Philosophy & Technology (2021) 34:923–940
https://doi.org/10.1007/s13347-021-00442-x

RESEARCH ARTICLE

Detecting Fake News: Two Problems for Content Moderation

Elizabeth Stewart[1]

Received: 29 May 2020 / Accepted: 18 January 2021 / Published online: 11 February 2021

Abstract

The spread of fake news online has far reaching implications for the lives of people offline. There is increasing pressure for content sharing platforms to intervene and mitigate the spread of fake news, but intervention spawns accusations of biased censorship. The tension between fair moderation and censorship highlights two related problems that arise in flagging online content as fake or legitimate: firstly, what kind of content counts as a problem such that it should be flagged, and secondly, is it practically and theoretically possible to gather and label instances of such content in an unbiased manner? In this paper, I argue that answering either question involves making value judgements that can generate user distrust toward fact checking efforts.

Keywords Social media · Content moderation · Fake news

1 Introduction

In our current information age with its abundance of freely available content, there is an increasing problem with people forming erroneous, and often harmful, beliefs after encountering "fake news" online. These impacts have led to calls for tech companies to do more to monitor content shared on their platforms. As social media platforms have responded to these calls, they have faced further criticism that their content moderation efforts amount to censorship. Given that the goal of content moderation is to decide what content should and shouldn't be made widely available, thereby influencing what people do and do not believe, there must be a concerted effort to ensure that what moderators flag as "fake" content is actually fake. However, accurately identifying fake news requires that interested parties agree on what makes fake content problematic, such that it merits removal, and that content moderators can reliably distinguish this content from non-problematic content. There are thus

✉ Elizabeth Stewart
eks2@email.sc.edu

[1] Department of Philosophy, University of South Carolina, Columbia, SC, USA

two sources of disagreement related to detecting fake news: (1) disagreement regarding what content should be subject to moderation and (2) disagreement regarding whether that content is categorized accurately.

These sources of disagreement represent two distinct challenges for the task of successfully detecting and flagging or removing fake news from online platforms while maintaining user trust. One is a policy challenge requiring online platforms, platform users, and government policymakers to determine which kinds of content a platform's moderation efforts should or should not target. The other is a labeling challenge requiring that content moderators accurately identify instances of targeted content. The purpose of this paper is twofold. Firstly, I argue that the process of resolving these disagreements is necessarily biased in that it requires prioritizing one set of values before others. While the inability to proceed with neutrality does not mean that platforms should cease all content moderation, it does mean that we should proceed with caution because content moderation may be counter-productive unless users trust the process. Thus, the second purpose of this paper is to better understand how these two sources of disagreement generate different challenges to user trust. I do not herein argue that one resolution of these disagreements is better than another; instead, I suggest that content moderation is not a silver bullet for our fake news problem. However, there are several steps that platforms can take that will facilitate user trust in the content moderation process.

In this paper, I focus only on moderation aimed at addressing informational content such as news articles or social media posts that make truth claims, not moderation aimed at other types of problematic content such as violent content, terrorism-related content, or child pornography. I first outline three different kinds of problematic content that platforms and regulatory bodies might variously identify as the appropriate target of moderation efforts and identify how disagreement regarding this target contributes to user distrust. Secondly, I argue that even if this policy challenge is met through reaching agreement regarding the appropriate target of moderation efforts, fact checkers cannot identify instances of this content without making value judgements. Fact checkers face both easy problems and hard problems when moderating content.[1] Easy problems may be theoretically simple to resolve, although perhaps practically difficult. Issues such as verifying whether an event occurred or whether an individual said what was attributed them are instances of such "easy" problems. However, checking the facts is not always as straightforward as verifying whether a quote is reported correctly. Hard problems arise when facts involve thick moral concepts, are partially true or partially misleading, or when experts disagree over the truth of a given fact. Resolving these hard problems requires that fact checkers make value judgements that result in biased outcomes. Finally, I suggest several measures that might facilitate user trust in content moderation efforts.

[1] My use of the easy/hard problem terminology is meant only to echo Chalmers' (1996) distinction between two different kinds of problems that arise in studying consciousness. The problems facing fact checkers are hard for different reasons than the problems Chalmers' discusses.

2 The Policy Challenge: What Content Should Moderation Target?

Using information theory as a framework, we can identify three broad categories of content that might be suitable candidates for content moderation. In information theory, an information system is comprised of some *target system* or information source in the world, a *transmitter* that represents the target system as *message* and conveys it via a *channel* to a *receiver* (Shannon 1948). In the context of news, the target system is the world, the transmitter is some journalist or content writer, and the message is an article conveyed via a website or social media to users of social media. There are three interfaces at which three distinct kinds of problematic content arise: *misinformation* arises at the interface between the world and the journalist, *disinformation* between the journalist and the content they produce, and *misleading content* between an article and its recipients.

