Media and Communication
Open Access Journal | ISSN: 2183-2439
Volume 9, Issue 1 (2021)

Dark Participation in Online Communication: The World of the Wicked Web

Editor
Thorsten Quandt

COGITATIO

Media and Communication, 2021, Volume 9, Issue 1
Dark Participation in Online Communication: The World of the Wicked Web

Published by Cogitatio Press
Rua Fialho de Almeida 14, 2º Esq.,
1070-129 Lisbon
Portugal

Academic Editor
Thorsten Quandt (University of Münster, Germany)

Available online at: www.cogitatiopress.com/mediaandcommunication

Table of Contents

Table of Contents

Media and Communication (ISSN: 2183–2439)
2021, Volume 9, Issue 1, Pages 84–87
DOI: 10.17645/mac.v9i1.4020

Editorial

Can We Hide in Shadows When the Times are Dark?

Thorsten Quandt

Department of Communication, University of Münster, 48143 Münster, Germany;
E-Mail: thorsten.quandt@uni-muenster.de

Submitted: 17 January 2021 | Published: 3 February 2021

Abstract
The editorial discusses the relevance of analyzing some problematic aspects of online participation in consideration of events that happened during the preparation of this thematic issue. It critically challenges the eponymous 'dark participation' concept and its reception in the field, and calls for a deeper exploration of epistemological questions — questions that may be uneasy and difficult to answer, as they also refer to the issue of balance and scientific positioning in the face of threats to public communication and democratic ideals.

Keywords
dark participation; disinformation; duality; epistemology; participatory journalism; public communication; online communication

Issue
This editorial is part of the issue "Dark Participation in Online Communication: The World of the Wicked Web" edited by Thorsten Quandt (University of Münster, Germany).

1. The Season of Light, the Season of Darkness

It was the best of times, it was the worst of times,
it was the age of wisdom, it was the age of foolishness,
it was the epoch of belief, it was the epoch of incredulity,
it was the season of Light, it was the season of Darkness.
(Dickens, 1859, p. 1)

The famous opening paragraph of Dickens' *A Tale of Two Cities* masterfully describes the major conflicts and extremes of a chaotic time of social and political upheavals. Set in the years leading to the French revolution, the historical novel is referring to a specific period and location (to be more precise, two specific locations: the eponymous two cities London and Paris)—but the opening paragraph has gained a life of its own in public, (pop) culture and science. I am not a native speaker, but it always struck me with awe: It's an ingenious way of expressing the duality of revolutionary times, and in some ways, also of how some of our current times feel like.

Dickens' opening paragraph transcends the specifics of the novel's plot and localization, and that is probably why so many people since the original publication in 1859 could connect to its deeper meaning, especially so if they found themselves in periods of profound social change. Indeed, his magnificent lines sound more current than ever, and they also resonate with this thematic issue's topic—especially as they literally refer to the duality of light and darkness as two opposing positions and potentials.

A figurative understanding of light and dark, referring to a larger duality of the social, is a seed concept of this current issue. When being approached by *Media and Communication* to serve as the editor of a thematic issue, I had just published an article in said journal on "dark participation" (Quandt, 2018), focusing on the "bleak flip side" (p. 18) of citizen participation in online environments, including phenomena like trolling, bullying, strategic disinformation and hate campaigns. Based on the strong, and sometimes even quite emotional reactions to this piece (which are certainly not the standard for a publication in a scientific journal), its core topic looked like a perfect candidate for deeper exploration. Little did I know what would happen in the short time between the call for papers, the subsequent review

process of the amazing pieces the journal received, and the release of the issue—and how this would make both the topic (and the introductory Dickens quote) more current than ever.

In barely 18 months, the world witnessed the outbreak and rapid development of a pandemic, paralleled by a confusing cacophony of voices, conspiracy theories and disinformation regarding the Coronavirus, some of it originating from dubious sources on the web and dark participation in online forums. Early on in the crisis, the WHO labeled this an 'infodemic' (WHO, 2020)—a highly contested (but yet popular) term to describe the social and communicative situation in the pandemic. Naturally, the critics are correct: As a start, it would need to be called a 'disinfodemic,' if the core problem is intentional falsehoods and not just an exponential growth of information in a short time—and there are other limitations of this concept that cannot be discussed in a brief editorial. Yet, it still underlines the timeliness of this issue's core ideas, which seemed to gain even more urgency in the Coronavirus crisis.

Shortly before release date, the world's longest-standing democracy—the United States of America—were shaken by pictures of a mob storming the Capitol, incited by a president who had lost the election, but did not accept the results of the election and called it rigged on multiple occasions (without presenting substantial proof for these claims). For that infamously 'twittering' president, traditional media were primarily 'fake news,' and many of his supporters organized themselves via social media and online platforms—the most radical ones just trusting their own sources and stirring up themselves in a rather hermetic information environment. While communication studies has rightfully questioned the existence of "filter bubbles" in general (Bruns, 2019), it became apparent that there are related issues on the extreme edges of an increasingly polarized society, where opinion formation is (self-)organized in radical pockets of a rather 'wicked' web. Indeed, public observers identified some forms of 'dark participation' in online environments as a danger to democracies, and numerous politicians around the world called for action against populism, hate and disinformation.

These incisive developments call for a short moment of reflection regarding the conceptual core of this thematic issue, its changing context and resulting epistemological questions. Therefore, this editorial slightly deviates from the expectation of giving an overview of the enclosed articles. Luckily, my colleague Oscar Westlund—editor of the journal *Digital Journalism*—was asked to comment on the thematic issue, and he does a much better job at an introduction than me here (Westlund, 2021). Reading his commentary before working through the issue is highly recommend, and then re-reading it after that procedure as well. Further, the current president of the International Communication Association Claes de Vreese adds some crucial contextual thoughts, putting some of the arguments in this thematic issue in (disciplinary) perspective (de Vreese, 2021). Reading his commentary as a concluding remark will certainly widen the scope of how the issues at stake can be discussed.

2. Darkness, Debates and Discipline

As mentioned above, the idea for this thematic issue had its origin in an earlier, quite personal exploration of dark participation in online environments, published in this very journal roughly two years ago. Like this editorial, I chose to partially deviate from a traditional article format there. If you haven't read it and still plan to do so—then please stop reading exactly here →☆←, as the following will include 'spoilers'!

The piece itself was, on the surface, an exploration of the concept of 'dark participation,' which was depicted as a counter-concept to a 'naïve,' abundantly positive and 'pure' concept of user participation discussed in communication studies and journalism roughly around the turn of the millennium and the subsequent decade. The 'dark participation' concept was developed in a systematic, yet intentionally generic way in the middle section of the piece. However, this systematic discussion of dark participation was also meant as a device to lead the reader astray: The plan was to get the reader nodding her or his head and agreeing with the argument. The reader should fully embrace the focus on dark participation as an innovative and convincing concept. Then, in the last third of the article, it was revealed that such a one-sided debate of the 'dark' side would be equally misguided as the overly optimistic and normatively narrow expectations regarding participation, and that some crucial and balancing counter-arguments were left out on purpose to get her or him agreeing with the intended position. So the article was actually a call for balance in the discussion, despite its title and core concept: Just focusing on dysfunctional effects and being fascinated by the doom and gloom of the dark side would be as wrong as naïvely expecting every user in online environments to be a heroic, liberal savior of democracy. Metaphorically speaking, the pitch black of 'dark participation' in the piece was poured into the crystal white of some earlier approaches to end up with a more fitting grey.

As noted above, the article lead to some surprisingly emotional, even visceral reactions, which are uncommon for a scientific journal article: Some readers loved it, and some really hated it. And in both groups, there were people who just referred to the dark participation concept itself without the proper 'balancing' contextualization—maybe overlooking the mirror trick this article really is. Now I mention this article and its history not for self-reflection, but to point out the issues of doing research on participation in general, and how personal and emotional it can be: This is not an 'empty' concept by definition, as the participation in public communication and social processes logically refers to democratic ideals—and therefore ideas that may be close to our heart. It can be theoretically argued that citizen participation always

entails an ethical component, therefore something like 'dark participation' does not exist or is a "perversion" (Carpentier, Melo, & Ribeiro, 2019) of participation—and indeed, the 'pure' form of citizen participation is a shimmering star in the sky that may be needed as guidance for our actions. On the other hand, there is ample of empirical proof that there are grave issues with some forms of participation in online environments—and we as social scientists cannot ignore the fact that parts of the political elite and public in many countries regard some of these a danger to society (and even call for measures to restrict dark participation).

In that sense, a discussion of such phenomena leads to difficult epistemological questions, and calls for a reflection on our positioning and perspective as individual scholars and our discipline(s) at large.

3. Out of the Grey Zone, Back into the Light?

Following the above argument, can we as scientists stay in the secure space of the ivory tower and observe these issues from afar, with the impartial gaze of a neutral observer, painting the world in a diffuse grey? Or do we even stand on top of it, and observe through a normative lens with conceptual nobility, far above the lowlands of confusing empirical contradictions? And all of this while we receive alarming evidence of concepts like participation being turned into a dark counter-image of what we hoped they would be?

As noted above, the initial piece on dark participation argued for balance—based on my personal perception of a dominant one-sidedness both in the early debate on participation and its much darker counterpart as of recently. However, such a call for balance may also lead to a situation where scientists hide in a hazy and shapeless 'grey zone,' where no position is taken, everything appears value free—and everything looks similar. Given the events that happened during the production of this thematic issue, and based on the findings and approaches assembled here, I have some nagging doubts. Figuratively speaking: Can we as scientists hide in shadows when the times are getting darker and darker—and won't the safe grey zone disappear with the fading light? Maybe we cannot be fully neutral here, as open science itself is also part of open debates and open societies? And therefore, shouldn't we have a vital interest in their success?

Naturally, this refers to well-known epistemological questions of social sciences and the dispute between normative positions and a (arguably) 'value free' critical-rationalist position, and between neutral and activist research. While some of these questions have been discussed in great depth in other fields, and while they were always somewhat present in communication studies, I feel that they need to be discussed in a more substantive way, given the current challenges we observe in online communication and the social alike. It is no coincidence that difficult times of social change often breed epistemological questions and paradigmatic changes in science as well—as change and the related anomalies reveal fractures in existing paradigms. The events we saw unfold during the preparation of the thematic issue may be partially cause and effect of such change. Some of it is related to the evolution of online communication and the transformation of society in an increasingly 'digitized' world, where information flows do not adhere to the logics of traditional media and journalism. Dark participation (or whatever label you prefer) is certainly not its sole cause, but part of this.

In this short editorial, I could only hint at these deeper, epistemological issues that parallel the fine pieces of research in this thematic issue. Naturally, there is notable variance here: The authors come from different world regions—Europe, Asia and the United States—they analyse multiple forms of 'dark participation' in online communication, and they favor various empirical and theoretical approaches. However, they are united by their deep interest in the given phenomena, often driven by an implicit or explicit goal: exploring dark participation and delineating it from its light counterpart. And by doing so, they may be helping in saving citizen participation from the destructive 'doom and gloom' that seems to be so pervasive these days. This may also be an answer to my concerns that an occupation with the dark side may result in a diffuse gray—researching dark participation in such a way may contribute to a better understanding of other forms of participation as well, and therefore help in identifying factors that protect these from being dragged into the dark. So instead of ending up in a diffuse grey zone, such research may result in a much sharper contrast between light and dark.

The introductory quote from Dickens' *A Tale of Two Cities* brilliantly expresses this duality where the light and dark coexist in all their variety. Applied to the many negative observations our field has recently made in relation to forms of dark participation and their dangers to society, this also holds a hopeful promise: that if we observe chaos and foolishness in democratic crisis, then there is also the potential for stability, elegance and wisdom.

Acknowledgments

I would like to thank Oscar Westlund for his thoughtful comments that helped to further develop and refine the editorial.

Conflict of Interests

The author declares no conflict of interest.

References

Bruns, A. (2019). *Are filter bubbles real?* Cambridge: Polity.

Carpentier, N., Melo, A., & Ribeiro, F. (2019). Rescuing participation: A critique on the dark participa-

tion concept. *Communicação e Sociedade*, *36*, 17–35. Retrieved from https://journals.openedition.org/cs/1284

de Vreese, C. (2021). Beyond the darkness: Research on participation in online media and discourse. *Media and Communication*, *9*(1), 215–216.

Dickens, C. (1859). *A tale of two cities*. London: Chapman & Hall.

Quandt, T. (2018). Dark participation. *Media and Communication*, *6*(4), 36–48.

Westlund, O. (2021). Advancing research into dark participation. *Media and Communication*, *9*(1), 209–214.

WHO. (2020). *Novel Coronavirus (2019-nCov)* (Situation report No. 13). Geneva: WHO. Retrieved from https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200202-sitrep-13-ncov-v3.pdf

About the Author

Thorsten Quandt is a Professor of Online Communication at the University of Münster, Germany. His research fields include online communication, digital games and (online) journalism. Quandt is particularly interested in the societal changes connected to the Internet and new media, and the question of how human beings have evolved in sync with these changes. His earlier works on participatory journalism and online newsroom production have been widely cited in the field of (digital) journalism research.

Media and Communication (ISSN: 2183–2439)
2021, Volume 9, Issue 1, Pages 88–98
DOI: 10.17645/mac.v9i1.3419

Article

Uninvited Dinner Guests: A Theoretical Perspective on the Antagonists of Journalism Based on Serres' Parasite

Gerret von Nordheim * and Katharina Kleinen-von Königslöw

Department of Journalism and Mass Communication, University of Hamburg, 20146 Hamburg, Germany;
E-Mails: gerret.vonnordheim@uni-hamburg.de (G.v.N.), katharina.kleinen@uni-hamburg.de (K.K-v.K.)

* Corresponding author

Submitted: 30 June 2020 | Accepted: 23 August 2020 | Published: 3 February 2021

Abstract
In the digital age, the crisis of journalism has been exacerbated by antagonistic actors infiltrating the journalistic system without adhering to its norms or logic. Journalism itself has been ill-prepared to respond to this challenge, but journalism theory and research have also had trouble in grasping these phenomena. It is thus the aim of this article to propose a theoretical perspective on a specific set of antagonists characterized by its paradoxical nature. It is 'the excluded third, included' as described by Serres, the parasite that is both part of the system and its antagonist. From the perspective of systems theory, the parasite is a subsystem that threatens the integrity of the primary system. Thus, the parasite is defined by the relations that describe its position, its behaviour towards the host system. Due to these peculiarities—this contradiction, this vagueness—it evades a classical bivalent logic. This may be one reason why the paradoxical nature of the antagonist from within, the 'uninvited dinner guest,' has not been described as such until now. The parasitic practices follow the logic of the hacker: He is the digital manifestation of Serres' parasite. Accordingly, parasitic strategies can be described as news hacks whose attack vectors target a system's weak points with the help of specific strategies. In doing so, they not only change the system output but also compromise its values and exploit its resources.

Keywords
antagonists; attack vector; hacking; news hacks; parasite; Serres; systems theory

Issue
This article is part of the issue "Dark Participation in Online Communication: The World of the Wicked Web" edited by Thorsten Quandt (University of Münster, Germany).

1. Introduction

From a journalistic perspective, the digital age can be described as a phase of differentiation and de-differentiation (Wang, 2020). The clash of old and new media logics in the hybrid media system (Chadwick, 2017) creates a need to engage in boundary work (Carlson & Lewis, 2019). In our participatory media world, these processes of translation and synchronization do not take place exclusively within the journalistic system. Its blurred, fluid, and permeable boundaries also allow third parties to make an impact without having to subject to system norms. This is the type of actor this article is about—according to Serres, it is "the excluded third, included" (Serres, 2007, p. 76), or also: the parasite.

These actors are distinguished from other antagonistic actors by four particular characteristics: (1) They do not act from outside, they do not try to combat journalists through physical violence or oppression, such as censorship—they act from within; (2) by doing so unilaterally they avail themselves of journalistic resources; (3) they take advantage of the freedoms of a democratic public in pursuit of their strategic goals, compromising the very values on which these freedoms are based; and (4) they are an inherent part of the system (while at the same time being alien to the system) and therefore

cannot be eliminated without restricting the freedoms (or values) of the system itself.

In recent years, central terms have been coined and frameworks established, which now allow us to analytically describe and classify these actors and their strategies: The term 'information disorder' (Wardle & Derakhshan, 2017), for example, focuses on various forms of false information and its effects in the hybrid media system (see also Bayer et al., 2019); 'dark participation' (Quandt, 2018) denotes the destructive potential inherent in every participatory technology (see also 'data craft,' Acker, 2018; or 'digital influence machine,' Nadler, Crain, & Donovan, 2018; or 'alternative influence,' Lewis, 2018); 'source hacking' (Donovan & Friedberg, 2019) refers to strategies to manipulate digital journalism (see also Phillips, 2018). Despite this great wealth of different concepts, we still lack models to describe these relationships on a theoretical level. To close this gap, we draw on *The Parasite*, a work by French philosopher Michel Serres (2007). In addition to the biological meaning of the term parasite as an organism living off a host animal (p. 50), Serres (2007) uses an inductive-iterative approach to explore various linguistic levels of meaning of the term 'parasite'—as an uninvited house guest (p. 50) or as disruptive noise (p. 47), called a 'signal parasite' in French. Serres' analyses usually depart from an analogy (for example, a parable, see Section 2.2) and are then projected onto a more general context. Through this abstraction, he is able to overcome the classical bivalent logic, introducing the parasite as an inevitable third party. Serres' trivalent logic has also been incorporated by Luhmann into his systems theory (Luhmann, 2008). Here the parasite is the actor that undermines the functional differentiation of social systems and weakens the boundaries to their environments (Leanza, 2014).

As we study the dynamics of the digital world, the datafication of human behaviour (Couldry & Yu, 2018), the platformisation of the web ecosystem (Nieborg & Poell, 2018), and hacking as "digital parasitism" (Aradau, Blanke, & Greenway, 2019, p. 2548), it becomes clear how relevant and on-point Serres' and Luhmann's theoretical foundations are today. They allow us to classify the above-mentioned sample cases of antagonistic behaviour. For instance, source hacking practices such as "keyword squatting" (Donovan & Friedberg, 2019, p. 37) or "evidence collages" (Donovan & Friedberg, 2019, p. 26) can be described as parasitic news hacks that operate on an attack vector (see Section 4.3), in which they apply strategies, exploit vulnerabilities, manipulate output, compromise values, and appropriate resources.

The model of the news hack thus not only defines the constellation of actors in relation to journalism but also describes the process of a one-way transfer of resources and the gradual undermining of the system's values. It thus associates these parasitic practices with findings on the erosion of trust in the media (Newman, Fletcher, Schulz, Andı, & Nielsen, 2020, p. 14) or the normalization of right-wing rhetoric in the mainstream (Larson & McHendry, 2019, p. 518).

2. Framework: Theory of Parasites

2.1. Paradoxes

In order to underscore the desideratum of a new terminology for antagonistic actors, we will take up the figure of the "rogue actor" (Entman & Usher, 2018, pp. 302–303), a notion that suffers from an unresolved theoretical contradiction: According to Entman and Usher (2018), rogue actors violate norms while acting outside of them—they are actors who attack journalism from outside. At the same time, they use the codes of the system, for example, by spreading "entirely baseless false information formatted like traditional news" (Entman & Usher, 2018, p. 303). Their success is therefore also based on the fact that they do not act exclusively outside the norms, but rather partially harness them whenever it suits their need. We are facing the logical problem of the 'excluded third,' which Serres (2007) describes as follows: "Which is the third part? Or who or what is the third, in this logic of the trenchant decision? Is the third excluded or not? Here we have a trivalent logic where we expected only a bivalent one" (p. 78).

In the concept of 'dark participation,' this trivalence resonates as "negative, selfish or even deeply sinister contributions" (Quandt, 2018, p. 40) to the news-making process. It describes the alien intruder that negates the systemic process, in Serres' words: "The dark side of the system" (Serres, 2007, p. 61). The present study takes up Quandt's concept and adds a relational dimension to it. By describing actors as 'parasitic,' we do not attribute an ontological quality, but rather assign a relational position to antagonistic actors (Leanza, 2014, p. 37) within the journalistic system.

2.2. Analogies

The meaning and functioning of a parasitic element is best described via relationships, as Serres shows in his powerful analogies. In his rat parable, he draws on a traditional fairy tale to describe a chain of parasitic constellations (Serres, 2007, pp. 47–65). In keeping with the literal meaning of the Greek word parasite, which refers to an organism that feeds on another (from *pará*, 'beside,' and *sitos*, 'food'), the parable is about a city rat who is hosting his visiting cousin, a country rat, at a meal under his landlord's table. The metaphor of the uninvited dinner guests describes the transfer of resources, not only of food but also of social capital—the embarrassment of the rat-infested landlord contrasts with the country rat's admiration of his sophisticated cousin. On the other hand, the parable illustrates the reflexive mechanism (i.e., a mechanism that applies to itself, Luhmann, 1970) of a parasitic relationship: One parasite opens the door to the next by creating access.

In his systems theory, Niklas Luhmann specifically references the information-theoretical elements of Serres' concept of parasites. Like Serres, albeit in a less abstract way, he uses a parable to illustrate his concept of a parasitic subsystem, which is based on Serres. This is how he describes the commercialization of art:

> The opportunity to sell becomes a seduction, quickly attracting parasites who take advantage of this problematic relationship between art and the market, offering advice and mediation services: Please don't make the artwork too large; yes, make it original, because in the language of the market, originality means scarcity, but don't be too eccentric in your execution; perhaps it would be best if you 'redescribed' previous art styles, quote them, parody them, break common boundaries, but make sure the market can still recognize it and relate it to existing art. (Luhmann, 2008, p. 389, authors' translation)

2.3. Position

The parasite can be described as a subsystem, an intermediary that mediates between the inside and the outside of the system. Its actions are paradoxical in that it "confirms and rejects the operational closure of its host system" (Leanza, 2014, p. 28). Serres describes it as a "semiconductor" (Serres, 2007, p. 341), its relationships are "one-way streets" (Serres, 2007, p. 106). To this end, the parasite positions itself "in the most profitable positions, at the intersection of relations" (Serres, 2007, p. 107)—like highwaymen at trade route junctions. Parasites thus also highlight the fault lines of existing orders, their vulnerabilities: "The places where you can shake up a system" (Schmitt, 2011, p. 45, authors' translation).

2.4. Relation

The interaction between the embedded parasite and the system can be described by two relations. The first one simply is the parasite's draining of resources without providing any service in return (Serres, 2007, p. 59). This relation of imbalance inevitably leads to the destruction—of both the host and the parasite that only exists in relation to it. The second relation is less one-sided—it describes how the presence of the parasite affects the system. According to Serres, a parasite is also the "simplest and most general operator on the variability of systems" (Serres, 2007, p. 324). If we imagine the system as a network, a parasitic node "disidentifies from the network but continues to be appended to it" (Mejias, 2010, p. 615). These nodes introduce noise, interfering with the network "while forcing it to adjust to [their] presence" (Mejias, 2010, p. 615). For Serres, the parasite is, therefore, a disruption "that changes the order, and thus its meaning" (Serres, 2007, p. 313). We call this interdependence the 'compromising relation.'

2.5. Reflection

Lastly, the paradox of the parasite is also reflected in the fact that it creates something new. By acting in a different mode of relationality: "The parasite invents something new….He establishes an unjust pact; relative to the old type of balance, he builds a new one" (Serres, 2007, p. 95). The primary parasite thus enables the next parasite to position itself in a similar relation to the system: "Parasitic orders are parasite-enabling orders" (Leanza, 2014, p. 37, authors' translation). The reflexive mechanism of the parasite manifests itself as a chain.

In summary, a parasitic element is characterized by four points: (1) The parasite positions itself as an intermediary at system boundaries, more precisely, at existing fault lines. Its interdependence with the system can be described by two relations, which we call (2) a relation of imbalance and (3) a relation of compromise. (4) Its mechanism is reflexive, i.e., it acts as a catalyst for other similar parasitic relations.

3. Parasites of Journalism: A Relational Approach

3.1. The Eufunctional Relation

Serres' statement about parasites, in general, has always applied to the media system as well: "Real production is undoubtedly rare, for it attracts parasites that immediately make it something common and banal. Real production is unexpected and improbable; it overflows with information and is always immediately parasited" (Serres, 2007, p. 49). The interdependence between media is therefore always fraught with imbalance. In pure newspaper markets, this tendency towards parasitic mimicry used to be compensated for by the value of topicality—information loses its value over time and an information parasite who steals from the printing press would always face a certain time delay. However, the more this time lag was reduced, the more each wave of media innovation was accompanied by clamouring voices accusing the newcomer of parasitically profiting from the original production, with those who were once considered parasites later accusing the next crop of parasites of exploiting their one-sided advantage: Representatives of the printing press successively denounced telegraph companies and their ticker services (Kielbowicz, 2015, p. 27), then the radio (Patnode, 2011, p. 87), and later television (Davies, 1998, p. 25), as parasitic media.

These waves aptly illustrate the phenomenon of differentiation and de-differentiation of systems triggered by parasitic disruption. The evolution of the media ecology's complexities can be described as a re-calibration of system boundaries and a constant re-definition of what is 'inside' and what is 'outside' (see 'boundary work' of Carlson & Lewis, 2019).

It becomes clear that the system is capable of inclusion up to a certain point—that every differentiation

is accompanied by a parasitic stage of imbalance that lasts until a new equilibrium is established. This inclusion, however, presupposes that the host and parasite have certain structural similarities. In this context, Schneider distinguishes between parasites "that latch on to the reproductive process of a functional system in either a eufunctional or a dysfunctional way" (Schneider, 2014, p. 100, authors' translation). While eufunctional parasites push into the existing order, dysfunctional parasites don't strive for inclusion, but rather profit from their intermediate status.

3.2. The Dysfunctional Relation

The classical dysfunctional parasite, similar to Luhmann's example of the art market, is the actor that commodifies communication—for example through advertising or public relations. Here too, there is a transfer of resources: As McAllister (1996) writes, ads "leech credibility, like a parasite" (p. 140)—he further describes the one-sidedness of the relation as a promotional contamination of the host. The parasitic logic of advertising thus constantly demands new hosts providing ever new, fresh resources. The current stage of this escalation consists of para-journalistic forms that profit from a steady de-differentiation of journalistic system boundaries. Schauster, Ferrucci, and Neill (2016, p. 1416) emphasize that trust in journalism and its credibility suffer particularly from these new forms of advertising (e.g., Native Advertising or Brand Journalism).

The decisive difference between eu- and dysfunctional parasitism is therefore that the latter functions within the system without being fully integrated into it; it follows a logic that is foreign to the system, implanting this logic into the system in the form of noise, thus forcing it to adapt. This not only leads to a loss of resources but also to its values being compromised.

Similarly, we can locate other forms of persuasive communication in a dysfunctional relationship to these values and their vehicles. For instance, the often-observed paradoxical relationship between populism and journalism—Haller and Holt speak of "paradoxical populism" (Haller & Holt, 2018, p. 1665)—also fits the inherently paradoxical notion of the parasite. In fact, populism has repeatedly been described as a parasite of democratic values (Fournier, 2019; Urbinati, 2014). Fournier (2019) also points out its paradoxical nature when he writes that populism is "inherent to the features of constitutional democracy" (p. 381) while at the same time pursuing the "objective to destroy the same constitutional system" (p. 364). Urbinati (2014) stresses that once populism attains its goal of dominating a democratic state, "it can modify its figure radically" (p. 135). This mechanism of destructive modification from within is what we mean by the 'compromising relation.' Bayer et al. (2019) describe this process with the help of the populist method of strategic disinformation:

> False information in itself (if it does not violate others' reputation, for example) enjoys the protection of freedom of expression, but when the whole environment of public discourse becomes occupied and dominated by falsehood, it frustrates the primary purpose of freedom of expression. (Bayer et al., 2019, p. 79)

The parasite thus uses the freedoms of the system to compromise the values from which these very freedoms are derived. The strategy of compromising becomes particularly clear in the example of the "lying demagogue" (Hahl, Kim, & Zuckerman Sivan, 2018, p. 1), who violates various norms (in particular, truthfulness) so obviously that the establishment is compelled to distance itself from this actor: "But this very need by the establishment to distance itself from the lying demagogue lends credibility to his claim to be an authentic champion for those who feel disenfranchised by that establishment" (Hahl et al., 2018, p. 8). As a consequence, anyone who abides by the norms, who is not ostracized by the establishment as a pariah, appears "less obviously committed to challenging it" (Hahl et al., 2018, p. 8). This also, and primarily, concerns the authority of journalistic actors who act as watchdogs within the norms of a competitive democracy (Strömbäck, 2005, p. 332).

This pattern can be applied to various strategies employed by antagonistic actors. For example, the strategy of "leak forgery" (i.e., a politically instrumentalized data breach; see Donovan & Friedberg, 2019, p. 18) is parasitic to the value of anonymity and the journalistic freedom of source protection. The strategy of "pseudoscience" (Hartzell, 2018, p. 17), i.e., camouflaging strategic information as a study, etc., is parasitic to the deliberative value of rationality. The value of representation is compromised by dark participation, by "sock puppets" or "deep cover" strategies (Acker, 2018, p. 14). The freedom of identity and self-development is undermined by 'hate spins' (George, 2016), the value of political agonism (Mouffe, 2000) is undermined by "strategic controversy" (Lewis, 2018, p. 31), etc.

It remains unclear, however, what changes to the journalistic system enabled these parasitic chains. As was the case with earlier disruptions, the decisive impulse for differentiation is technological innovation. In contrast to earlier differentiation waves, however, this one has spawned an actor who opens the system boundaries to the outside, positioning itself as an intermediary on the system boundaries: the platform.

4. Parasites of the Digital Realm

The relationship between journalism and the major information intermediaries (Helberger, Kleinen-von Königslöw, & van der Noll, 2015)—i.e., the platforms Google, YouTube, Facebook, Instagram, Twitter, etc.—can be described as a process of differentiation and mutual dependence, attraction, and repulsion, aptly culminating in the paradoxical term 'frenemies' (Bell, 2015),

which means the same thing as Serres' notion of the 'included third.' In fact, in a relatively short period of time, platform companies managed to create a situation where the direction of the parasitic relation is not always evident. The example of the platforms makes it all too clear that this relation is a matter of perspective: In the polygon between user, platform, content provider, advertiser, etc., anyone can be host or parasite. In fact, the platform completely dissolves the bivalent distinction between parasite and host.

In the digital sphere, "life-processes must be converted into streams of data inputs for computer-based processing" (Couldry & Yu, 2018, p. 4473). The parasitic medium—a by-product of the actual interaction—materializes in this process of datafication. Nowhere is this more obvious than in the digital world, where any instance of communication opens a door to parasitic third-party use—datafication is a "legitimate means to *access*…people's behavior" (Van Dijck, 2014, p. 198, emphasis in the original). In a second logical step, the parasite itself acts as host, enabling and catalysing the very type of interaction whose secondary product it desires—to use another one of Serres' analogies (2007, p. 160): The "farmer parasites the fauna." In the digital world, this evolution of a parasitic logic is described as platformisation: "The penetration of *economic*, *governmental*, and *infrastructural* extensions of digital platforms into the web and app ecosystems, fundamentally affecting the operations of the cultural industries" (Nieborg & Poell, 2018, p. 4276, emphasis in the original). The parasite becomes an infrastructure in its own right, striving for a monopoly and exponentially increasing its influence.

Assuming the perspective of journalism, we quickly grasp the extent of the parasitic behaviour of the actor 'platform' towards the system. The relation of imbalance is most clearly reflected in the "parasitic relationship to news production" (Siapera, 2013, p. 17), along with a radical redistribution of advertising revenues in favour of digital platforms (Bell & Owen, 2017). They combine and scale up the above-described logics of advertising communication and competing media. At the same time, they leverage their position as a parasitic host to disruptively de- and recode system norms. For the Facebook algorithm, Caplan and Boyd (2018) describe this compromising effect as an isomorphism. They observed: "How algorithms structure disparate businesses and aims into an organizational field, leading them to change their goals and adopt new practices" (Caplan & Boyd, 2018, p. 2). In the tradition of isomorphism and bureaucracy research, algorithms are therefore to be considered an "extension of bureaucratic tools such as forms" (Caplan & Boyd, 2018, p. 3). Similarly, Serres describes the bureaucratic power "to edit the laws and to withdraw knowledge from the greatest number" (Serres, 2007, p. 98). This particular "theft of information" (Serres, 2007, p. 97) is also characteristic of the platforms' algorithms—they act as unproclaimed laws without having to be transparent. For journalism, this means that "by defining and redefining the concept of relevance or 'value' of information and news media, Facebook increasingly writes the rules, or code, that defines which content succeeds or fails" (Caplan & Boyd, 2018, p. 5).

4.1. Parasitic Infrastructure

But platforms are not the only parasites feeding on journalism. Due to the reflexive nature of parasitism described above, they also function as "opportunity structures" (Ernst, Esser, Blassnig, & Engesser, 2019, p. 170) for other actors to act co-parasitically alongside. They form a powerful sub-system that opens the boundaries of the journalistic system to various forms of attackers who use the platform logic as a parasitic infrastructure.

Bayer et al. (2019) note that "the interests of the technology providers (online platforms, social networks, and digital advertisers) and the actors behind [the] disinformation campaign[s] are to some extent aligned" (p. 31). Both seek the users' attention, yet rather than competing for this resource, they support each other symbiotically: The actors behind disinformation campaigns have a "full suite of services" (Bayer et al., 2019, p. 32) at their disposal, made available to them by the platforms. Central tools for this are monitoring, profiling, targeting, and automatic optimization of target publics, as well as having the timing, placement, and content of influence campaigns based on consumer data and real-time feedback (Nadler et al., 2018, p. 11). Many of these strategic communication techniques are not new—but the parasitic infrastructure "accelerates their reach, hones their precision, and offers the means to evade detection and penalties" (Nadler et al., 2018, p. 27). This behaviour contrasts starkly with the platforms' projected self-image as champions of democratic values such as neutrality and equality (Gillespie, 2010, p. 352). They "largely deny responsibility for quality and accuracy of the frames they disseminate and profit from, thereby giving rogue actors and ideological media power to distort democracy" (Entman & Usher, 2018, p. 306).

4.2. Border Crossers

The current omnipresence of political-strategic actors who position themselves mid-way between journalistic mainstream discourse and extremist ideologies can only be understood in the context of this co-parasitic synergy. Here too, parasitic relations are at play, following the paradoxical logic of the included third. From their intermediary position, they are able to draw attention away from the democratic discourse (relation of imbalance), to mobilize and recruit; while at the same time invoking the values of democratic discourse to claim a role as a legitimate spokesperson, thus successively undermining it (compromising relation).

They benefit from their "lack of ideological cohesion, leadership and organization" (Fielitz & Marcks,

2019, p. 7)—a blurry strategy in whose haze they can cross back and forth between the mainstream discourse, dominated by journalism, and the undemocratic outside. Accordingly, the 'alt-right' functions "as a rhetorical bridge between white nationalism and the mainstream public" (Hartzell, 2018, p. 24). As Hartzell (2018) explains, this intermediary position allows right-wing actors to either act in proximity to extremist ideology or to temporarily distance themselves from it (e.g., after a terrorist attack or right-wing riots), whatever is most opportune in a given situation (p. 24).

A parasite defines itself through its relation, which also means that it cannot exist without a reference point or host. Here, this need to act in relation to a host is reflected in the euphemistic claim to represent an alternative (see also Holt, Ustad Figenschou, & Frischlich, 2019; for alternative media; Lees, 2018; for 'Alternative for Germany,' etc.). As Mészáros (2005) aptly puts it, the strategy of these "alternative" groups consists in "dismissing their adversary with an *aprioristic negativity*, remaining thus entirely dependent (i.e., intellectually parasitic) on the arguments" (p. 257, emphasis in the original) of the other side.

The 'alternative' aspect is thus primarily a pose intended to legitimize the demand for visibility in the democratic discourse. Right-wing actors consequently present themselves as marginalized and discriminated against by the mainstream. On social media, they share stories about their 'coming out' as conservatives, demanding "ideological diversity" in the mainstream (Lewis, 2018, pp. 21–22). As Lewis (2018) points out, this countercultural positioning is misleading: "These influencers are adopting identity signals affiliated with previous countercultures, but the actual content of their arguments seeks to reinforce dominant cultural racial and gendered hierarchies" (p. 24).

It is therefore important to distinguish this mimicry from the phenomenon of counterpublics (Fraser, 1990), which can be classified as eufunctional parasites: Although they do not always strive for symbiosis with the dominant public, they share the basic values of the dominant system. In contrast, Larson and McHendry (2019) describe the alleged 'alternative' publics as "parasitic publics…that feed off of oppressive conditions in the public sphere by articulating with dominant discourses to exploit dominant publics' centripetal force" (p. 519). Here, too, the compromising relation becomes evident: The value of diversity, the inclusive centripetal force of the democratic public sphere, is abused in order to weaken it and thus cause a "societal norm shift" (Quandt, 2018, p. 43).

4.3. Parasitic Strategies

Contrary to the public image projected by its representatives (as 'alternative' or 'countercultural'), a parasitic element is not capable of creating something original on its own—its innovative power consists solely in putting existing things into new relations, channelling away resources, and compromising values. In a network-like information environment, the parasitic element is the node that generates noise, thus stirring up 'information disorder' (Wardle & Derakhshan, 2017): It disidentifies from the network by means of disinformation (false information shared to cause harm) or malinformation (genuine information shared to cause harm), thus forcing other nodes to adapt (or at least irritating them). One result, for instance, is the spread of misinformation (spreading false information without malicious intent; for the distinction between dis-, mal-, and misinformation, see Wardle & Derakhshan, 2017, p. 5).

