1. As concisely as you can, please tell me what is: (10 minutes)
   a. Misinformation
   b. Disinformation
   c. Propaganda
   d. Fake news
2. To what extent do you agree with the following statements? (20 minutes)
   Answer: (Scale 1-10, 1 = Strongly Disagree, 10 = Strongly Agree)
   Why?
   a. Misinformation and disinformation are a major problem for society.

   b. Social media platforms are causing “information disorder” and disrupting democracies.
   c. The design of social media platforms means they are likely to promote sensational claims over sober claims, regardless of the truth of those claims.
   d. Tech companies should be forced to make their algorithms and data transparent so that researchers can interrogate them.
   e. Trained journalists possess skills and practices unique to their profession that allow them to find out verifiably true things.
   f. The mainstream media (by which I mean newspapers, television and websites with large audiences who produce original reporting and claim to adhere to journalistic codes of practice) generally acts in the “public interest” and has adequate mechanisms in place to make sure that truth is promoted over lies.
   g. It is more important to protect free speech than it is to protect people from “harmful content”.
3. If you were put in charge of fixing the problem of “information disorder” in a liberal democracy, what would be your top priority or priorities? (10 minutes)

Journalism-specific questions (30 minutes)

1. Does your publication allow user generated content (e.g. user comments)? If so, what policies do you have to govern what users can or cannot say, and do any concern misinformation?
2. Outside of UGC, does your publication(s) take steps to combat misinformation or disinformation? What approaches do you use?
3. What do you know about AI systems? If you have sufficient knowledge, how do you think journalists could best contribute to an AI system set up to detect misinformation?
4. Should platforms consult journalists when they are writing content moderation policy or building AIs that present users with news and information? If so, how would journalists add value?
5. Determining what is true is a tricky notion. What standard does your organization set for determining “practical truth”? At what point is a fact sufficiently “verified” for publication?
6. If you were to import this notion into a system dealing with millions of competing claims, how could you make such an approach work at scale?
7. Do you agree that journalism has strong "public interest" ethos? If so, what sorts of ideas from journalism could platforms benefit from adopting / exploring further?
8. Do you think that platforms should be scrutinised by self-regulatory bodies or government regulators, in the mold of the Press Council or the ACMA? If so, what sorts of issues should be scrutinised? How would this help?

Technologist-specific questions (30 minutes)

1. Do you commonly see misinformation and disinformation on your platform? If so, where are you most likely to see it, and what form does it take?
2. Policies
   a. What policies and guidelines do you have for misinformation and disinformation? Why did you choose to take this approach? Please be mindful not to disclose anything sensitive or confidential to me.

    b. When you created your policies and guidelines, did you ask for input from journalists or other experts? What was your reasoning? Do you think this could be helpful?

3. Labelling.
    a. What's the best way that journalists / newsrooms can contribute to labelling efforts improve machine learning?
    b. What's the best approach to labelling for complex topics like misinformation?
    c. Labelling can be boring and laborious. But a large volume of labels is essential to make AI perform well. How can you keep high-quality labellers like professional journalists engaged?
4. Practical truth.
    a. Determining what is true is a tricky question. How does your platform handle deciding what is true?
    b. How do you (or your partners) reach a threshold where you decide something is verified?
    c. How could the verification work that journalists do be put to use in your moderation systems?
5. Public interest.
    a. To what extent does the tech industry consider "public interest" when creating platforms / products? Do you think product managers should be forced to consider "public interest" when creating new products or algorithms?
    b. What would a "public interest" approach to designing social media products look like?
6. Regulation
    a. Do you think that the tech industry has sufficient regulation (whether self- or government) in place?
    b. Media often organise self-regulation according to certain principles, making themselves answerable to third parties with binding agreements to, for instance, publish prominent corrections to mistakes. Is this a model that would be helpful to apply to platforms?

References

Acker, A., & Donovan, J. (2019). Data craft: a theory/methods package for critical internet studies. *Information, Communication & Society*, *22*(11), 1590-1609. https://doi.org/10.1080/1369118X.2019.1645194

Aigne Roy, E. (2020). New Zealand media put Christchurch gunman in his place with focus on victims. *The Guardian*. https://www.theguardian.com/world/2020/aug/27/new-zealand-media-put-christchurch-gunman-in-his-place-with-focus-on-victims

Alba, D. K., Ella, & Silver, J. (2021). What happened when Trump was banned on social media. *The New York Times*. https://www.nytimes.com/interactive/2021/06/07/technology/trump-social-media-ban.html#:~:text=When%20Facebook%20and%20Twitter%20barred,a%20risk%20to%20public%20safety.

Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2020). Scaling up fact-checking using the wisdom of crowds. *Preprint at https://doi. org/10.31234/osf. io/9qdza*.

Altay, S., Berriche, M., Heuer, H., Farkas, J., & Rathje, S. (2023). A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field. *Harvard Kennedy School misinformation review.*, *4*(4). https://doi.org/10.37016/mr-2020-119

Amakoh, K. O. (2020, 2020//). Maintaining Journalistic Authority. Disinformation in Open Online Media, Cham.

Anderson, C. W., Bell, E., & Shirky, C. (2015). POST-INDUSTRIAL JOURNALISM: ADAPTING TO THE PRESENT. *Geopolitics, History and International Relations*, *7*(2), 32-123.

Andi, S. (2021). How and why do consumers access news on social media? In N. Newman (Ed.), *Reuters Institute Digital News Report 2021*. Reuters Institute & University of Oxford. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021/how-and-why-do-consumers-access-news-social-media

Apps, L. (1990). Media Ethics in Australia. *Journal of Mass Media Ethics*, *5*(2), 117-135. https://doi.org/10.1207/s15327728jmme0502_4

Asr, F. T., Mokhtari, M., Taboada, M., Maci, S. M., McGlashan, M., Seargeant, P., & Demata, M. (2024). Misinformation Detection in News Text: Automatic methods and data limitations. In (1 ed., pp. 79-102). Routledge. https://doi.org/10.4324/9781003224495-7

Asr, F. T., & Taboada, M. (2019). Big Data and quality data for fake news and misinformation detection. *Big Data & Society*, *6*(1), 2053951719843310.

Attard, M., Davis, M., Lee, K., & Wilding, D. (2024). Inquiry into the Communications Legislation Amendment (Combatting Misinformation and Disinformation Bill) 2024: Submission from UTS Centre for Media Transition to the Senate Environment and Communications Legislation Committee. In C. f. M. Transition (Ed.).

Attorney General's Department. (2019). *Abhorrent Violent Material Act Fact Sheet*. Retrieved from https://www.ag.gov.au/crime/publications/abhorrent-violent-material-act-fact-sheet

Bago, B., Rand, D. G., & Pennycook, G. (2020). Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General*, *149*(8), 1608-1613. https://doi.org/10.1037/xge0000729

Barbera, D. L., Maddalena, E., Soprano, M., Roitero, K., Demartini, G., Ceolin, D., Spina, D., & Mizzaro, S. (2024). Crowdsourced Fact-checking: Does It Actually Work? *Information processing & management*, *61*(5), 103792. https://doi.org/10.1016/j.ipm.2024.103792

Barlow, J. P. (1996). *A Declaration of the Independence of Cyberspace*. Retrieved 26/2/2022 from https://www.eff.org/cyberspace-independence

Barot, T. (2014). Verifying Images. In C. Silverman (Ed.), *Verification Handbook: An Ultimate Guideline on Digital Age Sourcing for Emergency Coverage* (pp. 35-41).

Bauer, A. J., Nadler, A., & Nelson, J. L. (2022). What is Fox News? Partisan Journalism, Misinformation, and the Problem of Classification. *Electronic news (Mahwah, N.J.)*, *16*(1), 18-29. https://doi.org/10.1177/19312431211060426

Bélair-Gagnon, V., Larsen, R., Graves, L., & Westlund, O. (2023). Knowledge work in platform fact-checking partnerships.

Benegal, S. D., & Scruggs, L. A. (2018). Correcting misinformation about climate change: the impact of partisanship in an experimental setting. *Climatic change*, *148*(1-2), 61-80. https://doi.org/10.1007/s10584-018-2192-4

Beres, D. (2024). Generative AI's Slop Era. *The Atlantic*. https://www.theatlantic.com/newsletters/archive/2024/08/ai-search-bots-war/679429/

Bernays, E. L. (1928). *Propaganda, by Edward L. Bernays*. H. Liveright, 1928.[i.e.1930].

Bernstein, J. (2021). Bad news: Selling the story of disinformation. Retrieved 5/9/2021, from https://harpers.org/archive/2021/09/bad-news-selling-the-story-of-disinformation/

Bing. (n.d.). *Bing News PubHub Guidelines for News Publishers*. Retrieved 5/1/2022 from https://pubhub.bing.com/Home/Help

Bossio, D. (2024). News bargaining incentive: the latest move in the government's 'four-dimensional chess' battle with Meta. Retrieved 28/12/2024, from https://theconversation.com/news-bargaining-incentive-the-latest-move-in-the-governments-four-dimensional-chess-battle-with-meta-245838

