Combating disinformation with AI: Epistemic and ethical challenges

Benjamin Lange
Faculty of Philosophy
Ludwig Maximilian University of Munich
Munich, Germany
benjamin.lange@lrz.uni-muenchen.de

Theodore M. Lechterman
Institute for Ethics in AI
University of Oxford
Oxford, UK
theodore.lechterman@philosophy.ox.ac.uk

***Abstract*—AI-supported methods for identifying and combating disinformation are progressing in their development and application. However, these methods face a litany of epistemic and ethical challenges. These include (1) robustly defining disinformation, (2) reliably classifying data according to this definition, and (3) navigating ethical risks in the deployment of countermeasures, which involve a mixture of harms and benefits. This paper seeks to expose and offer preliminary analysis of these challenges.**

***Keywords*—disinformation, ethics, epistemology, artificial intelligence, social media, fake news**

I. Introduction

The spread of disinformation on online media and communication channels presents an increasingly serious global problem. Disinformation campaigns have already catalyzed genocides, lynchings, insurrections, pandemics, and suicides, as well as economic destruction and reputational damage [16][6][9][2].

Concern about the ongoing threat of disinformation has spawned various initiatives designed to combat it by govern-ments, firms, NGOs, journalists, and researchers. These initia-tives increasingly draw on artificial intelligence to detect disin-formation and facilitate the deployment of countermeasures. AI-based tools include ActiveFence, Blackbird AI, BotSentinel, ClaimBuster, Cyabra, and Graphika, among others. But efforts to identify and combat disinformation face a litany of epistemic and ethical challenges. These include (1) robustly defining disinformation, (2) reliably classifying data according to this definition, and (3) navigating ethical risks in the deployment of countermeasures, which involve a mixture of harms and bene-fits. This paper seeks to expose and offer preliminary analysis of these challenges.

In what follows, we offer an initial philosophical analysis of the risks involved in combating disinformation with AI-assisted tools. We focus on identifying significant ethical risks, under-stood, for simplicity, as a function of the magnitude of a harm multiplied by its prevalence and likelihood. We follow the con-ventional definition of "harm" in moral and legal philosophy as the setback of interests [13]. Significant ethical risks therefore can involve subjecting many people to major harms, many peo-ple to minor harms, or few people to major harms.

Much neighboring scholarship on these issues specifically addresses the ethics of content moderation on social media plat-forms, with primary attention to the responsibilities of platform hosts [19][20]. Our approach differs in two main ways. First, our focus is on the phenomenon of disinformation itself and not particular channels of transmission, which raise many more specific issues. Second, although our account certainly has implications for the responsibilities of channel owners and managers, they are not our primary targets. As opportunities for disinformation detection and counteraction become more widely available to governments, firms, and NGOs, we seek to expose and analyze epistemic and ethical challenges that pertain to any party engaged in advanced practices of disinformation detection or counteraction.

The paper is structured as follows. Section II notes the importance of a robust definition of disinformation and evaluates three candidate definitions. Section III discusses the ethical risks associated with diagnosing disinformation once a definition is adopted. Section IV then turns to the potential ethical risks related to counteraction measures. Section V concludes with a discussion of issues for further research.

II. Defining Disinformation

To effectively identify and combat disinformation, AI-based approaches need to be based on a sound definition of disinformation and operationalizable for the desired deployment context. We propose that a candidate definition should meet the following two related desiderata:

> **Accuracy:** The definition needs to be as conceptually sound and exhaustive as possible.
>
> **Operationalizability:** The definition needs to be practically applicable by enabling observers to distinguish disinformation with available resources and at tolerable cost.

The first desideratum seeks to ensure that the candidate definition classifies instances of disinformation in a way that can maximally justify and explain observers' considered judgments [24]. A conceptually sound definition will precisely distinguish disinformation from neighboring concepts, including misinformation, information, parody, and satire. Despite the obvious value of a conceptually sound definition, we note a potential tradeoff between conceptual accuracy and operationalizability in practical contexts. We think it infeasible to endorse a definition of disinformation that is impossible to apply in practice because, for example, it relies on criteria that cannot be meaningfully assessed by external observers. Conversely, employing a practical definition that is very easy to apply may come at the

cost of its accuracy. Successfully combating disinformation in practice is hence subject to an inherent tension between conceptual and practical considerations that may in turn amplify epistemic and ethical risks.