Misinformation arises when a journalist does not accurately understand the target system that they are writing about. This may happen because of an honest journalistic mistake or because they simply don't care about accuracy, as in the case of those who write fake news content for a living. Misinformation is characterized by a lack of veracity, that is, the message does not accurately convey information about the target system. Disinformation, on the other hand, is characterized by the author's intentions and arises when a journalist intends their message to deceive or mislead their audience (Wardle and Derakhshan 2018). Such a journalist passes along information that they know is false or perhaps is literally true but implies something false. Disinformation can include deception regarding both content and authorial status. That is, a purveyor of disinformation can deceive an audience about the veracity of a news item as well as deceiving an audience about their own status as a journalist. They can, for example, present content as though it were the product of genuine journalistic practices when, in fact, it was not (Fallis and Mathiesen 2019; Pepp et al. 2019). Finally, misleading content is characterized by the effect that it has on the audience. Content can mislead an audience for a number of reasons. An audience may misunderstand an article if it is too vague or leaves out key information. Alternatively, an audience may misunderstand an article if they fail to properly understand the author's intentions. Satirical writing, for example, often misleads readers who fail to understand that the intent was to entertain rather than inform (Garrett et al. 2019). Additionally, content can mislead when people treat it as though it were news produced through traditional journalistic methods when, in fact, it was not (Pepp et al. 2019). There are obvious overlaps between these different categories. For example, much disinformation is also misleading, although it need not be as a journalist may intend to deceive their audience, but fail. However, while overlapping, each is characterized by a different problem which might be a potential target for fact checkers.

Among the various platforms and websites that fact check content, there is little consistency in what problems they wish to address through fact checking. Snopes, as a fact checking website, does not value free expression in the same way that Facebook or Twitter, as social media platforms, does. Instead, Snopes values identifying misleading content, regardless of whether it is the result of misinformation, disinformation, or simply a misunderstanding. Thus, Snopes investigates a much wider range of content when checking facts than Facebook or Twitter, addressing content that might mislead an audience, regardless of whether the content was intended to do so. Facebook, on the other hand, doesn't want to flag content such as opinion pieces as fake, although opinion articles are often misleading, but it does want to flag disinformation and unintentional misinformation. Twitter takes a more targeted approach, flagging tweets that it deems harmful. This includes tweets that threaten physical, psychological, or informational harms. These informational harms include misinformation or disinformation that threatens public health or civic engagement, such as election information.

The decisions regarding which problems to address are not without controversy. Facebook's decision to exempt satire, political ads, and opinion pieces from fact checking drew criticism from those concerned about how these kinds of content often mislead audiences (Owen 2019). Critics also argued that the decision to exempt political ads and opinion pieces created a loophole for certain kinds of disinformation (Horwitz 2019).

My purpose here is not to argue for which problem platforms should prioritize. Rather, I wish to point out that the way in which the question is resolved has important consequences for whether users will trust the results. Without a uniform understanding of what counts as problematic content, i.e., "fake news," across platforms, content moderation can generate user distrust in two ways. Firstly, if a platform adopts a narrow focus on fact checking when users expect a broad focus, users are likely to develop false assurances, which when disappointed can generate distrust. Suppose the platform's goal is to flag only false content, or misinformation, but the user thinks that the platform flags true disinformation as well. If the content moderation works as the platform intends, content that is perhaps literally true, but is intentionally misleading will not be flagged as fake. If the users, however, think that the platform flags all kinds of problematic content, then they may assume that whatever content they see is straightforwardly true and thus are susceptible to attempts to mislead them. This mismatch is doubly harmful. On the one hand, users are likely to develop false beliefs if they rely on the platform to flag true disinformation when it doesn't. On the other hand, if they realize that the content they are viewing is intentionally misleading, they are likely to distrust the platform's content moderation efforts, viewing them as ineffective or biased.

Secondly, if a platform adopts a broader focus regarding fact checking than users think appropriate, users may find the resulting content moderation patronizing or consider it as unethical censorship leading to distrust. Over the course of the 2020 US election cycle and the Covid-19 pandemic, both Twitter and Facebook have taken an increasingly proactive and targeted approach to misinformation. Twitter has widened the scope of what counts as an informational harm to include medical misinformation and disputed election information. In addition to targeting Covid-19 and election misinformation for fact checking purposes, Facebook has set up dedicated resource centers for users to access reliable information and has set alerts regarding updated information at the top of users' newsfeeds.

While this proactive approach has been celebrated by many, it has also drawn criticism. Users eventually tired of Facebook's daily notifications regarding voter registration information in the November 2020 election, which sparked a number of memes mocking what users found as patronizing reminders to engage in their civic duties. The proactive approach has also drawn accusations of censorship and has generated user distrust toward the platforms' content moderation efforts. This distrust has prompted many users to move to alternative platforms which they perceive as more friendly to free speech. In the week following the 2020 election, Parler, a social media platform that casts itself as a platform for free expression, gained so many new users that it had technical problems accommodating the sudden influx. Parler does not fact check any posts and only removes posts that it is legally required to, such as child pornography, copyright violations, and content promoting terrorism. It also removes content "in order to prevent our services from being used by someone in the commission of a crime or civil tort, especially when these interfere with our mission of providing a welcoming, nonpartisan Public Square" (Parler 2020).

In order to address these sources of distrust, platforms and government regulatory bodies must reach agreement regarding what kinds of content, if any, social media platforms are responsible for moderating. Currently, the USA has relatively little regulatory guidance for online content moderation. While the USA has free speech protections encoded in its Constitution, these protections do not extend to the private sphere where online platforms reside. Online platforms are granted further discretion in content moderation in section 230 of the 1996 Communications Decency Act, which protects online platforms from legal liability for user generated content and allows platforms to remove content if that removal is done in good faith. Criticism regarding Section 230 is mixed. Some want to see it changed because they believe that it fosters the spread of problematic content because it protects platforms from the legal ramifications of shared content. Others, however, believe that it fosters partisan censorship because platforms have wide ranging control over the kinds of problematic content they choose to address. Accusations of censorship triggered a Senate Judiciary meeting with the heads of Twitter, Facebook, and Google regarding Section 230 and content moderation.[2] Given this mixed criticism of Section 230, it is unclear whether current platform practices do not adequately address problematic content or whether their efforts go too far and violate users' free expression. Without stakeholders first reaching agreement on what counts as problematic content, platforms are unable to develop moderation policies that address stakeholders' concerns. In order to address these concerns, legislative bodies and platforms must reach agreement regarding what kinds of content platforms are legally obliged to address and what they are legally prevented from addressing. This would provide a benchmark against which platforms, users, and policymakers could evaluate content moderation efforts.