Bossio, D., Flew, T., Meese, J., Leaver, T., & Barnet, B. (2022). Australia's News Media Bargaining Code and the global turn towards platform regulation. *Policy & Internet*, *14*(1), 136-150. https://doi.org/https://doi.org/10.1002/poi3.284

Bounegru, L., Gray, J., Venturini, T., & Mauri, M. (2018). *A Field Guide to "Fake News" and Other Information Disorders*. Public Data Lab. https://doi.org/http://doi.org/10.5281/zenodo.1136272

Brashier, N. M., Pennycook, G., Berinsky, A. J., & Rand, D. G. (2021). Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences*, *118*(5), e2020043118. https://doi.org/10.1073/pnas.2020043118

Budak, C., Nyhan, B., Rothschild, D. M., Thorson, E., & Watts, D. J. (2024). Misunderstanding the harms of online misinformation. *Nature*, *630*(8015), 45-53. https://doi.org/10.1038/s41586-024-07417-w

Butler, J. (2024). Labor dumps misinformation bill after Senate unites against it. https://www.theguardian.com/australia-news/2024/nov/24/labor-dumps-misinformation-bill-after-senate-unites-against-it

Caplan, R., Hanson, L., & Donovan, J. (2018). *Dead reckoning: Navigating content moderation after fake news*. https://datasociety.net/library/dead-reckoning/

Carson, A., Gravelle, T. B., Phillips, J. B., Meese, J., & Ruppanner, L. (2023). Do Brands Matter? Understanding Public Trust in Third-Party Factcheckers of Misinformation and Disinformation on Facebook. *International journal of communication (Online)*, *17*, 6051.

The ClaimReview Project. Retrieved 25/6/2023 from https://www.claimreviewproject.com/

Coaston, J. (2018). YouTube, Facebook, and Apple's ban on Alex Jones, explained. *Vox*. **https://www.vox.com/2018/8/6/17655658/alex-jones-facebook-youtube-conspiracy-theories**

Combatting Misinformation and Disinformation (Cth). (2024).

Common, M. F. (2020). Fear the Reaper: how content moderation rules are enforced on social media. *International review of law, computers & technology*, *34*(2), 126-152. https://doi.org/10.1080/13600869.2020.1733762

Community Notes Guide. (2023). Twitter. Retrieved 18/6/2023 from https://communitynotes.twitter.com/guide/en

Demartini, G., Mizzaro, S., & Spina, D. (2020). Human-in-the-loop Artificial Intelligence for Fighting Online Misinformation: Challenges and Opportunities. *IEEE Data Eng. Bull.*, *43*(3), 65-74.

DIGI. (2021). *Australian Code of Practice on Disinformation and Misinformation*. https://digi.org.au/disinformation-code/

DiResta, R. (2018). Computational propaganda: If you make it trend, you make it true. *The Yale review*, *106*(4), 12-29. https://doi.org/10.1111/yrev.13402

Donovan, J. (2020). Social-media companies must flatten the curve of misinformation. *Nature (London)*. https://doi.org/10.1038/d41586-020-01107-z

Dorsey, J. [@jack]. (2020). Twitter.

Douek, E. (2021a). Governing Online Speech: From "posts-as-trumps" to proportionality and probability. *Columbia Law Review*, *121*(3), 759-834.

Douek, E. (2021b). More Content Moderation Is Not Always Better *Wired*. https://www.wired.com/story/more-content-moderation-not-always-better/

Eldridge, S. A. (2019). Where Do We Draw the Line? Interlopers. *Media and communication (Lisboa)*, *7*(4), 8. https://doi.org/10.17645/mac.v7i4.2295

Ellul, J. (1957). Information and Propaganda. *Diogenes*, *5*(18), 61-77. https://doi.org/10.1177/039219215700501805

European Commission, a. (2022). Strengthened code of practice on disinformation. https://digital-strategy.ec.europa.eu/en/library/strengthened-code-practice-disinformation-2022

Facebook. (2021). *Facebook response to the Australian disinformation and misinformation industry code*. https://digi.org.au/wp-content/uploads/2021/05/Facebook-commitments-under-disinfo-and-misinfo-code-final-report.pdf

Facebook. (n.d.). *COVID-19 and Vaccine Policy Updates & Protections*. Facebook. Retrieved 23/9/2021 from https://www.facebook.com/help/230764881494641

Flew, T., & Wilding, D. (2020). The turn to regulation in digital communication: the ACCC's digital platforms inquiry and Australian media policy. *Media, Culture & Society*, *43*(1), 48-65. https://doi.org/10.1177/0163443720926044

Franceschi-Bicchierai, L. (2022). Facebook Doesn't Know What It Does With Your Data, Or Where It Goes: Leaked Document. Retrieved 8/3/2024, from https://www.vice.com/en/article/akvmke/facebook-doesnt-know-what-it-does-with-your-data-or-where-it-goes

Frenkel, S., & Kang, C. (2021). *An Ugly Truth: Inside Facebook's Battle for Domination*. Harper.