A. Three Definitions of Disinformation

We examine three possible definitions of disinformation in light of the suggested desiderata. The first contender is the so-called

> **Intention Definition:** Disinformation is misleading information that is intended to cause harm [15][12, p. 260].

"Misleading information" in this definition refers to information that has a "propensity to cause false beliefs" [15, 12: p. 260]. This definition characterizes disinformation by focusing on the quality of the will behind it. This means that it covers ineffectual disinformation scenarios: disinformation that fails to cause harm even though it was intended to do so. As we discuss further below, the ability to assess intent is an important criterion for justifying countermeasures with a punitive dimension.

This definition is endorsed by several experts [18]. However, one challenge with this definition lies in uncovering the intentions of a disinformation outbreak when the originator of the disinformation is unclear. Even when the originators are identifiable, observers may have limited access to their mental states, which will often inhibit efforts to ascribe intentionality in a meaningful way. During a pandemic, for instance, people may concoct or promulgate various misleading claims about the origins or treatment of a disease without having malicious intentions. While some may indeed have ulterior motives, the basis for these motives may not always be obvious.

In recognition of these challenges, a second candidate definition is the so-called

> **Harm Definition:** Disinformation is misleading information that causes or risks causing systematic harm to some agent for the benefit of another [5, p. 80].

This definition characterizes disinformation in a way that focuses on the likelihood that the misleading information will lead to bad outcomes for a victim (regardless of the intentions that may lie behind it). As before, misleading information refers to information that has a propensity to cause false beliefs.

While intentions are implicitly characterized in this definition insofar as the disinformation (expectedly) benefits someone, it avoids "intention" terminology and speaks only of harms and benefits, which may be easier to measure in practice. This definition also captures the possibility that disinformation can be spread without malicious intent, such as when people simply share harmful and misleading content posted by others.

One challenge this definition faces is capturing ineffectual disinformation: disinformation that fails to harm a victim even though it was intended to do so. Even if a malicious disinformation campaign intended to harm someone but, for some reason, no harm manifests, identifying and combating the campaign may still be warranted.

The last contender is the

> **Hybrid Definition:** Disinformation is misleading information that either (a) intends to harm or (b) systematically harms or risks harming someone [14].

This hybrid definition is a combination of the previous two definitions and characterizes all cases of disinformation as either intending to harm or de facto harming. It thereby covers both the input and output dimensions of disinformation. This definition also excludes other categories of misinformation, such as parody and satire, if and when they are neither intended to cause harm or likely to do so. However, the definition would indeed classify instances of parody and satire as disinformation where such instances are significantly harmful in intent or effect.

This definition is the most exhaustive. It covers ineffectual, malicious, and inadvertent promulgated untruths. This best mitigates the risk of failing to classify some misleading information as disinformation. However, this definition might be harder to operationalize than the previous two, since a heuristic aimed at filtering for disinformation would have to focus on all of the key variables contained in the previous two definitions. If resources for data gathering and analysis are scarce, the additional explanatory power of this definition may come at some costs.

B. Assessment of Definitions of Disinformation

The Hybrid Definition is the most conceptually sound of the definitions we have considered because it avoids all intuitively problematic cases of misclassification. However, this definition needs to operationalize both harms and intentions. A tradeoff thus exists between conceptual soundness and operationalizability. Moreover, the definition needs to rely on a measure of intentions that is distinct from its measure of harms. Insofar as such a measure cannot be deployed, employing the Hybrid Definition may be futile.

The Harm Definition will be sufficient if the primary tactics to be deployed do not involve punitive measures. If the goal is merely to correct the public record through public service announcements, a disinformation mitigation campaign has no need to trace the intentions of those who have distorted the public understanding of an issue. Confirming the intentions behind an action, however, is generally an important criterion for justifying punitive responses to it. If a disinformation mitigation campaign wishes to neutralize or deter individuals who spread false claims, it must be able to distinguish culpable from innocent disinformation spreading, which generally requires knowledge of intentions. We return to this idea in Section IV.