The EU has taken several steps toward identifying the appropriate target of moderation efforts and the various responsibilities of relevant stakeholders. The High Level Expert Group (HLEG) on fake news and online disinformation specify that the target of their efforts is disinformation defined as "false, inaccurate, or misleading information designed, presented, and promoted to intentionally cause public harm or for profit" (High level expert group on fake news and online disinformation 2018, p. 10). With this definition in mind, they suggest five key areas for intervention

[2] https://www.c-span.org/video/?476686-1/social-media-content-moderation Date of Access:11/09/2020

to target disinformation: (a) increasing transparency, (b) promoting media literacy, (c) empowering users and journalists, (d) safeguarding diversity and sustainability of news media, and (e) continuing research on disinformation and the efficacy of interventions. With respect to the responsibilities of platforms, they do not recommend actions that amount to censorship. However, they do suggest that platforms must collaborate with independent fact checking organizations, prioritize trustworthy information in ranking algorithms, and link users to trustworthy content where appropriate, especially in cases of trending news items. In a July 2020 follow-up report on the Code of Practice on Disinformation, the efforts of Facebook, Twitter, and other large companies to label items as “false” or “misleading” are noted as positive achievements (European Commission 2020).

Even though many large platforms, including Facebook, have voluntarily accepted the Code of Practice, disinformation remains an issue (AVAAZ 2020) and accusations of censorship persist. Despite agreement on the target of content moderation, identifying instances of disinformation is incredibly challenging and the results are not uncontroversial. This challenge is due to disagreement over what counts as information versus disinformation. In order to prioritize trustworthy information, platforms must determine what counts as trustworthy and who is qualified to identify it as such. Even if all stakeholders agree on what kind of content poses a problem, the challenge of determining whether a particular news item counts as an instance of that problem remains.

3 The Labeling Challenge: What Tasks Are Involved in Categorizing Content?

In order to effectively address all of the various problems under the fake news umbrella, fact checkers must reliably distinguish between true content (information), false content (misinformation), intentionally misleading content (disinformation), and actually misleading content. Distinguishing information from misinformation requires identifying the veracity of the content. Distinguishing disinformation from content that is not intended to mislead requires understanding the author’s intentions. Finally, distinguishing misleading content from non-misleading content requires determining the truth value of the content and whether a typical person is likely to believe something other than the truth after reading the article. Sorting content into these categories might be ideal; however, determining the veracity of content, understanding authorial intentions, and determining whether a news item is likely to mislead the average reader are not straightforward tasks. Unfortunately, humans import their personal values in a variety of ways throughout the fact checking process, thus compromising their assessments of veracity, authorial intent, and an item’s potential for misleading an audience. There are aspects to these tasks that are theoretically easy to resolve, although sometimes practically difficult. I call these easy problems, whereas hard problems are problems that theoretically challenging to resolve. While both kinds of problems pose a serious challenge to maintaining user trust in content moderation, hard problems may be insurmountable.

3.1 Determining Veracity

It is difficult to determine the veracity of a statement due to limited access to the reported events. In order to accurately flag fake news and thereby foster user trust, fact checkers must “get their facts straight.” This requires two things: (1) that a fact checker checks verifiable claims, which I shall call “facts” and (2) that a fact checker verifies the claim correctly. There are a number of problems involved in verifying facts, some of which are easy to resolve while others are hard to resolve. Easy problems involved in verifying facts include the need to travel to far away places or find reliable footage in order to observe events. Verifying facts might also involve gaining specialized training in a certain area or performing experiments to validate another person’s findings. While these practical problems certainly limit what a fact checker can accomplish, they are theoretically straightforward problems to address. Given enough time and resources, a fact checker could travel to the necessary places or receive whatever training is required to verify a claim. Unfortunately, however, even if these practical problems were resolved, hard problems still remain.

Consider the following article titles, published leading up to the 2020 Democratic primaries, all from sources that claim to tell the truth. Potential candidate Beto O’Rourke proposed a policy changing what entities would be granted tax-exempt status. The National Review published an article regarding this policy entitled “Beto Proposes To Oppress Church With State” (Editors 2019). ABC News, covering the same story, published an article entitled “Beto O’Rourke said he would revoke tax-exempt status from religious organizations that oppose same-sex marriage” (Cook 2019). Meanwhile, LGBTQ Nation reported, “Beto O’Rourke is ahead of his time when it comes to anti-LGBTQ religious institutions” (Bollinger 2019).

There are two ways that someone could take issue with these articles. The first involves disputing the facticity of these articles; arguing instead that these are opinions or interpretations of events and thus are not the kinds of things that have truth values that should be fact checked. Politifact chooses to fact check claims that are verifiable, highlighting the importance of distinguishing factual claims from other kinds of claims, such as opinions. However, distinguishing factual claims from non-factual claims is not always straightforward, especially when evaluating questionable news items. Another person might very well argue that these are not mere opinions, but these claims describe certain facts. What counts as opinion and what counts as a factual claim is not entirely clear. In a case that coincided with Facebook adjusting its fact checking policy, two scientists published an opinion article challenging the reliability of computational models of climate change which was flagged by Facebook’s fact checkers as false news (Michaels and Rossiter 2019). The authors challenged this decision, which resulted in the removal of the false news label on the basis that it was an opinion article (Horwitz 2019). However, while technically an opinion piece, the authors clearly intended that their readers accept as true certain claims which are rejected by other scientists (Dessler et al. 2019).