The parasitic strategies stem from a variety of origins: Propaganda techniques (Lukito et al., 2020) work synergistically with (online) marketing tools (Donovan & Friedberg, 2019) and the insider knowledge of former journalists who now publish in partisan or alternative media (Phillips, 2018, p. 12). In parasitic chains, the function of such 'alternative' journalists is to "launder" (Donovan & Friedberg, 2019, pp. 16, 24), i.e., lend legitimacy to slogans and narratives from social media for populist actors to pick up and carry into the political discourse.

Moreover, many parasitic techniques can be traced back to the digital subculture of hacking or its subforms "social hacking" (Kerr & Lee, 2019, p. 11) or trolling (Phillips, 2018, p. 19). Like "a parasite, hacking draws all its strength, strategies and tools from the system on which and in which it operates" (Gunkel, 2001, p. 6). At the same time, hacking does not introduce anything new into the system: "It derives everything from the host's own protocols and procedures" (Gunkel, 2001, p. 6). According to Aradau et al. (2019), hacks are thus "'acts of digital parasitism,' which create parasitic interferences by working beside or alongside digital technologies and assembling collectives of coders and non-coders'' (p. 2548). In this sense, hacking can be understood as an overarching umbrella term for a wide variety of parasitic practices in the digital world.

The news hackers' strategies—called 'exploits' in IT security lingo—target the weak spots of the journalistic system or, to use the term from earlier, its predetermined fault lines. It is often difficult to identify whether such vulnerabilities are due to a lack of journalistic diligence, a lack of competence, or digital naivety ("anything x.0"; see Quandt, 2018, p. 38), or research practices that are prone to manipulation (Lukito et al., 2020; McGregor, 2019)—or whether it is simply a lack of financial resources to defend against attack. News hacks aim to manipulate editorial output, i.e., distort the tone, volume, and journalistic agenda (cf. media bias of Eberl, Boomgaarden, & Wagner, 2017). However, as we have seen, the attack vector does not end here—it also compromises values and exploits resources (Figure 1).

Thus, certain news hacks such as "keyword squatting" (co-opting of keywords or accounts related to breaking news events, social movements, etc., to manipulate

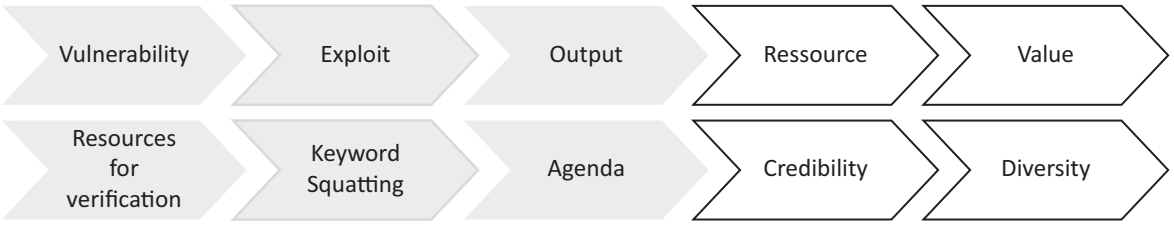


Figure 1. Parasitic attack vector. Note: The starting point is identifying a predetermined fault line or vulnerability (e.g., lack of resources, lack of competence etc.), in the case described here a lack of resources for verification of user-generated content; then an exploit is applied (keyword squatting, leak forgery, strategic controversy etc.), manipulating an output (volume, agenda or tonality of reporting) leading to resources being appropriated by the parasitic relations (e.g., information, reputation, credibility, trust, authority, attention) and values being compromised (e.g., truthfulness, rationality, representation, identity, agonism).

search traffic; Donovan & Friedberg, 2019, p. 37) or "evidence collages" (compiling a blend of verified and unverified information; Donovan & Friedberg, 2019, p. 26) not only prey on weaknesses in journalistic practices (e.g., lack of resources/competencies for verification, homophilia, metrics orientation, partisan bias, false balance) and distortions of journalistic content (e.g., manipulated saliences in the agenda). The relational-parasitic angle highlights that beyond this, news hacks unfold their real impact by also diminishing journalistic resources (credibility, authority, attention, etc.) and damaging values (participation, diversity, truthfulness, etc.).

5. Conclusions

The antagonist of the journalistic system described in this article is characterized by its paradoxical nature. It is the 'the excluded third, included' as described by Serres, the parasite that is both part of the system and its antagonist. From the perspective of systems theory, the parasite is a subsystem that threatens the integrity of the primary system.

The notion of the parasite is not new to the media system—as we have retraced in this article, the differentiation and integration of new media or logics that are alien to the system (e.g., advertising) and the resulting tensions can be described as parasitic relations. While these tensions between the host and structurally similar, eufunctional parasites can be resolved by way of de-differentiation or inclusion, the presence of dysfunctional parasites will successively weaken the system. These parasites do not strive for inclusion but benefit from their intermediate position, from which they drain resources (relation of imbalance) and introduce noise into the system to force a de-coding of system values (compromising relation).

Thus, the parasite is not defined by characteristics in the ontological sense, but mainly by certain relations that describe its position, its behaviour towards the host system. Due to these peculiarities—this contradiction, this vagueness—it evades a classical bivalent logic (Serres, 2007, p. 275). This may be one reason why the paradoxical nature of the antagonist from within, the 'uninvited dinner guest,' has not been described as such until now. The present work seeks to help close this gap by adding a relational perspective to concepts such as 'rogue actors' (Entman & Usher, 2018) or 'dark participation' (Quandt, 2018).

Intermediary platforms were identified as a central parasitic infrastructure in relation to which strategic communicators behave co-parasitically in the sense of a reflexive mechanism. Their practices follow the parasitic logic of the hack, which draws all its "strength, strategies and tools from the system" (Gunkel, 2001, p. 6). The hacker is the digital manifestation of Serres' parasite. Accordingly, parasitic strategies can be described as news hacks whose attack vectors target a system's weak points with the help of specific exploits. In doing so, they not only change the system output, but also compromise its constitutive values and exploit its vital resources.

This systematization of parasitic relations can serve as a starting point for future analyses of antagonistic actors and their practices. In this article, we were only able to merely allude that from the perspective of the journalistic system, various forms of strategic communication can be described as dysfunctional parasites, for example populist rhetoric, demagogy, but also practices of the alt-right. Follow-up case studies could be conducted to further deepen these analyses. For a comprehensive view, research projects could use the attack vector model, starting by questioning how journalists behave in the digital space and which journalistic practices make them vulnerable to attack (analytical focus: journalistic practice, as seen, for example, in McGregor, 2019; McGregor & Molyneux, 2020). The second step would be to identify and describe the strategies news hackers use to exploit these previously identified vulnerabilities, with a special focus on the technical infrastructure that makes the attack possible in the first place (analytical focus: parasitic practice and infrastructure, as seen, for example, in Donovan & Friedberg, 2019; Phillips, 2018). At the output level, the question is how content is manipulated, how parasitic strategies change reporting (analytical focus: content, as seen, for example,

in Lukito et al., 2020). These three levels and their associated mechanisms are put into an analytical context by their parasitic relations. For example, the relation of imbalance could be operationalized as a loss of news credibility, caused, for instance, by a lack of civility in user comments (for example in Prochazka, Weber, & Schweiger, 2018). At the same time, following the notion of a one-sided exchange, the question would be which actors benefit from this loss of resources; whether, for example, the loss of credibility on the journalistic side goes hand in hand with an increased willingness to believe in alternative media (see 'displacement effect,' for example, in Omar & Ahrari, 2020).

Lastly, the relation of compromise allows us to evaluate parasitic practices in the context of democratic values. News hacks are always based on liberal system norms, degrees of freedom that can be leveraged with various intentions—and not always in accordance with their original rationale. Parasites exploit this ambivalence, thus compromising the value of the norm. Measures taken against these harmful practices are therefore often directed against the freedoms and values themselves. This restrictive backlash can be observed at various levels: Newsrooms shut down their comment sections (Quandt, 2018, p. 37), platforms delete harmless content in response to manipulation attempts (Acker, 2018, p. 4), journalists' rights are restricted in favour of secret services' scope of authority (cf. the discussion on the German 'state Trojan' and digital source protection in Meister, 2020). Measures like these are a contradiction of the very system, which in turn can lead to loss of credibility and legitimacy. Referring to Popper, Fielitz and Marcks (2019) claim that this dilemma is an inevitable consequence, a "reloaded 'paradox of tolerance,'" "being intolerant of (liberal) structures producing intolerance" (p. 3). Destructive measures, however, can only be legitimate as a last resort and must be critically questioned if presented as the only alternative. Preference should be given to constructive measures that identify vulnerabilities and apply 'patches' to eliminate them (for example in journalistic training or in terms of resources).

One limitation of these perspectives is the fact that they are retrospective, meaning that they are only partially suited for prevention. If research does not want to be reduced to taking stock of damage done and reconstructing events that have already happened, it must find ways to anticipate future developments. As a possible solution to this communication science dilemma, Schäfer and Wessler (2020) suggest considering sociotechnical innovations "that are (potentially) relevant to public communication" (p. 308), and to thus identify potential risks at an early stage as "interventionist innovation research" (p. 309). In addition, it could be useful to look into IT security strategies (simulated attacks, honeypot scenarios, sensitization) or organizational and design principles for risk management, such as diversification, to see whether they could be adapted for preventive research.

Acknowledgments

The authors would like to thank the four anonymous reviewers as well as the Academic Editors for their valuable and constructive feedback on the manuscript.

Conflict of Interests

The authors declare no conflict of interests.

References

Acker, A. (2018). *Data craft: The manipulation of social media metadata*. New York, NY: Data & Society's Media Manipulation Research Initiative. Retrieved from https://datasociety.net/wp-content/uploads/2018/11/DS_Data_Craft_Manipulation_of_Social_Media_Metadata.pdf

Aradau, C., Blanke, T., & Greenway, G. (2019). Acts of digital parasitism: Hacking, humanitarian apps and platformisation. *New Media & Society*, *21*(11/12), 2548–2565. https://doi.org/10.1177/1461444819852589

Bayer, J., Bitiukova, N., Bard, P., Szakács, J., Alemanno, A., & Uszkiewicz, E. (2019). *Disinformation and propaganda: Impact on the functioning of the rule of law in the EU and its member states*. Brussels: European Parliament's Committee on Civil Liberties, Justice and Home Affairs (LIBE). Retrieved from https://www.europarl.europa.eu/RegData/etudes/STUD/2019/608864/IPOL_STU(2019)608864_EN.pdf

Bell, E. (2015, April 30). Google and Facebook are our frenemy. Beware. *Columbia Journalism Review*. https://www.cjr.org/analysis/google_facebook_frenemy.php

Bell, E. J., & Owen, T. (2017). *The platform press: How Silicon Valley reengineered journalism*. New York, NY: Tow Center for Digital Journalism.

Caplan, R., & Boyd, D. (2018). Isomorphism through algorithms: Institutional dependencies in the case of Facebook. *Big Data & Society*, *5*(1), 1–12. https://doi.org/10.1177/2053951718757253

Carlson, M., & Lewis, S. C. (2019). Boundary work. In K. Wahl-Jorgensen & T. Hanitzsch (Eds.), *The handbook of journalism studies* (2nd ed., pp. 123–135). London: Routledge.

Chadwick, A. (2017). *The hybrid media system*. Oxford: Oxford University Press.

Couldry, N., & Yu, J. (2018). Deconstructing datafication's brave new world. *New Media & Society*, *20*(12), 4473–4491. https://doi.org/10.1177/1461444818775968

Davies, D. R. (1998). From ridicule to respect: Newspapers' reaction to television, 1948–1960. *American Journalism*, *15*(4), 17–33. https://doi.org/10.1080/08821127.1998.10739140

Donovan, J., & Friedberg, B. (2019). *Source hacking: Media manipulation in practice*. New York, NY: Data

& Society's Media Manipulation Research Initiative. Retrieved from https://datasociety.net/wp-content/uploads/2019/09/Source-Hacking_Hi-res.pdf

Eberl, J.-M., Boomgaarden, H. G., & Wagner, M. (2017). one bias fits all? Three types of media bias and their effects on party preferences. *Communication Research*, *44*(8), 1125–1148. https://doi.org/10.1177/0093650215614364

Entman, R. M., & Usher, N. (2018). Framing in a fractured democracy: Impacts of digital technology on ideology, power and cascading network activation. *Journal of Communication*, *68*(2), 298–308. https://doi.org/10.1093/joc/jqx019

Ernst, N., Esser, F., Blassnig, S., & Engesser, S. (2019). Favorable opportunity structures for populist communication: Comparing different types of politicians and issues in social media, television and the press. *The International Journal of Press/Politics*, *24*(2), 165–188. https://doi.org/10.1177/1940161218819430

Fielitz, M., & Marcks, H. (2019). *Digital fascism: Challenges for the open society in times of social media*. Berkeley, CA: Berkeley Center for Right-Wing Studies. Retrieved from https://escholarship.org/uc/item/87w5c5gp

Fournier, T. (2019). From rhetoric to action, a constitutional analysis of populism. *German Law Journal*, *20*(3), 362–381. https://doi.org/10.1017/glj.2019.22

Fraser, N. (1990). Rethinking the public sphere: A contribution to the critique of actually existing democracy. *Social Text*, *25*. https://doi.org/10.2307/466240

George, C. (2016). *Hate spin: The manufacture of religious offense and its threat to democracy*. Cambridge, MA: The MIT Press.

Gillespie, T. (2010). The politics of 'platforms.' *New Media & Society*, *12*(3), 347–364. https://doi.org/10.1177/1461444809342738

Gunkel, D. J. (2001). *Hacking cyberspace*. Boulder, CO: Westview Press.

Hahl, O., Kim, M., & Zuckerman Sivan, E. W. (2018). The authentic appeal of the lying demagogue: Proclaiming the deeper truth about political illegitimacy. *American Sociological Review*, *83*(1), 1–33. https://doi.org/10.1177/0003122417749632

Haller, A., & Holt, K. (2018). Paradoxical populism: How PEGIDA relates to mainstream and alternative media. *Information, Communication & Society*. https://doi.org/10.1080/1369118X.2018.1449882

Hartzell, S. L. (2018). Alt-white: Conceptualizing the "alt-right" as a rhetorical bridge between white nationalism and mainstream public discourse. *Journal of Contemporary Rhetoric*, *8*(1/2), 6–25.

Helberger, N., Kleinen-von Königslöw, K., & van der Noll, R. (2015). Regulating the new information intermediaries as gatekeepers of information diversity. *Info*, *17*(6), 50–71. https://doi.org/10.1108/info-05-2015-0034

Holt, K., Ustad Figenschou, T., & Frischlich, L. (2019). Key dimensions of alternative news media. *Digital Journalism*, *7*(7), 860–869. https://doi.org/10.1080/21670811.2019.1625715

Kerr, E., & Lee, C. A. L. (2019). Trolls maintained: Baiting technological infrastructures of informational justice. *Information, Communication & Society*, 1–18. https://doi.org/10.1080/1369118X.2019.1623903

Kielbowicz, R. B. (2015). Regulating timeliness: Technologies, laws, and the news, 1840–1970. *Journalism & Communication Monographs*, *17*(1), 5–83. https://doi.org/10.1177/1077699014566380

Larson, K. R., & McHendry, G. F. (2019). Parasitic publics. *Rhetoric Society Quarterly*, *49*(5), 517–541. https://doi.org/10.1080/02773945.2019.1671986

Leanza, M. (2014). Grenzrauschen: Zur Figur des Parasiten in der Systemtheorie [Border noise: On the figure of the parasite in systems theory]. *BEHEMOTH: A Journal on Civilisation*, *Das Andere der Ordnung*, *28*. https://doi.org/10.6094/BEHEMOTH.2014.7.1.771

Lees, C. (2018). The 'Alternative for Germany': The rise of right-wing populism at the heart of Europe. *Politics*, *38*(3), 295–310. https://doi.org/10.1177/0263395718777718

Lewis, R. (2018). *Alternative influence: Broadcasting the reactionary right on YouTube*. New York, NY: Data & Society's Media Manipulation Research Initiative. https://datasociety.net/wp-content/uploads/2018/09/DS_Alternative_Influence.pdf

Luhmann, N. (1970). Reflexive Mechanismen [Reflexive mechanisms]. In N. Luhmann (Ed.), *Soziologische Aufklärung 1: Aufsätze zur Theorie sozialer Systeme* [Sociological education 1: Essays on the theory of social systems] (pp. 92–112). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-322-96984-2_5

Luhmann, N. (2008). *Schriften zu Kunst und Literatur* [Texts on art and literature]. Berlin: Suhrkamp.

Lukito, J., Suk, J., Zhang, Y., Doroshenko, L., Kim, S. J., Su, M.-H., Xia, Y., Freelon, D., & Wells, C. (2020). The wolves in sheep's clothing: How Russia's Internet research agency tweets appeared in U.S. news as vox populi. *The International Journal of Press/Politics*, *25*(2), 196–216. https://doi.org/10.1177/1940161219895215

McAllister, M. P. (1996). *The commercialization of American culture: New advertising, control, and democracy*. London: Sage.

McGregor, S. C. (2019). Social media as public opinion: How journalists use social media to represent public opinion. *Journalism*, *20*(8), 1070–1086. https://doi.org/10.1177/1464884919845458

McGregor, S. C., & Molyneux, L. (2020). Twitter's influence on news judgment: An experiment among journalists. *Journalism, 21*(5), 597–613. https://doi.org/10.1177/1464884918802975

Meister, A. (2020, June 18). Staatstrojaner für Geheimdienste: "Tritt die Regelung in Kraft, werden wir dagegen klagen" [State Trojan for intelligence ser-

vices: "If the regulation is enacted, we will appeal it"]. *Netzpolitik*. Retrieved from https://netzpolitik.org/2020/staatstrojaner-fuer-geheimdienste-tritt-die-regelung-in-kraft-werden-wir-dagegen-klagen

Mejias, U. A. (2010). The limits of networks as models for organizing the social. *New Media & Society*, *12*(4), 603–617. https://doi.org/10.1177/1461444809341392

Mészáros, I. (2005). *The power of ideology*. London: Zed Books.

Mouffe, C. (2000). *The democratic paradox*. London: Verso.

Nadler, A., Crain, M., & Donovan, J. (2018). *Weaponizing the digital influence machine: The political perils of online ad tech*. New York, NY: Data & Society's Media Manipulation research initiative. Retrieved from https://datasociety.net/library/weaponizing-the-digital-influence-machine

Newman, N., Fletcher, R., Schulz, A., Andı, S., & Nielsen, R. K. (2020). *Reuters Institute digital news report 2020*. Oxford: Reuters Institute for the Study of Journalism.

Nieborg, D. B., & Poell, T. (2018). The platformization of cultural production: Theorizing the contingent cultural commodity. *New Media & Society*, *20*(11), 4275–4292. https://doi.org/10.1177/1461444818769694

Omar, B., & Ahrari, S. (2020). Mainstream and non-mainstream media in Malaysia: Does lack of credibility lead to displacement? *Newspaper Research Journal*, *41*(2), 127–145. https://doi.org/10.1177/0739532920919825

Patnode, R. (2011). Friend, foe, or freeloader? Cooperation and competition between newspapers and radio in the early 1920s. *American Journalism*, *28*(1), 75–95. https://doi.org/10.1080/08821127.2011.10678182

Phillips, W. (2018). *The oxygen of amplification: Better practices for reporting on extremists, antagonists, and manipulators*. New York, NY: Data & Society's Media Manipulation Research Initiative. Retrieved from https://datasociety.net/wp-content/uploads/2018/05/1_PART_1_Oxygen_of_Amplification_DS.pdf

Prochazka, F., Weber, P., & Schweiger, W. (2018). Effects of civility and reasoning in user comments on perceived journalistic quality. *Journalism Studies*, *19*(1), 62–78. https://doi.org/10.1080/1461670X.2016.1161497

Quandt, T. (2018). Dark participation. *Media and Communication*, *6*(4), 36–48. https://doi.org/10.17645/mac.v6i4.1519

Schäfer, M. S., & Wessler, H. (2020). Öffentliche Kommunikation in Zeiten künstlicher Intelligenz: Warum und wie die Kommunikationswissenschaft Licht in die Black Box soziotechnischer Innovationen bringen sollte [Public communication in times of artificial intelligence: Why and how communication science should crack open the black box of sociotechnical innovations]. *Publizistik, 65*(3), 307–331. https://doi.org/10.1007/s11616-020-00592-6

Schauster, E. E., Ferrucci, P., & Neill, M. S. (2016). Native advertising is the new journalism: How deception affects social responsibility. *American Behavioral Scientist*, *60*(12), 1408–1424. https://doi.org/10.1177/0002764216660135

Schmitt, M. (2011). Parasitäre Strukturbildung: Einsichten aus System—und Netzwerktheorie in die Figur des Parasiten [Parasitic structure formation: Insights from system and network theory into the figure of the parasite]. In B. P. Priddat & M. Schmid (Eds.), *Korruption als Ordnung zweiter Art* [Corruption as an order of the second kind] (pp. 43–59). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-93011-4_3

Schneider, W. L. (2014). Parasiten sozialer Systeme [Parasites of social systems]. In B. Heintz & H. Tyrell (Eds.), *Interaktion—Organisation—Gesellschaft revisited* [Interaction—Organization—Society revisited] (pp. 86–108). Berlin: De Gruyter.

Serres, M. (2007). *The parasite*. Minneapolis, MN: University of Minnesota Press.

Siapera, E. (2013). Platform infomediation and journalism. *Culture Machine*, *14*, 1–28.

Strömbäck, J. (2005). In search of a standard: Four models of democracy and their normative implications for journalism. *Journalism Studies*, *6*(3), 331–345. https://doi.org/10.1080/14616700500131950

Urbinati, N. (2014). *Democracy disfigured: Opinion, truth, and the people*. Cambridge, MA: Harvard University Press.

Van Dijck, J. (2014). Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology. *Surveillance & Society*, *12*(2), 197–208. https://doi.org/10.24908/ss.v12i2.4776

Wang, Q. (2020). Differentiation and de-differentiation: The evolving power dynamics between news industry and tech industry. *Journalism & Mass Communication Quarterly*, *97*(2), 509–527. https://doi.org/10.1177/1077699020916809

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making* (Report No. 27). Strasbourg: Council of Europe.

About the Authors

Gerret von Nordheim is a Postdoctoral Researcher in the field of communication science at the University of Hamburg. His research focuses on intermedia effects in the hybrid media system, especially between journalism, social media, and populist actors. He is specialized in the field of computational methods.

Katharina Kleinen-von Königslöw is a Professor for Journalism and Communication Studies, with a focus on Digital Communication and Sustainability at the Universität Hamburg. Her research focuses on the impact of digitalisation and technological innovations on political communication and, in particular, the role of social network platforms and their use by citizens and political actors.

COGITATIO

Media and Communication (ISSN: 2183–2439)
2021, Volume 9, Issue 1, Pages 99–109
DOI: 10.17645/mac.v9i1.3418

Article

Communities of Darkness? Users and Uses of Anti-System Alternative Media between Audience and Community

Christian Schwarzenegger

Department of Media, Knowledge and Communication, University of Augsburg, 86159 Augsburg, Germany;
E-Mail: christian.schwarzenegger@phil.uni-augsburg.de

Submitted: 30 June 2020 | Accepted: 5 September 2020 | Published: 3 February 2021

Abstract
The hopes regarding the positive impact of the Internet and digital participation in civic society have faded in recent years. The digital realm is now increasingly discussed regarding its role in putting democracy in jeopardy and polarizing public debate by propagating extremist views and falsehoods. Likewise, the perception of so-called alternative media as beneficial carriers of counter-public spheres and as important complements to mainstream positions in social debate has flipped. Alternative media are now often associated with the "Wicked web" of disinformation, political populism, or even radicalization. Following Quandt's (2018) notion of 'dark participation' and Phillips and Milner's (2017) description of the Internet as ambivalent, this article asks, whether the same holds true for the users of alternative media: a segment of the audience traditionally discussed in terms of community, engagement, participation, and strong ideological identification with progressive political causes. Do users of 'dark' alternative media bond with their media in similar ways to constitute communities of darkness? Based on interviews with 35 users of alternative media from a left-leaning, right-wing, Russian-tied and/or conspiracy spectrum users, uses of alternative media are pictured as grey rather than black or white. The findings illuminate the ambivalences within alternative media users as audiences and communities. Ambivalences are found regarding the use of alternative sources as audience or community members, regarding a shared attitude of criticality and anti-systemness, which connects trans-medially and trans-ideologically, as well as the experienced comfort of community, which can become a main motivation for use.

Keywords
alternative media; anti-system; audience; community; dark participation; populism

Issue
This article is part of the issue "Dark Participation in Online Communication: The World of the Wicked Web" edited by Thorsten Quandt (University of Münster, Germany).

1. Introduction: Into Darkness

The tide has turned. The high hopes regarding the positive impact of the Internet and digital participation in civic society have faded over the last couple of years. We do live in times 'after the hype' (Kaun & Uldam, 2017) and optimistic aspirations that democracy would be enhanced by the blooming of social movements and digital media activism, allegedly enabled by the Internet, have sobered. The digital realm is now increasingly discussed in terms of its role in putting democracy in jeopardy and polarizing public debate by propagating extremist views and falsehoods. The dark side of digital media technologies is that they can also be means of suppression rather than tools for empowerment (Treré, 2016). Likewise, the perception of so-called alternative media as beneficial carriers of counter-public spheres giving voice to minority positions and critique as important correctives to mainstream positions in social debate has flipped. Previously romanticized oases of voice and deliberation providing a fresh breeze for democratic progress are now suspected to represent foul swamps of disinformation ecologies, vile tools for political agitation, or even drivers of radicalization.

But with alternative media now increasingly being discussed in the nexus of populist or extremist politics

(Holt, 2020) and regarding political agitation, disinformation dissemination, and its allowing propaganda to bypass the checks and balances of professional journalism, what does this reveal about their audiences? Traditionally, these users—a gravely under-researched and hardly known species—have been discussed in terms of community, engagement, participation, and their high levels of identification with alternative media products. When alternative media are considered tools for ill intent, does this suggest that their users are shrouded by darkness? And do they knowingly or unknowingly support vile causes through their participation? This article investigates the users, user communities, and usage of alternative media to better comprehend their relationship with sinister goals and anti-democratic tendencies.

Quandt (2018, p. 44) when proposing his concept of 'dark participation' emphasized that although there "is a large variety of participation behaviours that are evil, malevolent, and destructive," the future of digital communication "is not all doom and gloom," as the past was not all bright. By "adding some black to the pearly-white idealism of citizen engagement" (Quandt, 2018, p. 37) in digital contexts, one "might end up with a more appropriate grey," instead of the high hopes or sombre sorrows concerning the digital realm's role for democracy and communication. Other authors have also contributed to the "deconstruction of earlier, naïve ideas" (Quandt, 2018, p. 37) regarding the web as a happy place for the betterment of society. Phillips (2015) argues that malignant online practices, e.g., trolling, widely condemned as obscene and deviant, are not that deviant but must be understood as integral elements of digital culture, supported and nourished by a responsive environment. In this sense, practices may arguably be offensive, weird or obscene, but they are 'normal' and a characteristic of the online ecosystem nonetheless. Phillips and Milner (2017, p. 5) describe the Internet as ambivalent, and argue that digital communication practices can be "simultaneously antagonistic and social, creative and disruptive, humorous and barbed, [and hence] are too unwieldy, too variable across specific cases, to be essentialized as this as opposed to that." Considering digital culture as ambivalent "collapses and complicates binaries within a given tradition" (Phillips & Milner, 2017, p. 11), e.g., between alternative and mainstream, abhorrent and admirable, odd and normal, light and dark participation.

In order to approach users and user communities around alternative media, the article first provides a glimpse at the rich history of competing understandings of alternative media and impeding characteristics and features. This is followed by addressing the profound lack of research into the audiences and users of alternative media and discussing the little knowledge we have about it. Then, the findings of an exploratory and theory-generating interview study with users of "dark alternative media" are presented. The results aim to highlight that the users of alternative media cannot simply be classified based on the orientation or content of the platforms they tend to use, rather, their motives, practices, and identification with the alternative media are varied and ambivalent.

2. Alternative Media and Their Respective Mainstreams

There is a lot of 'conceptual confusion' (Holt, 2020) around the notion of alternative media, and we can look back on decades of rich debate on what constitutes alternative media and in how far they pose alternatives to exactly what. "Endless discussions about its key features and practices" (Hájek & Carpentier, 2015, p. 365) in academia have reinforced a conceptual binary between alternatives and their respective mainstream counterparts. The notion of 'alternative' implies that it must be a complementary, substitutional, additional, or simply different version to something else. As Holt, Ustad Figenschou, and Frischlich (2019) have described, the active positioning of so-called alternative media vis-à-vis an alleged mainstream is a key dimension for approaching and understanding alternative news media. Instead, this juxtaposition is also crucial for identity management and self-perception of (some) alternative media outlets, which feast on their status as a corrective or even as being an explicit opposition to the mainstream, including purportedly one-sided or incomplete representations of social reality proposed by legacy media (Figenschou & Ihlebæk, 2019; Holt, 2019). Furthermore, nurturing such a collocation of alternative and mainstream as opposing blocks, can also contribute to essentializing either side as uniform and suggest homogeneity and negate actual diversity and pluralism in the discussed area.

To help dissolve the binary between mainstream and alternative and find a more nuanced and sensitive understanding of the notion, Downing (2001) preferred to speak of alternative media with a dedicated political agenda as radical media. He described such radical media as being "generally small-scale and coming in many different forms, but with the common characteristic of presenting, proposing and providing alternative visions to hegemonic policies, priorities and perspectives'' (Downing, 2001, pp. v). While Downing foregrounded the political and 'resistance potential,' Atton (2002) in his take, emphasized a broader understanding of their transformative potential as a key characteristic of radical alternative media. The transformative potential can also extend beyond more narrow political contexts. Also, Fuchs has emphasized the potential of alternative media to resist and reform, when he modelled them as critical media (Fuchs, 2010). For Fuchs, alternative media as critical media question domination, express the positions of the oppressed, and dispute "for the advancement of a co-operative society" (Fuchs, 2010). Alternative media has traditionally been seen as associated with the progressive left and somewhat idealized by scholarship (Holt et al., 2019). This development was historically consistent with the hopes

and expectations regarding the democratic potential of digital participatory culture of the coming Internet.

'Alternative media in the service of repression' (Downing, 2001), on the contrary, were rather sidelined or neglected. While their existence and perilous potential were acknowledged, they mostly remained out of focus, and 'rebellious media,' fighting for more positively perceived causes in the reform of culture and society were brought to the mainstage of scholarship. In recent years, this has (slightly) changed and the other alternatives (Atton, 2006), typically depicted as darker alternatives to their positive counterparts, became more prominent in research (Haller, Holt, & de la Brosse, 2019). These darker alternatives are alternative media, which are often described as linked to political extremes, mostly far-right populism (Holt, 2019), to conspiracy theorists, or having ties to Russia. Instead of a potential, they are regarded as a peril for democracy and this recent research on alternative media has emphasized their role in the spread of disinformation and as drivers of political polarization or even radicalization. Studies on alternative media are hence likely to actually speak of very diverse phenomena using a similar vocabulary. Some authors such as Hájek and Carpentier (2015) hence advocate that media theory should better protect the 'alternative media signifier' against being too widely applied, lest it also be applied to those who have simply claimed the alternative label for themselves and who do not meet the requirements and features to be defined as such.

The apparent diversity of alternative media and the many versions of what they can be alternatives to suggest considering alternative and mainstream as a shifting continuum rather than absolute categories (Holt et al., 2019; Kenix, 2011; Rauch, 2016). For instance, based on her analysis of audience's understandings of the mainstream alternative dialectic, Rauch (2015) has described—similar to others (e.g., Atton, 2002)—how being an alternative can either relate to the product or the process. She speaks of organizational alternatives (for instance amateurism vs. professionalism, commercialism vs. non-commercial orientation) or of content alternatives (offering other views, other topics, voicing critique). With changing media landscapes, changing political systems, and evolving public debates, what is alternative at one point can become mainstream and vice versa (Kenix, 2011, p. 17). Holt et al. (2019) argue that alternative and mainstream hence must be interpreted as strictly contextual and relational. Speaking of alternative or mainstream then only makes sense 'in regard to' something:

> By considering alternative news media as an "alternative" and "in regard" to—allows to put them into context. It accommodates alternative news media inspired by diverse political (left as well as right wing), religious (e.g., fundamentalist or extreme liberal) or philosophical (e.g., animal rights) ideologies that outspokenly describe themselves as counter-hegemonic correctives to mainstream news media or are described as such by their audience or third parties. (Holt et al., 2019, p. 866)

I will follow this relational understanding and focus on such alternative media, typically regarded as dark alternatives, which are alternative regarding their political position and characterized by what Holt (2018) speaks of as anti-systemness. These alternative media outlets position themselves as opposed to the alleged mainstream media, which are regarded as representatives of the system and hence accomplices to the political establishment, distorting or concealing reality for their interest (and against the manipulated people). Such a kind of anti-systemness, combined with an anti-elitist and anti-establishment attitude has also been identified as a characteristic feature of populism (Krämer, 2018; Mazzoleni, 2008). Anti-systemness, however, is not necessarily populist, and traditionally alternative media as carriers of counter-public spheres have also featured this stance. Like alternative media, counter-public spheres (Fenton & Downey, 2003) also have a history of being romanticized in media and communication scholarship as inherently progressive and pro-democratic.

The anti-system stance supports the idea that alternative media platforms can blend well with populist politics and provide mutual sustenance (Holt, 2020). These media "do not have to follow commercial logic, journalistic conventions, or ethical principles: they can be as radical and polemical as they wish" (Noppari, Hiltunen, & Ahva, 2019, p. 26). A lack of commercial orientation has traditionally been seen as a characteristic of alternative media. However, not following journalistic conventions and being highly polemic and polarizing as well as partisan can also be part of a flourishing business model. Right-wing platforms such as *Breitbart* and *InfoWars* in the US; the Austrian *unzensuriert*, and the German *KenFM* espouse their alternative (political) views with ideological as well as financial interests.

Additionally, alternative media platforms can be attractive for users susceptible to political populism. For instance, studies have shown that alternative media are strongly featured and referenced in the (social media) communication by populist political actors (Bachl, 2018). Several studies have found similar strong links and mutual referencing between alternative media platforms and populist political parties or populist politicians (Haller & Holt, 2019). Yet, while new alternative media in the digital realm can cater to problematic causes and help pursue darker political goals, this cannot automatically be assumed for their users. The audience and users of alternative media remain widely unknown.

3. In Search of the Virtually Unknown—The Audiences and Users of Alternative Media

17 years ago, Downing (2003, p. 625) argued that there was "a distinctly disturbing gulf between our currently

fragmentary knowledge or debates concerning how audiences and readers use alternative media" in comparison to "the mass of descriptions and theorizations of alternative media," their potential impact and their role for social movements or in critical counter-public spheres. Back then, he described the research into audiences and usage of the myriad and ever-growing numbers and outlets within the alternative media spectrum as a "minimally developed" area of research. The "virtual absence" (Downing, 2003, p. 626) of audiences and users has since not changed drastically and audiences still remain the neglected foster child of research into alternative media. For instance, among the 50 chapters in the *Routledge Companion to Alternative and Community Media* (Atton, 2015), not one is specifically dedicated to the audience of alternative news media. There is, however, a section on the communities and identities, which form around the practices of producing, contributing to, and being part of an alternative news media cosmos. Especially with alternative media associated with or linked to particular social movements, people in the vicinity of alternative media are regarded as part of distinct or amorphic activist groups constituting 'interpretive communities' (Rauch, 2007), rather than news audiences. But the people who are served by and use alternative media, mostly remain out of the picture.

Especially regarding the users of the aforementioned darker alternatives, Noppari et al. (2019, p. 24) note that their users "have stereotypically been labelled as misguided and as having insufficient media literacy." The authors further see such characterizations as being mostly based on assumptions, as little empirical research has been done on users (and producers) of such media offerings. One of the rare exceptions is provided by Müller and Schulz (2019, p. 3) who also describe research on the users of alternative media as "scarce." In their own take on the audiences of "alternative media with an affinity for populism," Müller and Schulz (2019) focused on political attitudes and patterns of media use as predictors for the likeliness of exposure to alternative media. While they differentiated between occasional and frequent, recurring users of alternative media, details of how people make use of alternative sources and how this might play out in their media use or media repertoire over time, were not under scrutiny.

A study that looked more closely into the how and why of using alternative media platforms, or as they call them 'counter-media,' was presented by Noppari et al. (2019). Based on their interviews with users of Finnish right-wing alternative media platforms, they distinguish three different types of users. According to them, system sceptics, as a first type, can be described as societal outsiders, with strong political or ideological beliefs. This type of users shows high hopes into the counter-public sphere in which they actively participate and hope for a change of the system. They often actively support the alternative media they consume or try to share their views on their own social media profiles. Instead of being generally sceptical, the distrust and scepticism of the second type, agenda critics, was aimed at specific topics, in which they believed legacy media would push its own particular agenda and be hostile to their personal opinions. This type of user would share alternative media content and often belong to social media groups around them. For the first and second type, active contribution and association with the alternative media they use were relevant. The third type, the casually discontent, are critical of certain journalists or topics and seek alternative positions in alternative media but use them only sporadically. They did not show ideological commitment to the partisan alternative media but rather were characterized by what the authors call savvy scepticism and constant irony. A crucial takeaway from this study is that for dedicated users, seeking contact to and comfort in a community of likeminded people was fundamental. Antagonism against legacy media and their allegedly lopsided agenda was a vital building block of community efforts.

Particularly in today's high-choice digital media environments, fresh contact and first-time exposure to alternative media outlets can easily occur incidentally and without further intention or attention of would-be users. Due to algorithmic curation of content selection and content presentation, follow-up interaction can become more likely after the initial contact with a particular source (O'Callaghan, Greene, Conway, Carthy, & Cunningham, 2015). Initial contact can play a role in the process of red pilling, i.e., making fresh contact with alternative news media and then being attracted into a sphere of increasingly extremist content. Red pilling, as Marwick and Lewis (2017) describe, often begins with contact to a content which appears attractive (a topic, an idea they can identify with, or a style of presentation, e.g., dark humour) and eventually spreads from there. However, people who occasionally have contact with alternative media platforms are not automatically drawn into a spiral of radicalization (Munn, 2019).