III. Diagnosing Disinformation

Once a robust definition of disinformation has been chosen, the next step is to operationalize that definition for detecting instances of disinformation. Here is where AI offers numerous potential advantages. For instance, natural language processing can be used to summarize text, identify the author's ideological stance, and assess whether content exhibits manipulative rhetoric [21]. AI can detect doctored content (i.e., "deepfakes") by analyzing abnormal features of a subject's physiology, scanning for technical traces of content editing, or using deep neural networks to separate authentic from inauthentic content on the basis of numerous parameters [22]. AI can be used to detect fake users

("bots"), which are frequently implicated in the spread of disinformation [23]. None of these methods on its own is sufficient for detecting disinformation according to the definition we propose, as none includes an assessment of harm. However, these and related methods provide enhanced capabilities for broader disinformation detection efforts. They are also implicated in the ethical risks that follow.

Two main risks in disinformation detection are producing false positives and false negatives.

> **False positives:** False positives occur when diagnosing material as disinformation that is not in fact disinformation.
>
> **False negatives:** False negatives occur when incorrectly classifying disinformation content as legitimate information.

False positives can entail subjecting falsely-identified spreaders to harmful countermeasures. False negatives risk allowing harmful untruths to continue spreading undetected. Both failures, when publicly revealed, can undermine the credibility of disinformation detectors and set back progress on efforts to reduce the spread of disinformation.

Both risks become drastically more likely when the definition of disinformation is itself problematic. Thus, one of the main challenges for AI-based approaches to combating disinformation is to make their conceptualization of disinformation as watertight as possible.

Further problems occur when attempting to operationalize a strong definition with available tools. For instance, an intention-based definition may be hard to apply, as observers have limited insight into actors' mental states. Proxies may be insufficiently reliable. A harm-based definition requires speculation about the potential magnitude of harm and the urgency of intervening, not to mention further choices about specifications of harm and urgency. Often, harm magnitude and urgency of intervention only become clear after the fact.

Even when these challenges have been overcome, disinformation detectors face additional challenges of determining whether a given piece of information fits the disinformation definition. These challenges include lack of verifiable data, highly judgment-sensitive topics, topics where there is significant expert disagreement, and situations where information is dynamically evolving. The COVID-19 pandemic exhibits several of these features at once, which helps to explain why researchers have struggled to develop accurate models for detecting pandemic-related disinformation [17].

Disinformation detection systems may also be biased in certain ways. They may be biased toward certain ideological positions or scientific theories. They may be compromised by conflicts of interest between different stakeholders. Human agents may fall prey to automation bias by trusting too readily in assessments offered by AI systems [25]. And the AI components may be biased themselves. For instance, if these systems draw on natural language processing, they may treat certain communities' linguistic conventions as the norm for truth-telling and misclassify messages that rely on alternative linguistic conventions from marginalized communities [26].

IV. Counteracting Disinformation

Efforts to combat suspected disinformation will naturally inherit and compound any limitations of defining or diagnosing information as disinformation. But the deployment of mitigation measures also introduces further ethical risks. As discussed presently, these include risks in inoculating victims and debunking false content, risks in neutralizing the source of disinformation, and risks involving the political legitimacy of different participants in information warfare. The discussion that follows is largely speculative, as many of the risks we foresee have not been extensively studied or documented. It also focuses on targeted post-hoc interventions, in contrast to more general efforts to promote media literacy [29], which contain fewer risks of harm.

A. Debunking and Inoculation

The most obvious tactic a disinformation combatant may wish to use is broadcasting what it believes to be the truth of the matter—to "set the record straight." This can backfire, obviously, if the combatant is in the grip of false beliefs. The combatant may have incomplete information. The information may be in flux. Current expert consensus may be subsequently disproven. The combatant may have access to the relevant facts but misinterpret their significance. And so on. These challenges are evident from the COVID-19 pandemic response, during which many experts initially rejected calls for face masks and warned of transmission via contact with contaminated surfaces, only to walk back these positions later [7].

Attempts to set the record straight can also backfire if the combatant possesses the truth but fails to explain it in a way that people fully understand and appreciate. Climate change communication provides a cautionary tale. Numerous studies have suggested that a contributor to climate change skepticism is a systematic misunderstanding of what uncertainty means in scientific research, where few claims are ever definitively proven [4]. These studies have spurred the scientific community to reexamine how to communicate effectively to the public.