A second issue involves individuals who agree that the articles express facts, but disagree on the truth values of those facts. There are several different forms of disagreement that are particularly problematic for fact checkers. One form involves disagreements regarding morally or ethically thick concepts. We might imagine one person who, upon reading the first article about Beto's tax policy, nods their head in agreement and upon reading the second article argues that it is a half-truth. They might argue that it leaves out the fact that revoking tax exempt status oppresses churches. This person responds to the final article with a cry of "Fake news!! Beto is *not* ahead of his time with respect to religious institutions; in fact, he doesn't understand the position of religious institutions at all." We might just as easily imagine two other individuals who come to very different conclusions about these articles. One might scoff at the claim regarding the oppression of churches and write it off as fake news. Another might scoff at both the claim that Beto is ahead of his time and the claim about church oppression. The point here is that whether these claims count as verifiable facts as opposed to opinions, as well as whether the article is labeled as false or misleading, involves whether the individual agrees with the values expressed in the article. Disagreements on whether it is factually true that Beto intends to oppress churches that don't endorse same sex marriage involve disagreements over what constitutes oppression, the correct relationship between the church and the state, the extent of civil liberties, and what rights those liberties guarantee for citizens.

This highlights a hard problem with determining the veracity of news articles: more evidence is not always sufficient to resolve disagreements. We might wish that the news would "just state the facts." However, factual claims expressed in the news often involve "thick ethical concepts," which have a descriptive use and a normative use, that are often entangled. Words like "cruel," "brave," "just," or, in our examples, "oppress" and "ahead of the times" have such entangled uses. They are entangled in the sense that when someone uses them to describe a situation, they also express a value judgement regarding it and vice versa. Hilary Putnam describes a thick ethical concept as one which "simply ignores the supposed fact/value dichotomy and cheerfully allows itself to be used sometimes for a normative purpose and sometimes as a descriptive term" (Putnam 2002, p. 35). In expressing that Beto intends to oppress the church or that he is ahead of his time, the journalists use language both descriptively and normatively. While some might wish that journalists would avoid normative language altogether, descriptions of the world often presuppose a set of values such that describing events is impossible without relying on those presupposed values. For example, even the seemingly straightforward and innocuous phrasing in the second article regarding "same-sex marriage" relies on a value judgement about the nature of marriage, namely that it can exist between members of the same sex. When thick ethical concepts are used in the news, all the conflicting parties may consider themselves as stating true facts. However, their assessment of the truth of those facts depends not on the observable events that occurred but on how those events are interpreted and evaluated.

Another form of disagreement that is particularly relevant when considering how to label data involves disagreement among experts. Consider groups critical of climate change or vaccines. The majority of scientists and the public agree that climate change is happening due to human activity and that vaccines do improve public health. However, there are people, including scientists and other experts, who are critical of current research on these topics and it is important that their voices not be discounted on the basis of their dissent alone. This problem is exacerbated when experts disagree over emerging events, such as the novel coronavirus SARS CoV-2 and the associated Covid-19 disease. In such emerging events, the problem is less about whether experts disagree over the veracity of a claim and more about whether the methods used to arrive at that claim were methodologically and statistically sound. While experts themselves may be unsure of what is true or false in emerging events, they do know legitimate ways to find the answers to their questions. Non-experts, however, are disadvantaged because they lack the relevant specialized scientific or statistical expertise to evaluate the ways in which experts arrive at their claims. Determining what makes someone credible is often impossible for those who are not domain experts, and even within a community of domain experts, there is disagreement on who is credible, who is not, and what differentiates the two. Thus, an individual labeling content regarding such issues, who is often not a domain expert, must be careful not to label content critical of these as misinformation or disinformation on the basis of its disagreement with majority opinion alone. Not only does this risk silencing critical voices that have legitimate concerns or constructive criticism, but, in doing so, it risks undermining the credibility that the project of fact checking is seeking to establish.

While "just stating the facts" sounds like a worthy goal, it is impossible to do so without importing one's values. A person's values are involved in both evaluating whether content counts as a verifiable fact and, if it does count, whether that fact is true or false. Furthermore, an individual's values are involved in determining who to trust when experts disagree over the truth of a claim. Thus, any effort to detect fake news will reflect the values of fact checkers in both the content subject to checking and the label applied. Users who do not share those values are thus likely to disagree with the results, regardless of what additional evidence is offered in support of the fact checkers' claims.

3.2 Determining Authorial Intent

Distinguishing disinformation from information and misinformation requires understanding what an author intended regarding both the content of their writing and the response of their audience. A writer of satire might produce similar content as a writer of false disinformation, but the satirical piece is considered socially acceptable and the other is not because in the first case the author's intentions are to entertain or humor whereas in the second author intends to deceive or manipulate. More difficult to distinguish, perhaps, is the author of true disinformation from the author of information not intended to mislead or deceive.