The argument that people who do distrust legacy news media are likely to refrain from accessing alternative news sources and search for alternative and allegedly independent sources has been supported by multiple authors at different times (e.g., Jackob, 2010; Newman, Fletcher, Kalogeropoulos, Levy, & Nielsen, 2018); albeit with broad and diverse understandings of alternative media. According to Leung and Lee (2014), using Social Networking Sites for the purpose of news consumption is strongly related to coming across alternative media and being exposed to their content. Users in social media settings often struggle to differentiate the trash from the treasure and to identify the valuable or reliable sources from the dubious (Edgerly, 2017; Schwarzenegger, 2020; Tandoc et al., 2017); strategies of news authentication or verification tend to be either unsuccessful or pursued only until users' pre-existing assumptions about a topic are confirmed. In an exploratory study, Schwarzenegger (2020) compared digital news navigation and information verification strategies of users of alternative media

with non-users. He found that users of alternative media, in particular, made paradoxical calls for unbiased reporting and perceived legacy media coverage as biased or not neutral. They also thought highly of themselves as critical thinkers and competent at detecting wrong information as well as balancing the biases, when using alternative media as complementary sources. While they typically pictured themselves as sceptical against all sides, their media practices revealed that they were highly selective in their criticality and did not doubt alternative sources in the same way as legacy media.

As mentioned above, the users of alternative media are often rather discussed with regards to community building and identity work (see respective chapters in Atton, 2015) than in terms of audience research. Alternative media, like other digital communities in general, can become ideological touchstones for their users and provide them with a sense of belonging and community (Rauch, 2007) which might have a particular allure for those considered as 'societal outsiders' elsewhere. While the foundation for this sense of belonging can be found in ideological beliefs or coherent worldviews, the appeal of an online community can also be based on shared practices or expectations. For instance, Topinka (2018) described how users can sustain communities based on their shared sense of a twisted and somewhat abusive humour and the lustful breach of societal taboos. In this trajectory users of alternative media platforms can also find comfort by perceiving themselves as members of (at least loosely connected) collectives rather than being alone. Whether it is the political direction of the content, the support of being sceptical together, or the thrill of following something that is deemed harmful in the public eye is an empirical question. The community perspective can be beneficial for understanding what attracts users to alternative media and how users socialize and bond (both online and offline) even if they do not fully identify with media ideologically. But as Dagron (2007) has criticized, the image of alternative media communication as small, isolated, and pure forms of community communication does not correspond to reality anymore, if it ever did. Following Postill (2008), the dominant role of community as the all-encompassing gaze on users of activist and alternative online media can obscure the fact that the uses can be impure (Dagron, 2007) and ambivalent or casual and incidental. Digitalization has supported an unprecedented increase in news media platforms, by legacy media and alternative sources alike. In high-choice media environments, alternative media can become part of a media repertoire without their users even being aware of or giving weight to their alternativeness, as they encounter them embedded in social media environments, as they do with legacy media. In this sense, not assuming that users of alternative media can be classified as a close community of likeminded people, but discussing them in terms of audiences, "as the people who, in their capacity as social actors, are attending to, negotiating the meaning of, and sometimes participating in the multimodal processes initiated or carried out by institutional media" (Schrøder, 2019, p. 160). Audiences are typically researched regarding their motives and criteria regarding selection and interpretation of media and its contents. Hence, applying an audience perspective on alternative media users can help to unravel the granular motives and uses as well as the 'nuanced gratifications' (Sundar & Limperos, 2013) that can be found in the use of alternative media.

4. Method

This article is based on 35 guided interviews with users of alternative media. Alternative media titles were generally differentiated as left-leaning, right-wing, Russian-tied, or as belonging to the conspiracy/esoteric spectrum based on previous research (e.g., Bachl, 2018; Schweiger, 2017) which guided the recruitment of participants. Interviews were conducted with German, Austrian, and Suisse alternative media users (aged 22 to 63) in 2019 by the author and a team of student assistants in a research seminar; the analysis was performed by the author alone. The interviews were semi-structured and conversational, lasting between 43 and 113 minutes. Most interviews took place face-to-face, some via Skype. With permission, each interview was digitally recorded in its entirety and transcribed subsequently. Pseudonyms have been used for quoted material to protect the privacy of participants. The recruitment process for the study was organised in both consecutive and parallel steps. First, a series of media outlets related to the four different camps discussed here were included based on the previous research, self-positioning of the media, and public debate. The left-leaning outlets included *NachDenkSeiten* and *Rubikon*; the right-wing outlets comprised, for example, *Compact* and *PI-News*; the Russian-tied outlets contained, among others, *RT Deutsch* and *Sputnik News*; and the outlets belonging to the conspiracy/esoteric spectrum were e.g., *KenFM* and *kla.tv*. In a second consecutive step, users of these outlets were identified and contacted in two ways: Student assistants in the research seminar functioned as 'mediators' (Kristensen & Ravn, 2015), and were asked if they knew someone who was a dedicated, visible, or self-declared user of the identified alternative media outlets. Additionally, active users on the alternative media outlets' social media sites were identified based on their online practices and participation in the online comment sections and contacted via the respective social media channels. This was difficult, as the effort to make contact would often be filtered out, simply not seen, or wilfully ignored. As a third and complementary strategy, we followed a snowball principle (Goodman, 2011), asking interview partners who were already participating in the study whether they could recommend other users they knew who might be interested in participating in the study. The final sample included

users from different countries, which is rather an effect of a shared German-speaking mediascape and transnational connectivities among alternative media users. For instance, RT and Sputnik in their German versions cater to the different countries. Austrian right-wing alternative media find some resonance among German users (e.g., *unzensuriert*, *Wochenblick* and *Alles Roger*), whereas right-leaning German platforms are also used and shared in the neighbouring countries. Subjects of the interviews were: Participants' general media use and repertoires in the everyday now, and as remembered, their media ideologies and media beliefs including trust and criticism of the media, and their political orientation. Further, interviews addressed how the respondents initially came in contact with alternative media, and the role played by alternative media outlets for them both in their media repertoire and regarding their general worldview and outlook on (political) issues. Of particular importance in the interviews was the topic of community participation. Hence interviews addressed the question of how far and in which way respondents were partaking in community activities around alternative media and in how far they perceived their online activities as community-related. Data analysis followed a qualitative content analysis coding-scheme and was supported with QDA-Software f4-analysis. The analysis was guided by the deductive categories reflected in the interview guidelines (e.g., community participation, media ideologies and beliefs), which were applied to the material. In the process of coding, these deductive categories were inductively refined, and additional categories, as well as new subcategories, were introduced whenever new themes and issues emerged in the data. In the last step, the material was coded once again with the final coding scheme and theses were formulated as a result of the coding.

5. Findings: Users of Alternative Media beyond Dark and Light

The findings of the study are organized in three theses, each addressing a particular shade of grey and ambivalence among the communities of users and audiences of dark alternative media.

5.1. 1st Ambivalence: Sometimes Users are Just Users and Rather Audience Members Than Community Members

Users of alternative media—when addressed in research—are typically discussed in terms of identification and community. Following the study presented here, this is important for some, but is far from being relevant to all informants. On the contrary, users of all types of alternative sources in the research spectrum expressed that alternative media are one type of sources they turn to, in order to complete their picture or add additional layers to their information spectrum:

> In my opinion, no address or source always reports authentically and always brings the facts in the necessary depth. In my opinion there is no such thing. Instead, you have to get the information on the subject from diverse sources. (Ralf, 30)

The search for alternative viewpoints can be motivated by distrust or scepticism towards legacy media. As in previous research, the informants expressed that they frequently find mainstream coverage to be incomplete, biased, and omitting or concealing relevant aspects of the issue, partially to fit their own agenda or, equally important, for commercial interests. Besides political partisanship, blatant sensationalism and poor reporting also deter users. However, scepticism rarely turns into hostility towards the media or reproaches of wilful manipulation.

Rather than being hostile media illiterates, some of the informants display a rather sophisticated understanding of affordances and challenges of news reporting, e.g., time constraints for reporting and limited capacities that demand selection of topics and perspectives presented. Alternative media as news sources are then sometimes sought out as the other part and complementary voices, but without necessarily considering this information more trustworthy or complete. Quite the opposite, some of the users of *RT* or *Sputnik* argued that they would, of course, consider these platforms openly biased and driven by Russian agendas. However, as users assumed to know what the biases are, they thought they were able to balance or filter them on reception. Additionally, other media were believed to also carry biases, but in an obscured way. Similar observations could be made with right-wing media platforms, which were sometimes seen as amplifiers of German populist party AfD politics, but users would trust their individual competence to cope with this partisanship. Probably, this is an indicator of users overestimating their capability to verify information and check sources.

An interesting example of the variety of uses and motives that can result in recurrent consumption of alternative media sources is represented by Sabrina, a 53-year-old flight attendant and avid user of right-wing alternative media. She doesn't identify with their political views but recurrently frequents such platforms to check out "what they are up to next" while often thinking "this must be bollocks." Sabrina frequently engages in online-discussions and aims to debunk misinformation and advocate for a civilized discourse among online users. To this end, she wants to know what "the other camp" is currently discussing and sharing, to brace herself for arguments to expect and misinformation that is likely to be referred to in debates. A similar practice could also be observed in a different direction. Some media sceptical alternative users, with low levels of trust in legacy media, would even intensify their use of public service broadcasting news, to unravel "manipulations": "It's not that what they show didn't happen, but as soon as the interpretative framing begins, when they move away

from plain and simple reporting, there are other powers and interests at play" (Anton, 43).

Although some users will buy into virtually everything the alternative media offers and doubt the mainstream positions with similar intensity, it was quite common that also alternative media were handled with care and distanced scepticism. Informants like electronic engineering student Theodor expressed that they trust legacy media overall, which constitute by far the largest portion of their media diet. Yet, they garnish and complement it with alternative news media. Still, the expectation is that they will only occasionally find an opposing (ideological) counterweight to the mainstream, rather they access such sources for variation and nuance: "One is not necessarily more balanced informed I would say, but more varied….In any case, you get a larger overview, which address others or concern others or something like that" (Theodor, 26).

Some informants have rather sporadic contact with particular alternative media titles, while others follow them continuously as part of their media repertoires. But, and this is important, some treat them as sources among other sources, neither privileged nor condoned, others rather trust legacy media and use alternative sources for cross-referencing, while others rather believe alternative sources and use them to check and challenge mainstream reporting. Users do not necessarily identify with them but become a—frequently critical, amused, and oppositional—audience.

5.2. 2nd Ambivalence: Alternative Media Users are Diverse but Anti-Systemness Connects Trans-Ideologically and Trans-Medially

Although informants were recruited based on their use of at least one 'dark' alternative media title, the interviews demonstrate, that users hardly ever remained exclusive users of just one platform. As they reported, the share of alternative media they frequented grew over time. This modification of the composition of media repertoires was partly furthered by digital recommender systems but also following the personal recommendations of other users and online commentators. One informant explains:

> It's the thirst for knowledge, the curiosity, that stimulates it and I think understanding. If I may go back to the picture of the puzzle, the puzzle grows exponentially. So it is growing, in comparison, if I only focus on the mainstream media, then I might just have the edge together…through the alternative media and the mixture with the established media, but I also manage to put the inner part together slowly and leisurely and part by part. At least I am convinced of that [laughing]. (Bettina, 38)

The example of Luise (59) demonstrates that the growth of alternative news media rarely follows a clear-cut political agenda nor happens entrenched in ideologies or political camps. Hence, classifying users as either left-wing or right-wing sympathizers based on the particular platforms they use is problematic. Luise's initial contact with alternative media was with rather esoteric titles, which are also crosscutting (partially cloaked) right-wing narratives. From there, her diet expanded via Facebook and YouTube to include alternative media allocated to the conspiracy and left-leaning spectrum. This pattern was very common, with the committed users of right-wing alternative media being least likely to expand beyond their own segment and the left-leaning and conspiracy titles being most commonly part of a combined diet. Luise's user type represents an ideological bricoleur, taking fragments of different ideological and political camps as long as they fit an overall orientation of discontent, anti-systemness, and critique. The platform *KenFm*, regarded as a conspiracy platform, was popular with users who would otherwise either prefer left-leaning or right-wing platforms. Also, media with links to Russia were popular across political orientations or did blend with all other types of alternative media in the study. This orientation-dominated use of alternative media outlets became also apparent in users from different countries who could relate to the anti-system reasoning provided by the alternative media and transferred stances to their own particular original political contexts.

In sum, the same alternative media platforms are used by users with a variety of political and ideological orientations and different motivations. At the same time, users with highly different backgrounds and orientations may pick their arguments and pieces of information from the same or similar alternative media outlets. This ideological flexibility was not limited to picking from a wide array of ideologies from alternative media, legacy media content, e.g., public broadcasting satire format Die Anstalt was also highly acclaimed for its anti-systemness. These findings foster the importance of media repertoire or media ecology perspectives, which do not analyse particular platforms and their users in isolation, but in a relational perspective, and with regard to the use of other news sources, both legacy and alternative media.

Users of alternative media may cut across different ideological areas and not belong to coherent communities around them, yet still have common attitudes and features. Across all types of alternative media under scrutiny users typically thought of themselves as being very critical media users and critical thinkers in general, better informed and more knowledgeable than regular media audiences and their acquaintances:

> Much more critical than others. I notice that in conversations with others. They are clearly less informed than me, they have less general knowledge and yes, but some of them don't want to, they are less interested. I can say that quite clearly, I am obviously more critical than the average. (Michael, 52)

These third-person effects of users of alternative media, who would consider themselves savvier and more competent to identify disinformation are well in line with previous research. Schwarzenegger (2020) has shown, however, that high confidence in one's competence to make sense of and assess the veracity of information does not necessarily reflect actual skill but may keep users from actively challenging information.

5.3. 3rd Ambivalence: The Comfort of Community Can Outshine Ideologies, Bringing Light to the Darkness

So far, I have shown that users of dark alternative media do not necessarily convert to a community. But there are also cases in which the cosy experience of belonging and sharing commonalities is crucial, and can become even more important than the alternative news per se. The comfort of not being alone in their scepticism, the feeling of being understood by likeminded people and not considered weirdos, loonies, or conspiracy theorists for their divergent views—an imminent fear of some of the informants—is an illuminating experience for users within their alleged darkness:

> You can discuss about everything and you are not looked at crookedly if you have a different opinion about the news from the traditional media. So that you can exchange views and also say that what the other media say is often nonsense. So that you also know that others also recognize this and you are not alone. (Felix, 30)

Some users expressed that they rather refrain from trying to convince others of their alternative views or to openly convert them. Partly, because they want to avoid confrontation or objection, partly because 'waking up' is someone everyone needs to do on their own terms. In this case, participants seek and find support, social interaction, and validation in the social media communities built around alternative media outlets. The significance of the community for individual users is best illustrated by the case of 38-year-old Bettina. She is an avid user of a platform which belongs to the conspiracy spectrum of the sample and is notorious for its anti-systemness. Since Bettina started following this platform and its YouTube channel, she also engaged in their Facebook fan group. The group name indicates it is open 'for system critics' only and Bettina since starting out with a few user-comments she made it to the rank of group administrator. Outside the web, Bettina is a trans-woman. In the interview, she described that since her transition and due to her current personal situation—working lots of night shifts, being alone at work and mostly alone during the days—for her, engaging in online discussions and alternative media related online groups is one of a few "chances to talk to people." Through her personal background, she is sensitive regarding gender-related issues, which she thinks are blatantly addressed badly in mainstream and alternative media alike. In the group, however, she only rarely engages in discussion when "gender stuff" is addressed. Bettina does not want to be "outed" in the group as trans and does not want to endanger her status within the group. At least occasionally, for Bettina, the sense of community and belonging she experiences in the group can outshine the ideology. The personal situation and experience of Bettina is certainly exceptional. But her case is nonetheless a focal glass for other informants' experiences. Peaceful co-existence in like-mindedness is also highlighted as an important community feature by Luise and Marianne, who are also part of an alternative media fan community on Facebook. "It is important that we are on the same page, share similar viewpoints, are on the same level. Neither too far to the left, nor to the right" (Marianne, 62). However, while the demand for non-radical and balanced positions was very common and positions too extreme were considered out of place, it was the users of right-wing alternative media in particular who would object to a qualification of the media outlets they used as far-right, and would rather reframe them as conservative or "how the common people think." They were hence generalizing and normalizing the worldviews they found in these media. Communities provided them with an environment that helped imagine right-wing worldviews as commonsensical.

6. Conclusions

Over the last couple of years, a growing interest in alternative media, especially those in the digital realm and catering to an anti-system stance can be observed in media and communication research. However, the notoriously "weak appetite for user research within alternative media" diagnosed by Downing (2003, p. 627) was not stimulated in the same fashion as the general attention for alternative media has increased. This study set out to generate an appetite for further investigative research into audiences and users of allegedly dark alternative media and their communities of darkness.

The findings sketched here suggest that users of alternative media are not a homogenous lot: their practices, motives, and orientations are more nuanced than simply aligning a populist media with a populist audience. However, at this point, this study does not aim to provide a renewed definition of the audience of the wide diversity of platforms and outlets referred to as alternative media. Instead, the empirical part of the study set out to question some purported certainties and to advocate for greater ambivalence when researching users of alternative media. Treré (2020) highlights the potential of recognizing the ambivalent nature of digital media and communication practices. According to him, this does not equal simply acknowledging that technologies can be used for the good or the bad. Instead, it means critically charting the social, cultural, and political conditions under which certain kinds of media practices, technological

appropriations, and media imaginaries were generated, combined, and implemented by concrete individual and collective actors in specific historical contexts. Thinking of the users of alternative media as both audiences and communities enable the highlighting and foregrounding of particular moments of their engagement with the media in their repertoire, which can be caused by not only various but even contradictory impulses. At the same time, it also prevents hasty universalism based on an overemphasis or neglect of community aspects.

The three theses presented here aimed to illuminate the ambivalences within alternative media audiences and open avenues for future research. First, when alternative media use is not necessarily linked to ideological identification and community participation, but can simply be part of diversified media diets in high-choice media environments, it is important to learn more about the effect and long-term impact that the use of alternative media can have for the composition and interplay of media in media repertoires over time. Second, the findings invite further investigation of the commonalities and differences across alternative media and users from different political camps and ideologies in trans-ideological and transmedial combinations. The cross-references, entanglements, relations, interdependencies, and mutual influences that the use of alternative media has on the media repertoire, the information horizon, and political participation and orientation can only be understood in the long run through a transmedia or media ecology perspective.

This can help prevent claims of exceptionalism and universalism alike. It further helps to understand the potential problematic impact that alternative media and their anti-systemness can have on public discourse at large, beyond the sometimes irrelevant niches in which they circulate and reinforce their positions (Holt, 2020). At the same time, it can help deconstruct superficial views of "pure" users, that qualify them according to their assumed political ideologies, as conspiracy-loons with insufficient media literacy or even prone for radicalization based on an alleged impact the alternative media may have on their users—an overemphasis of effects and identification supported by the community view, which would be considered an overcome position in audience research. Third, the results suggest that the role of belonging and community require further attention. Media literacy programs and initiatives which raise awareness of the perils of online disinformation and propaganda need take into account that fact-checking and literacy can't be effective antidotes when their side effect would be the dissolution of users' important personal social bonds.

As with every kind of self-reported data, one has to be careful about particular elements of the users' accounts. It would be naïve to simply take the self-presentation as critical thinkers, well informed and open-minded citizens equally wary against falsehoods from all sides for granted. Further, it could be said that the findings of this study are influenced by the sample composition. In all steps, the recruitment process was tedious and not without setback: Besides a rather typical low response rate when recruiting via social media, several people who were approached as potential participants declined for a variety of research-related reasons. For instance, people feared that the findings would be used to further discredit alternative media, which in the view of some respondents were treated unfairly in public debate. Another line of reasoning when declining to participate can be explained by the anti-system stance described above: University researchers would then be seen as associated with the system if not even representatives of what they saw themselves in opposition to. A third line of decline was due to the EU's General Data Protection Regulation and the requirement to fill in official University forms to express consent, as users did not want to be associated and/or recorded in some cases. In general, the responses of those who actively declined the invitation to participate suggest that those users who were especially highly entrenched in the ideological camps behind the respective alternative media and with a strong anti-systemness were likely not to participate and hence rather moderate users may have been open to contributing to the study.

The lack of political hardliners, the absence of stubborn ideologists and of incompetent media-illiterates may signal too reflexive, too openminded, and too few misguided, ill-informed views. However, there were some of these negatives in the sample, and probably even more so among the users who refused to participate in this study. But at the same time, this limitation is also a main takeaway from the study. Users and uses of alternative media are ambivalent, not a secluded community of darkness, but diverse people with diverse backgrounds and motivations, who happen to use alternative media as part of their media repertoires. They are neither black nor white. There is a whole lot of grey to discover around them.

Acknowledgments

This research has greatly benefited from the support with recruitment and interviews by the 19 participants of my MA research Seminar 2019 in Augsburg, and my three student assistants Aliscia Albani, Julia Bartsch and Sabrina Zierer. I would also like to thank the Academic Editor and all four anonymous reviews for their meticulous feedback and rich suggestions in finalizing this article.

Conflict of Interests

The author declares no conflict of interests.

References

Atton, C. (2002). *Alternative media*. London and Thousand Oaks, CA: Sage.

Atton, C. (2006). Far-right media on the Internet: Culture, discourse and power. *New Media & Society*, *8*(4), 573–587. https://doi.org/10.1177/1461444806065653

Atton, C. (Ed.). (2015). *The Routledge companion to alternative and community media*. London and New York, NY: Routledge.

Bachl, M. (2018). (Alternative) media sources in AfD-centered Facebook discussions. *Studies in Communication | Media*, *7*(2), 256–270.

Dagron, A. G. (2007). Call me impure: Myths and paradigms of participatory communication. In L. K. Fuller (Ed.), *The power of global community media* (pp. 197–207). New York, NY: Palgrave Macmillan US. https://doi.org/10.1007/978-1-137-01625-6_18

Downing, J. (2001). *Radical media: Rebellious communication and social movements*. Thousand Oaks, CA: Sage.

Downing, J. D. H. (2003). Audiences and readers of alternative media: The absent lure of the virtually unknown. *Media, Culture & Society*, *25*(5), 625–645. https://doi.org/10.1177/01634437030255004

Edgerly, S. (2017). Seeking out and avoiding the news media: Young adults' proposed strategies for obtaining current events information. *Mass Communication and Society*, *20*(3), 358–377. https://doi.org/10.1080/15205436.2016.1262424

Fenton, N., & Downey, J. (2003). Counter public spheres and global modernity. *Javnost: The Public*, *10*(1), 15–32. https://doi.org/10.1080/13183222.2003.11008819

Figenschou, T. U., & Ihlebæk, K. A. (2019). Challenging journalistic authority: Media criticism in far-right alternative media. *Journalism Studies*, *20*(9), 1221–1237. https://doi.org/10.1080/1461670X.2018.1500868

Fuchs, C. (2010). Alternative media as critical media. *European Journal of Social Theory*, *13*(2), 173–192. https://doi.org/10.1177/1368431010362294

Goodman, L. A. (2011). Comment: On respondent-driven sampling and snowball sampling in hard-to-reach populations and snowball sampling not in hard-to-reach populations. *Sociological Methodology*, *41*(1), 347–353. https://doi.org/10.1111/j.1467-9531.2011.01242.x

Hájek, R., & Carpentier, N. (2015). Alternative mainstream media in the Czech Republic: Beyond the dichotomy of alterative and mainstream media. *Continuum*, *29*(3), 365–382. https://doi.org/10.1080/10304312.2014.986061

Haller, A., & Holt, K. (2019). Paradoxical populism: How PEGIDA relates to mainstream and alternative media. *Information, Communication & Society*, *22*(12), 1665–1680. https://doi.org/10.1080/1369118X.2018.1449882

Haller, A., Holt, K., & de la Brosse, R. (2019). The 'other' alternatives: Political right-wing alternative media. *Journal of Alternative and Community Media*, *4*(1), 1–6. https://doi.org/10.1386/joacm_00039_2

Holt, K. (2018). Alternative media and the notion of anti-systemness: Towards an analytical framework. *Media and Communication*, *6*(4), 49. https://doi.org/10.17645/mac.v6i4.1467

Holt, K. (2019). *Right-wing alternative media*. London and New York, NY: Routledge.

Holt, K. (2020). Populism and alternative media. In B. Krämer & C. Holtz-Bacha (Eds.), *Perspectives on populism and the media* (pp. 201–214). Baden-Baden: Nomos.

Holt, K., Ustad Figenschou, T., & Frischlich, L. (2019). Key dimensions of alternative news media. *Digital Journalism*, *7*(7), 860–869. https://doi.org/10.1080/21670811.2019.1625715

Jackob, N. (2010). No alternatives? The relationship between perceived media dependency, use of alternative information sources, and general trust in mass media. *International Journal of Communication*, *4*, 589–606.

Kaun, A., & Uldam, J. (2017). Digital activism: After the hype. *New Media & Society*, *20*(6), 2099–2106. https://doi.org/10.1177/1461444817731924

Kenix, L. J. (2011). *Alternative and mainstream media: The converging spectrum*. London: Bloomsbury.

Krämer, B. (2018). Populism, media, and the form of society. *Communication Theory*, *28*(4), 444–465. https://doi.org/10.1093/ct/qty017

Kristensen, G. K., & Ravn, M. N. (2015). The voices heard and the voices silenced: Recruitment processes in qualitative interviews studies. *Qualitative Research*, *15*(6), 722–737. https://doi.org/10.1177/1468794114567496

Leung, D. K. K., & Lee, F. L. F. (2014). Cultivating an active online counterpublic: Examining usage and political impact of Internet alternative media. *The International Journal of Press/Politics*, *19*(3), 340–359. https://doi.org/10.1177/1940161214530787

Marwick, A. E., & Lewis, R. (2017). *Media manipulation and disinformation online*. New York, NY: Data & Society Research Institute. Retrieved from https://datasociety.net/pubs/oh/DataAndSociety_MediaManipulationAndDisinformationOnline.pdf

Mazzoleni, G. (2008). Populism and the media. In D. Albertazzi & D. McDonnell (Eds.), *Twenty-first century populism* (pp. 49–64). London: Palgrave Macmillan UK. https://doi.org/10.1057/9780230592100_4

Müller, P., & Schulz, A. (2019). Alternative media for a populist audience? Exploring political and media use predictors of exposure to Breitbart, Sputnik, and Co. *Information, Communication & Society*. Advance online publication. https://doi.org/10.1080/1369118X.2019.1646778

Munn, L. (2019). Alt-right pipeline: Individual journeys to extremism online. *First Monday*, *24*(6).

Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D., & Nielsen, R. K. (2018). *Reuters institute digital news report 2018*. Oxford: Reuters Institute for the Study

of Journalism.

Noppari, E., Hiltunen, I., & Ahva, L. (2019). User profiles for populist counter-media websites in Finland. *Journal of Alternative and Community Media*, *4*(1), 23–37.

O'Callaghan, D., Greene, D., Conway, M., Carthy, J., & Cunningham, P. (2015). Down the (white) rabbit hole: The extreme right and online recommender systems. *Social Science Computer Review*, *33*(4), 459–478. https://doi.org/10.1177/0894439314555329

Phillips, W. (2015). *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. Cambridge, MA, and London: The MIT Press.

Phillips, W., & Milner, R. M. (2017). *The ambivalent Internet: Mischief, oddity, and antagonism online*. Cambridge: Polity.

Postill, J. (2008). Localizing the Internet beyond communities and networks. *New Media & Society*, *10*(3), 413–431.

Quandt, T. (2018). Dark participation. *Media and Communication*, *6*(4), 36-48. https://doi.org/10.17645/mac.v6i4.1519

Rauch, J. (2007). Activists as interpretive communities: Rituals of consumption and interaction in an alternative media audience. *Media, Culture & Society*, *29*(6), 994–1013. https://doi.org/10.1177/0163443707084345

Rauch, J. (2015). Exploring the alternative-mainstream dialectic: What "alternative media" means to a hybrid audience. *Communication, Culture & Critique*, *8*(1), 124–143. https://doi.org/10.1111/cccr.12068

Rauch, J. (2016). Are there still alternatives? Relationships between alternative media and mainstream media in a converged environment: Reconceiving alternative and mainstream media. *Sociology Compass*, *10*(9), 756–767. https://doi.org/10.1111/soc4.12403

Schrøder, K. C. (2019). Audience reception research in a post-broadcasting digital age. *Television & New Media*, *20*(2), 155–169. https://doi.org/10.1177/1527476418811114

Schwarzenegger, C. (2020). Personal epistemologies of the media: Selective criticality, pragmatic trust, and competence–confidence in navigating media repertoires in the digital age. *New Media & Society*, *22*(2), 361–377. https://doi.org/10.1177/1461444819856919

Schweiger, W. (2017). *Der (des)informierte Bürger im Netz* [The disinformed citizen in the web]. Springer: Wiesbaden.

Sundar, S. S., & Limperos, A. M. (2013). Uses and Grats 2.0: New gratifications for new media. *Journal of Broadcasting & Electronic Media*, *57*(4), 504–525. https://doi.org/10.1080/08838151.2013.845827

Tandoc, E. C., Ling, R., Westlund, O., Duffy, A., Goh, D., & Zheng Wei, L. (2017). Audiences' acts of authentication in the age of fake news: A conceptual framework. *New Media & Society*, *20*(8), 2745–2763. https://doi.org/10.1177/1461444817731756

Topinka, R. J. (2018). Politically incorrect participatory media: Racist nationalism on r/ImGoingToHellForThis. *New Media & Society*, *20*(5), 2050–2069. https://doi.org/10.1177/1461444817712516

Treré, E. (2016). The dark side of digital politics: Understanding the algorithmic manufacturing of consent and the hindering of online dissidence. *IDS Bulletin*, *41*(1), 127–138. https://doi.org/10.19088/1968-2016.1

Treré, E. (2020). *Hybrid, media, activism: Ecologies, imaginaries, algorithms*. London and New York, NY: Routledge.

About the Author

Christian Schwarzenegger (PhD) is a Researcher and Lecturer (Akademischer Rat) at the University of Augsburg, Germany, and from October 2020 to June 2021 a Visiting Professor at the University of Salzburg, Austria. His research interests include the impact of digital transformation on culture and society with an emphasis on everyday life and participation in the digital society. Christian has recently published his research in journals such as *New Media and Society*, *Digital Journalism*, and *Convergence*.

Media and Communication (ISSN: 2183–2439)
2021, Volume 9, Issue 1, Pages 110–119
DOI: 10.17645/mac.v9i1.3331

Article

What Is (Fake) News? Analyzing News Values (and More) in Fake Stories

Edson C. Tandoc Jr. [1,*], Ryan J. Thomas [2] and Lauren Bishop [2]

[1] Wee Kim Wee School of Communication and Information, Nanyang Technological University, 637718, Singapore; E-Mail: edson@ntu.edu.sg
[2] Missouri School of Journalism, University of Missouri, Columbia, MO 65211, USA; E-Mails: thomasrj@missouri.edu (R.J.T.), laurenkbishop@mail.missouri.edu (L.B.)

* Corresponding author

Submitted: 9 June 2020 | Accepted: 9 August 2020 | Published: 3 February 2021

Abstract
'Fake news' has been a topic of controversy during and following the 2016 U.S. presidential election. Much of the scholarship on it to date has focused on the 'fakeness' of fake news, illuminating the kinds of deception involved and the motivations of those who deceive. This study looks at the 'newsness' of fake news by examining the extent to which it imitates the characteristics and conventions of traditional journalism. Through a content analysis of 886 fake news articles, we find that in terms of news values, topic, and formats, articles published by fake news sites look very much like traditional—and real—news. Most of their articles included the news values of timeliness, negativity, and prominence; were about government and politics; and were written in an inverted pyramid format. However, one point of departure is in terms of objectivity, operationalized as the absence of the author's personal opinion. The analysis found that the majority of articles analyzed included the opinion of their author or authors.

Keywords
content analysis; disinformation; fake news; inverted pyramid; news values; objectivity; traditional news

Issue
This article is part of the issue "Dark Participation in Online Communication: The World of the Wicked Web" edited by Thorsten Quandt (University of Münster, Germany).

1. Introduction

'Fake news' became a topic of controversy during and following the 2016 U.S. presidential election. False stories, such as those reporting that the Catholic Pope had endorsed Donald Trump or that Hillary Clinton had authorized the sale of weapons to a terror group, went viral on social media. The phenomenon has called into question the responsibilities of social media giants like Facebook in providing a platform where misinformation can spread quickly (Carlson, 2018; Johnson & Kelling, 2018) while several governments across the world have considered legislative interventions to address the spread of fake news (Haciyakupoglu, Hui, Suguna, Leong, & Rahman, 2018; Katsirea, 2018; Tambini, 2017).

Fake news stands in contrast to 'real news,' which is produced by journalists who have long commanded an important gatekeeping role in deigning events as newsworthy and in separating fact from falsehood. Journalism's normative standing does not appear out of thin air but is the result of norms and routines built up over time, into which new entrants are socialized. Such norms and routines help maintain journalism's epistemic authority as a reliable arbiter of what is true and what is not (Carlson, 2017). Simply, this is what helps journalism *be believed*. It follows, then, that 'fake news' producers would imitate the conventions of 'real news' to leech off of journalism's authority and convince readers that the material presented to them is an authentic account.

This seems a logical presumption, but does it hold true? There has already been ample research on fake news (for a review, see Tandoc, 2019) and, in particular, its 'fakeness,' taking into account the motivations of its producers, its conceptual contours, and its relation-

ship with other forms of deceptive communication (see, e.g., Allcott & Gentzkow, 2017; Finneman & Thomas, 2018). However, there are far fewer works looking at the *content characteristics* that make 'fake news' look like 'real news.' Mourão and Robertson (2019) analyzed articles published during the 2016 election season in the U.S. by 50 American websites that have been labeled as fake news sites, finding that such articles, many of which were not outright falsehoods, generally "employed moderate levels of sensationalism, clickbait, misleading content and partisan bias" (p. 2090). But what about the extent to which fake news comports with established journalistic conventions? While many studies have focused on the 'fakeness' of fake news, fewer have examined its 'newsness.' Such an inquiry would illuminate the extent to which 'fake news' stories incorporate elements of 'news' into their design and thus draw on those elements as markers of authority. Therefore, this study examines the content characteristics of fake news stories through a content analysis of articles from identified fake news sites in the U.S.

2. Literature Review

2.1. Fake News in Context

Legacy journalism has been beset by a series of intersecting challenges to its legitimacy, from the diffusion of technologies of content creation to economic tumult to collapsing public trust (Carlson, 2020; Tong, 2018). Meanwhile, a growing number of people consume news via social media (Gottfried & Shearer, 2016), eroding the relationship between traditional journalism organizations and their audiences due to the insertion of a mediator, allowing for the rapid diffusion of information with little regard to its veracity. Fake news, then, is emblematic of a collapse of journalistic sensemaking authority and "highlights the erosion of long-standing institutional bulwarks against misinformation in the Internet age" (Lazer et al., 2018, p. 1094). In the U.S., these events occur against a backdrop of political polarization, where partisanship influences how people respond to messages (Nyhan & Reifler, 2010). This is itself layered onto a political culture characterized by relatively easy uptake of conspiracy theories (Oliver & Wood, 2014). The result of these trends, it has been argued, is that "we are now facing a situation in which a large share of the populace is living in an epistemic space that has abandoned conventional criteria of evidence, internal consistency, and fact-seeking'' (Lewandowsky, Ecker, & Cook, 2017, p. 360).

2.2. Fakeness and Newsness

The term 'fake news' is not new. It has, for example, been used to refer to the political satire of figures like Jon Stewart and Stephen Colbert, who have approximated the conventions of broadcast news (and, in Colbert's case, partisan punditry) for comedic effect (Baym, 2005; Borden & Tew, 2007). The underlying concept of deceptive mass communications is also, of course, not a new phenomenon. From the colonial era to the 'professionalization' period of journalism in the early 20th century, the journalists of the day would routinely use hoaxes, sensationalism, and exaggeration as a means of selling newspapers (Fedler, 1989; Finneman & Thomas, 2018).

Fake news is a complex and somewhat controversial concept due to wide variation in the way it is used in public discourse. It is notoriously difficult to define, drawing hoaxes, conspiracy theories, state-sponsored propaganda, partisan-slanted information, manipulated content, satire, and parody into its orbit (Tandoc, Lim, & Ling, 2018). Scholars have attempted to navigate this terrain by offering definitions of the concept. Such definitions have included "the intentional deception of a mass audience by nonmedia actors via a sensational communication that appears credible but is designed to manipulate and is not revealed to be false" (Finneman & Thomas, 2018, p. 358); "information that has been deliberately fabricated and disseminated with the intention to deceive and mislead others into believing falsehoods or doubting verifiable facts" (McGonagle, 2017, p. 203); and "news articles that are intentionally and verifiably false, and [that] could mislead readers" (Allcott & Gentzkow, 2017, p. 213).

In their analysis of the definitions provided to date, Tandoc et al. (2018) demonstrate how the different conceptualizations offered vary in two dimensions: level of facticity and intent. With regard to facticity, while satire and parodies use deception for the main purpose of humor, propaganda and manipulation mainly seek to deceive. Satire mimics and makes fun of the news but ultimately still depends on facts, while parodies rely on fictitious accounts for humor. The intent behind the production of fake news also vary. When outrageous headlines trick readers into clicking a story or if they get drawn to a particular story and visit the page, their clicks get converted into advertising dollars; this is a *financial* motive. This seems to be what motivated some Macedonian teenagers to create fake news websites and produce fake news articles; their earnings from their fake election stories dwarfed the Macedonian average monthly salary (Subramanian, 2017). By contrast, an *ideological* motive would be to intentionally muddy public discourse or discredit particular personalities or institutions in order to advance (or prevent) particular political outcomes. For example, Russian forces marshalled a sophisticated disinformation operation in fulfilment of their strategic aims (Haigh, Haigh, & Kozak, 2018).