One variant of broadcast counter-messaging is content labeling on social media, where a platform attaches a label to posts containing sensitive content. During the COVID-19 pandemic, for instance, Facebook automatically applied a label to any post that mentioned the topic of vaccines [1]. The label contained a link to official information sources. Although this approach generally avoids the difficulty of adjudicating the complexities of the debate, one challenge with Facebook's approach is that it also casts doubt on posts that contained no questionable content. Raising doubts about legitimate posts may be counterproductive to the larger aims of content moderation. It is also a potentially objectionable restriction on expressive liberty.

In addition to broadcasting the truth, combating disinformation can involve selective inoculation of disinformation victims [30]. A combatant might directly communicate with segments of a population believed to be particularly susceptible to deceptive messages or likely to suffer most from false beliefs about a given topic. In a pandemic, this might be health care workers and medically vulnerable groups, such as residents of

assisted living facilities. Here, there may be risks of discrimination and condescension, or the appearance thereof. Those who are targeted for disinformation inoculation may wonder why they have been targeted, and they may infer that targeting results from a judgment of their cognitive inferiority. How the message is conveyed can also make a difference. A condescending tone—by, for instance, overstating the obvious, ridiculing, or admonishing—may insult recipients' intelligence or dignity.

Combatants may be tempted to deploy insidious or manipulative tactics, such as subliminal advertising or secretly employing influential personalities to promulgate certain talking points, as Minneapolis attempted to do during the trial of the murder of George Floyd [8]. Tactics that bypass human reasoning threaten the value of autonomy, as they indicate disrespect for their targets' rational agency [27, pp. 90, 2]. As the Minneapolis case shows, such tactics are also likely to provoke backlash when they are discovered, which can cause irreparable harm to com-batants' credibility.

B. Neutralizing Sources and Spreaders

Perhaps the most significant ethical risks in disinformation mitigation involve attempts to neutralize disinformation propagators, i.e., those who disseminate misleading information with harmful intentions or effects. Neutralization techniques may involve attempts to identify and publicly expose these individuals and groups, subjecting them to public censure and sanctions. Neutralization techniques may involve pressing charges with law enforcement and/or seeking to have social media accounts or content removed. Or, they may involve measures to disrupt communication directly through hacking or cyberattacks.

The kinds of aggressive tactics discussed above carry particular risks. Many of these are associated with the use of force and well handled with the *ius in bello* principles from just war theory, which provide well-established guidelines for combat activities [3, 12]. Chief among the risks of aggressive tactics is target misidentification. If the combatant mistakenly identifies someone as a disinformation propagator and subjects them to harsh treatment, the combatant has subjected an innocent person to potentially serious harm. Collateral damage represents a second major concern. That is, even if the combatant has accurately identified the guilty parties, the countermeasures deployed may involve harming other innocent people. Attempts to expose the suspects or disrupt their communications channels may result in unintended harms to the suspects' family, colleagues, or neighbors, or to others who use the channels for legitimate purposes. Related to this is the risk of disproportionality, that the tactics chosen may accomplish their intended goal but do so with excessive damage to the suspected individual. An attempt to publicly shame a suspected disinformation propagator may create opportunities for others to make violent threats, vandalize their property, or worse.

Culpability makes a difference to the justification of countermeasures. An individual or group that willfully creates or promotes disinformation may be culpable in ways that those who innocently share disinformation content may not be. Some perpetrators may be willful but lack components of moral responsibility due to reasons of age or cognitive condition. While certain neutralization measures may be warranted regardless of culpability if the associated harms are severe enough, punitive measures may be entirely unjustified for excusable wrongdoing [28]. The fact that it may often be difficult to determine whether the suspects are culpable creates additional grounds for caution.

Twitter's approach to disinformation during the 2020 U.S. election presents an interesting combination of targeted inoculation and neutralization elements. Twitter sought to combat disinformation by labeling posts that it deemed to contain false or misleading claims and limiting options for commenting and retweeting [10]. The accuracy of these methods is not clear. However, one challenge with this approach is that it can be counterproductive for certain users. For those who believe Twitter is ideologically biased against their side, Twitter's content labels may simply reinforce beliefs in the credibility of that content. Another risk, of course, is that this method may drive aggrieved users toward less regulated channels, where disinformation can proliferate more freely. These are merely some of the risks that must be balanced against the overall reduction in the spread of disinformation that such a method may achieve.