Consider the following set of claims made in one of Donald Trump's political ads. The ad claimed (1) that Joe Biden pressured Ukraine to fire its prosecutor, (2) the prosecutor was investigating a company Biden's son was involved with, (3) Democrats wanted to impeach Trump for discussing these facts with the President of Ukraine, and finally (4) that Democrats wanted to overturn Trump's fairly won election. While the claims appear at least mostly true, the ad implied at least two additional, unspoken, claims: (5) that Biden pressured Ukraine to fire its prosecutor in order to protect his son and (6) that the Democrats' impeachment efforts were unfair. These latter two claims, however, are questionable. If the author genuinely believed that everything spoken and implied is true and only intended that the audience gain information, then the ad would not be a case of disinformation, even if the implied content is actually false. However, if the author intended that the audience accept as true something that they themselves believe is either false or has an unknown truth value, then the ad would be a case of disinformation.

There are several ways that an author's intentions may be obvious to a reader. The website that publishes an author's work often reveals the author's intentions regarding both the content and reader. For example, publishing on a known satirical website makes it obvious that their intent is not to try to persuade people to believe the content of their work. If a journalist works for a reliable news source, however, then it is likely that the journalist intends that their audience should accept the content of their article as true. Additionally, authors may make their intentions obvious through the use of evidential discourse markers. Evidentials provide the reader information regarding the speaker's commitment to the content they are sharing (Ifantidou 1994). When an author includes phrases such as "I think," "perhaps," or "maybe," they alert their readers that they themselves are not certain of the information that they are sharing and thus that readers should not take the content as certainly true.

However, while the website on which an article appears and the use of distancing language often provide a clear indication of an author's intentions for their audience, such as entertaining or informing their audience, it does not help identify cases of disinformation, where the author intentionally hides their true intentions. These cases represent a hard problem because deceptive authors frequently exploit their target audiences' trust networks in such a way that accusations of intentional deception fall on deaf ears. Suppose a journalist for a traditional, non-satirical news outlet publishes misinformation. It is difficult to assess whether or not the journalist intended to include false content or whether it was an accidental reporting error. If a fact checker flags their article as disinformation, a journalist could either deny that their report was false or reply that they didn't *intend* to pass on misinformation; they simply got the facts wrong. In both cases, their target audience may believe the journalist over a fact check label due to confirmation bias, partisanship, or other cognitive biases (Gelfert 2018).

While an author may claim that including misinformation was merely accidental or not intended, perhaps their writing might reveal otherwise. There is some research that has used machine learning to try to find stylistic differences between fake and legitimate news. However, this research relies on datasets of fake news labeled by some person or group of people, which leads back to the initial problem of identifying fake news (Horne and Adali 2017). Rather than analyzing textual features of fake news in general, we might instead focus on features of deceptive text in order to identify disinformation. Humans are, in general, not particularly good at detecting deception, especially without cues such as speaker prosody and body language (Rubin and Conroy 2012). There is research that suggests that deceptive text expresses distinctive patterns that a machine can identify, such as evasive, unclear, or impersonal language (Zhou et al. 2004; Zhou and Zhang 2008).

However, it is unclear that the features that distinguish deceptive texts from truthful texts remain stable across different situations. For example, some research suggests that deceptive texts include more words indicative of negative emotions, while other research suggests the opposite (Ali and Levine 2008). Thus, it isn't clear that using textual analysis will reliably identify authors who intend to deceive their audiences, especially when the situations in which the deception arises differ. For example, deception research often uses either data from experiments where participants are asked to lie to each other or real-world data such as transcriptions of criminal interrogations. It isn't clear that the kind of deception involved in creating fake news would exhibit the same kind of linguistic features as the high stakes kind of deception displayed by criminals under interrogation or low stakes deception displayed by experiment participants in a contrived situation in a research lab. Journalists spreading disinformation have more time to craft their articles than individuals in real-time dialogue. Additionally, while such journalists are not at risk of being convicted of crime, they are also not simply playing a game in a lab; they have incentive to create convincing content, but they won't go to jail if they fail.

Given the difference in stakes and time to prepare, it is likely that the linguistic cues that detect deception in other areas may not generalize to fake news. Even narrowing the scope of deceivers from criminals or experiment participants to journalists does not guarantee that various authors find themselves in similar situations such that their writing will exhibit the same deceptive features. Some authors may be writing deceptive content because they are personally invested in deceiving their audience, others may be simply trying to sell a story or fulfill an editor's request. For example, authors intending to manipulate their audience into believing a particular false proposition may write differently from authors generating fake news in order to generate revenue from ads. Both generate disinformation but have different motivations and different levels of personal investment regarding the deception of their audience (Gelfert 2018).

Without a reliable method of determining an author's intentions when they haven't made their intentions obvious, distinguishing disinformation from less nefarious forms of information and misinformation is a real challenge. Humans are generally bad at detecting deception and statistical methods are, as yet, unreliable. Even if the statistical methods were reliable, such methods rely on human efforts to identify fake news and are thus subject to the same problem that they are trying to solve. At some point, detecting disinformation bottoms out in some person or group of persons making a judgement regarding whether they think the journalist has good intentions, such as informing or entertaining their audience, or bad intentions, such as deceiving or misleading their audience. When faced with such a dispute, platform users predisposed to trusting the journalist or idea posed in the article may not be dissuaded by a fact check label.

3.3 Determining the Potential for Misleading an Audience

As determining the veracity and authorial intent of an article is challenging, we might wonder whether we can dispense altogether with the tripartite differentiation of news into information, misinformation, and disinformation which requires these features. Is it possible to identify fake news on some basis that does not rely on veracity and authorial intent? Pepp et. al. (2019) propose a definition of fake news in an effort to do just this. They argue that fake news is content that is treated as though it were produced through standard journalistic methods when, in fact, it was not. This locates the "fakeness" of content in its status as a news article.