García-Marín, D., Elías, C., & Soengas-Pérez, X. (2022). Big data and disinformation: Algorithm mapping for fact checking and artificial intelligence. In *Total Journalism: Models, Techniques and Challenges* (pp. 123-135). Springer.

Garcia, L., & Shane, T. (2021). A guide to prebunking: a promising way to inoculate against misinformation. Retrieved 19/4/2025, from https://firstdraftnews.org/articles/a-guide-to-prebunking-a-promising-way-to-inoculate-against-misinformation/

Gibbons, A., & Carson, A. (2022). What is misinformation and disinformation? Understanding multi-stakeholders' perspectives in the Asia Pacific. *Australian journal of political science*, *57*(3), 231-247. https://doi.org/10.1080/10361146.2022.2122776

Gillespie, T. (2018). *Custodians of the internet : platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gleicher , N. (2018). *Coordinated Inauthentic Behavior Explained*. Facebook. https://about.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/

Goldberg, M. (2024). What Trump Did to the G.O.P., Musk Did to Twitter. *The New York Times*. https://www.nytimes.com/2024/09/20/opinion/trump-elon-musk-twitter-x.html

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. https://doi.org/10.48550/arxiv.2301.04246

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Google. (2021). *Australia Code of Practice on Disinformation and Misinformation: Google Initial Report, May 2021*.

Google. (2024a). *Machine Learning Glossary*. Retrieved 1/12/2024 from https://developers.google.com/machine-learning/glossary

Google. (2024b). Submission on the Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2024.

Google. (n.d.). *Fact check (ClaimReview) structured data*. https://developers.google.com/search/docs/appearance/structured-data/factcheck#rating

Gorwa, R. (2024). *The Politics of Platform Regulation: How Governments Shape Online Content Moderation*. Oxford University Press. https://doi.org/10.1093/oso/9780197692851.001.0001

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 2053951719897945. https://doi.org/10.1177/2053951719897945

Graves, L., & Anderson, C. W. (2020). Discipline and promote: Building infrastructure and managing algorithms in a "structured journalism" project by professional fact-checking groups. *New media & society*, *22*(2), 342-360. https://doi.org/10.1177/1461444819856916

Grimmelmann, J. (2015). The Virtues of Moderation. *Yale Journal of Law & Technology*, *17*, 42-109.

Guess, A. M., Lyons, B. A., Tucker, J. A., & Persily, N. (2020). Misinformation, Disinformation, and Online Propaganda. In (pp. 10-33). Cambridge University Press.

Halliday, J. (2012). Twitter's Tony Wang: 'We are the free speech wing of the free speech party'. *The Guardian*. https://www.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech

Hemanus, J. (2021). From Yellow Journalism to Internet Echo Chambers – Exploring the History of "Fake News". *The Gale Review*. https://review.gale.com/2021/05/18/exploring-the-history-of-fake-news/?utm_source=rss&utm_medium=rss&utm_campaign=exploring-the-history-of-fake-news

Hermida, A., Lewis, S. C., & Zamith, R. (2014). Sourcing the Arab Spring: A Case Study of Andy Carvin's Sources on Twitter during the Tunisian and Egyptian Revolutions*. *Journal of Computer-Mediated Communication*, *19*(3), 479-499. https://doi.org/10.1111/jcc4.12074

Higgins, E. (2021). *We Are Bellingcat: An Intelligence Agency for the People*. Bloomsbury Publishing.

Ireton, C. (2018). Truth, Trust and Journalism: Why it Matters. In C. Ireton & J. Posetti (Eds.), *Journalism, fake news & disinformation: handbook for journalism education and training*. UNESCO.

Isaac, M., & Schliefer, T. (2025). Meta to End Fact-Checking Program in Shift Ahead of Trump Term. *The New York Times*. https://www.nytimes.com/2025/01/07/technology/meta-fact-checking-facebook.html

Jahangir, R. (2025). The EU's Code of Practice on Disinformation is Now Part of the Digital Services Act. What Does It Mean? *Techpolicy.Press*. Retrieved 19/4/2025, from https://www.techpolicy.press/the-eus-code-of-practice-on-disinformation-is-now-part-of-the-digital-services-act-what-does-it-mean/

Jenkins, H. W. (2021). The Hunter Biden Laptop Is Real. *Wall St Journal*. https://www.wsj.com/articles/the-hunter-biden-laptop-is-real-11625868661

Jericho, G. (2019). The Trouble with Journalism. *Meanjin*(Summer 2019). https://meanjin.com.au/essays/the-trouble-with-journalism/

Jiang, S., Baumgartner, S., Ittycheriah, A., & Yu, C. (2020, 2020). Factoring Fact-Checks: Structured Information Extraction from Fact-Checking Articles. New York, NY, USA.