V. Legitimate Authority

Specific issues arise depending on whether the combatant or the suspected disinformation propagator is a state or non-state entity. If the combatant is a non-state entity acting on its own accord, it risks charges of vigilantism and extrajudicial punishment if it seeks to apply sanctions to suspected disinformation propagators. The state is the default authority for law enforcement and criminal punishment, and while private actors may be justified in acting in certain cases, these decisions require exceptional justification. One potential justification is that the combatant owns the channel on which disinformation is flowing. In such cases, the combatant would be justified in creating or enforcing terms of service to limit disinformation. Without terms of service, however, depriving users of the service may constitute an arbitrary breach of contract. And in neither case would a channel owner be justified in taking additional punitive measures beyond removing content or limiting access.

If the state is the combatant in question, worries may arise over state silencing of speech and the protection of expressive liberty. Governments are generally under a stronger duty than private entities to tolerate offensive or misleading speech, given the state's coercive power and the importance of expressive liberty to government accountability.

Conditions change when the disinformation propagator is itself a state entity. When this is an agent of a democratic state, the normal recourse is to expose and prosecute this behavior publicly, rather than engage in surreptitious countermeasures. When this is an agent of a foreign state, engaging in information warfare with a foreign adversary is generally a prerogative of national militaries and counterintelligence services and not appropriate for private organizations. State defense forces may choose to engage in information warfare with a foreign adversary and also to employ private contractors in these efforts. In both cases, these decisions are subject to the law and ethics of warfare.

VI. Conclusion

This paper has sought to provide a preliminary analysis of epistemic and ethical challenges of combating disinformation with AI. Although we have focused on significant risks, precisely which risks apply depends on the particular focus and methods of a mitigation campaign. Tools that merely provide fact-checking services do not face the risks inherent in efforts to neutralize disinformation at its source. However, such services will indeed face definitional and diagnostic risks. And any secondary efforts to use these tools for debunking or neutralization will of course encounter the risks we have identified in these areas.

Our analysis has focused selectively on significant ethical risks in different stages of disinformation mitigation. We have not attempted a comprehensive analysis of all pertinent ethical risks, nor have we attempted to identify the risks specific to any particular method or approach to AI-driven disinformation mitigation. Additionally, our intention has been to identify and describe key ethical risks rather than to defend particular principles for balancing or resolving them. Despite these limitations, we believe the initial analysis offered here can provide firm foundations for future work on this topic.

Acknowledgment

We thank Bradley J. Strawser, Michael Skerker, David Whetham, and two anonymous reviewers for discussion on these topics and/or comments on an earlier draft.

References

[1] "Facebook to label vaccine posts to combat COVID-19 misinfo," *Associated Press*, Mar. 15, 2021. [Online]. Available: apnews.com/article/business-misinformation-coronavirus-pandemic-mark-zuckerberg-55583cb7cee4e6966f49f8604d76a081

[2] "How WhatsApp helped turn an Indian village into a lynch mob," *BBC News*, Jul. 19, 2018. [Online]. Available: www.bbc.com/news/world-asia-india-44856910

[3] D. Whetham, "The just war tradition: A pragmatic compromise," in *Ethics, law, and military operations*, D. Whetham, Ed. Houndmills, Basingstoke, Hampshire ; New York, NY: Palgrave Macmillan, 2011.

[4] S. van der Linden, "The social-psychological determinants of climate change risk perceptions: Towards a comprehensive model," *Journal of Environmental Psychology*, vol. 41, pp. 112–124, Mar. 2015, doi: 10.1016/j.jenvp.2014.11.012.

[5] B. Skyrms, *Signals: evolution, learning, & information*. Oxford ; New York: Oxford University Press, 2010.

[6] P. Mozur, "A genocide incited on Facebook, with posts from Myanmar's military," *New York Times*, Oct. 15, 2018. [Online]. Available: www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html

[7] D. Lewis, "COVID-19 rarely spreads through surfaces. So why are we still deep cleaning?," *Nature*, vol. 590, no. 7844, pp. 26–28, Feb. 2021, doi: 10.1038/d41586-021-00251-4.

[8] M. Levenson, "Minneapolis will pay influencers to fight misinformation during officers," *New York Times*, Feb. 26, 2021. [Online]. Available: www.nytimes.com/2021/02/26/us/minneapolis-influencers-chauvin-george-floyd.html

[9] M. Garrett, "'It was a drug': Capitol riot exposes reach of QAnon disinformation," *CBS News*, Jan. 31, 2021. [Online]. Available: www.cbsnews.com/news/qanon-capitol-riot-reach/

[10] V. Gadde and K. Beykpour, "An update on our work around the 2020 US Elections," *Twitter Company Blog*, Nov. 12, 2020. blog.twitter.com/en_us/topics/company/2020/2020-election-update.html

[11] L. Floridi and M. Taddeo, Eds., *The ethics of information warfare*. Heidelberg: Springer, 2014.