While this definition would not require content moderators to determine truth values or read minds, there are several reasons why this approach fails to adequately guide content moderation decisions. The first is that the authors require that such items must be spread broadly in order to count as fake news. If only a handful of individuals wrongly treat an article as legitimate news, then this does not render it fake news. Ideally, however, content moderation involves removing problematic content *before* it has a chance to spread widely. Additionally, the requirement that items must be broadly treated as real news means that an article's status as fake news is dynamic. An article that counts as fake news at time *t* may not count as fake news at time *t'* if people cease to treat it as the product of standard journalistic practice. Similarly, an article wrongly treated as the product of traditional journalism in one community counts as fake news in that community, but may not count as fake news in a different community. Thus, the same news item can simultaneously count as both fake and legitimate. The authors consider this an advantage as it focuses on the historical development of an article and the various properties it possesses along the way. A dynamic view of fake news is problematic for content moderation, however, as warning labels would have to be targeted to specific communities and continuously updated. Not only would this be practically difficult for moderators but it would also be quite confusing to platform users.

This leads to the most important objection against utilizing this definition for content moderation purposes. It is not clear that wrongly treating content as though it was produced according to standard journalistic methods is necessarily problematic. If someone treats an article in this way and the article is not clearly false nor intended to be so, flagging it as fake news seems unnecessary. It could, perhaps, alert readers to approach the article with caution because the article may not meet traditional journalism's standards. However, failing to meet journalistic standards is only a problem insofar as it contributes to the development of platform users' erroneous beliefs. Flagging content that is unlikely to do so may confuse platform users as uncontroversially true articles as well as false or misleading articles would be similarly labeled. As demonstrated by Snopes' and Politifact's methods of choosing content to fact check, what is important when fact checking is whether the article is likely to *mislead* people. The content that is most likely to mislead people is content that is either false or intentionally misleading. Thus, for the purposes of content moderation, the tripartite distinction between information, misinformation, and disinformation remains relevant.

Some have argued that intentional misleadingness is the fundamental feature of disinformation, rather than simply an intent to mislead (Fallis 2015; Soe 2018). This distinction is important because intentional misleadingness suggests that the author's intent to mislead must be realized, whereas I have suggested that authorial intent alone is enough to classify content as disinformation. Making misleadingness a defining feature of disinformation has two unpleasant consequences. The first is that this definition rules out content that was intended to mislead an audience, but failed. According to this definition, two authors, sharing the same desire to mislead their audience, may produce similar content, but one may produce disinformation and the other may produce misinformation. Perhaps the second author is very incompetent and doesn't know how to write in a convincing manner. Alternatively, however, they may write very convincingly, but their audience is highly educated regarding the topic the author has written about.

This leads to the second problem with viewing misleadingness as central to disinformation: whether content counts as disinformation depends on the audience likely to encounter it. Thus, the very same content might, given one audience, count as disinformation, but given another audience, count as either information or misinformation. Additionally, in order to identify disinformation, one would have to know not only the author's intent but also what audience will encounter the news item and the likelihood that they will come to believe something false after reading it, both of which are difficult to ascertain in advance. Like the Pepp et al. definition, this means that fake news could not be identified prior to broad dissemination and that content could simultaneously count as fake and legitimate, resulting in inconsistent application of labels and subsequent user confusion. Rather than flagging disinformation based on propensity for misleading an audience, which is not fixed prior to the dissemination of the message, it is more straightforward to focus on predicting authorial intent, which does not change depending on who encounters the message.

If actual misleadingness is not a fundamental feature of disinformation or misinformation, why should we care about labeling content as misleading? Flagging such content may help authors or publishers who didn't intend to mislead people address potential misunderstandings or clarify what they intended their audience to understand. Additionally, it provides further grounds for flagging or removing disinformation. However, any attempt to identify misleading content faces three challenges. As previously mentioned, whether content is misleading is inherently audience specific. Whether a piece of information is likely to mislead a person depends on what they already know or believe that they know. Donald Trump's aforementioned political advert is not likely to mislead an individual familiar with both Biden and Trump's dealings in Ukrainian politics. Yet, such an individual might still say that the advert is misleading.

This leads us to the second challenge for identifying misleading content: whether a person labels an item as misleading depends on what they think that other people already know or believe. If a person labeling content believes that the audience knows very little about the topic, they may be more likely to label it as misleading. Alternatively, if they think their audience highly educated, they may suppose that content is less likely to mislead the audience.

The final challenge in identifying misleading content is that people are quite surprising in what they are willing to believe, regardless of their knowledge or education. Snopes, in answering why they cover satire and humor, says, "Quite evidently nothing can be put online—no matter how preposterous in concept or plainly labeled it might be—that some people won't believe to be true (or at least allow might be true). And since everything put online has the potential to reach billions of people, even if only a very small percentage of the global audience misunderstands it, that percentage may still represent a very large number of people (Mikkelson 2019)."

These challenges do not make it impossible to label content as "misleading," however, any such labeling will be subject to the following constraints. Those judgements will always reflect what the person labeling the content knows, what they judge that others know and what they judge that others may come to accept as true. This raises another worry about bias entering the dataset. A fact checker may, for example, suppose the audience is less educated or more gullible than in reality, leading to more content labeled as misleading. The audience, however, might find an excess of content flagged as misleading confusing, insulting, or patronizing and disregard the results of the fact checker. Fact checkers must, then, be careful about what they assume that people are likely to believe lest they ostracize the people that they are trying to protect.