What is noteworthy is the extent to which existing definitions of fake news focus on its 'fakeness'—that is, its degree of facticity and the intent behind its production—while far less attention has been afforded to its 'newsness'—the extent to which it imitates established journalistic conventions, using them to convey

truthfulness. For example, Finneman and Thomas (2018) note that fake news "appears credible" (p. 358) but *how*, precisely, does it appear to be so? If fake news is "fabricated information that mimics news media content in form but not in organizational process or intent" (Lazer et al., 2018, p. 1094) then what exactly is this form? The literature to date has not explored these questions. To understand the newsness of fake news, we need to first examine the newsness of real news.

2.3. 'Real News' and its Routines

Journalism is identifiable by its adherence to a set of routines, which are "patterned, repeated practices, forms, and rules that media workers use to do their jobs" and "practical responses to the needs of media organizations and workers" that "optimize the relationships between an organization and its environment" (Shoemaker & Reese, 2014, pp. 165, 168). These routines not only make the process of newswork more efficient but also help maintain journalism's authority as a reliable arbiter of what is true and what is false (Carlson, 2017).

Though scholarship emphasizes how fake news stories are built on falsehoods, the assumption that fake news stories mimic real news is often taken for granted. This is an important assumption to test, as it has implications for how we understand and deal with fake news as a social problem. One way to examine to what extent fake news articles mimic real news is to compare them based on attributes that characterize traditional news, on top of being based on facts. Defining news is not easy. Schudson (2018) defines it as "usually" referring to "novel information about relatively recent affairs" (p. 999). News writing textbooks also usually refer to news as an account of a recent, significant, extraordinary, and interesting event (e.g., Harcup, 2015; Kershner, 2005; Richardson, 2007). But aside from what news is about, conceptualizations of what news is have also included conventions on how it is produced and communicated; for example, news is marked by several content-related conventions, such as the use of an inverted pyramid format (e.g., Harcup & O'Neill, 2017; Thomson, White, & Kitley, 2008; Vos, 2002).

In comparing real news with fake news based on some content markers of real news, this study is modelled on prior work that has looked at emergent journalistic actors that, in producing news, adopt the norms and routines of the 'mainstream,' rather than actively departing from or challenging them. For example, Tandoc (2018) examined the extent to which *BuzzFeed*, a relative newcomer to the journalistic field, abided by the same set of rules as *The New York Times*, long regarded as an industry leader and national newspaper of record in the U.S., finding that, with regard to the presence of particular news values, the dominant topic, news format, and use of objectivity, *BuzzFeed* behaved very much like a traditional news organization. This current study adopts this framework and considers the following as representative (albeit not definitive) markers of 'real news,' which can be used to measure the 'newsness' of fake news: *News values*, *news topic*, *news format*, and *objectivity*.

2.3.1. News Values

One way that journalism can be distinguished from other forms of writing is through journalists' use of news values, which refer to journalists' "shared operational understanding" that informs "the mediated world that is presented to news audiences" (Harcup & O'Neill, 2017, p. 1470). These criteria for determining newsworthiness are "passed down to new generations through a process of training and socialization" (Harrison, 2006, p. 153). Though different news organizations may prioritize different news values according to market orientation, national context, and the degree of journalistic autonomy, it suffices to say that the identification or combination of any mix of news values serves as a cue to the newsworthiness of a story, and the more criteria an event satisfies, the more likely it is to become news (Harcup & O'Neill, 2017).

No taxonomy of news values can be definitive. We focus here on four specific news values that recur in empirical studies of traditional news content and in meta-analyses of the literature (see, e.g., Harcup & O'Neill, 2017; Shoemaker & Reese, 2014). These are the news values of *timeliness*, *negativity*, *prominence*, and *impact.* Though proximity is a commonly studied news value, it is operationalized in terms of the issue or event's proximity to the newsroom; since most fake news sites come and go and operate anonymously, it is difficult to ascertain their respective geographic locations.

The news value of *timeliness* pertains to the recency of the information and "responds to the impetus of the news being recent and up-to-date" (Kilgo, Lough, & Riedl, 2020, p. 270). Timeliness is not only a long-established news norm, but also one that is embraced by emergent journalistic actors, such as *BuzzFeed*, that incorporate it into their reporting (Tandoc, 2018). *Negativity* refers to 'bad news' stories possessing unpleasant undertones that disrupt the normal state of affairs, such as those involving conflict or tragedy (Harcup & O'Neill, 2017). The news value of *prominence* refers to the involvement of prominent individuals or organizations and has been found to be key to the 'shareworthiness' of content on social media (García-Perdomo, Salaverría, Kilgo, & Harlow, 2018; Kilgo et al., 2020). Finally, *impact* relates to the significance, magnitude, or effects of the issue or event at hand in terms of their scale, reflecting how news stories spotlight "the most severe storms, the most damaging fires, the most deadly accidents, the most important speeches, and the most interesting organizations because these are likely to affect the most readers and viewers and have the most serious consequences" (Fedler, Bender, Davenport, & Drager, 2001, p. 110). Working from the premise that fake news tries to mimic traditional news, we ask:

RQ1: What percentage of articles published by fake news sites contain the following news values: a) timeliness, b) negativity, c) prominence, and d) impact?

2.3.2. Topic

The news production process is also traditionally characterized by the classification of stories by topic, such as politics, crime, business, health, or entertainment. This classification affects how stories are organized in a newspaper or online (Dick, 2011) and shapes stories' shareability (García-Perdomo et al., 2018). There are normative judgments associated with this, where stories about politics and government are treated as possessing greater normative import than other kinds of journalism (Schultz, 2007). Indeed, research by Tandoc (2018) found that both *BuzzFeed* and *The New York Times* published stories about government or politics most frequently, followed by crime or terrorism stories. Thus, we also ask:

RQ2: What news topics do fake news sites write most frequently about?

2.3.3. News Format

Analyses of journalistic content have also focused on the format of journalistic prose. The inverted pyramid style, where the most important information is placed at the top, dominates mainstream news reporting. This is likely due to its normative purchase, being associated with objectivity due to the way it standardizes the presentation of news content as an authoritative account of events (Thomson et al., 2008; Vos, 2002). Though alternatives exist, such as the narrative style common to literary journalism (see Johnston & Graham, 2012) or emergent forms such as the 'listicle' (see Tandoc, 2018), the inverted pyramid persists in the U.S. as a dominant format of organizing news, so much so that the format itself can trigger heuristics that affect the perceived credibility of a message (Sundar, 2008). It follows, therefore, that fake news producers would attempt to imitate this style of narrative to appear authoritative and thus be believed. Thus:

RQ3: What percentage of articles published by fake news sites use the inverted pyramid format?

2.3.4. Objectivity

Finally, a common marker of traditional journalism in the U.S. is the use of objectivity, which casts journalists in the role of impartial scientists pursuing the evidence wherever it leads and demonstrates "faith in 'facts,' a distrust of 'values,' and a commitment to their segregation" (Schudson, 1978, pp. 4–5). As a signaling mechanism, objectivity implies that journalists have obtained "all relevant information" and vetted it to "determine why accounts conflict and which more accurately reflect reality" (Ryan, 2001, p. 4). Objectivity is frequently cited as being at the root of journalism's epistemic authority—that is, its credibility as the arbiter of what is factual (Carlson, 2017).

Though opinion remains a prominent part of U.S. journalism (e.g., opinion columns, public affairs talk shows), within the output of mainstream news organizations it is compartmentalized away from news and given its own section. This is both a literal and symbolic separation, reinforcing the norm that opinion ought not intrude into news reporting (Thomas, 2018). Accordingly, empirical studies of news content have operationalized objectivity as the absence of opinion (see, e.g., Lawrence, Molyneux, Coddington, & Holton, 2014; Molyneux, 2015; Tandoc, 2018; Tandoc & Thomas, 2017). It follows, then, that fake news producers may seek to take advantage of this journalistic credibility in order to be believed, by putting on a semblance of objectivity. Therefore:

RQ4: What percentage of articles published by fake news sites exclude personal opinion of the author?

3. Method

3.1. Sampling

Seeking to examine the extent to which fake news mimics real news, this study is based on a content analysis of 886 articles from 23 fake news sites. Sampling took several stages. First, we built a list of fake news sites by relying on lists published by news and entertainment site *BuzzFeed* and fact-checking sites *PolitiFact* and *FactCheck*. *BuzzFeed*'s list was based on "the top-performing Facebook content from 96 fake news websites…built up over the past two years of covering this topic" and cross-referenced against a chart by Hoaxy (a tool that visualizes the spread of articles online) resulting in "a more comprehensive list of pure fake news sites" (Silverman, 2016, para. 4). *PolitiFact*'s list was based on "every website on which [they had] found deliberately false or fake news stories since we started working along with Facebook" (PolitiFact, 2017, para. 5). *FactCheck* offered "a list of websites that have posted deceptive content" (*FactCheck*, 2017, para. 1). As has been noted (e.g., Mourão & Robertson, 2019), differences in the ways that fake news is conceptualized and measured result in challenges in this line of empirical work. None of the lists we draw upon claimed to be exhaustive but are indicative of efforts at legitimate news and fact-checking organizations to catalog fake news sites. A combined list based on these three sources included 230 fake news sites.

Next, we randomly selected 23 sites from the list, representing 10% of the listed fake news sites. Then, for each randomly selected site, we collected links to all their published articles between February 28, 2017, and February 28, 2018, using BuzzSumo, a social media marketing online tool that allows tracking of online content and their social media engagement metrics that has

been used by studies on social media content (see, e.g., Cadman & Galvin, 2019; Sommariva, Vamos, Mantzarlis, Đào, & Martinez Tyson, 2018; Waszak, Kasprzycka-Waszak, & Kubanek, 2018). This yielded 9,915 articles, from which we randomly selected 992 articles, again representing 10% of the sample. Some of the links, however, were no longer active when we conducted the study, leaving the final study with a total of 886 articles for content analysis. We analyzed the selected articles by reading them on the actual webpage where they were published—that is, we clicked on the links we collected from BuzzSumo to access the articles. We focused on coding the article's main text and excluded any complementary materials, such as accompanying visuals. Thus, the unit of analysis for this study is the article's main text. Due to the transitory nature of fake news sites, our sample is only representative of the fake news sites represented in these lists and not of the continuously evolving fake news ecosystem in the U.S.

3.2. Variables

Two coders were trained using a content analysis manual adapted from an earlier study (Tandoc, 2018). The manual included measures of what previous studies have considered as markers of traditional news. Following three training sessions, two practice coding sessions, and acceptable intercoder reliability values, the actual coding began with each coder independently coding half of the sample. The first practice coding involved 20 recent fake news articles collected from a fake news site not included in the sample. The purpose of the first practice coding was to introduce the coders to the process of coding as well as obtain initial feedback on the coding manual. The second practice coding involved 20 randomly selected fake news articles from the population of articles where the actual sample came from. These 20 articles were excluded from the final sample.

3.2.1. News Values

The articles were coded for the presence or absence of four news values common in the literature: timeliness, negativity, prominence, and impact. *Timeliness* refers to whether the article was about something recent, timely or seasonal. *Negativity* refers to whether the article focused on the negative aspects of the issue or event. *Prominence* refers to whether the article involved well-known or elite sources, either individuals or organizations. Finally, *impact* refers to whether the issue or event in the article has high significance in terms of its effects or consequences to the population. The coders consistently coded for news values (*Krippendorf's* $\alpha = 0.74$).

3.2.2. News Topics

The articles were also coded for their main story topic. Drawing on an integrated list of recurrent topics identified in the literature (Becker, 2009; Becker, Lowrey, Claussen, & Anderson, 2000; Magin & Maurer, 2019; Maguire, 2014; Schierhorn, Endres, & Schierhorn, 2001; Sjøvaag, 2015; Tandoc, 2018), stories were coded if the main story topic was about: government and politics; crime or terrorism; economy and business; education; environment and energy; transportation and public works; accidents and disasters; science, health, and technology; religion; social problems and human rights; human interest; sports; or entertainment. Intercoder reliability agreement was initially low (*Krippendorf's* $\alpha = 0.22$), which prompted additional coder training on this measure until both coders had a common and confident understanding of how the topic categories were to be coded. Subsequent intercoder testing showed the intercoder reliability score to be close to the acceptable range (*Krippendorf's* $\alpha = 0.64$).

3.2.3. News Format

The articles were coded for their format, or how the story was written or presented. It could be any of these commonly used news formats, as deduced from the literature: inverted pyramid, listicle, chronology, reversed chronology, or narrative. Since this was a straightforward measure, the coders achieved perfect agreement for this category.

3.2.4. Objectivity

Finally, the articles were coded for the presence or absence of the journalist's opinion on the subject matter or issue. For an article to be coded as having the opinion of the journalist, the inclusion of personal opinion must be explicit. For example, a fake news article wrongly claimed eight witnesses of the Las Vegas mass shooting in October 2017 had suspiciously died, talked about how "the official narrative stinks so badly" and described as "staggering" the number of witnesses who "have died in suspicious circumstances"; these claims have been debunked by fact-checking organizations, such as Snopes.com. In this article, the use of value-laden adjectives in sentences not attributed to any source explicitly includes personal judgments of the author. Implicit inclusion of opinion, such as the choice of particular sources over others, is therefore not captured in this variable. This was also coded as a binary nominal variable (opinion present/opinion absent), with an acceptable intercoder reliability score (*Krippendorf's* $\alpha = 0.87$).

4. Findings

4.1. News Values

RQ1 asked what percentage of articles published by fake news sites contain the news values of a) timeliness, b) negativity, c) prominence, and d) impact. The analysis showed that 98.6% of the articles analyzed included the

news value of timeliness; 89.2% included the news value of negativity; 79.7% included the news value of prominence; but only 32% included the news value of impact. In comparison, a previous study that analyzed the content of *The New York Times* (Tandoc, 2018), whose framework we have adopted for this study, had found that majority of its news articles included the news values of timeliness (72.7%), negativity (74.5%), prominence (64.2%), and impact (59%). Therefore, in terms of the news values of timeliness, negativity, and prominence, articles from fake news sites seem to mimic real news articles (see Table 1). However, most of the articles analyzed do not have the news value of impact, focusing on trivial things, such as a fake news article reporting that a woman was hospitalized after she was beaten with dildos.

Table 1. News values (%).

	Yes	No
Timeliness	98.6	1.4
Negativity	89.2	10.8
Prominence	79.7	20.3
Impact	32.3	67.7

4.2. News Topic

RQ2 asked what topics are most frequently written about by fake news sites. The analysis found that 51.6% of the articles analyzed were about government or politics (see Table 2). This was followed by crime or terrorism (19.5%) and by science, health or technology (10.3%). In comparison, among the most common topics reported by *The New York Times* based on a previous study were government or politics (31.6%); crime or terrorism (27.1%); and science, health, or technology (8.2%; Tandoc, 2018). However, many of political or crime-related stories we analyzed focused on trivial matters, potentially aimed at fanning political polarization rather than disseminating important information. For example, a fake news article reported that a leaked email revealed that "Michelle Obama admits she hates Hillary Clinton." While this is considered a story about politics, it is an 'issue' that does not involve or affect the population.

Table 2. News topic.

Topic	Percentage
Government/Politics	51.6
Crime/Terrorism	19.5
Science/Health/Technology	10.3
Sports/Entertainment/Arts	6.9
Accidents/Disasters	3.7
Economy/Business	2.5
Public Services	1.6
Religion/Churches	1.5
Environment/Climate Change	1.1

4.3. News Format

RQ3 asked about the most commonly used news format by fake news sites. The analysis found that 98.8% of the articles analyzed used the inverted pyramid format (see Table 3). In comparison, Tandoc (2018) found that 70.8% of the news articles published by *The New York Times* used inverted pyramid.

Table 3. News format.

Format	Percentage
Inverted pyramid	98.8
Listicle	.4
Chronology	.1
Narrative	.7

4.4. Objectivity

Finally, RQ4 asked what percentage of the articles published by fake news sites adhered to the standard of objectivity by excluding any personal opinion of the author. The analysis found that only 35.7% of the articles analyzed excluded personal opinion, while the majority, or 64.3%, included the personal opinion of the author or authors (see Table 4). For example, a trivial fake news article that claimed rapper Jay-Z was caught "shapeshifting" by passengers in a United Airlines flight included references to how the airline was desperate to avoid another scandal and engaged in steps to cover-up the incident. Such comment was not attributed to any source. By contrast, a content analysis of news articles published by *The New York Times* found that 75.8% of its articles excluded journalists' opinions (Tandoc, 2018).

Table 4. Objectivity.

Presence of Opinion	Percentage
No	35.7
Yes	64.3

5. Discussion

This study set out to explore the extent to which fake news content imitates the conventions of traditional, 'real' news. Where previous studies have focused on the 'fakeness' of fake news, this study focused on its 'newsness.' Guided by previous studies that mapped out markers of traditional news, this study analyzed articles published by fake news sites based on news values, topic, format, and objectivity. The study found that in terms of news values, topic, and format, the articles analyzed look very much like traditional news. The majority of the articles we studied included the news values of timeliness, negativity, and prominence; were about government and politics; and were written in an inverted pyramid format. However, one point of departure is in

terms of objectivity, operationalized as the absence of the author's personal opinion. The analysis found that the majority of the articles included the personal opinion of their author or authors. The news value of impact was also not very common among fake news sites, which seem to focus a lot on concocting trivial stories.

5.1. Implications

By identifying the content characteristics common across stories published by fake news sites, this study has provided empirical data to inform what may have previously been assumed. Our findings, overall, suggest that fake news producers *imitate the conventions of traditional news*. This mimicry leeches off journalism's epistemic authority for deceptive ends. Put another way, for fake news producers, news is simply the means, but deception is the ends. Overall, this reinforces the association between journalistic routines and content conventions and journalism's epistemic authority. By mimicking these content conventions, from writing style and format to news values, fake news producers exploit journalism's social standing. This lends support to the assumption that fake news, as a specific form of deliberate attempts at disinformation, refers to articles devoid of factual basis *deliberately packaged to look like news* in order to deceive.

The findings of this study help to illuminate what content characteristics of real news fake news producers are appropriating to give their outputs a semblance of truthfulness or even legitimacy. An underlying ideological motivation, such as sowing distrust on a government investigation of a mass shooting, can be propagated even in the absence of facts as fake news producers can package a false claim (e.g., suspicious deaths of witnesses that signal a cover-up) supporting their underlying motivation (e.g., sow distrust in the government) with content characteristics associated with real news (e.g., reference to a timely event, focus on a negative aspect, peg to a prominent topic, use of inverted pyramid, among others) to turn a false narrative into one that looks like a real, legitimate news story. Employed in a regular fashion, for both completely false as well as real but incomplete or sensationalized articles, the appropriation of content characteristics of real news can potentially don a website with a cloak of legitimacy, at least for those readers its articles are able to mislead. Thus, the 'newsness' of fake news helps not only specific fake news articles to deceive, but also potentially the websites and social media accounts that regularly publish them.

However, the analysis also uncovered some areas of departure, the most notable of which is in terms of objectivity. This study found that the majority of the articles analyzed included the opinion of their author or authors. It may well be that the *absence* of objectivity explains why fake news is so potent. By explicitly appealing to readers' existing predispositions through the inclusion of similar opinions by the author or authors, fake news articles increase their resonance, legitimacy, and believability among a group of readers, a phenomenon known as confirmation bias (Taber & Lodge, 2006). This could also reflect the prominence of partisan punditry and commentary in the media landscape (see, e.g., Levendusky, 2013) and the acceptance of opinion as a news value. In their study of what young adults consider news, Armstrong, McAdams, and Cain (2015) found that "consumers may have come to expect—and even seek out—subjective, opinion-laden news to help them make sense of prominent, impactful, and controversial events and issues" (p. 95). Given these conditions, it may be the case that fake news producers are cognizant of changes in how journalism is being produced, received, and evaluated and are taking advantage of such shifts.

Of particular interest is the finding that the articles we studied used the inverted pyramid style of prose while departing from the objectivity norm, as the two have typically been treated as congruent (Thomson et al., 2008; Vos, 2002). It *may* be the case that the association between the two is weakening, although it is beyond the scope of this study to establish this with empirical certainty. A plausible explanation may lie in the intent of the fake news producers, who may have observed that the inverted pyramid remains prevalent as a 'standard' way of organizing news presentation while the objectivity norm may be less salient to the goal of deepening partisan attitudes in targeted populations. Another explanation may lie in the nature of the sample, which focused on articles from fake news sites in the U.S., a country characterized by growing political polarization, declining trust in journalism along partisan lines, and the prominence of opinionated and partisan media content (Newman, & Fletcher, Schulz, Andı, & Nielsen, 2020). These intersecting and mutually reinforcing factors represent a context where fake news producers feel that content displaying a high degree of partisanship is likely to gain traction.

Fake news is a problematic term, and one could argue that persisting in its use—in other words, deigning such content as "news" to begin with—mistakenly deigns it with legitimacy. That the term has been taken up by politicians to describe unfavorable reporting (Lischka, 2019) makes this terrain yet more complicated. However, it remains a worthwhile endeavor to examine the extent to which those that pass off deceptive information as news mimic the conventions of real news for deceptive ends. This provides more precision in determining how fake news approximates 'newsness' in content if not in ethics.

5.2. Limitations and Directions for Future Research

The findings of this study have to be understood in the context of several limitations. To be sure, a content analysis can only analyze manifest content and not the motivations and routines behind content patterns. Future studies should look into practices that lead to the content

patterns this study has uncovered. If fake news looks like real news, what routines do fake news producers follow to construct fake news outputs and how do those routines compare with those of journalists? Granted, pursuing such a line of research may be replete with practical challenges.

While the lists of fake news sites we used for sampling are comprehensive, they are not exhaustive, since fake news sites come and go. Therefore, our findings cannot be generalized to the whole population of fake news in the U.S.; at most, our findings represent the fake news sites in the lists we used at the time of data collection. It is possible that fake news has evolved since then, and future studies can build on our findings to continue tracking how fake news evolves. We also focused our sampling on websites labeled as fake news sources, similar to what a previous study conducted (Mourão & Robertson, 2019), which had found that these sites do not exclusively published falsehoods but also truthful accounts. Our study, however, focused on examining the use of journalistic conventions rather than reliance on facts per se.

Finally, we focused on articles published by fake news sites identified in the U.S., and we should be wary of suggesting that what would pertain to one context would pertain elsewhere, given differences in political and media contexts across systems. Fake news is a global problem, and it is important to study it in other national contexts. If fake news packages fake information to look like real news, how does it look like in media contexts whose form and substance are different from that of U.S. journalism?

Despite these limitations, we hope our findings contribute to a more nuanced understanding of fake news. The findings of this current study not only provide empirical support for the assumption that fake news mimics real news to leech off journalism's social legitimacy and authority, but also raise questions for future studies. For example, an interesting finding is that after politics and crime, the topics of science, technology and health are the third most frequent subjects of fake news articles. This has implications for how the public understands, or misunderstands, already complex but important issues involved in science, technology, and health (such as climate change, vaccinations, and Covid-19 remedies). Furthermore, as this study showed how similar fake news is to real news when it comes to content structure, such content characteristics no longer suffice as demarcations between real and fake news. They should no longer be held as authenticity cues. Indeed, such content markers are what fake news producers *exploit* to deceive readers. For example, quoting prominent personalities has long been associated with newsworthiness, but it should not be used as an automatic measure of truth or indicator of trustworthiness, since fake news can also attribute made-up quotes to real people. Newsworthiness as a concept must be revisited, and its heuristic value for journalists and audiences questioned. Future studies should explore how real news can distinguish itself from fake news and how their results can be communicated to readers to equip them with skills to distinguish what is real from what is fake, and to value the former over the latter.

Acknowledgments

The authors are grateful for the work of the data coders. The first author's work on this project was also supported by a Tier 1 Academic Grant from the Singapore Ministry of Education.

Conflict of Interests

The authors declare no conflict of interests.

References

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236.

Armstrong, C. L., McAdams, M. J., & Cain, J. (2015). What is news? Audiences may have their own ideas. *Atlantic Journal of Communication*, *23*(2), 81–98.

Baym, G. (2005). The Daily Show: Discursive integration and the reinvention of political journalism. *Political Communication*, *22*(3), 259–276.

Becker, L. (2009). Employment. In C. H. Sterling (Ed.), *Encyclopedia of journalism* (pp. 514–519). Thousand Oaks, CA: Sage.

Becker, L., Lowrey, W., Claussen, D., & Anderson, W. (2000). Why does the beat go on? *Newspaper Research Journal*, *21*(4), 2–16.

Borden, S. L., & Tew, C. (2007). The role of journalist and the performance of journalism: Ethical lessons from "fake" news (seriously). *Journal of Mass Media Ethics*, *22*(4), 300–314.

Cadman, K., & Galvin, J. P. (2019). Graft versus host disease: An analysis of Internet and social network activity and engagement. *Biology of Blood & Marrow Transplantation*, *25*(3), 81–82.

Carlson, M. (2017). *Journalistic authority: Legitimating news in the digital era*. New York, NY: Columbia University Press.

Carlson, M. (2018). Facebook in the news: Social media, journalism, and public responsibility following the 2016 Trending Topics controversy. *Digital Journalism*, *6*(1), 4–20.

Carlson, M. (2020). Fake news as an informational moral panic: The symbolic deviancy of social media during the 2016 U.S. Presidential election. *Information, Communication, & Society*, *23*(3), 374–388.

Dick, M. (2011). Search engine optimization in UK news production. *Journalism Practice*, *5*(4), 462–477.

FactCheck. (2017). Misinformation directory. *FactCheck*. Retrieved from https://www.factcheck.org/2017/07/websites-post-fake-satirical-stories

Fedler, F. (1989). *Media hoaxes*. Ames, IA: Iowa State University Press.

Fedler, F., Bender, J. R., Davenport, L. D., & Drager, M. W. (2001). *Reporting for the media* (7th ed.). San Diego, CA: Harcourt College.

Finneman, T., & Thomas, R. J. (2018). A family of falsehoods: Deception, media hoaxes, and fake news. *Newspaper Research Journal*, *39*(3), 350–361.

García-Perdomo, V., Salaverría, R., Kilgo, D. K., & Harlow, S. (2018). To share or not to share: The influence of news values and topics on popular social media content in the United States, Brazil, and Argentina. *Journalism Studies*, *19*(8), 1180–1201.

Gottfried, J., & Shearer, E. (2016). News use across social media platforms 2016. *Pew Research Center*. Retrieved from https://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016

Haciyakupoglu, G., Hui, J. Y., Suguna, V. S., Leong, D., & Rahman, M. F. B. A. (2018). *Countering fake news: A survey of recent global initiatives*. Singapura: S. Rajaratnam School of International Studies—Nanyang Technological University. Retrieved from https://www.think-asia.org/handle/11540/8063

Haigh, M., Haigh, T., & Kozak, N. I. (2018). Stopping fake news: The work practices of peer-to-peer counter propaganda. *Journalism Studies*, *19*(14), 2062–2087.

Harcup, T. (2015). *Journalism principles and practice*. Thousand Oaks, CA: Sage.

Harcup, T., & O'Neill, D. (2017). What is news? News values revisited (again). *Journalism Studies*, *18*(12), 1470–1488.

Harrison, J. (2006). *News*. New York, NY: Routledge.

Johnson, B. G., & Kelling, K. (2018). Placing Facebook: 'Trending,' 'Napalm Girl,' 'fake news' and journalistic boundary work. *Journalism Practice*, *12*(7), 817–833.

Johnston, J., & Graham, C. (2012). The new, old journalism: Narrative writing in contemporary newspapers. *Journalism Studies*, *13*(4), 517–533.

Katsirea, I. (2018). 'Fake news': Reconsidering the value of untruthful expression in the face of regulatory uncertainty. *Journal of Media Law*, *10*(2), 159–188.

Kershner, J. W. (2005). *The elements of news writing*. Boston, MA: Pearson Allyn and Bacon.

Kilgo, D. K., Lough, K., & Riedl, M. J. (2020). Emotional appeals and news values as factors of shareworthiness in ice bucket challenge coverage. *Digital Journalism*, *8*(2), 267–286.

Lawrence, R. G., Molyneux, L., Coddington, M., & Holton, A. (2014). Tweeting conventions: Political journalists' use of Twitter to cover the 2012 Presidential campaign. *Journalism Studies*, *15*(6), 789–806.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., . . . Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096.

Levendusky, M. (2013). *How partisan media polarize America*. Chicago, IL: University of Chicago Press.

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory & Cognition*, *6*(4), 353–369.

Lischka, J. A. (2019). A badge of honor? How The New York Times discredits President Trump's fake news accusations. *Journalism Studies*, *20*(2), 287–304.

Magin, M., & Maurer, P. (2019). Beat journalism and reporting. In J. F. Nussbaum (Ed.), *Oxford research encyclopedia of communication*. Retrieved from https://doi.org/10.1093/acrefore/9780190228613.013.905

Maguire, M. (2014). *Advanced reporting: Essential skills for 21st century journalism*. New York, NY: Routledge.

McGonagle, T. (2017). 'Fake news': False fears of real concerns. *Netherlands Quarterly of Human Rights*, *35*(4), 203–209.

Molyneux, L. (2015). What journalists retweet: Opinion, humor, and brand development on Twitter. *Journalism*, *16*(7), 920–935.

Mourão, R. R., & Robertson, C. T. (2019). Fake news as discursive integration: An analysis of sites that publish false, misleading, hyperpartisan and sensational information. *Journalism Studies*, *20*(14), 2077–2095.

Newman, N., & Fletcher, R., Schulz, A., Andı, S., & Nielsen, R. K. (2020). *Reuters Institute digital news report 2020*. Oxford: Reuters Institute for the Study of Journalism. Retrieved from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330.

Oliver, J. E., & Wood, T. J. (2014). Conspiracy theories and the paranoid style(s) of mass opinion. *American Journal of Political Science*, *58*(4), 952–966.

PolitiFact. (2017, April 20). PolitiFact's guide to fake news websites and what they peddle. *PolitiFact*. Retrieved from https://www.politifact.com/article/2017/apr/20/politifacts-guide-fake-news-websites-and-what-they

Richardson, B. (2007). *The process of writing news: From information to story*. Boston, MA: Pearson.

Ryan, M. (2001). Journalistic ethics, objectivity, existential journalism, standpoint epistemology, and public journalism. *Journal of Mass Media Ethics*, *16*(1), 3–22.

Schierhorn, A. B., Endres, F. F., & Schierhorn, C. (2001). Newsroom teams enjoy rapid growth in the 1990s. *Newspaper Research Journal*, *22*(3), 2–15.

Schudson, M. (1978). *Discovering the news: A social history of American newspapers*. New York, NY: Basic Books.

Schudson, M. (2018). News. In T. P. Vos & F. Hanusch (Eds.), *The international encyclopedia of journalism studies* (pp. 999–1005). Hoboken, NJ: John Wiley & Sons.

Schultz, I. (2007). The journalistic gut feeling: Journalistic doxa, news habitus, and orthodox news values. *Journalism Practice*, *1*(2), 190–207.

Shoemaker, P. J., & Reese, S. D. (2014). *Mediating the message in the 21st century: A media sociology perspective* (3rd ed.). New York, NY: Routledge.

Silverman, C. (2016, December 30). Here are 50 of the biggest fake news hits on Facebook from 2016. *BuzzFeed News*. Retrieved from https://www.buzzfeednews.com/article/craigsilverman/top-fake-news-of-2016

Sjøvaag, H. (2015). Hard news/soft news: The hierarchy of genres and the boundaries of the profession. In M. Carlson & S. C. Lewis (Eds.), *Boundaries of journalism: Professionalism, practices, and participation* (pp. 101–117). New York, NY: Routledge.

Sommariva, S., Vamos, C., Mantzarlis, A., Đào, L. U. L., & Martinez Tyson, D. (2018). Spreading the (fake) news: Exploring health messages on social media and the implications for health professionals using a case study. *American Journal of Health Education*, *49(*4), 246–255.

Subramanian, S. (2017, February 2). Inside the Macedonian fake news complex. *Wired*. Retrieved from https://www.wired.com/2017/02/veles-macedonia-fake-news

Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger & A. J. Flanagin (Eds.), *Digital media, youth, and credibility* (pp. 73–100). Cambridge, MA: The MIT Press.

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50*(3), 755–769.

Tambini, D. (2017). *Fake news: Public policy responses*. London: The London School of Economics and Political Science. Retrieved from http://eprints.lse.ac.uk/73015

Tandoc, E. C. (2018). Five ways BuzzFeed is preserving (or transforming) the journalistic field. *Journalism*, *19*(2), 200–216.

Tandoc, E. C. (2019). The facts of fake news: A research review. *Sociology Compass*, *13*(9), 1–9.

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining 'fake news': A typology of scholarly definitions. *Digital Journalism*, *6*(2), 137–153.

Tandoc, E. C., & Thomas, R. J. (2017). Readers value objectivity over transparency. *Newspaper Research Journal*, *38*(1), 32–45.

Thomas, R. J. (2018). Advocacy journalism. In T. P. Vos (Ed.), *Journalism* (pp. 391–413). Berlin: Walter de Gruyter.

Thomson, E. A., White, P. R. R., & Kitley, P. (2008). 'Objectivity' and 'hard news' reporting across cultures: Comparing the news report in English, French, Japanese, and Indonesian journalism. *Journalism Studies*, *9*(2), 212–228.

Tong, J. (2018). Journalistic legitimacy revisited: Collapse or revival in the digital age? *Digital Journalism*, *6*(2), 256–273.

Vos, T. P. (2002). Newswriting structure and style. In W. D. Sloan & L. M. Parcell (Eds.), *American journalism: History, principles, practices* (pp. 296–305). Jefferson, NC: McFarland & Co.

Waszak, P. M., Kasprzycka-Waszak, W., & Kubanek, A. (2018). The spread of medical fake news in social media: The pilot quantitative study. *Health Policy & Technology*, *7*(2), 115–118.

About the Authors

Edson C. Tandoc Jr. (PhD) is an Associate Professor at the Wee Kim Wee School of Communication and Information at Nanyang Technological University, Singapore. His studies have focused on the impact of journalistic roles, new technologies, and audience feedback on the news gatekeeping process. He has also looked at how readers make sense of critical incidents in journalism and take part in reconsidering journalistic norms; and how changing news consumption patterns facilitate the spread of fake news.

Ryan J. Thomas (PhD) is an Associate Professor of Journalism Studies in the Missouri School of Journalism at the University of Missouri. His research program addresses the intersection of journalism ethics and the sociology of news, focusing on journalism amid processes of change: the forces shaping journalism, how journalists make sense of them, and how these changes affect journalism's institutional obligations and role in public life.

Lauren Bishop works at the *Grants Pass Daily Courier* in Oregon as the City Hall reporter. She graduated from the University of Missouri in 2020 with a BA in Journalism and a BA in Arts in Political Science.

COGITATIO

Media and Communication (ISSN: 2183–2439)
2021, Volume 9, Issue 1, Pages 120–133
DOI: 10.17645/mac.v9i1.3519

Article

You're Definitely Wrong, Maybe: Correction Style Has Minimal Effect on Corrections of Misinformation Online

Cameron Martel [1,*], Mohsen Mosleh [1,2] and David G. Rand [1,3]

[1] Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02142, USA;
E-Mails: cmartel@mit.edu (C.M.), mmosleh@mit.edu (M.M.), drand@mit.edu (D.G.R.)
[2] Science, Innovation, Technology, and Entrepreneurship Department, Business School, University of Exeter, Exeter, EX4 4PU, UK
[3] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

* Corresponding author

Submitted: 30 July 2020 | Accepted: 8 September 2020 | Published: 3 February 2021

Abstract
How can online communication most effectively respond to misinformation posted on social media? Recent studies examining the content of corrective messages provide mixed results—several studies suggest that politer, hedged messages may increase engagement with corrections, while others favor direct messaging which does not shed doubt on the credibility of the corrective message. Furthermore, common debunking strategies often include keeping the message simple and clear, while others recommend including a detailed explanation of why the initial misinformation is incorrect. To shed more light on how correction style affects correction efficacy, we manipulated both correction strength (direct, hedged) and explanatory depth (simple explanation, detailed explanation) in response to participants from Lucid ($N = 2{,}228$) who indicated they would share a false story in a survey experiment. We found minimal evidence suggesting that correction strength or depth affects correction engagement, both in terms of likelihood of replying, and accepting or resisting corrective information. However, we do find that analytic thinking and actively open-minded thinking are associated with greater acceptance of information in response to corrective messages, regardless of correction style. Our results help elucidate the efficacy of user-generated corrections of misinformation on social media.

Keywords
cognitive reflection test; corrections; dark participation; debunking; fake news; misinformation; social media

Issue
This article is part of the issue "Dark Participation in Online Communication: The World of the Wicked Web" edited by Thorsten Quandt (University of Münster, Germany).

1. Introduction

An estimated 3,6 billion people use social media as of 2020, with this number expected to only increase in the next decade (Clement, 2020). Furthermore, people are increasingly utilizing social media platforms as a primary source of news consumption—indeed, it has been estimated that about two-thirds of American adults at least occasionally get news via social media, despite apprehensions about the accuracy of such news (Shearer & Matsa, 2018). The advent of social media as a means of news dissemination has led to widespread concern over the spread of misinformation and 'fake news' ("fabricated information that mimics news media content in form but not in organizational process or intent"; Lazer et al., 2018, p. 1094). Although fake news comprises a relatively small proportion of Americans' daily media diet (0.15%; see Allen, Howland, Mobius, Rothschild, & Watts, 2020), it may still be harmful. For instance, in the months leading up to the 2016 USA Presidential election, false news stories favoring Trump were shared about 30 million times on Facebook; those favoring

Clinton were shared 8 million times (Allcott & Gentzkow, 2017). More recently, misinformation and disinformation about Covid-19 has spread quickly on social media (Frenkel, Alba, & Zhong, 2020), potentially with fatal consequences. As a result, there is great interest in identifying approaches to combat misinformation.