Jordan, J. C., Doug, Buck, K., Gaetz, M., & Steube, G. (2020). *Reining in big tech's censorship of conservatives*.

Kleinman, Z. (2022). Twitter boss Elon Musk keeps conspiracy theorist Alex Jones off platform. *BBC News*. https://www.bbc.com/news/technology-63701423

Klonick, K. (2018). The New Governors: The People, Rules, and Processes Governing Online Speech. *Harvard law review*, *131*(6), 1599.

Konger, K. (2023). Twitter Begins Removing Check Marks From Accounts. *The New York Times*. Twitter Begins Removing Check Marks From Accounts

Kovach, B., & Rosenstiel, T. (2014). *The elements of journalism : what newspeople should know and the public should expect* (Revised and updated third edition. ed.). Three Rivers Press.

Kruger, A., Beaman, L., & Attard, M. (2021). Ready or not: A survey of Australian journalists covering mis- and disinformation during the coronavirus pandemic. *Global Media Journal Australian Edition*, *15*(1).

Leingang, R. (2025). Meta's factchecking partners brace for layoffs. *The Guardian*. https://www.theguardian.com/technology/2025/jan/08/meta-layoffs-factchecking-partners

Lewis, S. C. (2012). THE TENSION BETWEEN PROFESSIONAL CONTROL AND OPEN PARTICIPATION: Journalism and its boundaries. *Information, Communication & Society*, *15*(6), 836-866. https://doi.org/10.1080/1369118X.2012.674150

LinkedIn. (n.d.). *LinkedIn Professional Community Policies*. https://www.linkedin.com/legal/professional-community-policies

Liu, I. J. (2023). *How we're fighting misinformation across Asia Pacific* https://blog.google/around-the-globe/google-asia/how-were-fighting-misinformation-across-asia-pacific/

MacCarthy, M. (2022). Transparency is essential for effective social media regulation. Retrieved 26/6/2023, from https://www.brookings.edu/blog/techtank/2022/11/01/transparency-is-essential-for-effective-social-media-regulation/

Marantz, A. (2019). Facebook and the "free speech" excuse. *The New Yorker*. https://www.newyorker.com/news/daily-comment/facebook-and-the-free-speech-excuse

Markson, S. (2023). Australian government's plan for 'Orwellian style' misinformation laws. *Sky News*. https://www.skynews.com.au/opinion/sharri-markson/australian-governments-plan-for-orwellian-style-misinformation-laws/video/31e67ac0db7a5e43e46c31888b2cce7c

Masnick, M. (2023). The Latest Dangerous Conspiracy Theory: That Conspiracy Theory Research Is Part Of A Big Conspiracy. *Techdirt*. Retrieved 17/11/2024, from https://www.techdirt.com/2023/06/27/the-latest-dangerous-conspiracy-theory-that-conspiracy-theory-research-is-part-of-a-big-conspiracy/

McIntyre, N., Bradbury, R., & Perrigo, B. (2022). Behind TikTok's boom: A legion of traumatised, $10-a-day content moderators. *The Bureau of Investigative Journalism*. https://www.thebureauinvestigates.com/stories/2022-10-20/behind-tiktoks-boom-a-legion-of-traumatised-10-a-day-content-moderators

Meese, J. (2023). *Digital Platforms and the Press* (NED - New edition ed.). Intellect Books.

Meta. (2021). *How Meta's third-party fact-checking program works*. Retrieved 25/6/2023 from https://www.facebook.com/formedia/blog/third-party-fact-checking-how-it-works

Metz, C. (2023). 'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead. *The New York Times*. https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html

Morris, E.-J., & Fonrouge, G. (2020, 14/10/2020). Smoking-gun email reveals how Hunter Biden introduced Ukrainian businessman to VP dad. *New York Post*. https://nypost.com/2020/10/14/email-reveals-how-hunter-biden-introduced-ukrainian-biz-man-to-dad/

Moschella, D. (2022). *It's Not Just Facebook—"Old Media" Spreads Misinformation, Too*. Retrieved 2/5/2024 from https://itif.org/publications/2022/01/10/its-not-just-facebook-old-media-spreads-misinformation-too/

Musk, E. [@elonmusk]. (2022). X. Retrieved 8/3/2022 from https://twitter.com/elonmusk/status/1519036983137509376?lang=en

Myers, S. L., & Frenkel, S. (2023). G.O.P. Targets Researchers who Study Disinformation Ahead of 2024 Election. *The New York Times*. https://www.nytimes.com/2023/06/19/technology/gop-disinformation-researchers-2024-election.html

Napoli, P. M. (2019). *Social media and the public interest : media regulation in the disinformation age*. Columbia University Press.