4 Potential Solutions to the Labeling Problem

Determining what kind of content has a truth value and what that truth value is, identifying authorial intent and calculating the potential for misleading an audience are clear challenges for content moderators. Each requires that the people checking content make value judgements about the truth of what is reported, how honest or dishonest an author is, and how likely an audience is to believe something false. The resulting content moderation runs the risk of labeling information, misinformation, and disinformation in ways which discriminate against groups that hold different values than the people labeling the content. Such discrimination may have the undesired affect of reinforcing distrustful attitudes toward trustworthy sources. These problems are to some extent unavoidable in the process of detecting fake news. There are individuals caught in echo chambers who, regardless of how content is labeled, will likely reject the results of fact checking exercises. What can be done to repair trust with members of such echo chambers is beyond the scope of this paper. Yet, there are some steps that platforms can take that could bolster typical users' trust through minimizing bias and maximizing transparency regarding the process and values involved in fact checking.

4.1 Diverse Moderators

Employing a diverse group of individuals to label each news item is one way to mitigate the undue influence of a single individual or group in the labeling process. This would address at least two problems. First, it would limit accusations of censorship through representing competing viewpoints. There is, however, a question regarding how many viewpoints to include or if there is some standard a view must meet in order to be included. If everyone's viewpoint is included, then moderation will simply mirror the same problems as the platforms that the moderators are tasked with assisting. Help in identifying qualified, but diverse, viewpoints may come through applying standards from journalism, scientific practices, or other information gathering fields. Using such standards would filter out outlier views, while still leaving room for diversity.

Unfortunately, having multiple groups label each news item is often practically impossible as the amount of content uploaded is simply too large relative to the limited number of available moderators. One potential solution to this problem involves relying on algorithms to assist in content moderation, which many platforms already do. However, these algorithms need datasets for training purposes that contain content pre-labelled as misinformation, disinformation, etc. As humans label this training data, the efficacy of automated fake news detection is subject to many of the same limitations as human fact checking and more besides. Algorithms will reflect the same biases as the people who labeled the training data. However, one advantage that automated efforts do have over human fact checkers is that the data needed for training does not need to be labeled in real time. This grants time for gathering diverse opinions regarding particular news items. In one training dataset, Credbank, researchers employed Amazon Turk workers as fact checkers to overcome this challenge (Mitra and Gilbert 2015). Researchers collected 60 million tweets and divided the tweets into 1049 events. Amazon Turk workers, after looking at tweets related to a given event, rated each event as credible or not. For each event, 30 Amazon Turk workers gave a credibility rating depending on how accurate/inaccurate they found the event and how confident they were regarding their rating.

This method is a step in the right direction; however, it may not help in cases of extreme polarization of opinion where moderators may not reach agreement. This, however, is not necessarily a problem and highlights the second benefit of employing diverse annotators: doing so would also make salient the reality that values are an inevitable aspect of content moderation. Rather than aiming for value neutrality, moderators could instead aim at identifying where their values align and where, specifically, they differ. In cases where agreement cannot be reached, perhaps the target article could simply be given a "content is disputed" label with an explanation of why moderators disagreed. This would help platform users identify not only what information is suspect but also why it is potentially problematic.

4.2 Limiting Scope and Employing Experts

Another means of mitigating the risk of bias while promoting user trust is through limiting the scope of the fact checking project to one type of problem. This would allow researchers to employ people who may be more qualified to identify one kind of problem versus another. For example, if the scope of the project is limited to distinguishing information from misinformation, then researchers could employ experts from various topic domains. While the problem of human judgement still exists in determining who counts as an expert on a topic and which experts should consult on a topic, this method does, at least, mitigate the problem that crowdsourced fact checking raises. At the very least, domain experts should be able to recognize and identify competing views on a topic and explain why they labeled a certain item as information or misinformation. Similarly, if the scope of the project is limited to identifying disinformation, experts regarding the psychology and language of deception might be employed to detect fake news. Finally, if the task is simply to identify what content an audience finds misleading, one could adopt Snope's method of letting people ask about content that they weren't sure about or found confusing.

In limiting the scope of what content fact checkers should address, platforms can identify those best qualified for identifying instances of that particular problem. This does not solve all of content moderation's problems, as there remains the issue of what qualifies a person for these sorts of tasks and the people that are chosen, regardless of qualification, still import their own personal values into labeling decisions.

4.3 Increased Transparency

Finally, platforms could increase trust in their content moderation efforts through increasing transparency in how and why content is flagged. Currently, when a news item is flagged on either Facebook or Twitter, it is not always clear what policy was violated or who determined that the item violated that policy. For example, Twitter removed a news article from the New York Post because the story featured suspected hacked materials, which violated Twitter's policy that prevents users from sharing content acquired through hacking. However, this reason was not immediately apparent and Twitter was accused of removing the article due to the information in the article, rather then due to the source of that information. While Twitter has since revised this particular policy, had there been greater transparency in the reasons for removal the situation could have been avoided entirely.

Similarly, increasing transparency in who flagged or removed particular items could also promote user trust in moderation decisions. In particular, it would be helpful for users to know whether an item was removed due to an algorithmic or human decision-maker because algorithms make different kinds of mistakes than humans. Early during the Covid-19 pandemic, Facebook blocked or flagged a number of posts from legitimate sources as spam, not due to the content of the posts but due to a bug in spam detection systems designed to remove links to harmful websites. While the problem was rectified fairly quickly, users were not immediately notified that posts were flagged by algorithms rather than by human moderators. The lack of transparency regarding who and why the content was flagged exacerbated confusion regarding what sources users could trust to dispense credible information regarding the pandemic.