1.1. Combatting Misinformation at the Platform-Level

One approach is to implement platform-level interventions (i.e., efforts implemented by social media platforms that may be applied to all users). The most widely implemented such approach, applying fact-check tags on disputed or false-rated news items, has substantial limitations. Professional fact-checkers cannot possibly keep up with the pace at which misinformation is produced. In addition to limiting the reach of fact-checking, this may promote increased perceptions of accuracy for unlabeled false headlines ('implied truth effect'; Pennycook, Bear, Collins, & Rand, 2020). Relatedly, general warnings instructing users to be cautious about the accuracy of news content they read and share may result in decreased belief in true news stories ('tainted truth effect'; Clayton et al., 2019). Approaches based on inoculation (i.e., preemptive exposure to and warnings of fake news; Roozenbeek & van der Linden, 2019) and accuracy nudges (i.e., reminding people to think about accuracy before consuming or sharing news content; Pennycook et al., in press; Pennycook, McPhetres, Zhang, Lu, & Rand, 2020), which induce people to be more discerning prior to their contact with misinformation, show substantial promise. So does utilizing layperson judgments (e.g., by harnessing the wisdom of crowds through users or contractors hired to provide quality ratings) to supplement machine learning approaches to misinformation detection (Epstein, Pennycook, & Rand, 2020; Kim, Tabibian, Oh, Schölkopf, & Gomez-Rodriguez, 2018; Pennycook & Rand, 2019a). However, it seems unlikely that platforms will ever be entirely able to control the misinformation problem.

1.2. Combatting Misinformation at the User-Level

In addition to interventions that can be applied by the platforms, it is therefore important to determine what kind of user-generated corrections may be most effective at combatting misinformation online. While correcting misinformation may ideally be a source of positive participation online, it may easily devolve into unproductive and even harmful discourse. This gives rise to the question of what type of corrective message most effectively combats dark participation, rather than gives way to it? One dimension by which corrective messages may differ is that of strength—how forcefully the corrective message corrects the shared misinformation. Less forceful, hedged messaging may lead to increased engagement with and acceptance of the corrective message. For instance, Lewandowsky, Ecker, Seifert, Schwarz, and Cook (2012) argue that effective corrections should attempt to affirm the worldview and identity of the individual being corrected—thus, a hedged correction may be less abrasive towards the corrected individual and their worldview and identity. Furthermore, Tan, Niculae, Danescu-Niculescu-Mizil, and Lee (2016) analyzed effective corrective discourse on Reddit, and found that hedging (i.e., a message that indicated uncertainty, such as "it could be the case"; Tan et al., 2016, p. 622) was more common in more persuasive arguments. This is perhaps because hedging makes an argument easier to accept through the use of a softer tone (Lakoff, 1975). However, it has also been suggested that hedging may add increased uncertainty to the corrective message, thus reducing its efficacy (see Rashkin, Choi, Jang, Volkova, & Choi, 2017). This would suggest that more direct, less hedged corrections of misinformation may provide the clearest form of correction. Alternatively, recent evidence suggests that the tone of a correction (uncivil, affirmational, or neutral) may not affect the effectiveness of corrections to misinformation (Bode, Vraga, & Tully, 2020). Ultimately, there remains limited causal evidence as to whether and how correction strength may impact correction engagement.

Another dimension by which corrections may vary is by explanation depth; for instance, whether the debunking message consists of a simple negation, or includes an alternative account to fill in the gap left by correcting misinformation (see Lewandowsky et al., 2012). In favor of brief refutations, it has been argued that simple rebuttals of misinformation are most effective (Lewandowsky et al., 2012). However, others have argued to avoid simple negations (Nyhan & Reifler, 2012) and instead provide detailed debunking messages (see Chan, Jones, Hall Jamieson, & Albarracín, 2017). Thus, it remains unclear whether corrections should be simple negations of truth, or if they should contain a more detailed explanation of why the shared misinformation is false.

1.3. Current Research

In the current study, we investigate the causal role of different corrective messaging strategies on engagement with corrections. Using a survey experiment in which participants are presented with a series of social media posts, we induce most participants to indicate that they would share a false headline. We then manipulate the style of corrective message participants receive in response to their shared article. Corrective messages varied by strength (direct correction, hedged correction) and depth (simple explanation, detailed explanation). All corrections also included a (non-functional) link to a purported debunking article on *Snopes*, which should also increase the efficacy of the corrective message (see Vraga & Bode, 2018; for related research on 'snoping,' see Margolin, Hannak, & Weber, 2018).

We first predict that (H1) hedged corrections will be perceived as less aggressive and more polite than direct

corrections, and that (H2) detailed corrections will be perceived as more informative and less unhelpful than simple corrections. We also predict that (H3) hedged, detailed corrections will elicit greater reply likelihood and (H4) predict greater acceptance of information, whereas direct and simple corrections will predict increased resistance of information. Finally, we anticipate that (H5) more analytic or actively open-minded individuals will have greater reply likelihood and acceptance of information in response to more detailed corrections.

The current research extends existing literature regarding debunking and corrections of fake news in three main ways. First, the existing literature assessing the effect of correction strength on correction engagement is primarily observational rather than causal (e.g., Tan et al., 2016). We seek to causally determine whether correction strength affects correction efficacy. Second, there is limited work assessing the interaction between various correction wording strategies. We assess not only whether there are main effects of correction strength and depth on engagement, but also if these correction styles may interact with one another. Third, we seek to explore the interaction between correction style and several key cognitive mechanisms which may impact the efficacy of certain forms of corrections. In particular, we utilize the cognitive reflection test (CRT; Frederick, 2005) to assess whether more analytic thinkers engage more with more detailed explanations. We also explore the role of actively open-minded thinking (Stanovich & West, 2007) in receptivity to various corrective messaging styles.

2. Methods

Our study was pre-registered at https://osf.io/eupwn/?view_only=cc6cd2cd0bae42788fcd28aacb505d9a. Furthermore, our full materials, data, and analysis code is available on the Open Science Framework (see https://osf.io/fvwd2/?view_only=cc6cd2cd0bae42788fcd28aacb505d9a).

2.1. Materials and Procedure

2.1.1. Participants

We recruited N = 2,228 participants (1,065 female, M_{age} = 44.84) via the online convenience sampling platform Lucid (Coppock & McClellan, 2019). Participants were first instructed to imagine they were currently on a social media platform such as Twitter or Facebook. Participants were then told they would be presented with a series of actual recent news headlines, as if they were appearing in their social media newsfeed.

2.1.2. News Headlines

Participants were randomly shown up to 28 actual headlines that appeared on social media, half of which were factually accurate (real news) and half of which were entirely untrue (fake news). Additionally, half of the headlines were favorable to the Democratic Party, and half were favorable to the Republican Party, based on pretest ratings (see Pennycook & Rand, 2019b). All fake news headlines were taken from *Snopes*. Real news headlines were selected from mainstream news sources (e.g., *NPR*, *The Washington Post*). Headlines were presented in the format of a social media post—namely, with a picture, headline, byline, and source (Figure 1).

After each headline, participants were asked whether or not they would share that article on social media publicly, such that other users could see and comment on it. If participants decided to share a real news article or decided not to share a fake news article, they were shown another headline. However, if participants decided to share a fake news article, then they proceeded to the rest of the study and saw no further headlines. Participants who did not share any fake news articles were not eligible to complete the correction message section of the study. This indication to share should simulate participants sharing such news articles as if they

Figure 1. Example news headline with picture, headline, byline, and source.

were actually on social media—indeed, recent research has found that self-reported willingness to share political news articles in online survey studies correlates with actual sharing on Twitter (Mosleh, Pennycook, & Rand, 2020).

2.1.3. Corrective Messages

Overall, 1,589 participants (71% of all participants) shared at least one fake news article, and thus completed the remainder of the study. After sharing a fake news article, participants were instructed to imagine receiving a public comment on their post. Participants were presented with one of four corrective messages, which varied by strength (direct, hedged) and depth (simple explanation, detailed explanation). These corrections were stylized as tweets from a fictional user. The first sentence of the message varied by strength—in the direct condition, the message read: "No way do I believe this article—it's definitely not true." In the hedged condition, the message read: "I'm not sure about this article—it might not be true." The second sentence of the message varied by depth—in the simple condition, the sentence read: "I found a link on Snopes that says this headline is false." In the detailed condition, the message read: "I found a link on *Snopes* that says this headline was created by a website that purposefully makes false stories." All messages ended with a stylized *Snopes* link (Figure 2).

2.1.4. Reply to Corrective Message

Next, participants were asked: "Would you reply to the above message?" 1 = "Yes, I would write a reply," 0 = "No, I would not write a reply." If participants indicated "Yes," they were asked to enter their reply via free response. If participants indicated "No," they were asked: "If you DID reply, what would you write?" and then allowed to enter their reply via free response.

2.1.5. Correction Motive

Participants were asked: "Why do you think the person wrote the message you received? Select all that apply." Participants could select from the following: "To inform me of valuable information," "To reinforce the image of themselves they'd like to present to me," "To develop a connection with me," "To achieve self-fulfillment for themselves," or "To get the word out about a specific cause.''

2.1.6. Evaluation of Corrector

After that, participants were asked to evaluate the trustworthiness of the person who wrote them the corrective message (Likert-scale: 1–7), as well as how positive and how negative their opinion was of the person who wrote them the message (Likert-scale: 1–7). Participants were also asked how much they agreed with the following statements: "The message I received on my social media post was [unhelpful/aggressive/informative/polite]." Likert-scale: 1 = Not at all agree, 7 = Strongly agree.

2.1.7. Self-Reported Belief-Updating

Then, participants were asked: "After viewing the comment on your shared article and replying to that comment, how do you view the accuracy of the article you

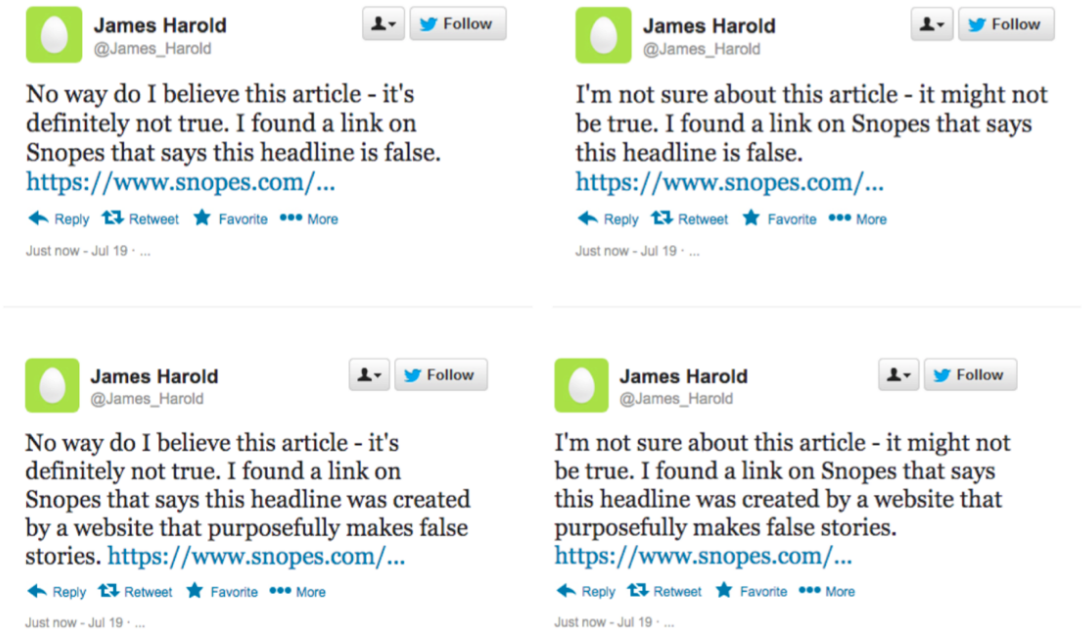


Figure 2. Corrective message conditions. Note: Clockwise, from top left: direct, simple; hedged, simple; hedged, detailed; direct, detailed.

shared?" Likert-scale: 1 = Much less accurate than initially thought, 2 = Slightly less accurate than initially thought, 3 = As accurate as initially thought, 4 = Slightly more accurate than initially thought, 5 = Much more accurate than initially thought.

2.1.8. Cognitive Reflection Test

The CRT is a brief task which measures participant tendencies to engage in analytic thinking. CRT items include an intuitive yet incorrect answer, which participants must override in order to answer correctly (e.g., "The ages of Mark and Adam add up to 28 years old. Mark is 20 years older than Adam. How many years old is Adam?" The common, intuitive answer is eight, whereas the correct answer upon reflection is four).

Participants completed a reworded version of the original CRT (Frederick, 2005; Shenhav, Rand, & Greene, 2012) and a four-item non-numeric CRT (Thomson & Oppenheimer, 2016).

2.1.9. Actively Open-Minded Thinking

Participants also completed a shortened version of the actively open-minded thinking scale (AOT; Stanovich & West, 2007). The AOT measures actively open-minded thinking, or the tendency to be open towards opinions or positions different from one's own (e.g., "A person should always consider new possibilities." 1 = I strongly disagree, 7 = I strongly agree).

2.1.10. Additional Measures

Participants next completed a brief political knowledge measure and standard demographics. Participants also were asked which social networking sites they use (Facebook, Twitter, Instagram, LinkedIn, Snapchat, Other). Participants were asked how often they are on social media and, if they indicated that they had either a Facebook or a Twitter account, how often they share content on Facebook or Twitter ("Never," "Less often," "Every few weeks," "1 to 2 days a week," "3 to 6 days a week," "About once a day," "2 to 5 times a day," "5 to 10 times a day," or "Over 10 times a day").

2.2. Analysis of Free Response Replies

Following our main study, we recruited 819 participants from Amazon Mechanical Turk to crowdsource coding of the free response replies we collected on several dimensions. Free response replies were rated an average of 5.14 times each ($SD = 2.24$), and each participant rated 10 replies on seven key dimensions. Participants were first given instructions carefully detailing each rating category. Participants then evaluated each headline on these dimensions, via a Likert-scale (1–7). These seven dimensions of evaluating replies were informed by categories of responding to corrective information as detailed by Prasad et al. (2009). Detailed explanations of these seven dimensions (denying original belief, belief updating, counter-arguing, attitude bolstering, selective exposure, disputing rationality, and inferred justification) may be found in our Supplemental Materials here: https://osf.io/fvwd2/?view_only=cc6cd2cd0bae42788fcd28aacb505d9a.

2.2.1. Rating Procedure

Participants first read instructions detailing how they will be asked to evaluate replies to corrective messages using seven different categories of response types. Participants then read the descriptions of these response types, and answered a reading comprehension check. Participants who failed this comprehension check were presented with the response type descriptions a second time. Finally, participants viewed 10 different replies and rated each response by all seven response type categories. Participants also were asked: "Overall, how positive is this reply?" and "Overall, how negative is this reply?" Likert-scale: 1 = Not at all [positive/negative], 7 = Very [positive/negative]. Participants also evaluated whether the replier indicated they only shared the fake article as a joke, or if the replier indicated that they plan on looking up more information about the article they shared.

All written replies from the Lucid study were rated by Amazon Mechanical Turk raters, except replies which were either blank or simply said "nothing." These replies were automatically coded as a 1 (not at all) across all categories. There were 320 such replies in total.

2.2.2. Intraclass Correlations

In order to assess the consistency of measurements made by our Amazon Mechanical Turk raters assessing the same replies, we computed intraclass correlations (ICC; descriptive measure of how strongly units in a group resemble one another) for each of the seven response type categories, plus ratings of overall positivity and negativity. In particular, we utilized a one-way random effects ICC model (since each reply was measured by a different set of randomly selected raters), as well as average measures, as our analyses ultimately utilize the average ratings for each reply (see Treventhan, 2017). Across all nine categories, our ICC1k was fair on average, meaning that reply ratings within response type categories adequately resembled one another, $ICC_{avg} = 0.46$ (common guidelines interpret greater than 0.40 as fair; Cicchetti & Sparrow, 1981), $ICC_{DenyOriginalBelief} = 0.33$, $ICC_{BeliefUpdating} = 0.57$, $ICC_{Counter\text{-}arguing} = 0.48$, $ICC_{AttitudeBolstering} = 0.36$, $ICC_{SelectiveExposure} = 0.38$, $ICC_{DisputingRationality} = 0.30$, $ICC_{InferredJustification} = 0.24$, $ICC_{Positive} = 0.73$, $ICC_{Negative} = 0.72$ (all $ps < .001$).

3. Results

3.1. Hedged Corrections Perceived as Less Aggressive, More Polite

In order to assess the effect of our correction style conditions on perceptions of corrections, we performed several analyses. We performed a linear regression model predicting how aggressive participants perceived the corrective messages they received, entering correction strength, depth, and their interaction as predictors. As expected, we found that participants who received hedged corrections perceived the correction as less aggressive, $b = -0.30$, $SE = 0.05$, $t(1558) = -6.19$, $p < .001$. There was no main effect of correction depth, nor interaction between conditions, $ps > .273$. Similarly, we found, as expected, that participants who received hedged corrections perceived the corrections as more polite, $b = 0.19$, $SE = 0.05$, $t(1557) = 4.00$, $p < .001$. Again, there was no main effect of correction depth nor interaction between conditions, $ps > .523$. Together, these measures suggest that there was a noticeable difference between direct and hedged corrective messages, such that hedged corrections were perceived as less aggressive and more polite, which supports our first hypothesis (H1). Indeed, these results suggest that our hedged condition was both definitionally manipulating hedging (i.e., via indicating uncertainty in wording by stating "I'm not sure"), as well as manipulating perceived aggressiveness and politeness of the correction. Additional analyses also suggest that hedged corrections promote slightly more positive perceptions of the corrector (see the Supplementary File).

We also performed a general linear model predicting how informative participants perceived the corrective message they received. Surprisingly, we found no main effects or interactions between correction conditions, $ps > .196$. We next performed a similar analysis substituting informativeness with unhelpfulness, but again found no main effects or interactions, $ps > .103$. Therefore, our results do not support our second hypothesis (H2), as we did not observe that participants evaluated corrections with more detailed explanatory depth as more informative or less unhelpful. These results suggest that while explanatory depth was definitionally manipulated in our design (i.e., the 'detailed explanation' correction contained information beyond a simple negation), it is not the case that explanatory depth manipulated the extent to which participants perceived the correction as informative or helpful.

3.2. No Meaningful Effect of Correction Strength and Depth on Reply Likelihood

The fraction of participants who said they would reply is shown in Figure 3.

For our main analysis, we then entered correction strength and depth, as centered dummies, plus their

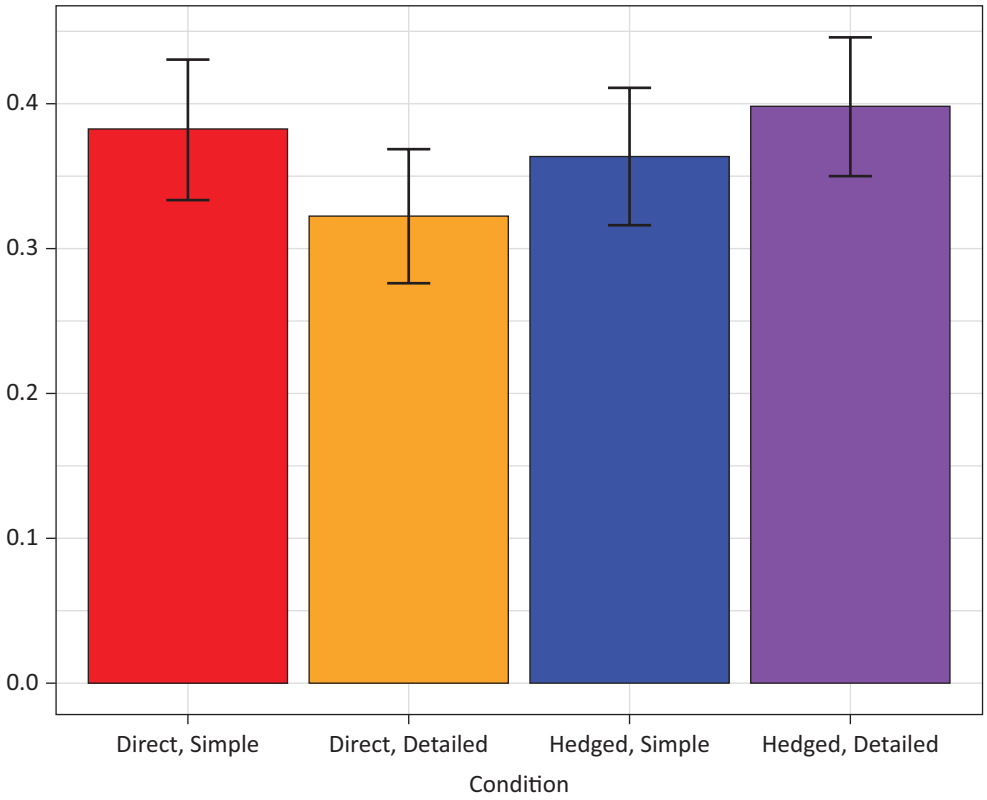


Figure 3. Likelihood of reply to the corrective message by condition. Notes: $N = 1{,}589$. Error bars reflect 95% confidence intervals.

interaction into a logistic regression to predict whether participants indicated that they would reply to the corrective message. Although we predicted that hedged and detailed corrections would elicit greater likelihood of replying, we found no main effect on reply likelihood of correction strength, $b = 0.06$, $SE = 0.05$, $z(1588) = 1.23$, $p = .219$, nor depth, $b = -0.03$, $SE = 0.05$, $z(1588) = -0.53$, $p = .598$ (Table 1).

Thus, our results do not support our third hypothesis (H3), that hedged, detailed corrections would elicit greater reply likelihood. We did find a (barely) significant interaction between correction strength and depth, $b = 0.10$, $SE = 0.05$, $z(1588) = 1.97$, $p = .049$, such that when the correction depth was detailed, hedged corrections elicited more responses than direct corrections. However, given our large sample and the fact that the p-value was only barely significant, this interaction should be interpreted with substantial caution.

3.3. No Meaningful Effect of Correction Strength and Depth on Reply Sentiment

As pre-registered, we averaged denying original belief and belief updating ratings to create a composite correction acceptance score (Figure 4).

We then entered correction strength and depth into a general linear model predicting correction acceptance score, allowing for an interaction between conditions. We found no significant main effects of correction strength, $b = 0.06$, $SE = 0.03$, $t(1588) = 1.96$, $p = .051$, or correction depth ($p = .600$) and no interaction between conditions ($p = .424$; Table 2). Our results did not support our fourth hypothesis (H4), which predicted that hedged, detailed corrections would elicit greater acceptance of information.

We next averaged the remaining five response categories (counter-arguing, attitude bolstering, etc.) into

Table 1. Logistic regression predicting likelihood of reply to corrective message.

	Estimate	Standard Error	*z*	*p*
Intercept	−0.55	0.05	−10.50	< .001***
Hedged	0.06	0.05	1.23	.219
Detailed Explanation	−0.03	0.05	−0.53	.598
Hedged*Detailed	0.10	0.05	1.97	.049*

Notes: * $p < .05$, ** $p < .01$, *** $p < .001$. Df = 1,588.

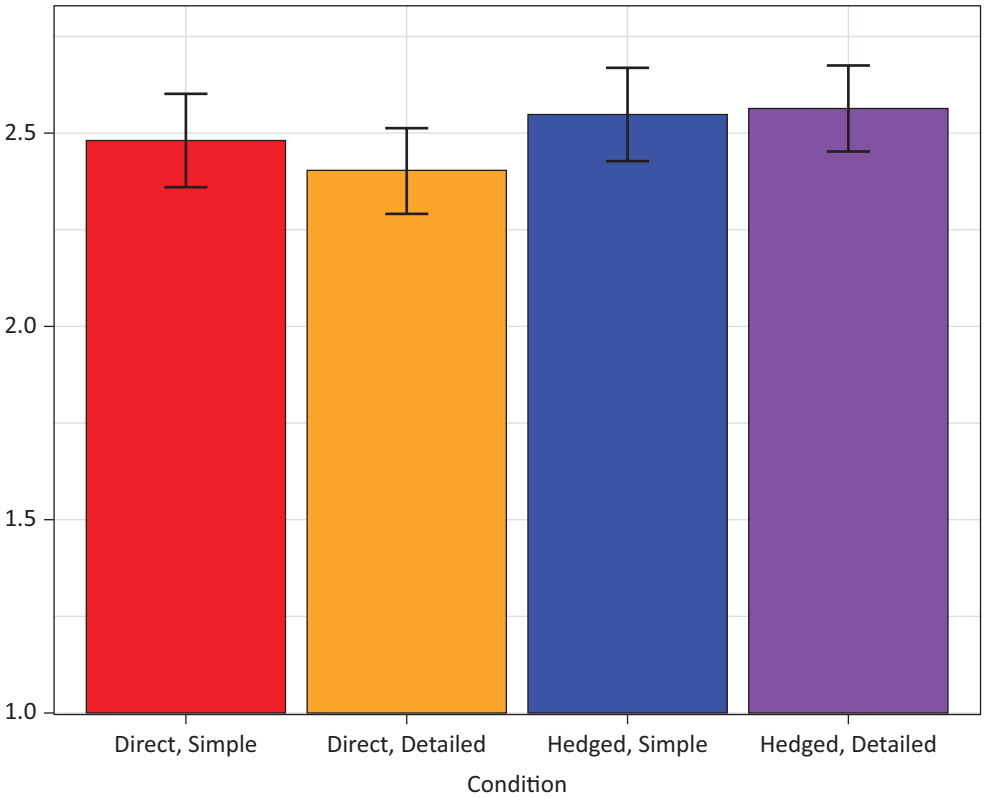


Figure 4. Average aggregated acceptance of corrective information (1–7 Likert-scale) by condition. Notes: $N = 1,589$. Error bars reflect 95% confidence intervals.

Table 2. General linear model predicting information acceptance by correction style condition.

	Estimate	Standard Error	*t*	*p*
Intercept	2.50	0.03	84.76	< .001***
Hedged	0.06	0.03	1.96	.051
Detailed Explanation	−0.02	0.03	−0.52	.600
Hedged*Detailed	0.02	0.03	0.80	.424

Notes: * $p < .05$, ** $p < .01$, *** $p < .001$. Df = 1,588.

an aggregated resisting information score (Figure 5), and predicted information resistance using a general linear model with correction strength and depth, allowing for an interaction.

We found no main effect of strength or depth, and no interaction between conditions, *p*s > .408 (Table 3).

We also performed five separate general linear models for each of the individual resisting information reply type categories. There were no significant main effects or interactions across all five linear models, *p*s > .146. Thus, our results again did not support our fourth hypothesis (H4), as direct and simple corrections did not predict increased resistance of information.

We also predicted overall positivity and overall negativity of reply using similar general linear models. Again, we found no significant main effects nor interactions, *p*s > .139.

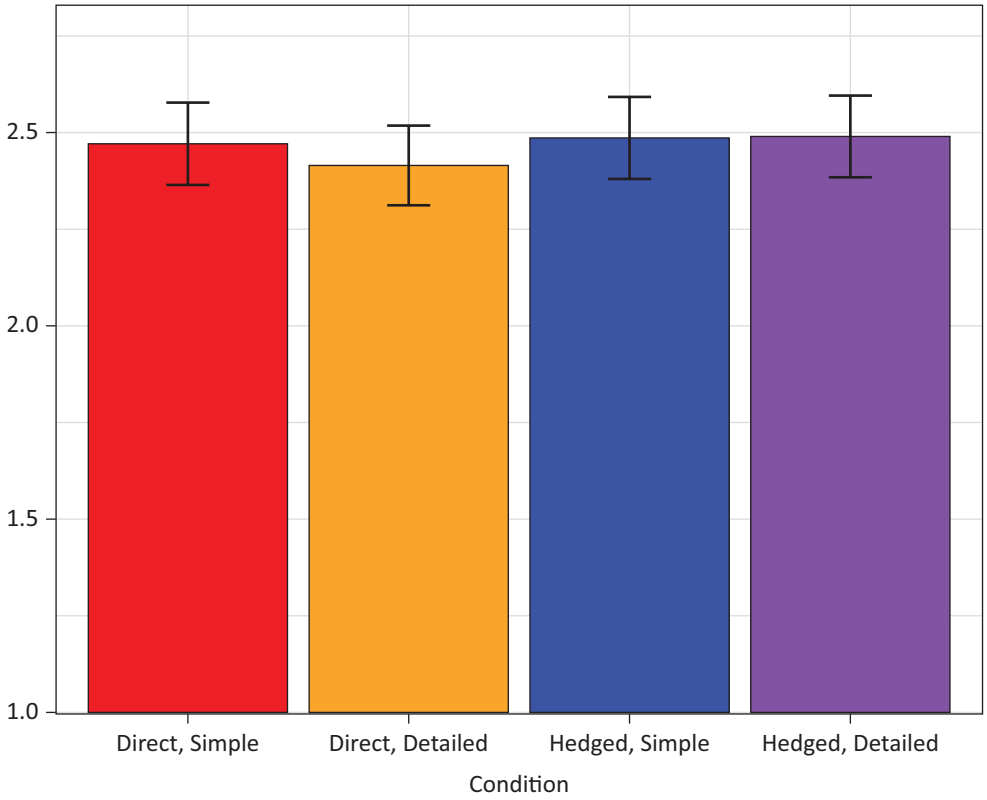


Figure 5. Average aggregated resistance of corrective information (1–7 Likert-scale) by condition. Notes: *N* = 1,589. Error bars reflect 95% confidence intervals.

Table 3. General linear model predicting information resistance by correction style condition.

	Estimate	Standard Error	*t*	*p*
Intercept	2.47	0.03	92.05	< .001***
Hedged	0.02	0.03	0.83	.408
Detailed Explanation	−0.01	0.03	−0.48	.632
Hedged*Detailed	0.01	0.03	0.56	.577

Notes: * $p < .05$, ** $p < .01$, *** $p < .001$. Df = 1,588.

3.4. No Effect of Correction Style on Self-Reported Belief Updating

We next predicted self-reported belief change (reverse-coded; 5 = Much less accurate than initially thought, 1 = Much more accurate than initially thought) using a general linear model, with correction strength and depth as predictors, allowing for their interaction (Figure 6). We found no significant main effect of correction strength, $b = 0.02$, $SE = 0.03$, $t(1551) = 0.75$, $p = .456$, or depth, $b = 0.03$, $SE = 0.03$, $t(1551) = 1.18$, $p = .240$, and no significant interaction between the conditions, $b = 0.05$, $SE = 0.03$, $t(1551) = 1.77$, $p = .078$.

3.5. Cognitive Reflection Predicts Increased Acceptance of Corrective Information

Next, we added CRT score as a predictor in our logistic regression predicting binary reply from correction strength and depth, allowing for all interactions. We found no significant main effect of, or interactions with, CRT score on reply likelihood, ps $> .132$. We then performed a similar analysis using a general linear model to predict aggregated acceptance of information. In this model, we found a notable main effect of CRT score, such that higher CRT score was associated with increased acceptance of corrective information, $b = 0.17$, $SE = 0.03$, $t(1587) = 5.77$, $p < .001$. We did not observe any significant interactions between CRT score and our correction conditions, contrary to our fifth hypothesis (H5) which predicted that more analytic participants would be more likely to accept detailed corrections (Table 4).

We next performed the same analysis except substituting accepting information with our composite resisting information score. Interestingly, we again found a positive main effect of CRT score, such that higher CRT score was associated with increased resistance of corrective information, $b = 0.06$, $SE = 0.03$, $t(1587) = 2.34$, $p = .019$; and no significant interactions between CRT score and our correction conditions (Table 5).

In order to further examine the relationship between CRT score and reply sentiment, we performed a z-test to compare the coefficient of CRT score on accepting information to the coefficient of CRT score on resisting information. We found that the coefficient of CRT score on accepting information was significantly greater than that of CRT score on resisting information, $z = 2.67$, $p = .008$. Our results thus suggest that on balance, participants with higher CRT scores are more accepting of corrective information.

3.6. Actively Open-Minded Thinking Predicts Increased Acceptance of Corrective Information

We again performed our main logistic regression model predicting binary reply, this time adding AOT score

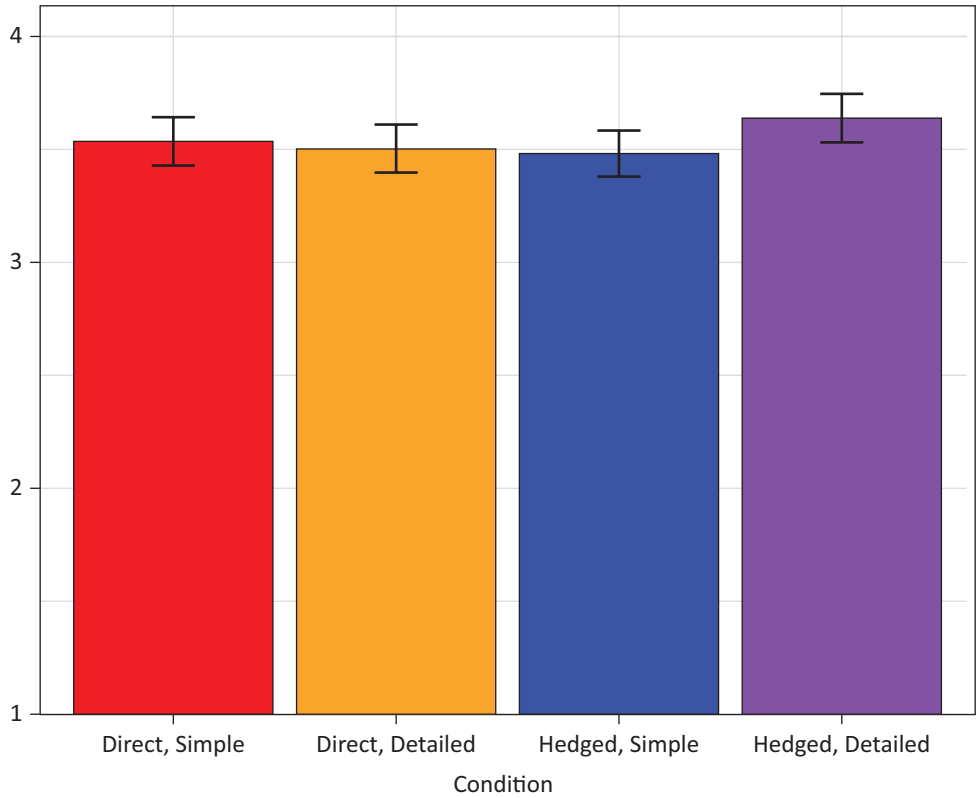


Figure 6. Self-reported belief updating (1–5 Likert-scale) by condition. Notes: $N = 1{,}552$. Error bars reflect 95% confidence intervals.

Table 4. General linear model predicting acceptance of information from correction style conditions and CRT as potential moderator.

	Estimate	Standard Error	*t*	*p*
Intercept	2.50	0.03	85.58	< .001***
Hedged	0.06	0.03	1.98	.048*
Detailed Explanation	−0.02	0.03	−0.59	.558
CRT Score	0.17	0.03	5.77	< .001***
Hedged*Detailed	0.02	0.03	0.74	.461
Hedged*CRT	0.04	0.03	1.53	.126
Detailed*CRT	−0.04	0.03	−1.20	.231
Hedged*Detailed*CRT	−0.02	0.03	−0.77	.443

Notes: * $p < .05$, ** $p < .01$, *** $p < .001$. Df = 1,587.

as a potential moderator, allowing for all interactions. We found a main effect of AOT on likelihood of reply, such that greater AOT score was associated with greater reply likelihood, $b = 0.15$, $SE = 0.05$, $z(1501) = 2.83$, $p = .005$. We found no interactions between correction conditions and AOT, ps $> .494$. We also ran a general linear model predicting aggregated acceptance of information from correction strength, depth, and AOT score, allowing for all interactions. We found a main effect of AOT score, such that greater AOT score was predictive of increased acceptance of corrective information, $b = 0.21$, $SE = 0.03$, $t(1501) = 7.00$, $p < .001$; and no significant interactions between correction conditions and AOT, ps $> .353$ (Table 6). These latter results thus do not support our fifth hypothesis (H5), as more actively open-minded participants were not more likely to accept information from detailed corrections.

We then performed the same analysis, substituting acceptance of information with aggregated resistance of information, but found no significant main effect of AOT ($p = .201$) and no significant interactions between correction conditions and AOT (ps $> .243$). Together, these results demonstrate that AOT is associated with increased acceptance, but not resistance, of corrective information.

We also performed several analyses looking at partisanship and social media use as potential moderators (see the Supplementary File).

4. Discussion

Our results suggest several conclusions about the effects of different styles of corrective messages on engagement with and replies to corrections of misinformation on social media. We find that hedged corrections are perceived as politer and less aggressive than direct corrections, and that hedged corrections result in a more positive perception of the corrector. Despite this, however, we do not find that hedged corrections are any more effective at eliciting replies to corrective messages, or promoting acceptance of corrective information. We consistently found no main effect of correction strength (direct, hedged) or explanatory depth (simple explanation, detailed explanation) on reply likelihood or reply sentiment. We did find some weak evidence of an interaction between correction strength and depth. This interaction was such that hedged, detailed corrections and direct, simple corrections yielded greater reply likelihood than direct, detailed corrections and hedged, simple corrections. This suggests that participants were perhaps sensitive to both correction strength and explanatory depth, yet neither correction style significantly impacted reply likelihood or the acceptance or rejection of the correction.