[12] L. Floridi, *The philosophy of information*. Oxford ; New York: Oxford University Press, 2011.

[13] J. Feinberg, *The moral limits of the criminal law volume 1: Harm to others*. New York: Oxford University Press, 1987. doi: 10.1093/0195046641.001.0001.

[14] D. Fallis, "A functional analysis of disinformation," presented at the iConference 2014, Mar. 2014. [Online]. Available: hdl.handle.net/2142/47258

[15] D. Fallis, "A conceptual analysis of disinformation," presented at the iConference 2009, Feb. 2009. [Online]. Available: hdl.handle.net/2142/15205

[16] P. Butcher, "COVID-19 as a turning point in the fight against disinformation," *Nat Electron*, vol. 4, no. 1, pp. 7–9, Jan. 2021, doi: 10.1038/s41928-020-00532-2.

[17] Herrmannova, Drahomira *et al.*, "Challenges in automated detection of COVID-19 misinformation," presented at the Workshop on Human Aspects of Misinformation Online at the 2021 ACM CHI Virtual Conference on Human Factors in Computing Systems, May 2021. doi: 10.5281/ZENODO.4697992.

[18] "Automated tackling of disinformation," European Parliament Research Unit, PE 624.278, Mar. 2019. [Online]. Available: www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_STU(2019)624278

[19] N. Syed, "Real talk about fake news: Towards a better theory for platform governance," *Yale Law Journal*, vol. 127, pp. 337–357, 2017-2018.

[20] É. Brown, "Regulating the spread of online misinformation," in *The Routledge handbook of political epistemology*, M. Hannon and J. de Ridder, Eds. London ; New York: Routledge/Taylor & Francis Group, 2021, pp. 214–225.

[21] Q. Su, M. Wan, X. Liu, and C.-R. Huang, "Motivations, methods and metrics of misinformation detection: An NLP perspective:," *NLPR*, vol. 1, no. 1–2, p. 1, 2020, doi: 10.2991/nlpr.d.200522.001.

[22] S. Lyu, "Deepfake detection: Current challenges and next steps," in *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, London, United Kingdom, Jul. 2020, pp. 1–6. doi: 10.1109/ICMEW46912.2020.9105991.

[23] O. Varol, E. Ferrara, C. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization," *ICWSM*, vol. 11, no. 1, pp. 280-289, May 2017.

[24] Y. Cath, "Reflective equilibrium," in *The Oxford handbook of philosophical methodology*, H. Cappelen, T. S. Gendler, and J. Hawthorne, Eds. Oxford University Press, 2016, pp. 213–230. doi: 10.1093/oxfordhb/9780199668779.013.32.

[25] "Algorithms are insufficient to curb disinformation," Oxford Analytica, Emerald Expert Briefings, Aug. 2021. doi: 10.1108/OXAN-DB263510

[26] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (technology) is power: A critical survey of 'bias' in NLP," *arXiv:2005.14050 [cs]*, May 2020, Accessed: Oct. 4, 2021. [Online]. Available: arxiv.org/abs/2005.14050

[27] T. M. Scanlon, *The difficulty of tolerance: Essays in political philosophy*, 1st ed. Cambridge: Cambridge University Press, 2003. doi: 10.1017/CBO9780511615153.

[28] P. F. Strawson, *Freedom and resentment and other essays*, London: Routledge, 2008. doi: 10.4324/9780203882566.

[29] T. Zerback, F. Töpfl, and M. Knöpfle, "The disconcerting potential of online disinformation: Persuasive effects of astroturfing comments and three strategies for inoculation against them," *New Media & Society*, vol. 23, no. 5, pp. 1080–1098, May 2021, doi: 10.1177/1461444820908530.

[30] S. Lewandowsky and S. van der Linden, "Countering misinformation and fake news through inoculation and prebunking," *European Review of Social Psychology*, pp. 1–38, Feb. 2021, doi: 10.1080/10463283.2021.1876983.