NewsGuard. (2021). *NewsGuard Extends and Expands Licensing Agreement with Microsoft* https://www.newsguardtech.com/press/newsguard-expands-agreement-with-microsoft/

Nielsen, R. K., & Fletcher, R. (2024). *Public perspectives on trust in news* (Digital News Report 2024, Issue.

Núñez-Mussa, E., Riquelme, A., Valenzuela, S., Aldana, V., Padilla, F., Bassi, R., Campos, S., Providel, E., & Mendoza, M. (2024). The Threat of Misinformation on Journalism's Epistemology: Exploring the Gap between Journalist's and Audience's Expectations when Facing Fake Content. *Digital Journalism*. https://doi.org/10.1080/21670811.2024.2320249

Nyaricki, E. (2024). International Fact-Checking Network awards $975,000 to fact-checkers serving 34 countries. *Poynter*. Retrieved 4/9/2024, from https://www.poynter.org/ifcn/2024/international-fact-checking-network-awards-975000-to-fact-checkers-serving-34-countries/

Nyhan, B., Reifler, J., & Ubel, P. A. (2013). The Hazards of Correcting Myths About Health Care Reform. *Medical Care*, *51*(2). https://journals.lww.com/lww-medicalcare/Fulltext/2013/02000/The_Hazards_of_Correcting_Myths_About_Health_Care.2.aspx

O'Connor, C., & Weatherall, J. O. (2019). *The Misinformation Age: How False Beliefs Spread*. Yale University Press.

Orlowski, J. (2020). *The Social Dilemma* Netflix.

Orth, T., & Bialik, C. (2024). Trust in Media 2024: Which news sources Americans trust — and which they think lean left or right. Retrieved 29/12/2024, from https://today.yougov.com/politics/articles/49552-trust-in-media-2024-which-news-outlets-americans-trust

Parliament of the Commonwealth of Australia, a. (2024). Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2024 Explanatory Memorandum. https://parlinfo.aph.gov.au/parlInfo/download/legislation/ems/r7239_ems_13b01a0b-4684-4e0e-b336-0028d4c0e3cd/upload_pdf/JC014003.pdf;fileType=application%2Fpdf

Paul, C., & Matthews, M. (2016). *The Russian "Firehose of Falsehood" Propaganda Model: Why It Might Work and Options to Counter It*. RAND Corporation. https://doi.org/10.7249/PE198

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science*, *31*(7), 770-780. https://doi.org/10.1177/0956797620939054

Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, *116*(7), 2521. https://doi.org/10.1073/pnas.1806781116

Picchi, A. (2022). Twitter Files: What they are and why they matter. *CBS News*. https://www.cbsnews.com/news/twitter-files-matt-taibbi-bari-weiss-michael-shellenberger-elon-musk/

Porter, J. (2021). Twitter bans 70,000 QAnon accounts as conservatives report lost followers. *The Verge*. https://www.theverge.com/2021/1/12/22226503/twitter-qanon-account-suspension-70000-capitol-riots

Posetti, J. (2018). News Industry Transformation: Digital Technology, Social Platforms and the Spread

of Misinformation and Disinformation. In C. Ireton & J. Posetti (Eds.), *Journalism, fake news & disinformation: handbook for journalism education and training*. UNESCO.

Questions and Answers: Digital Services Act. (2023). European Commission. Retrieved 25/6/2023 from https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2348

Rauch, J. (2021). *The Constitution of Knowledge*. Brookings Institution Press.

Roberts, S. (2021). *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.

Robison, K. (2024). Inside the shifting plan at Elon Musk's X to build a new team and police a platform 'so toxic it's almost unrecognizable'. *Fortune*. https://fortune.com/2024/02/06/inside-elon-musk-x-twitter-austin-content-moderation/

Roitero, K., Soprano, M., Portelli, B., De Luise, M., Spina, D., Mea, V. D., Serra, G., Mizzaro, S., & Demartini, G. (2023). Can the crowd judge truthfulness? A longitudinal study on recent misinformation about COVID-19. *Personal and ubiquitous computing*, *27*(1), 59-89. https://doi.org/10.1007/s00779-021-01604-6

Santa Clara Open Consultation Report. (2021).