Overall, given our consistently minimal effects of correction strength and depth on responses to corrections, our findings suggest that correction style and wording

Table 5. General linear model predicting resistance of information from correction style conditions and CRT as potential moderator.

	Estimate	Standard Error	*t*	*p*
Intercept	2.47	0.03	92.07	< .001***
Hedged	0.02	0.03	0.82	.414
Detailed Explanation	−0.01	0.03	−0.52	.604
CRT Score	0.06	0.03	2.34	.019*
Hedged*Detailed	0.01	0.03	0.52	.606
Hedged*CRT	0.03	0.03	1.02	.308
Detailed*CRT	−0.02	0.03	−0.68	.498
Hedged*Detailed*CRT	−0.01	0.03	−0.29	.769

Notes: * $p < .05$, ** $p < .01$, *** $p < .001$. Df = 1,587.

Table 6. General linear model predicting acceptance of information from correction style conditions and AOT as potential moderator.

	Estimate	Standard Error	*t*	*p*
Intercept	2.53	0.03	84.81	< .001***
Hedged	0.05	0.03	1.83	.068
Detailed Explanation	−0.02	0.03	−0.58	.563
AOT Score	0.21	0.03	7.00	< .001***
Hedged*Detailed	0.002	0.03	0.08	.935
Hedged* AOT	0.03	0.03	0.93	.353
Detailed* AOT	0.003	0.03	0.09	.930
Hedged*Detailed* AOT	−0.01	0.03	−0.19	.854

Notes: * $p < .05$, ** $p < .01$, *** $p < .001$. Df = 1,501.

do not have a substantial impact on how corrections of misinformation on social media are received. These findings are consistent with recent research on correction wording and tone, which found that correction tone did not substantially affect misperceptions (see Bode et al., 2020). Our current findings extend this research in several keyways. First, we demonstrate that correction strength (roughly analogous to tone) does not significantly affect engagement with corrections of political misinformation, whereas prior work has looked at apolitical misinformation (Bode et al., 2020). Second, we also show that differences in correction strength do not impact engagement or belief updating by the user who shared the corrected misinformation themselves. This in contrast with previous work, which has instead assessed the effect of corrections on third-party viewers (observational correction, i.e., how third-party users on social media are affected by corrections; see Vraga & Bode, 2017). Third, we also show that manipulating the explanatory depth of the correction also has a minimal effect on engagement with the corrective message.

Our research also extends previous research on cognitive style and misinformation, which has found that people who are more reflective are less likely to believe false news headlines (Bronstein, Pennycook, Bear, Rand, & Cannon, 2019; Pennycook & Rand, 2019b, 2020), and that deliberation causally reduces belief in false claims (Bago, Rand, & Pennycook, 2020)—regardless of their partisan alignment. Here we examine the relationship between cognitive style on the response to corrections (rather than the perceptions of the misinformation itself). We found that analytic thinking and actively open-minded thinking (as assessed by CRT and AOT scales) predicted increased acceptance of corrective information. This willingness to update one's beliefs in the face of corrective information may help to explain why more reflective individuals have more accurate beliefs. Importantly, our results also suggest that analytic and actively open-minded thinking relate to increased acceptance of corrective information regardless of correction style.

Finally, our findings suggest that attempts at misinformation correction are not doomed to simply further promote dark engagement and incite comment section 'flame wars' (Flame war, n.d.). Indeed, self-reported belief updating was positive on average ($M = 3.54$), and average belief updating in reply texts as scored by Amazon Mechanical Turk raters ($M = 2.63$) was greater than the individual averages of all forms of resisting information (max: $M_{\text{Counter-arguing}} = 2.60$). Thus, in line with previous research (e.g., Bode & Vraga, 2018), social media may not only serve as a medium for misinformation—online platforms may also enable and encourage user-generated corrections which, regardless of strength or explanatory depth, may be effective at combatting misinformation.

4.1. Limitations

The current research has several notable limitations. First, while we use a sample that is quota-matched to the American national distribution on age, gender, ethnicity, and geographic region, our findings may not generalize to other populations. Further research examining different countries and cultures, as well as underrepresented populations, is an important direction for future work.

Second, participants were not on their actual social media platforms, did not share fake news articles on their social media platforms, and knew that the correction they received was from a fictional account. Therefore, it is critical to test how the results of the current study generalize to more ecologically valid settings. Further research should examine the impact of manipulating corrective messages via a field experiment on a social media platform such as Twitter or Facebook.

Third, our study employs only one possible manipulation of hedging, and one possible manipulation of explanatory depth. Thus, it is plausible that other formulations of hedging or explanatory depth may yield differential engagement with corrective messages. For instance, our hedged message may be overly uncertain and perhaps more polite than other possible ways to hedge (e.g., "I'm not sure" vs. "It could be the case"). Thus, more certain and less polite hedged corrections may elicit greater engagement than the hedging manipulation we utilized. Furthermore, we definitionally

manipulated explanatory depth by utilizing one condition in which the explanation was a simple negation, and the other condition included generic details about the source of the misinformation (i.e., "created by a website that purposefully makes false stories"). Given that perceptions of informativeness and unhelpfulness did not differ based on explanatory depth condition, it may be the case that either more detailed or more specific explanations may also lead to higher or lower levels of engagement with the corrective message. Future research may explore these possibilities in greater depth.

Fourth, many of our null results were not that precisely estimated. Thus, our findings should not be interpreted as evidence of no difference between correction conditions. Rather, our minimal and null results should be interpreted as a lack of evidence suggesting correction style does affect correction engagement—and, given our pre-registered prior hypotheses regarding likely differences in correction outcomes based on prior research, this lack of evidence was both surprising and complements recent research also indicating that correction style does not substantially impact correction engagement. Nonetheless, our minimal and null results should be interpreted with caution—we do not claim to find evidence of no effect of correction style on responses to misinformation, but rather present our results suggesting that our experiment yielded an absence of any evidence showing an effect of correction style.

5. Conclusions

In sum, we do not find evidence that hedging corrections of misinformation or providing increased explanatory depth in corrections of misinformation had a meaningful impact on engagement with corrective messages on social media. Although we found differences in how these messages were perceived in terms of aggressiveness or politeness, we did not find any substantial difference in likelihood of replying, overall acceptance of corrective information, or overall resistance towards corrective information. Our results also suggest that more analytic individuals, and more actively open-minded individuals, are more likely to accept corrective information, irrespective of correction strength or explanatory depth. Ultimately, our current study suggests that corrective messages, regardless of precise style or wording, may nonetheless be used as a source of positive engagement and communication on social media in order to combat dark participation.

Acknowledgments

The authors thank Antonio Arechar for assistance with data collection, and gratefully acknowledge funding from the William and Flora Hewlett Foundation, the Ethics and Governance of Artificial Intelligence Initiative, and MIT Libraries. This material is also based on work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 174530.

Conflict of Interests

The authors declare no conflict of interests.

Supplementary Material

Supplementary material for this article is available online in the format provided by the author (unedited).

References

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–36.

Allen, J., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, *6*(14). https://doi.org/10.1126/sciadv.aay3539

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

Bode, L., & Vraga, E. K. (2018). See something, say something: Correction of global health misinformation on social media. *Health Communication*, *33*, 1131–1140.

Bode, L., Vraga, E. K., & Tully, M. (2020, June 11). Do the right thing: Tone may not affect correction of misinformation on social media. *Harvard Kennedy School Misinformation Review*. Retrieved from https://misinforeview.hks.harvard.edu/article/do-the-right-thing-tone-may-not-affect-correction-of-misinformation-on-social-media

Bronstein, M. V., Pennycook, G., Bear, A., Rand, D. G., & Cannon, T. D. (2019). Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of Applied Research in Memory and Cognition*, *8*(1), 108–117.

Chan, M. P. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science*, *28*(11), 1531–1546.

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, *86*(2), 127–137.

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., . . . Sandhu, M. (2019). Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief

in false stories on social media. *Political Behavior*, *42*, 1073–1095.

Clement, J. (2020, July 15). Number of global social network users 2017–2025. *Statista*. Retrieved from https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/#:~:text=Social%20media%20usage%20is%20one,almost%204.41%20billion%20in%202025.&text=Social%20network%20penetration%20is%20constantly,2020%20stood%20at%2049%20percent

Coppock, A., & McClellan, O. A. (2019). Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Research & Politics*, *6*(1). https://doi.org/10.1177/2053168018822174

Epstein, Z., Pennycook, G., & Rand, D. (2020). Will the crowd game the algorithm? Using layperson judgments to combat misinformation on social media by downranking distrusted sources. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–11). New York, NY: Association for Computing Machinery.

Flame war. (n.d.). In *Dictionary.com*. Retrieved from https://www.dictionary.com/browse/flame-war

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*, 25–42.

Frenkel, S., Alba, D., & Zhong, R. (2020, March 8). Surge of virus misinformation stumps Facebook and Twitter. *The New York Times*. Retrieved from https://www.nytimes.com/2020/03/08/technology/coronavirus-misinformation-social-media.html

Kim, J., Tabibian, B., Oh, A., Schölkopf, B., & Gomez-Rodriguez, M. (2018). Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the eleventh ACM international conference on web search and data mining* (pp. 324–332). New York, NY: Association for Computing Machinery.

Lakoff, G. (1975). Hedges: A study in meaning criteria and the logic of fuzzy concepts. In D. Hockney, W. Harper, & B. Freed (Eds.), *Contemporary research in philosophical logic and linguistic semantics* (pp. 221–271). Dordrecht: Springer.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., . . . Schudson, M. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096.

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131.

Margolin, D. B., Hannak, A., & Weber, I. (2018). Political fact-checking on Twitter: When do corrections have an effect? *Political Communication*, *35*(2), 196–219.

Mosleh, M., Pennycook, G., & Rand, D. G. (2020). Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter. *Plos One*, *15*(2). https://doi.org/10.1371/journal.pone.0228882

Nyhan, B., & Reifler, J. (2012). *Misinformation and fact-checking: Research findings from social science*. New York, NY: New America Foundation.

Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, *66*(11), 4921–5484. https://doi.org/10.1287/mnsc.2019.3478

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (in press). Shifting attention to accuracy reduces online misinformation. *Nature*.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting Covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, *31*(7), 770–780.

Pennycook, G., & Rand, D. G. (2019a). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, *116*(7), 2521–2526.

Pennycook, G., & Rand, D. G. (2019b). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50.

Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*, *88*(2), 185–200.

Prasad, M., Perrin, A. J., Bezila, K., Hoffman, S. G., Kindleberger, K., Manturuk, K., & Powers, A. S. (2009). "There must be a reason": Osama, Saddam, and inferred justification. *Sociological Inquiry*, *79*(2), 142–162.

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2931–2937). Copenhagen: Association for Computational Linguistics.

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), 1–10.

Shearer, E., & Matsa, K. E. (2018, September 10). News use across social media platforms 2018. *Pew Research Center*. Retrieved from https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in God. *Journal of Experimental Psychology: General*, *141*, 423–428.

Stanovich, K. E., & West, R. F. (2007). Natural myside bias

is independent of cognitive ability. *Thinking & Reasoning*, *13*, 225–247.

Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., & Lee, L. (2016). Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web* (pp. 613–624). Republic and Canton of Geneva: International World Wide Web Conference Committee (IW3C2).

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, *11*, 99–113.

Treventhan, R. (2017). Intraclass correlation coefficients: Clearing the air, extending some cautions, and making some requests. *Health Services and Outcomes Research Methodology*, *17*(2), 127–143.

Vraga, E. K., & Bode, L. (2017). Using expert sources to correct health misinformation in social media. *Science Communication*, *39*(5), 621–645.

Vraga, E. K., & Bode, L. (2018). I do not believe you: How providing a source corrects health misperceptions across social media platforms. *Information, Communication & Society*, *21*(10), 1337–1353.

About the Authors

Cameron Martel is a Graduate Student at Sloan School of Management, Massachusetts Institute of Technology. His research explores the cognitive processes underlying the belief and spread of misinformation. He is also broadly interested in social and political behavior online, particularly on social media platforms. In his research, he uses a variety of approaches from cognitive and social psychology, behavioral economics, and computational social science. He received his BS in cognitive science from Yale University. His work is supported by the National Science Foundation Graduate Research Fellowship.

Mohsen Mosleh is a Lecturer (Assistant Professor) at the Science, Innovation, Technology, and Entrepreneurship Department, University of Exeter Business School. Mohsen was a Postdoctoral Fellow at the MIT Sloan School of Management as well as the Department of Psychology at Yale University. Prior to his postdoctoral studies, Mohsen received his PhD from Stevens Institute of Technology in Systems Engineering with a minor in data science. Mohsen's research interests lie at the intersection of computational/data science and cognitive/social science. In particular, he studies how information and misinformation spread on social media, collective decision-making, and cooperation.

David G. Rand is the Erwin H. Schell Professor and Associate Professor of Management Science and Brain and Cognitive Sciences at MIT. His research combines behavioral experiments and online/field studies with mathematical/computational models to understand human decision-making. David focuses on illuminating why people believe and share misinformation and 'fake news,' understanding political psychology and polarization, and promoting human cooperation. He was named Poynter Institute Fact-Checking Researcher of the year in 2017, and received the 2020 FABBS Early Career Impact Award from the Society for Judgment and Decision Making.

COGITATIO

Media and Communication (ISSN: 2183–2439)
2021, Volume 9, Issue 1, Pages 134–143
DOI: 10.17645/mac.v9i1.3409

Article

From Dark to Light: The Many Shades of Sharing Misinformation Online

Miriam J. Metzger [1,*], Andrew J. Flanagin [1], Paul Mena [2], Shan Jiang [3] and Christo Wilson [3]

[1] Department of Communication, University of California, Santa Barbara, CA 93106, USA;
E-Mails: metzger@ucsb.edu (M.J.M.), flanagin@ucsb.edu (A.J.F.)
[2] Writing Program, University of California, Santa Barbara, CA 93106, USA; E-Mail: pmena@ucsb.edu
[3] Khoury College of Computer Sciences, Northeastern University, Boston, MA 02115, USA;
E-Mails: sjiang@ccs.neu.edu (S.J.), cbw@ccs.neu.edu (C.W.)

* Corresponding author

Submitted: 30 June 2020 | Accepted: 7 August 2020 | Published: 3 February 2021

Abstract
Research typically presumes that people believe misinformation and propagate it through their social networks. Yet, a wide range of motivations for sharing misinformation might impact its spread, as well as people's belief of it. By examining research on motivations for sharing news information generally, and misinformation specifically, we derive a range of motivations that broaden current understandings of the sharing of misinformation to include factors that may to some extent mitigate the presumed dangers of misinformation for society. To illustrate the utility of our viewpoint we report data from a preliminary study of people's dis/belief reactions to misinformation shared on social media using natural language processing. Analyses of over 2,5 million comments demonstrate that misinformation on social media is often disbelieved. These insights are leveraged to propose directions for future research that incorporate a more inclusive understanding of the various motivations and strategies for sharing misinformation socially in large-scale online networks.

Keywords
credibility; fake news; misinformation; news sharing

Issue
This article is part of the issue "Dark Participation in Online Communication: The World of the Wicked Web" edited by Thorsten Quandt (University of Münster, Germany).

1. Introduction

Few would dispute that the web has grown more wicked in recent years with the rise of 'fake news.' Indeed, the creation and spread of misinformation online for political or financial gain presents one of the greatest challenges to democratic society in modern history. Fake news and misinformation are said to increase political polarization, alter voters' perceptions of candidates and issues, and erode trust in critical democratic institutions (Allcott & Gentzkow, 2017; Ciampaglia, Mantzarlis, Maus, & Menczer, 2018; Hochschild & Einstein, 2015). Moreover, propelled largely through social media, fake news has been shown to spread faster and further than 'real' news (Lazer et al., 2018; Vosoughi, Roy, & Aral, 2018).

The danger of misinformation is that people will find such information to be credible. Fake news is thought to be dangerous because people are likely to believe the false information, and because it is spread through people's trusted contacts via social media, it will be similarly trusted by others. Hence, the presumed threat of fake news is that people will erroneously believe false information they encounter and that they will in turn propagate misinformation to others. Yet, there are likely many motivations for sharing information—including fake news—within one's social network that might impact both its spread and people's belief of it. For example, people might share misinformation for entertainment purposes, sarcastic reasons, or to illustrate a point counter to the message promoted in a false

news story. Under such circumstances, the danger of fake news may be less than feared or, perhaps, even mitigated or reversed.

This article interrogates common presumptions about sharing misinformation by analyzing people's motives for sharing fake news stories on social media. Drawing on literatures on news sharing, rumors, gossip, and urban legends suggests that a wide range of motivations exists for sharing misinformation, and that various motivations imply both positive and negative outcomes for democratic processes. Preliminary data from a research program that aims to understand dis/belief in fake news and misinformation in social media are then presented as proof-of-concept for some of the ideas about motives for sharing misinformation that are advanced in this article.

2. Motivations for Sharing News Information

Research on people's motivations to share news in general, and specifically via social media, is extensive. Motives for sharing news include: acting as an opinion leader, advocating for one's own beliefs, socializing, gaining social status, sharing experiences with others, and informing others. Although the news media decide which stories to publish, users make judgments about the value of what is published, which in turn affects news consumption. Through such processes of 'secondary gatekeeping,' whereby users share news stories with others (Shoemaker & Vos, 2009; Singer, 2014), information consumers propose what others should read, establishing "not only what is valuable to them as individuals but also what they believe will be important, interesting, entertaining, or useful to others" (Singer, 2014, p. 58). In this way, users filter, evaluate, and finally share news information.

This secondary gatekeeping capability has been enhanced by social media, where individuals are able to actively participate in the dissemination of news to large audiences. Indeed, users may experience a sense of agency when sharing news on social media by feeling that they have some control over what is posted on a social network (Oeldorf-Hirsch & Sundar, 2015). In this sense, the desire to act as a gatekeeper of information or an opinion leader may be a fundamental motivation for news sharing. Indeed, research has found that self-perceptions of opinion leadership influence users' news sharing intention (Ma, Sian Lee, & Hoe-Lian Goh, 2014) and that political expression enhances people's motivations to self-present as politically active on social media (Lane et al., 2019). Moreover, users may feel motivated to share news that supports their own views and contradicts dissenting beliefs since advocating for one's own beliefs can be a motivator of news sharing (Dafonte-Gómez, 2018), consistent with evidence of a relationship between news sharing and ideology spreading (Lottridge & Bentley, 2018).

In addition, studies have found that social gratifications and social status are motivators of news sharing on social media (Bright, 2016; Choi, 2016; Lee & Ma, 2012). A sense of connection with the online community is developed through news sharing (Lee & Ma, 2012) and sharing information with others contributes to satisfying the need for social interaction, which "helps people clarify their opinions, and gives them an opportunity to work out their personal positions relative to media content" (Weeks & Holbert, 2013, p. 215). Additionally, social standing can be achieved by sharing news deemed useful to those receiving it, which can "make the person passing it on appear well informed and intelligent" (Bright, 2016, p. 346). News sharing may thus help people gain status among peers (Bruns, 2018; Choi, 2016; Lee & Ma, 2012), particularly if they feel that they were informed earlier than others (Kubey & Peluso, 1990).

In some cases, sharing news may be a by-product of people's psychological need to share experiences with others (Harber & Cohen, 2005; Kubey & Peluso, 1990), consistent with the notion that people who have experienced disturbing events are emotionally compelled to share their experiences: "An unintended but often real consequence of their efforts to unburden themselves, we believe, is to inform members of their social networks of valuable news" (Harber & Cohen, 2005, p. 383). In this way, people may share news to relieve their own feelings (Kubey & Peluso, 1990). This emotional component of news sharing has been examined by researchers who found that emotional arousal can explain transmission of news (Berger & Milkman, 2012; Dafonte-Gómez, 2018), such that news evoking high-arousal positive or negative emotions is shared more than news evoking low-arousal emotions.

Finally, researchers have also identified altruistic motives for news sharing (Bruns, 2018; Chadwick & Vaccari, 2019; Dafonte-Gómez, 2018; Kümpel, Karnowski, & Keyling, 2015). For example, surveys among U.K. social media users found that "to inform others" and "to express my feelings" were the most important motivations for news sharing (Chadwick & Vaccari, 2019). Users feel motivated to serve others' information needs and share meaningful news with their online community, to such a degree that social media users are even prone to adapt news sharing to their audience's interests (Rudat, Buder, & Hesse, 2014).

3. Motivations for Sharing Misinformation

While plenty of research exists on why people share news online, only a handful of studies have examined people's motivations for sharing 'fake' news stories (e.g., Chadwick & Vaccari, 2019; Chen & Sin, 2013; Chen, Sin, Theng, & Lee, 2015; Duffy, Tandoc, & Ling, 2019; Talwar, Dhir, Kaur, Zafar, & Alrasheedy, 2019). Moreover, most studies assume that people do not realize the information they share is false—in other words, that the sharing of misinformation, including fake news, is unintentional.

The picture is more complicated with misinformation, however, because people may or may not know

the information is false when they share it, and thus may spread the misinformation either intentionally or unintentionally (Lawrie, 2019). Unintentional fake news sharing may be motivated by self-expression and socialization (Chen et al., 2015). For instance, Chen and Sin (2013) found that people share misinformation mainly to obtain other's opinions, to express their own opinions, and to interact with others. Similarly, researchers have suggested that the motivation to build relationships can lead people to unintentionally share fake news (Duffy et al., 2019). In fact, the factors that drive people to share real news, including emotional impact and relevance, are likely to be the same factors that make fake news highly shared as well (Duffy et al., 2019).

However, when the sharer knows the information is false, then other motives likely come into play. Researchers suggest that people knowingly share fake news because it conforms to their prior views and in order to maintain positive social relations (Duffy et al., 2019). Intentional fake news sharing has been positively associated with social media use behaviors such as self-disclosure, online trust, fear of missing out, and social media fatigue (Talwar et al., 2019). More specifically, people who tend to disclose more information in social media generally, those who receive a fake news item from a trusted source, people who seek popularity and a sense of belonging (i.e., fear social exclusion), and who experience information overload are more likely to pass fake news on to others in their social networks.

In addition, research shows that people who seek to entertain, troll, or debate with others are more likely to engage in intentional misinformation sharing, while those who seek to persuade or inform others are less likely to do so (Chadwick, Vaccari, & O'Loughlin, 2018). This suggests that some people share fake news as a way to disrupt political dialogue, while others with more civic incentives share fake news as part of legitimate political debate:

> Those motivated to debate may see sharing problematic news as a cultural norm; a practice that is simply part of 'what it takes' to engage politically on social media in order to attract attention and nudge others to take positions. (Chadwick et al., 2018, p. 4269)

Among participants in the study, 8.9% admitted to sharing news that they thought was made up when they shared it and 17.1% deliberately shared news that was exaggerated. Another study showed that 17.3% of British news sharers on social media knowingly shared news that they thought was made up (Chadwick & Vaccari, 2019). Similarly, a Pew Research Center survey found 14% of U.S. adults said they had shared a news story they knew was fake (Barthel, Mitchell, & Holcomb, 2016). And sharing fake news may be underreported due to social desirability biases that likely suppress self-reported data on misinformation sharing.

Finally, fake news sharing has some similarities to rumor, gossip, and urban legends. Research by Guerin and Miyazaki (2006), for instance, suggests "the primary function of telling rumors, gossip, and urban legends is not to impart information to the listener or alleviate listener anxiety about the topic, but to entertain or keep the listener's attention, thereby enhancing social relationships" (p. 23). Others have argued that both rumor sharing and fake news sharing fulfill three motivations: "To cope with uncertainty, build relationships, and for self-enhancement" (Duffy et al., 2019, p. 3). In this fashion, people engage in rumor as a way to collectively reduce uncertainty, enhance interpersonal relationships, and feel positive about themselves (Bordia & DiFonzo, 2005), as may be the case with sharing news misinformation as well.

4. Broadening the Range of Motivations for Sharing Misinformation

There is still much to be learned about why people share misinformation. Theoretical explanations are nascent, and only a few empirical studies have been conducted to date (Talwar et al., 2019). There is certainly a host of reasons in addition to those discussed above why people share fake news and misinformation. For example, intentionally sharing misinformation may be done to engage in 'collective fact-checking,' which likely derives from the more general motivation for uncertainty reduction in information sharing as mentioned earlier. Fake news stories can be sensationalistic and often contain other elements such as extreme partisan information that could arouse suspicion about their veracity (Mourão & Robertson, 2019). Consequently, recipients of such stories may share the suspected misinformation with their friends and followers on social media in an effort to crowdsource its truthfulness, and thus reduce uncertainty as to the information's credibility.

Another potential consequence of the sensationalism of many fake news stories is that their absurdity can provide fodder for ridicule. Sarcasm or mockery of such stories may therefore be another reason to share news that one knows is false. This reason could be conceived as a specialized subset of sharing for socializing and entertainment purposes. In such cases, the misinformation is shared for the purpose of highlighting its falsity to others. This is different type of social entertainment motive than passing a story to a friend you think would find it interesting or spreading a titillating rumor for its shock value, which have been discussed previously in the news sharing literature (Lee & Ma; 2012; see also Kümpel et al., 2015).

A motivation that ultimately shares the same purpose as sarcasm is sharing misinformation for educational purposes. Although 'informing others' is a well-known reason for sharing news information (e.g., Chadwick & Vaccari, 2019), our concept of education differs from prior notions of imparting information to others because it is not about informing recipients of the *content* of the story, but rather about informing

them about the *credibility* of the story itself. Such sharing thus functions to debunk misinformation, and serves as a warning to others not to believe the information. Alerting recipients to potential misinformation may stem from simple altruistic motives, or may be self-serving if it is done to demonstrate one's moral or intellectual superiority. If the latter, it may be a new variant of the 'status seeking' category of news sharing motives, as it casts the sharer as a well-informed opinion leader.

Thus, while social media users share fake news that they do not know or suspect to be false for the same reasons that they share any other type of news, there exist motivations that are unique to sharing misinformation that one knows to be false. An important issue that has not been explored is that a person's motivation for sharing information is likely to have a strong impact on the credibility beliefs of the recipient. If a sharer appears to believe the information they share, the recipient may also. Yet, if a sharer appears not to believe the information (for instance, if the information is shared for sarcastic or educational purposes), the recipient may not either. But how can the recipient know what a sharer believes?

Although another person's psychological motivation for sharing a piece of information and his or her credibility evaluation might be difficult for recipients to discern in many cases, the communication surrounding the shared information can offer useful clues. For example, a prominent fake news story shared on social media in 2016 said "Pope Francis Shocks World, Endorses Donald Trump for President." On social media, this post could be accompanied by various captions or comments that may signal different degrees of belief in the misinformation. For example, if the sharer were to write: "See, this proves I was right all along, I knew he would support Trump!" or even "Oh no, this is awful!" this might signal that the sharer finds the information to be credible. On the other hand, if the commentary instead read: "Oh yeah right, sure he did," or "Get a load of this, it's the funniest thing I've seen in a long time!" or "What kind of dope would believe this bull****???" it would indicate skepticism or disbelief on the part of the sharer (Figure 1).

Recent research supports the assertion that recipients of information notice cues regarding credibility from sharers. A study by Colliander (2019) found that exposure to others' comments that were critical versus supportive of a fake news story resulted in less favorable attitudes toward the news article, prompted participants to write more negative comments themselves about the article, and lowered their intentions to share the article in their social networks. Colliander concluded that "social media users seem to use the comments of other people as a guide for how to respond to disinformation online" (p. 208) based on human psychological tendencies toward conformity and a desire to maintain a positive self-concept. Winter, Bruckner, and Krämer (2015) observed similar social influence effects of negatively—versus positively-valenced mainstream news comments.

The foregoing discussion raises the possibility that the danger to democracy and to society due to misinformation and fake news circulating in social media may be mitigated at least partially to the extent that people spread fake news stories as a tool to engage with others in deliberative communication about its truthfulness or to signal its falsity. As U.S. Supreme Court Justice Louis Brandeis famously wrote, the remedy to 'bad' speech is not censorship, it is more speech, because through discussion and debate false information will be revealed as false. Sharing misinformation for purposes such as collective fact-checking, sarcasm, or education calls out the information as false, and thus may be good for deliberative democracy. In this way, and borrowing a term from van Heekeren (2020), social media commentary surrounding misinformation may function as a 'curative' to the problem of fake news, particularly over time as people reap the benefits of the collective sharing and debunking of misinformation.

To illustrate this potential, a preliminary study is next described in which the content of the communication surrounding misinformation is analyzed to gain insight into the range of reasons that people propagate such information, and how it may over time alter information interpretation. More specifically this research analyzed social media users' commentary on known fake news stories as a means to better understand public perceptions of dis/belief in, and credibility assessments of, misinformation shared online.

Figure 1. Example of a fake news post with examples of hypothetical commentary that signal a range of beliefs by the sharer. Source: "Pope Francis shocks world" (2016).

5. Preliminary Study of Dis/Belief in Misinformation Shared in Social Media

This study aimed to quantify the prevalence of dis/belief in misinformation at scale using a machine learning approach. The research was conducted in four steps. Step 1 involved identifying social media posts that contain misinformation and collecting user comments in response to those posts to analyze. Step 2 required reading the comments to determine if they reflect the commenter's belief or disbelief in the misinformation. Each comment was manually labeled by independent coders, and these data were used in Step 3 to examine language differences in comments expressing belief and disbelief, and to test if such differences could be used to build a classifier to identify dis/belief with reasonable accuracy. Finally, the classifier was leveraged in Step 4 to measure dis/belief at scale to answer the following RQs:

> RQ1: To what extent does the public believe misinformation shared via social media?

If misinformation is not believed to a large extent, that would discount much of its presumed danger to society. And if evidence of disbelief is found in comments shared by users, there is potential for those comments to influence others' beliefs in a cascading manner:

> RQ2: Is there is a time effect for expressed disbelief in misinformation, where the public gradually realizes the truth after a false claim is made, and therefore belief in false claims decreases over time?

This question is raised in light of possibilities of crowdsourcing for misinformation detection, including via formal reporting mechanisms (i.e., letting users flag fake news; see Tschiatschek, Singla, Gomez Rodriguez, Merchant, & Krause, 2018) or by the social influence processes proposed herein, whereby recipients are influenced by sharers' critical comments on fake news stories. While the technical details of the research are presented in full elsewhere (citation blinded for review), the methods employed and research results at each step are summarized in the next section.

5.1. Step 1

A sample of social media posts containing at least some misinformation was collected from a census of PolitiFact's fact-checked articles between January 1 to June 1, 2019. Posts were manually identified as containing misinformation that originated from Twitter, which was used because its API allowed the collection of user comments written in response to posts containing misinformation at the time of data collection, unlike other social media platforms. Next, comments (i.e., tweets) in response to the posts were collected. Using the fact-checked posts as seeds, a 1% sample of the tweet stream was queried to capture all comments to the seed posts. To reduce noise, only posts that had $>$ 50 comments were retained, which resulted in 6,809 comments to analyze in Step 2.

5.2. Step 2

Once the comments written in response to posts containing misinformation were collected, they were manually annotated with belief and disbelief labels. Two independent, trained coders read each comment and provided a binary label: either disbelief (i.e., the person who wrote the comment does not appear to believe the misinformation) or belief (i.e., the person who wrote the comment does appear to believe the misinformation). These two labels are mutually exclusive but not necessarily complementary, that is, although a tweet is not expected to show both belief and disbelief, it can show neither.

The inter-annotator percent agreement (i.e., the number of agreed labels over the total count) was used to evaluate intercoder reliability. 66.7% of the coder pairs were above 80% agreement, 88.9% were above 70% agreement, and only two were below 60% agreement, suggesting an acceptable level of agreement among annotators, especially for a relatively subjective task. To obtain a final label for each comment, a third independent coder read through all cases where the two original coders disagreed and provided a final tie-breaking judgement. Out of 6,809 tweets, 2,399 (35.2%) were labeled as expressing disbelief, 1,282 (18.8%) were labeled as expressing belief, 3,128 (45.9%) were labeled as neither, and none (0%) were labeled as both.

5.3. Step 3

Using the labeled dataset, a lexicon-based exploratory analysis of the language used across the comments expressing belief and disbelief was conducted. Two lexicons were employed: (a) LIWC (Tausczik & Pennebaker, 2010), the most widely-used lexicon for understanding psychometric properties of language, and (b) ComLex (Jiang & Wilson, 2018), a more contextual lexicon built from social media comments to misinformation, containing additional domain-specific categories, e.g., 'fake,' 'fact,' 'hate speech.' This technique provided a frequency for each category in the lexicon, which allowed a comparison of the distributions of language frequencies between comments expressing belief, disbelief, and neither by performing t-tests with Bonferroni corrections.

Results showed that comments expressing disbelief contained significantly more falsehood awareness language, including referrals to falsehood such as 'lie, propaganda' and 'fake, false'; referrals to the truth 'fact, research'; and negative character portraits such as 'liar, crook' and 'stupid, dumb.' Comments expressing disbelief also contained more negative emotions and negation (e.g., 'no, not') and less positive emotions and discrepancy words (e.g., 'should, would'). By contrast,

comments expressing belief contained fewer falsehood awareness signals (e.g., 'fake, false') and negative character portrait ('stupid, dumb'). Comments expressing belief also contained more exclamation (e.g., exclamation marks, '!, yay'); discrepancy words; and fewer negative reactions such as swearing (e.g., 'damn, fuck') and anger (e.g., 'hate, kill'). All t-tests were significant at least $p < .01$.

Next, natural language processing models were applied to build a classifier to automatically identify and label comments that express dis/belief in misinformation. Different types of classifiers were experimented with and the neural transfer-learning based classifier RoBERTa was found to be the best in terms of accuracy of classification, as determined by F1 scores (binary F1 scores were around 0.86 for disbelief and 0.80 for belief).

5.4. Step 4

The final step was to use the classifier on a larger dataset to label dis/belief expressed in responses to misinformation at scale. The dataset, collected previously by Jiang and Wilson (2018), contained 1,672,687 comments collected from Facebook, 113,687 from Twitter, and 828,000 from YouTube written in response to 5,303 fact-checked claims. These claims are drawn from the entire archive of Snopes and PolitiFact's articles between their founding and January 9, 2018. The dataset was fed to the RoBERTa-based classifier from Step 3 to answer the research questions.

RQ1 asked about the prevalence of dis/belief in misinformation. To analyze this, all of the original claims were parsed into three mis/information types: (a) true, if the claims were rated as 'true' by Snopes or PolitiFact—these claims contain no misinformation; (b) mixed, if the claims were rated as 'mostly true,' 'half true,' or 'mixed'—these claims contain some misinformation but also some truth; and (c) false, if the claims were rated as 'mostly false,' 'false,' or 'pants on fire!'—these claims contain mostly falsehood. Comments in the dataset that were posted after its corresponding factcheck article was published were then filtered out. Finally, the remaining comments were grouped by the mis/information type, and their dis/belief labels were averaged.

Results showed that as the veracity of claims decreased, disbelief increased while belief decreased (Figure 2). Specifically, 12%, 14%, and 15% of comments expressed disbelief in response to true, mixed, and false claims, respectively; whereas 26%, 21%, and 20% of comments expressed belief in response to true, mixed, and false claims, respectively.

Notably, the majority of comments indicated no sure indication of belief or disbelief and therefore yield no information regarding people's reactions to the claims spread via social media. Among comments indicating dis/belief, however, several implications emerge. First, an overall skepticism is evident from the data: not only do many people not believe fake information, but many people similarly disbelieve (12%) true information (and only 26% report believing true information). Thus, people appear generally distrustful of claims shared via social media in this study, suggesting a certain level of cautiousness that might guard against the deleterious effects of misinformation, consistent with the underlying rationale that not all misinformation is uniformly harmful. One potential explanation is that the partisan environment drives people to suspect any claims from the opposite ideological group regardless of veracity (Hochschild & Einstein, 2015). Another, though less likely, explanation is that people tend to be skeptical of all claims in social media, even when the claim is consistent with existing facts. Both explanations warrant further investigation.

Second, there is some indication that a portion of people commenting on misinformation are doing so to express their own disbelief in the content. This again supports the perspective that not all misinformation sharing implies tacit endorsement of the claims made therein, and provides evidence that at least some users are leveraging the power of social media in a fashion that can serve as a partial corrective to misinformation. If users are indicating that they do not believe the misinformation they see in social media, it suggests alternative, and perhaps useful, responses regarding content that is in turn shared with others.

That said, the difference in the prevalence of dis/belief across the mis/information types is small (e.g., 3% more people disbelieve versus believe false

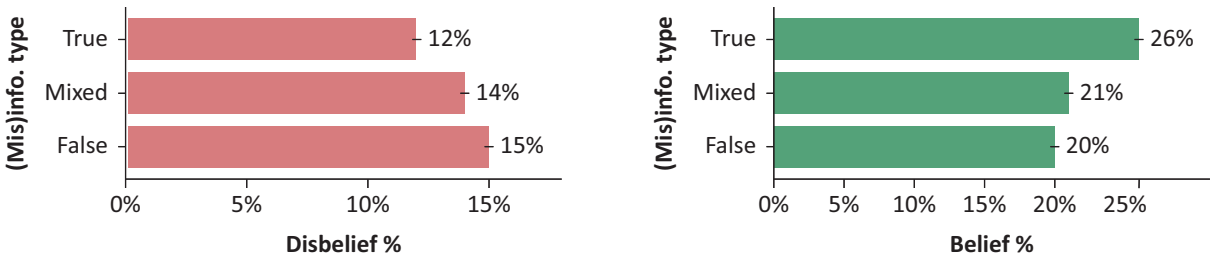


Figure 2. Prevalence of dis/belief for true, mixed, and false claims on social media. Notes: For disbelief (left graph), as the veracity of the claims decreases, the prevalence of expressed disbelief increases. For belief (right graph), as the veracity of the claims decreases, the prevalence of expressed belief also decreases.

statements). Yet, at scale, even small differences represent a substantial volume of occurrences where people are more likely to disbelieve versus believe false claims. Moreover, taken alone, the fact that in this study 15% of users report not believing misinformation is notable.