5 Conclusion

In order to promote user trust while moderating content online, platforms must strike a balance between honoring free expression of ideas and removing problematic content. Finding this balance requires that platforms clarify exactly which problems they want to address and on what basis decisions regarding particular news items are made. Decisions regarding which problems to address are not uncontroversial. It is thus important that platforms and other relevant stakeholders reach agreement regarding which problem moderation efforts should target and which they should ignore. However, even if stakeholders agree on which problems platforms should target, successfully addressing these problems requires that fact checkers reliably identify instances of those problems.

I've identified three potential varieties of problematic content that fall under the "fake news" umbrella; however, identifying instances of these problems requires that fact checkers determine the veracity of the content, the intent of the author, and/or whether content is likely to mislead an audience. Each of these tasks involves the values and judgement of the individual fact checkers and the resulting content moderation will reflect these values. For some, the risks involved in allowing these values or judgements to seep into content moderation may outweigh the value of fact checking at all.

However, there are ways to mitigate the problem, whether through using a diverse group of fact checkers to cross-validate labeling decisions, narrowing the scope of the project so that labeling decisions are made by people qualified in that particular domain, or increasing platform transparency. Regardless of the precautions taken when labeling content as fake or legitimate, bias will show up in the resulting moderation in ways that some people find agreeable, but others will dispute. Thus, while flagging fake news might help limit the spread of certain kinds of content, doing so will not necessarily stop the spread of misinformation or disinformation. Indeed, flagging certain kinds of content as misinformation may in fact have the adverse affect of reinforcing a belief in the opposite for individuals who don’t share the values of fact checkers.

Availability of Data and Material Not Applicable

Code Availability Not Applicable

Declarations

Conflict of Interest The author declares no competing interests.

References

Ali, M., & Levine, T. (2008). The language of truthful and deceptive denials and confessions. *Communication Reports*, *21*(2), 82–91.

AVAAZ (2020). How facebook can flatten the curve of the coronavirus infodemic. https://secure.avaaz.org/campaign/en/facebook_coronavirus_misinformation/.

Bollinger, A. (2019). Beto o’rourke is ahead of his time when it comes to anti-LGBTQ religious institutions. *LGBTQ Nation*, October 16.

Cook, J. (2019). Beto o’rourke said he would revoke tax-exempt status form religious organizations that oppose same-sex marriage. *ABC News*, October 11. Date Accessed: 11/19/19.

Dessler, A., Mears, C., Richardson, M., Po-Chedley, S., Venema, V. (2019). Washinton examiner op-ed cherry picks data and misleads readers about climate models. Date Accessed: 11/19/19.

Editors, N.R. (2019). Beto proposes to oppress church with state. *National Review*. Date of Access: 11/19/19.

European Commission (2020). Commission staff working document: assessment on the code of practice on disinformation - achievements and areas for further improvement.

Fallis, D. (2015). What is disinformation? *Library Trends*, *63*(3), 401–426.

Fallis, D., & Mathiesen, K. (2019). Fake news is counterfeit news. *Inquiry*, 1–20.

Garrett, R.K., Bond, R., Poulsen, S. (2019). Too many people think satirical news is real. https://theconversation.com/too-many-people-think-satirical-news-is-real-121666.

Gelfert, A. (2018). Fake news: a definition. *Informal Logic*, *38*(1), 84–117.

High level expert group on fake news and online disinformation (2018). A multi-dimensional approach to disinformation: report of the independent high level group on fake news and online disinformation.

Horne, B.D., & Adali, S. (eds.) (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In: Proceedings of the International AAAI Conference on Web and Social Media (vol. 11, no. 1).

Horwitz, J. (2019). Facebook to exempt opinion and satire from fact-checking. *The Wall Street Journal*, September 30.

Ifantidou, E. (1994). *Evidentials and relevance*. PhD thesis, University College London.

Michaels, P., & Rossiter, C.S. (2019). The great failure of climate models. *The Washington Examiner*, August 25. Date of Access: 11/19/19.

Mikkelson, D. (2019). Why we include humor and satire in snopes.com. Date of Access: 11/19/19.

Mitra, T., & Gilbert, E. (2015). Credbank: a large-scale social media corpus with associated credibility annotations. In *Ninth international AAAI conference on web and social media*.

Owen, L.H. (2019). Facebook is opening up a fact-checking loophole for satire creators. Hope all their motives are good!.

Parler, Inc. (2020). Community guidelines. https://legal.parler.com/documents/guidelines.pdf.

Pepp, J., Michaelson, E., Sterkin, R.K. (2019). What's new about fake news? *Journal of Ethics and Social Philosophy*, *16*(2), 67–94.

Putnam, H. (2002). *The collapse of the fact/value dichotomy and other essays*. Cambridge: Harvard University Press.

Rubin, V.L., & Conroy, N. (2012). Discerning truth from deception: human judgements and automation efforts. *FIMS Publications*, 64.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.

Soe, S.O. (2018). Algorithmic detection of misinformation and disinformation: Gricean perspectives. *Journal of Documentation*, *74*(2), 309–332.

Wardle, C., & Derakhshan, H. (2018). Thinking about 'information disorder': formats of misinformation, disinformation and mal-information. In Ireton, C., & Posetti, J. (Eds.) *Journalism, 'fake news', & disinformation: handbook for journalism education and training* (pp. 44–56). UNESCO.

Zhou, L., Burgoon, J.K., Nunamaker, J.F., Twitchell, D. (2004). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group Decision and Negotiation*, *13*(1), 81–106.

Zhou, L., & Zhang, D. (2008). Following linguistic footprints: automatic deception detection in online communication. *Communications of the ACM51*, *9*(September), 119–22.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.