Santos, F. C. C. (2023). Artificial Intelligence in Automated Detection of Disinformation: A Thematic Analysis. *Journalism and Media*, *4*(2), 679-687.

Satariano, A., & Isaac, M. (2021). The Silent Partner Cleaning Up Facebook for $500 Million a Year. *The New York Times*. https://www.nytimes.com/2021/08/31/technology/facebook-accenture-content-moderation.html

Schatto-Eckrodt, T., Boberg, S., Wintterlin, F., & Quandt, T. (2020, 2020//). Use and Assessment of Sources in Conspiracy Theorists' Communities. Disinformation in Open Online Media, Cham.

Schmitt, V., Villa-Arenas, L.-F., Feldhus, N., Meyer, J., Spang, R. P., & Möller, S. (2024, 2024). The Role of Explainability in Collaborative Human-AI Disinformation Detection. New York, NY, USA.

Shabani, S., & Sokhn, M. (2018). Hybrid Machine-Crowd Approach for Fake News Detection.

Shane, T. (2020). *The difference between the facts and the truth*. First Draft. Retrieved 27/2/2022 from https://medium.com/1st-draft/the-difference-between-the-facts-and-the-truth-59e23c6185d

Shapiro, I. (2024). *The Disputed Freedoms of a Disrupted Press* (1 ed., Vol. 1). Routledge. https://doi.org/10.4324/9781003223146

Shapiro, I., Brin, C., Bédard-Brûlé, I., & Mychajlowycz, K. (2013). Verification as a Strategic Ritual. *Journalism Practice*, *7*(6), 657-673. https://doi.org/10.1080/17512786.2013.765638

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. Newsl.*, *19*(1), 22–36. https://doi.org/10.1145/3137597.3137600

Silva, C. (2024). Content moderation in Trump's America is a political minefield. Retrieved 28/12/2024, from https://mashable.com/article/content-moderation-donald-trump

Sitek, P., Pietranik, M., Krótkiewicz, M., & Srinilta, C. (2020). Fake News Types and Detection Models on Social Media A State-of-the-Art Survey. In (Vol. 1178, pp. 562-573). Springer Singapore Pte. Limited. https://doi.org/10.1007/978-981-15-3380-8_49

Snapchat. (n.d.). *Community Guidelines*. https://www.snap.com/en-US/community-guidelines

Soprano, M., Roitero, K., La Barbera, D., Ceolin, D., Spina, D., Mizzaro, S., & Demartini, G. (2021). The many dimensions of truthfulness: Crowdsourcing misinformation assessments on a multidimensional scale. *Information processing & management*, *58*(6), 102710. https://doi.org/10.1016/j.ipm.2021.102710

Spitale, G., Biller-Andorno, N., & Germani, F. (2023). AI model GPT-3 (dis)informs us better than humans. *Science Advances*, *9*(26), eadh1850. https://doi.org/10.1126/sciadv.adh1850

Stocking, G., Wang, L., Lipka, M., Matsa, K. E., Widjaya, R., Tomasik, E., & Liedke, J. (2024). *America's News Influencers*.

Stokel-Walker, C. (2025). What does Meta's Oversight Board even do? *Fast Company*. https://www.fastcompany.com/91257773/whats-even-the-point-of-metas-oversight-board

Sunderland, A. (2019). A Partiality For The Truth. *Meanjin*. https://meanjin.com.au/blog/a-partiality-for-the-truth/

Suzor, N. (2019). *Lawless: The Secret Rules That Govern Our Digital Lives*. https://doi.org/10.1017/9781108666428

Taylor, J. (2023). ChatGPT's alter ego, Dan: users jailbreak AI program to get around ethical safeguards. *The Guardian*. https://www.theguardian.com/technology/2023/mar/08/chatgpt-alter-ego-dan-users-jailbreak-ai-program-to-get-around-ethical-safeguards

Taylor, S. (2013). *What is discourse analysis?* Bloomsbury Acadmemic.

The Oversight Board. (2024). Content Moderation in a New Era for AI and Automation. Retrieved 18/4/2025, from https://www.oversightboard.com/news/content-moderation-in-a-new-era-for-ai-and-automation/#:~:text=New%20generative%20AI%20models%20present,explain%20enforcement%20actions%20to%20users.

TikTok. (2020). *Community Guidelines*. https://www.tiktok.com/community-guidelines?lang=en#37

TikTok. (2024). Our response to the Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2024 [Provisions].

TikTok. (n.d.). *Combating Harmful Misinformation*. Retrieved 4/9/2024 from https://www.tiktok.com/transparency/en-us/combating-misinformation/

Tucker, J., Guess, A., Barbera, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., & Nyhan, B. (2018). Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3144139

Tutui, V. (2017). Some Reflections Concerning the Problem of Defining Propaganda. *Argumentum (Iași, Romania)*, *15*(2), 110-125.