The final analysis sought to answer RQ2, which asked whether belief in false claims decreases over time. The data revealed a very small-time effect, where disbelief increases 0.001% and belief decreases 0.002% per day after the initial false claim appeared. Notably, at scale this small-time effect would yield a large number of changes in dis/belief. This effect could be a result of seeing prior comments that are critical of the claim (i.e., a sharer's expressed disbelief in the misinformation), but further studies that move beyond machine learning techniques are needed to determine if this explanation is correct.

For example, social scientific methodologies such as surveys would be useful to elicit the variety of reasons why people share misinformation that they know to be false in social media, how often they accompany it with commentary signaling its veracity, and also the extent that social media users report receiving such information from their network contacts. Experiments similar to Colliander (2019) could shed light on the power of negative comments to alter recipients' credibility beliefs, especially if conducted unobtrusively where actual behavioral data (e.g., misinformation sharing) could be observed.

In sum, the results of our preliminary study show evidence that when people share misinformation via social media, they do not always take it at face value. For example, over a third of the comments analyzed in Step 2 reflected disbelief in tweets containing misinformation, and the wider analysis in Step 4 revealed that about one in six people disbelieved false information they encountered in various social media platforms. Moreover, our analyses found that belief in false claims decreases slightly over time, which might at least partly be a result of exposure to social comments critical of the misinformation, although future research will need to confirm this potential explanation. These data should prompt scholars to expand their thinking on why people share misinformation beyond unintentional sharing of (believed) misinformation, to suggest a healthy process of intentional social debunking of fake news that is rarely examined in the literature.

6. Conclusion

Although the creation and spread of misinformation online represent a serious challenge to democratic society, the nature and extent of the problem might to some degree be overstated. If, as our preliminary evidence suggests, misinformation on social media is often disbelieved, and to the extent that those sharing it are doing so for reasons that expose and help to stem the spread of misinformation, then shared misinformation is in fact not universally harmful and its propagation is not always and necessarily detrimental. By examining research on motivations for sharing news information generally, and misinformation specifically, we derive a range of motivations—including entertainment, sarcasm, and education—that broaden our understanding of the sharing of misinformation to include factors that may help to mitigate its danger.

Yet, because research to date has almost exclusively characterized information sharers as passive diffusers of misinformation, and information consumers as universally receptive to it, little is known about the prevalence and dynamics of information consumers' potentially defensive or constructive reactions to misinformation. To address this deficiency research must consider a host of factors that consider the sharing and consumption of misinformation as a more active, critical, and strategic process. For instance, future research is required to establish baseline data on the wide range of motivations invoked by those sharing misinformation, the extent to which they are aware of its in/accuracy, and the degree of intentionality in its sharing with others.

Relatedly, research must assess the extent to which misinformation that is shared with the intention of debunking it is in fact interpreted appropriately, and the degree to which social, psychological, and contextual factors affect such outcomes. Moreover, the complex network dynamics of social media information sharing (e.g., which network nodes are most influential, in what domains), coupled with the relational attributes that guide the mis/interpretation of shared misinformation (e.g., the relations among information sources and recipients), need to be examined to understand social influence processes in this context. Finally, paralinguistic cues (e.g., emojis, likes, etc.) should be examined as markers of endorsement, signals of alternative motivations (e.g., sarcasm), and cues demonstrating the extent of shared group identity, all of which may be interrelated in complex ways to affect misinformation belief.

In sum, insights into the wide range of motivations for sharing misinformation via social media suggest that the spread of misinformation should be viewed as a dynamic process among active users driven by a wide range of factors. Viewed this way, the negative aspects of misinformation may to some degree be mitigated and can only be fully understood as future research incorporates a more inclusive understanding of the various motivations and strategies for sharing misinformation socially in large-scale online networks.

Acknowledgments

This research was supported in part by NSF grant IIS-1553088. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

Conflict of Interests

The authors declare no conflict of interests.

References

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236.

Barthel, M., Mitchell, A., & Holcomb, J. (2016, December 15). Many Americans believe fake news is sowing confusion. *Pew Research Center*. Retrieved from http://www.journalism.org/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, *49*(2), 192–205.

Bordia, P., & DiFonzo, N. (2005). Psychological motivations in rumor spread. In G. A. Fine, C. Heath, & V. Campion-Vincent (Eds.), *Rumor mills: The social impact of rumor and legend* (pp. 87–101). New York, NY: Aldine Press.

Bright, J. (2016). The social news gap: How news reading and news sharing diverge. *Journal of Communication*, *66*, 343–365.

Bruns, A. (2018). *Gatewatching and news curation*. New York, NY: Peter Lang.

Chadwick, A., & Vaccari, C. (2019). *News sharing on UK social media: Misinformation, disinformation, and correction* (Report No. O3C 1). London: Loughborough University. Retrieved from https://repository.lboro.ac.uk/articles/News_sharing_on_UK_social_media_misinformation_disinformation_and_correction/9471269

Chadwick, A., Vaccari, C., & O'Loughlin, B. (2018). Do tabloids poison the well of social media? Explaining democratically dysfunctional news sharing. *New Media & Society*, *20*(11), 4255–4274.

Chen, X., & Sin, S. J. (2013). Misinformation? What of it? Motivations and individual differences in misinformation sharing on social media. *Proceedings of the American Society for Information Science and Technology*, *50*(1), 1–4. Retrieved from www.asis.org/asist2013/proceedings/submissions/posters/23poster.pdf

Chen, X., Sin, S. J., Theng, Y., & Lee, C. S. (2015). Why students share misinformation on social media: Motivation, gender, and study-level differences. *The Journal of Academic Librarianship*, *41*, 583–592.

Choi, J. (2016). Why do people use news differently on SNSs? An investigation of the role of motivations, media repertoires, and technology cluster on citizens' news-related activities. *Computers in Human Behavior*, *54*, 249–256.

Ciampaglia, G. L., Mantzarlis, A., Maus, G., & Menczer, F. (2018). Research challenges of digital misinformation: Toward a trustworthy web. *AI Magazine*, *39*(1), 65–74.

Colliander, J. (2019). "This is fake news": Investigating the role of conformity to other users' views when commenting on and spreading disinformation in social media. *Computers in Human Behavior*, *97*, 202–215.

Dafonte-Gómez, A. (2018). Audience as medium: Motivations and emotions in news sharing. *International Journal of Communication*, *12*, 2133–2152.

Duffy, A., Tandoc, E., & Ling, R. (2019). Too good to be true, too good not to share: The social utility of fake news. *Information, Communication & Society*, *23*(13), 1–15.

Guerin, B., & Miyazaki, Y. (2006). Analyzing rumors, gossip and urban legends through their conversational properties. *Psychological Record*, *56*, 23–34.

Harber, K., & Cohen, D. (2005). The emotional broadcaster theory of social sharing. *Journal of Language and Social Psychology*, *24*, 382–400.

Hochschild, J. L., & Einstein, K. L. (2015). *Do facts matter? Information and misinformation in American politics*. Norman, OK: University of Oklahoma Press.

Jiang, S., & Wilson, C. (2018). Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proceedings of the ACM on Human-Computer Interaction*, *2*, 1–23.

Kubey, R. W., & Peluso, T. (1990). Emotional response as a cause of interpersonal news diffusion: The case of the space shuttle tragedy. *Journal of Broadcasting and Electronic Media*, *34*, 69–76.

Kümpel, A. S., Karnowski, V., & Keyling, T. (2015). News sharing in social media: A review of current research on news sharing users, content, and networks. *Social Media + Society. 1*(2). https://doi.org/10.1177/2056305115610141

Lane, D. S., Lee, S. L., Liang, F., Kim, D. H., Shen, L., Weeks, B. E., & Kwak, N. (2019). Social media expression and the political self. *Journal of Communication*, *69*(1), 49–72.

Lawrie, L. (2019). *The fake news crisis of 2016: The influence of political ideologies and news trust on news consumer "innocent sharing"* (Unpublished doctoral dissertation). Wichita State University, Wichita, KA.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., . . . Rothschild, D. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096.

Lee, C. S., & Ma, L. (2012). News sharing in social media: The effect of gratifications and prior experience. *Computers in Human Behavior*, *28*(2), 331–339.

Lottridge, D., & Bentley, F. R. (2018, April 21). Let's hate together: How people share news in messaging, social, and public networks. In R. Mandryk & M. Hancock (Eds.), *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). New York, NY: Association for Computing Machinery.

Ma, L., Sian Lee, C., & Hoe-Lian Goh, D. (2014). Understanding news sharing in social media: An explanation from the diffusion of innovations theory. *Online Information Review*, *38*(5), 598–615.

Mourão, R. R., & Robertson, C. T. (2019). Fake news as discursive integration: An analysis of sites that publish false, misleading, hyperpartisan and sensational information. *Journalism Studies*, *20*(14), 2077–2095.

Oeldorf-Hirsch, A., & Sundar, S. S. (2015). Posting, commenting, and tagging: Effects of sharing news stories on Facebook. *Computers in Human Behavior*, *44*, 240–249.

Pope Francis shocks world, endorses Donald Trump for president, releases statement. (2016, July). *WTOE 5 News*. https://wtoe5news.com

Rudat, A., Buder, J., & Hesse, F. W. (2014). Audience design in Twitter: Retweeting behavior between informational value and followers' interests. *Computers in Human Behavior*, *35*, 132–139.

Shoemaker, P. J., & Vos, T. P. (2009). *Gatekeeping theory*. New York, NY: Routledge.

Singer, J. B. (2014). User-generated visibility: Secondary gatekeeping in a shared media space. *New Media & Society*, *16*, 55–73.

Talwar, S., Dhir, A., Kaur, P., Zafar, N., & Alrasheedy, M. (2019). Why do people share fake news? Associations between the dark side of social media use and fake news sharing behavior. *Journal of Retailing and Consumer Services*, *51*, 72–82.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*, 24–54.

Tschiatschek, S., Singla, A., Gomez Rodriguez, M., Merchant, A., & Krause, A. (2018). Fake news detection in social networks via crowd signals. In P.-A. Champin, F. Gandon, & L. Médini (Eds.), *Companion proceedings of The Web Conference* (pp. 517–524). New York, NY: Association for Computing Machinery.

van Heekeren, M. (2020). The curative effect of social media on fake news: A historical re-evaluation. *Journalism Studies*, *21*(3), 306–318.

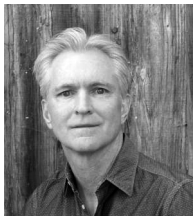
Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Weeks, B. E., & Holbert, R. L. (2013). Predicting dissemination of news content in social media: A focus on reception, friending, and partisanship. *Journalism & Mass Communication Quarterly*, *90*, 212–232.

Winter, S., Bruckner, C., & Krämer, N. C. (2015). They came, they liked, they commented: Social influence on Facebook news channels. *Cyberpsychology, Behavior, and Social Networking*, *18*(8), 431–436.

About the Authors

Miriam J. Metzger is a Professor of Communication at the University of California, Santa Barbara. Her research lies at the intersection of media, information technology, and trust, centering on how information technology alters humans' understandings of information credibility, privacy, and how networked communication technologies affect behavior.



Andrew J. Flanagin is a Professor in the Department of Communication at the University of California, Santa Barbara, where he is a former Director of the Center for Information Technology and Society. His work broadly considers processes of social influence in digitally-mediated environments, with emphases on the use of social media for information sharing and assessment; people's perceptions of the credibility of information gathered and presented online; and processes of collective organizing as influenced by the use of contemporary technologies.

Paul Mena teaches journalism and news writing at the University of California, Santa Barbara. He holds a PhD in Mass Communication from the University of Florida. His research interests include misinformation, credibility, fact-checking, journalism, and social media.

Shan Jiang is a PhD Student and Research Assistant in the Khoury College of Computer Sciences at Northeastern University. His research interests revolve around computational social science and natural language processing. His current work investigates how storytellers, platforms, and audiences interact in the online misinformation ecosystem. He has published award-winning papers and served on the program committees at WWW, CSCW, ICWSM, and AAAI.

Christo Wilson is an Associate Professor in the Khoury College of Computer Sciences at Northeastern University. He is a Founding Member of the Cybersecurity and Privacy Institute at Northeastern, and serves as Director of the Bachelors in Cybersecurity program. Professor Wilson's research focuses on online security and privacy, with a specific interest in using 'algorithm audits' to increase transparency and accountability of black-box online platforms.

COGITATIO

Media and Communication (ISSN: 2183–2439)
2021, Volume 9, Issue 1, Pages 144–157
DOI: 10.17645/mac.v9i1.3405

Article

Digital Civic Participation and Misinformation during the 2020 Taiwanese Presidential Election

Ho-Chun Herbert Chang [1,2,*], Samar Haider [2] and Emilio Ferrara [1,2]

[1] Annenberg School for Communication and Journalism, University of Southern California, Los Angeles, CA 90087, USA; E-Mails: hochunhe@usc.edu (H-C.H.C.), emiliofe@usc.edu (E.F.)
[2] Information Sciences Institute, University of Southern California, Los Angeles, CA 90087, USA; E-Mail: samarhai@usc.edu

* Corresponding author

Submitted: 30 June 2020 | Accepted: 5 September 2020 | Published: 3 February 2021

Abstract
From fact-checking chatbots to community-maintained misinformation databases, Taiwan has emerged as a critical case-study for citizen participation in politics online. Due to Taiwan's geopolitical history with China, the recent 2020 Taiwanese Presidential Election brought fierce levels of online engagement led by citizens from both sides of the strait. In this article, we study misinformation and digital participation on three platforms, namely Line, Twitter, and Taiwan's Professional Technology Temple (PTT, Taiwan's equivalent of Reddit). Each of these platforms presents a different facet of the elections. Results reveal that the greatest level of disagreement occurs in discussion about incumbent president Tsai. Chinese users demonstrate emergent coordination and selective discussion around topics like China, Hong Kong, and President Tsai, whereas topics like Covid-19 are avoided. We discover an imbalance of the political presence of Tsai on Twitter, which suggests partisan practices in disinformation regulation. The cases of Taiwan and China point toward a growing trend where regular citizens, enabled by new media, can both exacerbate and hinder the flow of misinformation. The study highlights an overlooked aspect of misinformation studies, beyond the veracity of information itself, that is the clash of ideologies, practices, and cultural history that matter to democratic ideals.

Keywords
2020 Taiwanese Presidential Election; digital civic participation; foreign interference; misinformation; Taiwan

Issue
This article is part of the issue "Dark Participation in Online Communication: The World of the Wicked Web" edited by Thorsten Quandt (University of Münster, Germany).

1. Introduction

Taiwan is one of the freest regions in Asia from a socio-political standpoint, and yet it receives some of the highest concentrations of online disinformation, due to its geo-political history with China (Monaco, 2017). With waning trust in the traditional media over the course of recent years (Hu, 2017), Taiwan has turned to grass-roots cyber-interventions, spearheaded by its community of civic 'hacktivists' (Fan et al., 2019; Rowen, 2015). The recent 2020 Taiwanese Presidential Election has thus presented a fierce battleground that re-examines democratic values.

Citizen participation in support of and detrimental to democratic ideals is not a new phenomenon. In *Dark Participation*, Quandt contrasted the utopian vision and dark realities of citizen news making (Quandt, 2018). Benight participation involves citizen-journalists who selflessly take part in democratic deliberation. Dark participation, in contrast, describes negative contributions to news production. This includes "trolling," piggy-backing off untruths, and the dissemination of disinformation. Recent studies have linked this with the growing populism in the West, a political trend also observed in Taiwan. Han Kuo-Yu, the presidential candidate running against the incumbent President Tsai, is frequently

compared to US President Trump in both his rise as a businessman-turned-politician and his use of politicized rhetoric (Cole, 2019).

Two years after dark participation has been first characterized, the political media ecosystem has evolved. The case of Taiwan encapsulates two gray areas in this light–dark dichotomy. First, both the diffusion and defense against misinformation are citizen-driven. Disinformation has been well-documented to arises from nationalized citizens from China (Yang, Chen, Shih, Bardzell, & Bardzell, 2017), instead of government-sponsored campaigns. Rather than light and dark participation as characterized by Quandt, the elections characterize the clash of divergent political ideologies rooted in seven decades of history.

Second, the use of digital tools to fight disinformation is a double-edged sword. The Anti-infiltration Act, passed two weeks prior to the elections, caused significant controversy, with its critics worrying it was too partisan. The former head of the National Communication Commission, who allegedly resigned over disagreements for this act, stated that although "disinformation is the enemy of an open and democratic society…they [might] lose that open society by fighting against it" (Aspinwall, 2020a). The use of technology to promote certain political discourses against foreign interference may appear positive, while simultaneously diminishing the vibrancy of domestic discourse.

1.1. Research Questions and Contributions

This article presents the case study of the 2020 Taiwanese Presidential Election, told through the lens of three widely adopted platforms: Line, Twitter, and Taiwan's Professional Technology Temple (PTT, Taiwan's equivalent of Reddit). Each platform reveals a unique dimension to the election's discourse. We draw primarily from two theoretical framings. First, we postulate the influx of disinformation as a threat to three democratic normative goods: self-determination, accountable representation, and public deliberation (Tenove, 2020). Second, we consider the modal actors of media regulation, specifically from political parties to grassroots volunteers.

As we will see in Section 2, disinformation has always played a part in Taiwanese elections. A walk through history shows interference techniques morphing from direct displays of military power to subtle digital manipulation efforts. However, pinpointing the sources of misinformation is a difficult task, which depends on each platform's accessibility. Instead, we focus on understanding the discourse topics, the users involved, and how they engage in discussion over these democratic, normative goods. Our research questions and hypotheses are as follows:

RQ1: How do discourse, user behavior, and political intent vary across and within platforms?

- H1a: Twitter will contain higher levels of foreign users, consistent with known percentages of platform usage.

RQ2: On Twitter, do we observe instances of geopolitical divisions and transnational solidarity?

- H2a: High levels of transnational support exist between Taiwan and Hong Kong.
- H2b: There will be higher levels of bot-like behavior from mainland Chinese users.

RQ3: Which democratic normative goods appear vulnerable to misinformation?

- H3: Posts about Tsai, and hence engagement with issues of accountable representation, will produce high levels of disagreement from Chinese users and rural areas.

RQ4: What is the role of the traditional media in spreading disinformation?

- H4: A sizable proportion of news articles will contain misinformation, consistent with the distrust in the traditional media.

By answering these questions, we aim to contribute critical literature about the next phase of the light and dark debate, specifically how citizens respond collectively to address dark participation. Like other Asian countries, Taiwan's social media ecosystem is dominated by chatroom-based communication, a distinction from the West. As misinformation spreads behind these closed doors, the power of government policies is limited. Disinformation regulation in these cases becomes less a matter of policy, and more a community norm, in line with recently proposed theoretical frameworks (Starbird, Arif, & Wilson, 2019). Taiwan's case study enables us to understand communal commitments to maintain the quality of public deliberation.

2. Background

2.1. The 2020 Taiwanese Elections: The Political Backdrop

In 1949, the conflict between Taiwan and China began. Facing defeat by the communist party, General Chiang Kai-Shek retreated to the island, and Taiwan has been a "de facto independent nation" (Monaco, 2017) ever since. Over the last four decades, a significant divide between the two countries has developed, in language, culture, and governance. According to the Freedom House in 2019, only Japan has a higher score for political freedom than Taiwan in Asia (Freedom House, 2019). Politically, two dominant parties have emerged:

- Democratic Progressive Party (DPP): The DPP is the nationalist and more liberal-leaning party in Taiwan. They are traditionally seen as more independence-leaning or holding a stronger sense of Taiwanese national identity.
- Kuo Ming Tang (KMT): The KMT is also a nationalist and liberal political party in Taiwan. They ruled the Republic of China between 1928 to 2000, after retreating to Taiwan in 1949. They traditionally advocate for closer economic ties with China.

These two parties provided the primary candidates during prior elections, and the 2020 elections were no exception. President Tsai Ing-wen from the DPP sought to defend her presidency, whereas Han Kuo-Yu was the opposing candidate nominated by the KMT. Tsai was, for the most part, an institutional candidate. She has law degrees from both Cornell and the London School of Economics and began her political career in independent governmental positions; she served as a trade negotiator for the WTO on behalf of President Lee Teng-hui. She joined the DPP in 2004 and was elected into leadership in 2008 as the first woman chairing a political party. She was defeated by Ma Ying-jeou in her first presidential run of 2012 and, in her second bid in 2016, she won by a landslide (Hsiao, 2016).

A year prior to the election Tsai was projected to lose to Han, due to a few factors. First, wage stagnation, public pension reform, and same-sex marriage led to general discontent toward her presidency. As a result, she suffered an astounding defeat during the 2018 local elections. However, the Hong Kong Anti-extradition Bill Protests triggered a change in sentiment across the island. After Chairman Xi Jinping gave a hardliner speech regarding the one-China policy, polls showed dramatic improvement to Tsai's campaign (Hsiao, 2019).

Han was different from previous candidates due to his unconventional background and rapid rise in political power. Starting a political career from unknown origins, he became the mayor of the third-largest city in Taiwan, Kaohsiung. No one expected him to win. Kaohsiung has been the DPP stronghold for more than 20 years, and the KMT chairman Wu Den-yih sent Han to contest Kaohsiung with no expectations of victory (Jansen, 2019). Yet, he won in a landslide, owing to a surge of popular support the media called the 'Han wave.' His slogan was simple—*Get Rich!* His iconic hairless head earned him the nickname 'the Bald Guy.' Within six months of his election, he declared his run for presidency (Reichenbach, 2020).

Importantly, the *Han Wave* bears many similarities to US President Trump's *Make America Great Again* movement (Cole, 2019). As the Han Wave swept across the island, he accrued a large group of dedicated 'Han fans' estimated at 1,2 million. He appealed to rural voters, employed economy-focused brash rhetoric, and most critically, he entertained. Similar to the way the media latched onto President Trump's tweets, Han appeared frequently on social media, the news, and in discourse led by supporters from China. It was against this backdrop—a dark horse candidate who had flipped the DPP's most supported city—that the 2020 Taiwanese Election was held.

2.2. Taiwan's History of Foreign Interference

Foreign interference from China is intimately tied with Taiwan's elections, first taking form as military exercises. Before Taiwan's first presidential election in 1996, the People's Liberation Army fired missiles in the water around the island, in a show of intimidation. In the form of information warfare, radio stations and large speakers project sound across the strait to influence the elections.

In recent years, interference from China has taken a different form. Chinese trolling has often been described as decentralized, arising from netizens (Internet citizens). *Diba*, a sizable group of Chinese nationalists is known to overcome China's Great Firewall to troll Taiwanese political leadership (Yang et al., 2017). Interestingly, *Diba* violates the People's Republic of China's legal norms for spreading pro-People's Republic of China messages on the Internet, in a manner ironically similar to movements on self-determination.

To contextualize what misinformation looks like in Taiwan, we present two recent cases. The first was after Typhoon *Jebi* hit Japan and knocked out Osaka's Kansai International Airport, a report from PTT said China had evacuated Chinese nationals from the airport. The report then said that if Taiwan citizens identified themselves as Chinese they would also be evacuated. Taiwan's Foreign Ministry Representative Su Chii-cheng, following waves of criticism that he failed to protect Taiwanese citizens during this natural disaster, committed suicide. After his death, it was revealed that the Chinese Communist Party (CCP) was also unable to evacuate Chinese citizens, and the original message, shared repeatedly online and amplified by legacy media outlets, was fabricated. The message was eventually traced back to Weibo (China's main microblogging site). The second case was during the 2018 mid-term elections, a widespread *ghost island* meme spread across social media, stoking fear of opportunity loss, economic stagnation, and government corruption. The term first arose on PTT, now used as self-deprecating criticism about Taiwan, but was successfully used by Chinese users to agitate feelings of emptiness and pessimism toward Taiwan's economic future.

While the source of false news may arise from mainland China, its amplification is often a direct result of Taiwan's traditional media. In these cases, although the CCP helped stoke fears by supporting these stories, the primary spread arose from sensational-oriented journalism practices in Taiwan itself.

Prior theories on the organization of disinformation campaigns show the modal actors of authoritarian regimes are the central governments, whereas

in democracies this is taken up by political parties (Bradshaw & Howard, 2018). Monaco (2017) delineates propaganda in Taiwan in two primary forms: 1) Internal campaigns—domestic political propaganda on social issues, usually between these two parties, where the modal actors are the political parties; 2) cross-strait campaigns—propaganda that originates from the mainland to promote unification discourse, where the modal actors are the central government (CCP).

As modal actors of regulation are also political parties, attempts to stymie disinformation may become internal campaigns of propaganda. In other words, the modal actor for defending against foreign disinformation may become the perpetrator domestically. Additionally, the case of *Diba* contradicts this framework, as it is not centralized and organized, but decentralized and spontaneous. To understand this gray area in greater depth, we review Taiwan's regulation of media platforms against misinformation.

2.3. The 2020 Elections: Working Together with Social Media Companies

On December 31, 2019, a highly controversial Anti-Infiltration Act was passed in the Legislative Yuan. The law regulated the influence of entities deemed foreign hostile forces in Taiwan (Aspinwall, 2020b). Containing 12 articles, it barred people from accepting money or acting on foreign instruction. Penalties were severe: violations include fines up to $10 million NTD ($333,167 USD) and five years in prison.

The passage of the law came with criticism. The KMT criticizing the incumbent DPP party for forcing it through legislation. As mentioned prior, the former director of the National Communications Commission believed it to negatively impact domestic free speech. However, although the nature and substance of misinformation were debated, both parties agreed foreign interference should be regulated on social media.

Information travels fast as Taiwan is one of the most technologically integrated countries, with an 86% Internet penetration rate and 78.8% smartphone penetration rate (Thomala, 2020). Moreover, around 60% of Taiwanese use social media to source news, particularly for civic and political engagement (Chen, Chan, & Lee, 2016). Table 1 shows the overall usage rates for platforms in Taiwan.

Leading up to the elections, Facebook and Line came under scrutiny for the different ways they function. Facebook is a more open, profile-based social network. Line is a chatroom service, therefore more 'private.' Dr. Puma Shen, a key member of the misinformation taskforce, categorized misinformation acting in three modalities (Hioe, 2020a):

1. Online and digital: On public social media platforms like Facebook.
2. Offline and digital: Apps such as Line disseminating messages directly from user to user.
3. Offline and physical: Local gangs, temples, and village leaders have for a long time taken illicit payments. As an example, many sources of payment were through off-shore, Chinese–Malaysian companies.

The importance of Facebook became apparent in the 2014 elections for the Taipei mayor. Ko Wen-ze, a physician with slight Asperger's, became the first alternative candidate to become elected mayor. As a believer in quantitative analytics, his campaign was driven by an in-depth analysis of *11–14 million* Facebook profiles, in a country of 23 million. In response, Facebook set up a 'war room' to help regulate content (Huang, 2019).

Due to this distrust in the traditional media, Taiwan has turned to third-party, cyber-solutions to help decide what sources of news are credible. Since chatrooms in Line are not available to the public moderation, misinformation flourishes. The Cofacts chatbot was created to counter chatroom-based misinformation (Han, 2018). Developed by g0v (gov-zero), a grassroots civic hacker group in Taiwan, users who receive questionable messages forward them to the Cofacts chatbot. The message is then added to a database, fact-checked by editors, before returned to the user. Future incidents of the same article are then automatically replied.

Taiwan is not unique in its attempts to fact-check, since Brexit and the 2016 US Presidential Election revealed the impact of misinformation. Ahead of the polls in July 2018, 90 media outlets in Mexico banded together to fact-check election misinformation, in collaborative, journalistic fact-checking (Terceros, 2018). Singapore has a state-run fact-checker called Factually, and Indonesia holds weekly misinformation briefings. However, the entirely citizen-driven approach is unique to Taiwan, though it exists along-side of governmental solutions and official resources. This crowd-sourced approach addresses centralized shortcomings and is consistent with advantages shown by Pennycook and Rand, particularly in regards to source credibility (Epstein, Pennycook, & Rand, 2020), quality (Pennycook & Rand, 2019), and publisher credibility (Dias, Pennycook, & Rand, 2020).

While Line and Facebook are conduits, PTT has emerged as an important source in Chinese misinforma-

Table 1. Social media platform usage in Taiwan.

Media Platform	Facebook	Line	Messenger	WeChat	Twitter
Usage Rate	89%	84%	55%	33%	27%

tion campaigns. Many PTT accounts are auctioned off Shopee, an auction website used frequently in Taiwan and Southeast Asia. They have also appeared on Taobao, China's auction site, with the most influential accounts being reportedly sold for $6,500 USD. As with the case of the Jibe typhoon, many journalists use PPT to source information, which causes false claims to be repeated via the traditional media.

2.4. How Disinformation Harms Democracy: Normative Goods Threatened by Disinformation

Here, we distinguish between misinformation and disinformation. The primary distinction is postulated upon intent. Misinformation denotes false information that is shared, regardless of an intent to mislead (Karlova & Fisher, 2013). It is generally accepted as a fallible aspect of human nature, in our propensity to misremember, mishear, and share sensational information. Disinformation denotes false information disseminated as a hostile act or political subversion. It is the intentional diffusion of misinformation.

When considering disinformation, there are vague assertions to how its spread is detrimental to democratic societies. It is valuable to discuss the specific *loci* it damages. Tenove typologies three democratic normative goods threatened by disinformation that require different policy responses (Tenove, 2020).

Self-determination refers to the ability of a democratic population to enact collective rules to govern themselves. Thus, they are primarily addressed through security policies at the international and domestic levels. This is perhaps most salient to Taiwan's governance. However, many contemporary democratic theorists maintain foreign influence is beneficial to self-determination. In a globalized world, the actions of one state influence another. Thus, policies of disinformation regulation draw the limit for which foreign actors can influence domestic policy.

Self-determination and Taiwan's sovereignty lie at the center of every election, with this time often emerging through Hong Kong. Solidarity between Taiwan and Hong Kong is not new, and modern support can be traced to the Sunflower movement in 2014. The slogan "Today's Hong Kong, Tomorrow Taiwan" emerged then, which showed Hong Kong as a constant measure of what happens if Taiwan loses its democratic freedom. This projection goes both ways: Hong Kong often frames Taiwan as a political utopia and is posed as a "lost historical possibility" (Hioe, 2020b). As we will see, the Hong Kong protests play a decisive role in shaping the discourse during the elections.

Accountable representation refers directly to the procedures of elections. In these cases, disinformation challenges citizen trust in elected representatives (European Commission, 2018). Classic examples include false claims as to where and when voting occurs, as demonstrated in the 2016 US Presidential election (DiResta et al., 2019). Another example includes false stories targeting specific candidates. In the 2020 Taiwanese Presidential election, two major stories emerged to discredit Tsai. According to these sources of false news Tsai faked her college degree and was a secret homosexual who wanted to corrupt Taiwanese children. The second false story stated that Tsai wanted to sell the country out to Japan and the US.

Public deliberation addresses the quality of public discourse. Rather than addressing actors themselves, as national security and election policies do, public discourse is protected via media regulation. According to theories of deliberative democracies, critical to well-informed public decision making requires communicative exchanges among citizens (Habermas, 1996). Here, disinformation threatens to undermine deliberative systems by increasing the quantity of false claims, diminishing engagement and opportunities to engage in public discussions.

The measures of our analysis are materializing: We wish to understand how these three normative goods emerge during the 2020 Taiwanese Elections. The same piece of misinformation can simultaneously act on all three goods. Next, we consider the specifics of the dataset and methods of our analysis.

3. Methods

3.1. Data

We use three main sources of data—Twitter, PTT, and Cofacts. First, we scraped Twitter using a keyword list pertaining to the elections, including the names of the three primary candidates (Tsai ing-wen, Han Kuo-Yu, and James Song and their parties). We also tracked terms about the election broadly, such as *Taiwan2020*, *ComeHomeAndVote*, and *TaiwanVote*.

As an overview of the dataset, Table 2 shows the general distribution of tweet languages. Since we have filtered using Taiwan as a necessary keyword, this data set is topically bound to discourse about the island. We observe a high level of Japanese and English tweets, which reflects the high Twitter usage in the West and Japan. In Japan, Twitter is used by 45 million monthly users (35%) and is the highest across all social media platforms (Yoshida, 2018). In comparison, Facebook only has 22% penetration ("Kokunai mau 2,800 man-ri toppa,'' 2017).

Second, we scraped PTT, often described as the Reddit of Taiwan. It was founded in 1995 by students at the National Taiwan University. With more than 1,5 million registered users, up to 150,000 users can be online during peak hours. Around 500,000 comments are posted every day. The structure of PTT is similar to that of Twitter, as users can like and reply to threads. However, reactions can be positive (推) or negative (虛).

In this article, we scraped 11,116 unique bulletin posts between November 1, 2019, and January 21, 2020, filtered on posts relating to the elections. The subset of

Table 2. Top 10 languages by the number of tweets.

Language	# of tweets	% of total
English	238,915	57.31%
Japanese	81,466	19.54%
Chinese	57,998	13.91%
German	5,792	1.39%
Undetected	4,807	1.15%
Spanish	4,651	1.12%
Thai	3,892	0.93%
French	3,095	0.74%
Portuguese	2,495	0.60%
Creole	2,494	0.60%

keywords included the three main candidates, and the words election and vote. For a cleaner subset, we vied to include only posts who included these words in their title. This totaled to 960,000 individual comments and replies on the posts, with IP, time, and date.

Third, and most importantly for misinformation, we analyzed discourse on Line. We use the Cofacts database—a public, crowd-sourced data set of misinformation—and we used the four, relevant data tables listed below:

1. Articles: Specific links, texts, or articles that users forward.
2. Article replies: A table that aggregates replies to an article, with a score of 1 or -1, indicating *True* or *False*.
3. Replies: An editor's reply to an article. There are four outcomes: a) Misinformation, the article contains potential misinformation or is unverified; b) opinion, the article contains information that is an opinion; c) non-rumor, the article is factual; d) not-an-article, the article does not pertain to Cofacts.
4. Reply requests: Includes the article ID, but also the *reason* why it was included.

Each of these data sources reveals a different aspect of misinformation during the 2020 Taiwanese Elections. Twitter shows the coverage of the elections from actors domestic and abroad. PTT shows the domestic discourse, and due to its provision of IP, geo-local distribution of discourse. Cofacts shows the types of posts that arouse suspicion, including fact-checked labels for whether they contain misinformation, and opinion, or fact. The primary form of our analysis consists of time-series and network analysis, with cross-sectional analysis in volume.

4. Results

Figure 1 gives an overview of participation volume across all three channels. Twitter is shown in blue, PTT in purple, and Line in green. Immediately, we observe a rise in volume as we approach the day of the election, January 11, 2020. However, on the day itself, levels depress in PTT and Line. The levels of Line are also more consistent throughout and increases after January 20th.

This is likely due to the combination of two reasons. First, it is against the law in Taiwan to post about the elections on the day ballots are counted. Since Line is closely tied to one's individual account, levels remained constant. We see a similar dip in PTT, though posts were still being written. This may reflect the semi-anonymous

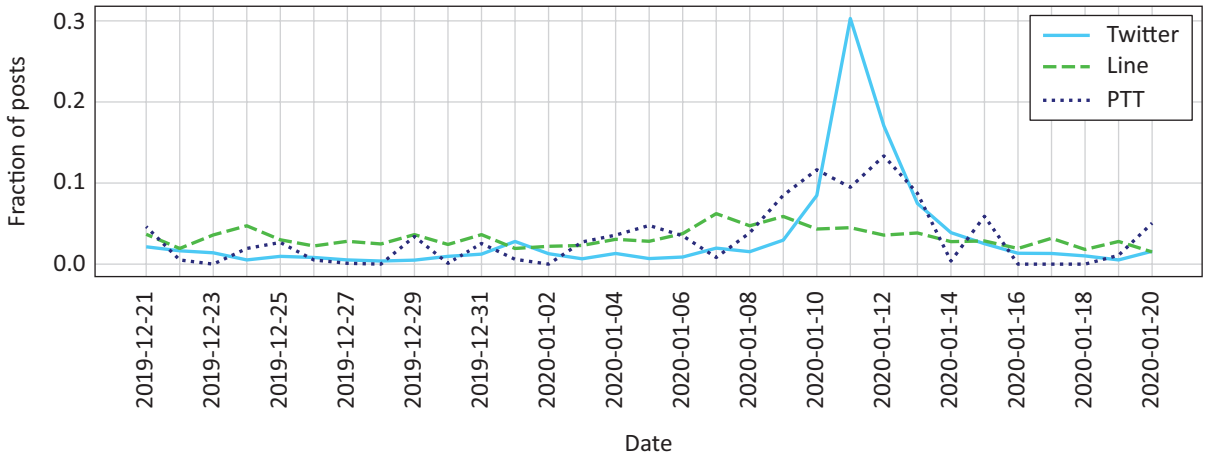


Figure 1. Fraction of posts per day on each platform during the collection timeframe.

nature of the platform. This suggests that electoral regulation was enacted unequally across platforms, and answers partially RQ1. Second, due to Twitter's low usage in Taiwan and higher penetration in Japan and the West, we observe a spike likely due to foreign coverage. Next, we consider each of these platforms in detail.

4.1. Twitter

We set to establish what fraction of the discussion on the Twitter platform is organic versus posted by automated accounts (a.k.a., bots; Ferrara, Varol, Davis, Menczer, & Flammini, 2016). We used a state-the-art bot detection tool designed for Twitter, Botometer (Davis, Varol, Ferrara, Flammini, & Menczer, 2016), to quantify bots' prevalence. In line with recent literature (Ferrara, 2020; Yang et al., 2019), we used the top 10 and bottom 10 percentiles of users to set apart likely human users from automated accounts, with Botscores of 0.06 and below for humans and 0.67 and above for bots. This yielded 14,948 human accounts (responsible for 30,365 tweets and 3.4% of the total tweets) and 14,929 bots (with 34,020 tweets representing 3.9% of the total tweets). The total number of accounts scored was 141,929 (with 389,851 tweets).