Twitter. (n.d.). *COVID-19 misleading information policy*. Twitter. Retrieved 23/9/2021 from https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy

Urquhart, C. (2013). Grounded Theory for Qualitative Research: A Practical Guide. In. https://doi.org/10.4135/9781526402196

US Congress. (2023). The weaponization of "disinformation" pseudo-experts and bureaucrats : how the federal government partnered with universities to censor Americans' political speech : interim staff report of the Committee on the Judiciary and the Select Subcommittee on the Weaponization of the Federal Government, U.S. House of Representatives. In: Committee on the Judiciary. Select Subcommittee on the Weaponization of the Federal Government.

van Assche, K., Beunen, R., Duineveld, M., & Gruezmacher, M. (2023). Adaptive methodology. Topic, theory, method and data in ongoing conversation. *International Journal of Social Research Methodology*, *26*(1), 35-49. https://doi.org/10.1080/13645579.2021.1964858

Vizoso, A., Vaz-Alvarez, M., & Lopez-Garcia, X. (2021). Fighting deepfakes: Media and internet giants' converging and diverging strategies against hi-tech misinformation. *Media and communication (Lisboa)*, *9*(1), 291-300. https://doi.org/10.17645/MAC.V9I1.3494

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146-1151. https://doi.org/10.1126/science.aap9559

Vu, H. T., & Saldaña, M. (2021). Chillin' Effects of Fake News: Changes in Practices Related to Accountability and Transparency in American Newsrooms Under the Influence of Misinformation and Accusations Against the News Media. *Journalism & Mass Communication Quarterly*, *98*(3), 769-789. https://doi.org/10.1177/1077699020984781

Wardle, C. (2014). Verifying User-Generated Content. In C. Silverman (Ed.), *Verification Handbook: An Ultimate Guideline on Digital Age Sourcing for Emergency Coverage* (pp. 25-32). European Journalism Centre.

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policymaking* (Council of Europe Report, Issue.

Weiss, B. [@bariweiss]. (2022, 9/12/2022). *7. What many people call "shadow banning," Twitter executives and employees call "Visibility Filtering" or "VF." Multiple high-level sources confirmed its meaning.*Twitter.

West, D. M., & White, N. (2024). Meta's Oversight Board is unprepared for a historic 2024 election cycle. *Brookings*. https://www.brookings.edu/articles/metas-oversight-board-is-unprepared-for-a-historic-2024-election-cycle/

Woolley, S. C. (2020). Bots and Computational Propaganda: Automation for Communication and Control. In N. Persily & J. A. Tucker (Eds.), *Social Media and Democracy* (pp. 89-110). Cambridge University Press. https://www.cambridge.org/core/product/A15EE25C278B442EF00199AA660BFADD

Yameogo, S. A., Araújo, J., Santos, M. Y., Assar, S., de la Vara, J. L., Assar, S., de la Vara, J. L., Santos, M. Y., Araújo, J., Araujo, J., Assar, S., DeLaVara, J. L., & Santos, M. Y. (2024). Improving Understanding of Misinformation Campaigns with a Two-Stage Methodology Using Semantic Analysis of Fake News. In (Vol. 514, pp. 121-130). Springer. https://doi.org/10.1007/978-3-031-59468-7_14

YouTube. (n.d.). *COVID-19 medical misinformation policy*. https://support.google.com/youtube/answer/9891785?hl=en

Zeng, F., Li, W., Gao, W., & Pang, Y. (2024). Multimodal Misinformation Detection by Learning from Synthetic Data with Multimodal LLMs. https://doi.org/10.48550/arxiv.2409.19656

Zeng, J., & Brennen, S. B. (2023). The Misinformation.

Zeng, X., Abumansour, A. S., & Zubiaga, A. (2021). Automated fact - checking: A survey. *Language and linguistics compass*, *15*(10), n/a. https://doi.org/10.1111/lnc3.12438

Zhang, S. (2021). *The unproven lab leak theory, Wuhan lab and virus origin: Reporting best practices*. First Draft. Retrieved 27/2/2022 from https://firstdraftnews.org/articles/best-practices-for-reporting-on-the-wuhan-lab-leak-theory/

Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., & Choudhury, M. D. (2023). *Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions* Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany. https://doi.org/10.1145/3544548.3581318

Zollmann, F. (2017). Bringing Propaganda Back into News Media Studies. *Critical Sociology*, *45*(3), 329-345. https://doi.org/10.1177/0896920517731134