Table 3 shows the differences in tweet types between humans and bots. Somewhat expectedly, humans post original tweets almost twice as much, and more quoted tweets. Bots on the other hand retweet without comment at almost 10% extra propensity. This is consistent with general characteristics of bot behavior (Davis et al., 2016).

A more pronounced difference can be observed with the language type. Table 4 shows the distribution of simplified and traditional Chinese. Simplified Chinese is used by China and traditional Chinese is used by Taiwan. We see only 7.4% of Chinese-humans users write in simplified, whereas 92.6% use in traditional Chinese. In contrast, for all Chinese-speaking bots, 31.5% use simplified Chinese, and 68.5% use traditional Chinese. This indicates a much stronger chance that a bot is adopting simplified Chinese. We corroborate this by considering the location of users. The bottom row shows that most of the tweets arise from non-local sources, which together affirms H1. We conclude much of the chatter on Twitter about Taiwan arises from outside of Taiwan.

The high level of English and Japanese in Table 2 over Chinese is of great interest. We find that around 50,000 out of 81,000 Japanese tweets are in response to President Tsai. While Tsai's dominant presence is largely expected, we note that Han in comparison has very few mentions, with no tweets from his account. Tsai also tweets frequently and intentionally in Japanese, such that common simplified Chinese users accused her of being "bought by Japan."

Since much of the discourse occurs amongst international users, the sites of democratic harm in this regard are primarily with *self-determination* and, to a lesser extent, *accountable representation*. This becomes clearer when we consider the network graph shown in Figure 2, which portrays the semantic space of our Twitter data through top hashtags. Here, nodes are hashtags, and edges are their co-occurrences. The network was produced by tabulating all co-occurrences, processed in NetworkX, and then plotted with Gephi (Bastian, Heymann, & Jacomy, 2009; Hagberg, Swart, & Chult, 2008).

The nodes in purple (center left) show the general discourse in traditional Chinese. Tsai takes up a large central role in setting the agenda. We also note the large cluster of Japanese responses (in orange) to Tsai. In contrast, Han and the Kuomingtang are mentioned much less (bottom left, in brown). To the left (in dark green), there is a cluster of hashtags that are supportive of the DPP and Tsai's camp, but does not take a central role in the semantic network. One possibility is these are pro-DPP campaign users that did not achieve traction. The tangible imbalance between the DPP and KMT in self-determination discourse answers RQ2.

The network also provides insight regarding transnational solidarity. We observe a distinct division of language in the network structure. The election's discourse

Table 3. Difference in tweet type distribution between humans and bots (top and bottom 10% by Botscore).

Type of tweet	Humans	Bots
Retweet without comment	63% (19,143)	69.6% (23,671)
Quoted tweet	15.1% (4,574)	12.5% (4,250)
Reply	10.9% (3,317)	11.7% (3,996)
Original tweet	11% (3,331)	6.2% (2,103)

Table 4. Proportion of simplified vs. traditional Chinese tweets and the location of tweets.

		Humans	Bots
Language	Traditional Chinese	92.6% (4,815)	68.5% (4,378)
	Simplified Chinese	7.4% (384)	31.5% (2,013)
Location	Local (Taiwan)	21.4% (6,513)	9.5% (3,222)
	Non-local	78.6% (23,852)	90.5% (30,798)

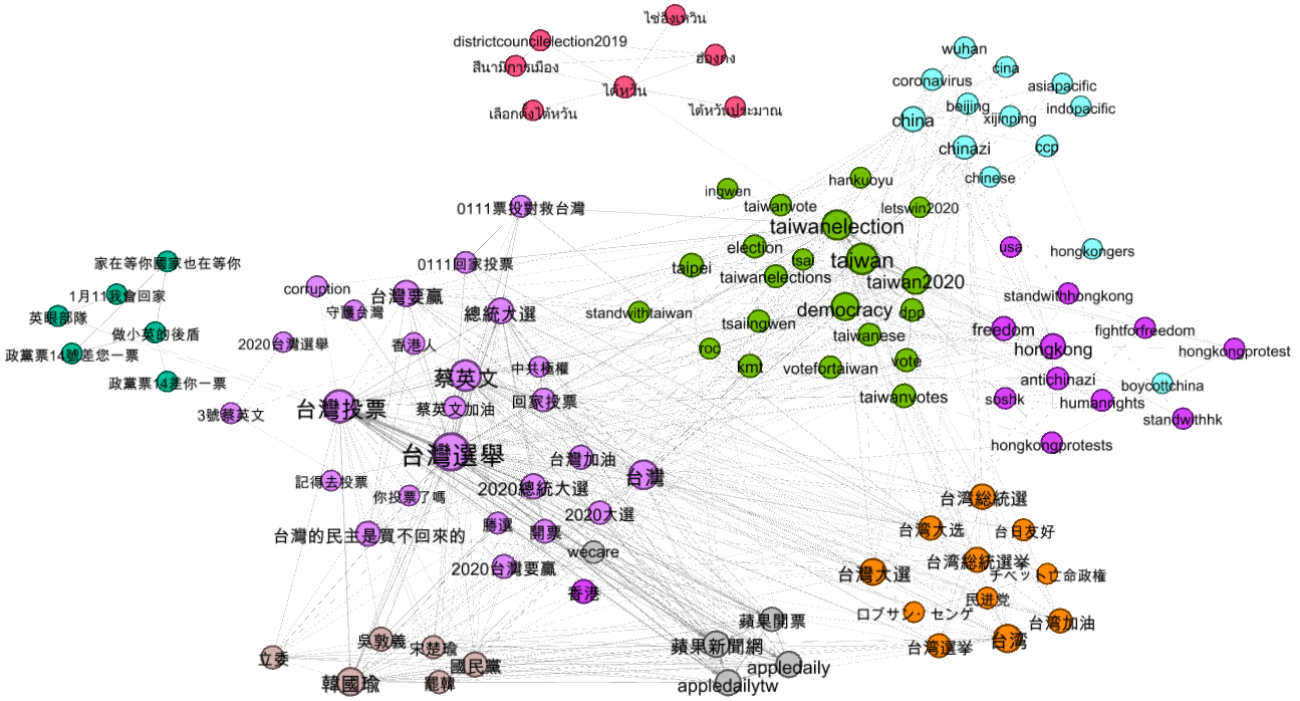


Figure 2. Semantic network of Taiwanese 2020 election Twitter discourse based on hashtags.

in English (green) is much better connected to other international themes, such as Hong Kong (purple, right), human rights issues in China (cyan), and by then, mentions of the novel coronavirus. The lack of trending hashtags in simplified Chinese and keywords indicates that while Chinese trolls may directly attack Tsai and her online campaign, their collective behavior on Twitter is decentralized. This is a shift away from Bradshaw and colleagues' characterization of centralized campaigns, and consistent with Yang et al.'s (2017) results.

Of note, the red cluster at the top denotes coverage from Thailand. The relationship between Taiwan, Hong Kong, and Thailand has been under the spotlight during the Covid-19 pandemic. In April 2020, after a celebrity drew outrage from Beijing viewers (McDevitt, 2020) and received large volumes of malicious trolling, users from Taiwan and Hong Kong began defending her online, along with the Thai users. The hashtag #nnevvy began trending, and an online community eventually known as the Milk Tea Alliance with users from Taiwan, Thailand, and Hong Kong was born. Looking at Thai coverage in the semantic network for Taiwan's election, the emergence of the Milk Tea Alliance is not sudden but a large trend of growing solidarity between the three user bases.

While the literature on transnational solidarity between Taiwan and Thailand has been sparse, activism between Hong Kong and Thailand can be traced back to 2016. When students in Thailand invited Hong Kong student activist Joshua Wong to share his experiences during the Umbrella Movement of 2014, as a speaker for the 1976 massacre of Thai student uprisings, he was detained at the Bangkok airport (Phoborisut, 2019). Protests calling for his release emerged across Hong Kong and Bangkok, which produced foundations for solidarity today.

We also note the important role of *Apple Daily*, a popular digital native newspaper in Hong Kong and Taiwan. Their co-occurrence with so many trending hashtags suggests, compared to other newspapers, that they disseminate their articles by carefully tracking the top trending keywords. In sum, results for Twitter suggest higher levels of automation, or bot-like behavior, in simplified Chinese accounts. However, the lack of trending terms suggests the lack of coordinated attacks, compared to discourse from Taiwan, largely set by Tsai and the DPP.

4.2. PTT

While Twitter provides insight into discourses of self-determination on the international front, it lacks details of public deliberation domestically. To recap, PTT is widely regarded as the 'Reddit' of Taiwan. New events are often posted here, to the extent that journalists have used it as a first-line source of information. Although competing platforms such as DCard have also risen in popularity in recent years, PTT has been a good representative of the local discourse, and its straightforward interface enables analysis of public discourse.

To get a sense of the discourse on PTT, the top terms are presented in Table 5, upon removing candidate names. We observe words that speak to a democratic process—*freedom*, *vote*, *democracy*, and *government*. The *Chinese Communist Party (CCP)* and *Hong Kong* are explicitly mentioned. Since this data set is conditional on being election-related, these keywords

Table 5. Top words found within the PTT bulletin board.

Original Text	Translation	Counts
自由	Freedom	1,258
投票	Vote	1,116
台北	Taipei	922
柯文哲	Ko Wen-Zhe	808
立委	Legislator	700
中央社	Central News Agency	479
香港	Hong Kong	447
水桶	Bucket	441
民主	Democracy	421
政府	Government	390
中共	Chinese Communist Party	344
主席	Chairman	300
八卦	Gossip	246
高雄市	Kaohsiung City	246

indicate the protests in Hong Kong, and shifting attitudes toward China played a large role in shaping discourse of self-determination. Common keywords in the comments section included the *elderly* and *sugarcane farmers*. Here, the tag 'sugarcane farmers' refers to the rural common folks. We also see PTT specific terms, such as *bucket* (水桶). The term 'cool down in a bucket of cold water' emerged as a euphemism for being suspended. 'Cockroach' and 'trash' are derogatory terms endemic to PTT's common vocabulary. Ko Wen-Zhe, the mayor of Taipei, is the fourth most mentioned term. Two major Taiwanese cities are mentioned—Taipei and Kaohsiung. As expected, Taipei is mentioned in conjunction with Ko, and Kaohsiung with Han.

Keywords only reveal a shallow interplay within online communication. Next, we consider the comments section, specifically we quantify the level of disagreement within each post. With P the number of commendations and N the number of dislikes, we define the disagreement score D as follows:

$$D = \frac{N}{N+P} \tag{1}$$

The choice of variable reflects negative (N) and positive (P) reactions. This measure scales with the number of disagreements (with respect to the initial post), while also capturing the diversity of commenting participants. For instance, a score of 0.5 indicates an equal number of users agreeing and disagreeing. Upon tagging the disagreement scores per article, we consider whether this is related to the specific discourse topics. We first subset all posts with disagreement scores greater than 0.5. Figure 3 shows the proportion of articles, subset on a specific topic. We observe that there is a disproportionate level of disagreement within the topics of Tsai and the DPP.

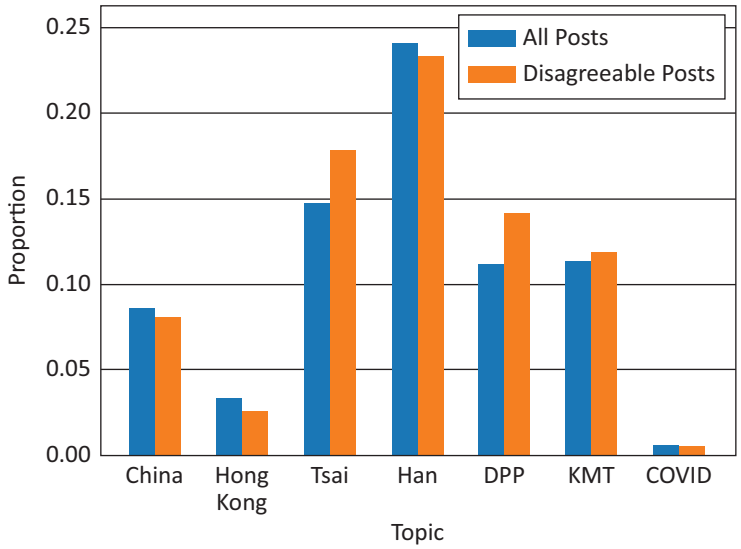


Figure 3. Proportion of article topics by the level of disagreement. Note: The figure shows that posts about Tsai and the DPP yield more disagreement in the discussion sections.

To understand where these disparities arise, we cross-section on the geo-local dimension of online engagement with certain topics. We leverage the given IP addresses within the comment sections to analyze across the urban-rural divide and between Taiwanese, Chinese, and other foreign commenting participants. Figure 4(a) shows urban and rural user participation on PTT. We observe that there is little variation across the two cohorts, with rural users engaging slightly more with Han and China-related discourse.

The discourse across international borders tells a much more compelling story. Figure 4(b) shows the topical distribution by Taiwanese, Chinese, and international IP addresses. Users with Chinese IPs disproportionately target posts about *China*, *Hong Kong*, *Tsai*, and the *KMT*, whereas they engage with *Han* at a much lower level. In contrast, there is little to no posting about the Covid-19 pandemic, relative to the domestic and international cohort.

Table 6 further shows that posts that involve Chinese users lead to higher levels of disagreement. This is pronounced in stories about Han, which produces high polarization across Chinese and Taiwanese users. Together, these answer RQ3 and confirm our H3 in regards to discourse about Tsai.

To summarize the results from PTT, we observe that accountable representation is a likely locus of disinformation. This also shows that individuals, rather than political parties, seem to be the target of choice. The selective engagement from Chinese citizens in these topics regarding self-determination—such as *Hong Kong*, *China*, and *Tsai*, while avoiding topics such as Han and Covid-19 shows the phenomenon of emergent coordination.

However, analyses of discourse gives us limited insight into disinformation directly. Next, we consider Line and the misinformation aggregated under Cofact's database.

4.3. Line and Cofacts

The Cofact's database includes user-forwarded Line posts and/or links that may contain misinformation. We similarly tagged the database with discourse topics, with an additional category for medicine.

The amount of misinformation is high compared to the number of factual claims. Incumbent President Tsai seems to attract the highest level of misinformation, both in proportion and in raw volume. This seems consistent with our observation of PTT and Twitter, where Tsai seems to attract the highest levels of controversy, and

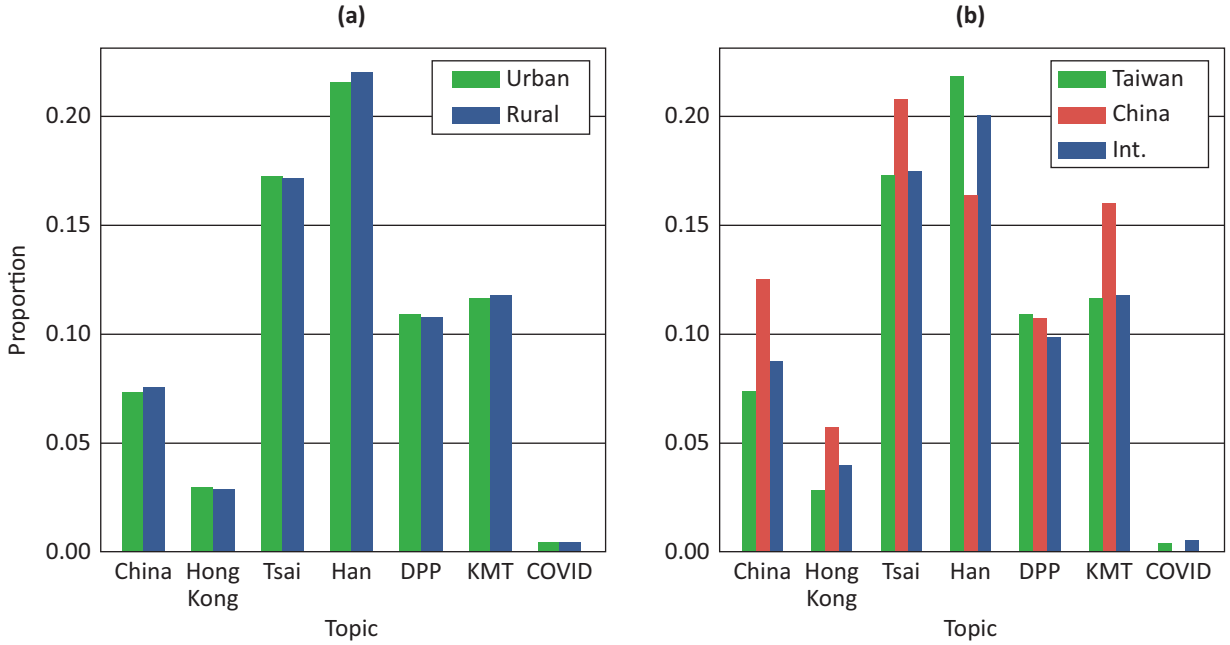


Figure 4. Level of disagreement across different user groups based on IP address: **(a)** across the urban-rural divide; **(b)** between users from Taiwan, China, and other international locations.

Table 6. Mean disagreement ratio between Chinese and Taiwanese IP addresses.

	Chinese IPs	Taiwanese IPs
All Stories	0.313±0.005	0.272±0.016
Tsai Stories	0.311±0.012	0.303±0.012
Han Stories	0.331±0.034	0.257±0.01

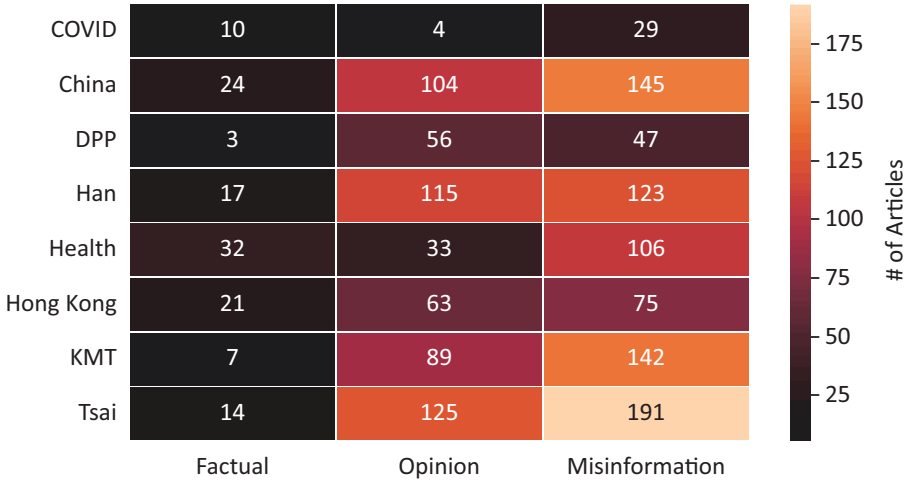


Figure 5. Heat map of misinformation classification by topic.

thus completes our addressing of RQ3 and confirmation of H3.

Interestingly, we observe a low correlation between the volume of reported cases for the DPP and Tsai. We offer two potential explanations. In the wake of Tsai's perceived failures during the midterm elections, fissures appeared between the DPP and Tsai. Thus, Tsai became the primary target of hoaxers, rather than the DPP itself. The decoupling became evident closer to the election. The second explanation follows a hypothesis presented earlier, where individuals are the more likely target of misrepresentation, at least in the case of foreign interference.

Finally, we consider the sources of misinformation, and attempt to answer RQ4 regarding the traditional media's role in spreading misinformation. Table 7 shows the top linked sources within the database and the percentage of misinformation.

We have two main takeaways. First, the primary inter-platform links are with social media and digital platforms such as Facebook and YouTube. These two together take up just over one-third of the reported links. Second, there is a high proportion of misinformation on dominant digital news platforms. For instance, hyperlinks for Google News alone contain almost 50% of all misinformation. Although it is technically difficult to ascertain the hosting domains in these cases, other digital news sources score poorly: *United Daily* (0.28), *KK News* (0.33), *Apple Daily* (0.28), and *ET Today* (0.25). Only *Liberty Times* scores low on misinformation (0.05). While it's true there may be selection bias—these are articles suspected of containing misinformation after all—the fact that verified news sources even contain misinformation is particularly concerning. Our findings confirm H4 and observations from the past (Monaco, 2017), that the traditional media is often responsible for amplifying misinformation.

Table 7. Top reported misinformation domains and their proportions of misinformation and opinion.

Web domain	Links	% Opinion	% Misinformation
Facebook	58	0.12	0.17
YouTube	55	0.04	0.38
Google	26	0.15	0.46
UDN (United Daily)	25	0.16	0.28
LTN (Liberty Times)	20	0.20	0.05
Kknews	18	0.06	0.33
Appledaily	18	0.06	0.28
Mygopen	18	0.06	0.39
Wikipedia	17	0.00	0.00
Bit	17	0.00	0.53
Line	13	0.15	0.23
Ettoday	12	0.17	0.25
g0v	9	0.00	0.22
Chinatimes	9	0.11	0.33
Twitter	7	0.57	0.14
Social media	133	0.11	0.26
Digital news	128	0.13	0.29

5. Conclusion

By 8 PM on January 11th, 2020 the results of the Taiwanese election were clear: Tsai had defended her presidency and won by the greatest margin in Taiwanese history. Despite large amounts of disinformation surrounding her candidacy, the outcome of the election seemed to indicate that was ineffective. An explanation may be dissonance. As Templeman recently postulated, due to the high levels of distrust in Chinese media, large levels of the population are inoculated against pro-Chinese sentiment (Templeman, 2020). However, despite the growing emphasis on domestic issues like wage growth and LGBT rights, elections never stray far from the China problem.

We began this study by discussing two gray areas regarding the frame of light and dark participation. The first and longer-standing issue is the clash of political ideologies between China and Taiwan. Second, and more importantly, the use of digital tools like bots and group removal to fight misinformation may limit the domestic diversity of political voices.

The first goal of this study was to understand the different facets of the elections communicated, using a thorough analysis of these three platforms. The second and more important goal, was to understand what the discourse and citizen participation say about the tension of employing digital tools to fight disinformation.

On Twitter, we found Tsai and the DPP's dominance in the digital campaign. Her engagement focuses on the international front, with users from anglophone countries and Japan. We observe more bot-like behavior coming from Chinese users and transnational solidarity between Hong Kong, Taiwan, and Thailand. The high volume of Tsai's content suggests counter-discourse against Chinese trolls is partisan.

On PTT, although Han is the most popular topic of discussion, it is Tsai and the DPP that elicited the most disagreement. A closer look at the geo-local origins reveals Chinese participation on issues such as Hong Kong, Tsai, and the KMT, while avoidance of Han and Covid-19. This also indicates that discussion surrounding Han arises predominantly domestically. These results suggest citizen participation from China focuses on discrediting Tsai and hence challenges accountable representation. We affirm this by considering Line. Stories about Tsai are the most reported stories. Lastly, a concerning level of misinformation arises from the traditional news media.

The high volume of Tsai-related misinformation, particularly from Chinese sources, may have justified the strong terms of the Anti-Infiltration Act. On December 13, 2019 alone, Facebook removed 118 fan pages, 99 groups, and 51 accounts that supported Han. One of these pages included 155,443 members. While some of these may have violated community standards or had traces of foreign interference, the hard-line approach certainly silenced legitimate support for Han. Perhaps Han would have done better had China not explicitly backed his campaign. Political bots can be used to promote democratic discourse, to washout foreign propaganda, but if the modal actors are political parties, the same technologies can stamp out the diversity of political opinion. This is especially dangerous in bipartisan situations, as in the case of Taiwan.

Digital tools alone do not determine the dark or light shade of a campaign; rather, it is whether their use violates the ideals of deliberative democracies. The case of Cofacts may provide a solution, in the domain of media regulation. To avoid partisan censorship of political information, crowdsourced solutions promise more equity and a diversity of voices. However, it is important to ensure that a representative committee of volunteers is present in fact-checking.

The case of Taiwan presents comparisons across the most salient axes in democratic theory: government vs. citizen-driven solutions, authoritarian control vs. self-determination. Peter Dalgren, in his canonical work, lays out four pillars for which civic culture rests upon: knowledge, loyalty to democratic values, practices and routines, identity as citizens (Dahlgren, 2000). Amidst rising populism, the ability for citizens to not only participate in news making, but to verify fact and build sociotechnical infrastructure, brings forth an optimism toward citizen-led democracy and public deliberation. For Taiwan, it is important to continuously refine its interpretation of free speech, not as a comparison with its neighbor across the strait, but a set of procedural and accountable standards.

Acknowledgments

We would like to thank Ashok Deb for providing data and Tom Hollihan for providing invaluable suggestions.

Conflict of Interests

The authors declare no conflict of interests.

References

Aspinwall, N. (2020a, January 10). Taiwan's War on Fake News Is Hitting the Wrong Targets. *Foreign Policy*. Retrieved from https://foreignpolicy.com/2020/01/10/taiwan-election-tsai-disinformation-china-war-fake-news-hitting-wrong-targets

Aspinwall, N. (2020b, January 3). Taiwan Passes Anti-Infiltration Act Ahead of Election Amid Opposition Protests. *The Diplomat*. Retrieved from https://thediplomat.com/2020/01/taiwan-passes-anti-infiltration-act-ahead-of-election-amid-opposition-protests

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In William W. Cohen & Nicolas Nicolov (Eds.), *Third international AAAI Conference on Weblogs and Social Media* (pp. 361–362). Califor-

nia, CA: AAAI Press.

Bradshaw, S., & Howard, P. N. (2018). *Challenging truth and trust: A global inventory of organized social media manipulation* (The Computational Propaganda Project No. 1). Oxford: University of Oxford. Retrieved from http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2018/07/ct2018.pdf

Chen, H.-T., Chan, M., & Lee, F. L. (2016). Social media use and democratic engagement: A comparative study of Hong Kong, Taiwan, and China. *Chinese Journal of Communication*, *9*(4), 348–366.

Cole, J. M. (2019, November 26). Taiwan's Han Kuo-Yu offers uncomfortable echoes of Donald Trump. *Nikkei Asian Review*. Retrieved from https://asia.nikkei.com/Opinion/Taiwan-s-Han-Kuo-yu-offers-uncomfortable-echoes-of-donald-trump

Dahlgren, P. (2000). The Internet and the democratization of civic culture. *Political Communication*, *17*(4), 335–340.

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016, April). Botornot: A system to evaluate social bots. In Jacqueline Bourdeau, Jim Hendler, Roger Nkambou Nkambou (Eds.), *Proceedings of the 25th International Conference Companion on World Wide Web* (pp. 273–274). Republic and Canton of Geneva: International World Wide Web Conferences Steering Committee.

Dias, N., Pennycook, G., & Rand, D. G. (2020). Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harvard Kennedy School Misinformation Review*, *1*(1).

DiResta, R., Shaffer, K., Ruppel, B., Sullivan, D., Matney, R., Fox, R., . . . & Johnson, B. (2019). *The tactics & tropes of the Internet Research Agency*. Lincoln, NE: University of Nebraska.

Epstein, Z., Pennycook, G., & Rand, D. (2020). Will the crowd game the algorithm? Using layperson judgments to combat misinformation on social media by downranking distrusted sources. In Regina Bernhaupt and Florian Mueller (Eds.), *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–11). New York, NY: CHI Steering Commitee.

European Commission. (2018). *Communication from the Commission: Securing free and fair European elections*. Brussels: European Commission.

Fan, F. T., Chen, S. L., Kao, C. L., Murphy, M., Price, M., & Barry, L. (2019). Citizens, politics, and civic technology: A conversation with g0v and EDGI. *East Asian Science, Technology and Society: An International Journal*, *13*(2), 279–297.

Ferrara, E. (2020). What types of COVID-19 conspiracies are populated by Twitter bots? *First Monday*, *25*(6). http://dx.doi.org/10.5210/fm.v25i6.10633

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96–104.

Freedom House. (2019). Freedom in the world. *Freedom House.* Washington, DC: Freedon House.

Habermas, J. (1996). *Between facts and norms: Contributions to a discourse theory of law and democracy*. Cambridge: MIT Press.

Hagberg, A., Swart, P., & Chult, D. (2008). *Exploring network structure, dynamics, and function using NetworkX* (No. LA-UR-08-05495; LA-UR-08-5495). Los Alamos, NM: Los Alamos National Lab (LANL).

Han, K. (2018, August 16). At Cofacts in Taiwan, volunteer editors and a time-saving chatbot race to combat falsehoods on Line. *Splice Media*. Retrieved from https://www.splicemedia.com/cofacts-fake-news-taiwan

Hioe, B. (2020a, January 6). Fighting fake news and disinformation in Taiwan: An interview with Puma Shen. *New Bloom*. Retrieved from https://newbloommag.net/2020/01/06/puma-shen-interview

Hioe, B. (2020b, July 22). Reorienting Taiwan and Hong Kong: New avenues for building power. *Lausan Magazine*. Retrieved from https://lausan.hk/2020/reorienting-taiwan-and-hong-kong

Hsiao, H. H. M. (2016). 2016 Taiwan elections: Significance and implications. *Orbis*, *60*(4), 504–514.

Hsiao, H. X. (2019, August 13). Fan song zhong manyan zhugong cai da xuan [Anti-infiltration act continues, Helps Tsai's Campaign]. *United Daily News*. Retrieved from https://web.archive.org/web/20200711180008/https://udn.com/news/story/11321/3985198

Hu, Y. H. (2017). Independent media, social movements, and the traditional news media in Taiwan. In Tong J., Lo SH. (Eds.), *Digital Technology and Journalism* (pp. 215–235). Cham: Palgrave Macmillan.

Huang, T.-T. (2019, December 21). Facebook to set up 'War Room' in Taiwan ahead of elections. *Taiwan News*. Retrieved from https://www.taiwannews.com.tw/en/news/3847551

Jansen, C.-T. (2019, July 12). Who is Han Kuo-yu: Could Kaohsiung's populist mayor be Taiwan's next president? *Post Magazine: South China Morning Post*. Retrieved from https://www.scmp.com/magazines/post-magazine/long-reads/article/3018052/who-han-kuo-yu-could-kaohsiungs-populist-mayor

Karlova, N. A., & Fisher, K. E. (2013). A social diffusion model of misinformation and disinformation for understanding human information behaviour. *Information Research*, *18*(1). Retrieved from http://InformationR.net/ir/18-1/paper573.html

Kokunai mau 2,800 man-ri toppa! Misshon sara-go no facebook, Nihon ni okeru kongo no tenbō [Over 28 million domestic monthly active users! Facebook after mission change, future prospects in Japan]. (2017, September 14). *Social Media Lab*. Retrieved from https://gaiax-socialmedialab.jp/post-53856

McDevitt, D. (2020, April 18). 'In milk tea we trust': How a Thai-Chinese meme war led to a new (online) Pan-Asia alliance. *The Diplomat*. Retrieved from https://thediplomat.com/2020/04/in-milk-tea-we-trust-how-a-thai-chinese-meme-war

Monaco, N. J. (2017). *Computational propaganda in Taiwan: Where digital democracy meets automated autocracy* (Working Paper 2017.2). Oxford: Oxford University Press.

Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, *116*(7), 2521–2526.

Phoborisut, P. (2019). East Asia in Action. Contesting Big Brother: Joshua Wong, Protests, and the Student Network of Resistance in Thailand. *International Journal of Communication*, *13*(23), 3270–3292.

Quandt, T. (2018). Dark participation. *Media and Communication*, *6*(4), 36–48. http://dx.doi.org/10.17645/mac.v6i4.1519

Reichenbach, D. (2020, March 18). The rise and rapid fall of Han Kuo-Yu. *The Diplomat.* Retrieved from https://thediplomat.com/2020/03/the-rise-and-rapid-fall-of-han-kuo-yu

Rowen, I. (2015). Inside Taiwan's Sunflower movement: Twenty-four days in a student-occupied parliament, and the future of the region. *The Journal of Asian Studies*, *74*(1), 5–21.

Starbird, K., Arif, A., & Wilson, T. (2019). Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human–Computer Interaction*, *3*(CSCW), 1–26. https://doi.org/10.1145/3359229

Templeman, K. (2020). How Taiwan stands up to China. *Journal of Democracy*, *31*(3), 85–99.

Tenove, C. (2020). Protecting democracy from disinformation: Normative threats and policy responses. *The International Journal of Press/Politics*. https://doi.org/10.1177/1940161220918740

Terceros, B. A. (2018, Jun 29). Ahead of Mexico's largest election, Verificado 2018 sets an example for collaborative journalism. *International Journalists' Network*. Retrieved from https://ijnet.org/en/story/ahead-mexico%E2%80%99s-largest-election-verificado-2018-sets-example-collaborative-journalism

Thomala, L.-L. (2020). Penetration of leading social networks in taiwan as of 3rd quarter 2019. *Statista*. Retrieved from https://www.statista.com/statistics/295611/taiwan-social-network-penetration

Yang, K. C., Varol, O., Davis, C. A., Ferrara, E., Flammini, A., & Menczer, F. (2019). Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, *1*(1), 48–61.

Yang, S., Chen, P.-Y., Shih, P. C., Bardzell, J., & Bardzell, S. (2017). Cross-strait frenemies: Chinese netizens VPN in to Facebook Taiwan. *Proceedings of the ACM on Human–Computer Interaction*, *1*(CSCW), 1–22.

Yoshida, H. (2018, December 26). Twitter No gekkan akutibuyuza-sū wa Nihon de 4500 man chō [Twitter has more than 45 million monthly active users in Japan]. *Tech Crunch*. Retrieved from https://jp.techcrunch.com/2018/12/26/twitter-2

About the Authors

Ho-Chun Herbert Chang is a Taiwanese-Canadian Doctoral Student at the Annenberg School for Communication and Journalism, University of Southern California. He studies how technological infrastructure drives collective change, large-scale digital behavior, and democracy in Taiwan. Herbert holds an MSc in artificial intelligence from the University of Edinburgh, and received his BA in math, quantitative social science, and a Senior Fellowship in creative writing from Dartmouth College.

Samar Haider is a PhD student in the Department of Computer and Information Science at the University of Pennsylvania. His research is at the intersection of computational social science, natural language processing, and machine learning, with a focus on analyzing social media and the news ecosystem. He received his MSc from the University of Southern California and his BSc from the University of Engineering and Technology, Lahore, both in computer science.

Emilio Ferrara (PhD) is Associate Professor of Communication and Computer Science at the University of Southern California. His research focuses on understanding the implications of technology and social networks on human behavior, and their effects on society at large. His work spans from studying the Web and social media, to collaboration systems and academic networks, from team science to online crowds. Ferrara has published over 150 articles on social networks, machine learning, and network science.

COGITATIO

Media and Communication (ISSN: 2183–2439)
2021, Volume 9, Issue 1, Pages 158–170
DOI: 10.17645/mac.v9i1.3421

Article

Investigating Visual Content Shared over Twitter during the 2019 EU Parliamentary Election Campaign

Nahema Marchal [1,*], Lisa-Maria Neudert [1], Bence Kollanyi [2] and Philip N. Howard [1]

[1] Oxford Internet Institute, University of Oxford, Oxford, OX1 3JS, UK; E-Mails: nahema.marchal@oii.ox.ac.uk (N.M.), lisa-maria.neudert@oii.ox.ac.uk (L.-M. N.), philip.howard@oii.ox.ac.uk (P.N.H.)
[2] Doctoral School of Sociology, Corvinus University of Budapest, 1093 Budapest, Hungary; E-Mail: bence.kollanyi@stud.uni-corvinus.hu

* Corresponding author

Submitted: 30 June 2020 | Accepted: 1 September 2020 | Published: 3 February 2021

Abstract
Political communication increasingly takes on visual forms. Yet, despite their ubiquity in everyday communication and digital campaigning, the use of these visuals remains critically understudied. In this article, we investigate the formats and modes of visual content deployed by Twitter users over a two-week period leading up to the 2019 EU Parliamentary elections and across two publics: those discussing the election at large and those discussing the more contentious issue of EU membership. Conducting a multilingual, cross-comparative content and thematic analysis of a sample of 1,097 images, we find that: (1) Visuals originating from traditional political actors prevailed among both Twitter discourses; (2) users shared substantial amounts of anti-EU, populist and, to a lesser extent, extremist images, though this content remained largely disjointed from the mainstream public debate; and (3) political humor emerged as a vector for anti-establishment and Eurosceptic themes, especially in discussions critical of the European project. We discuss the implications of our findings for the study of visual political communication and social media manipulation.

Keywords
elections; European politics; populism; social media; visual communication

Issue
This article is part of the issue “Dark Participation in Online Communication: The World of the Wicked Web” edited by Thorsten Quandt (University of Münster, Germany).

1. Introduction

We live in an age of visual communication: From the rise of selfies, memes and animated GIFs in digital culture to the surging popularity of visual-centric platforms like Instagram, Snapchat, and TikTok that reach billions of monthly users across the world. Visual content in the form of photos, videos, infographics and user-generated images are becoming central to our day-to-day interactions online, informing how we present ourselves (Senft & Baym, 2015; Thomson & Greenwood, 2020), communicate and understand the world around us (Highfield & Leaver, 2016; Pearce et al., 2020). Image sharing has recently seen a surge in popularity, not only on visual-centric platforms but also on Twitter, where over 50% of tweet impressions in 2019 were associated with images or other visual media (Meeker, 2019).

Visuals are also starting to take center stage in online political communication. While political parties and campaign managers have traditionally relied on leaflets, posters, and TV spots to rally support, the advent of digital technologies, and social media specifically, has seen political actors integrate new visual media strategies into their everyday communicative practices. Retouched Twitter profile pictures, vlogs from the campaign trail, Instagram livestreams, and staged photo ops capturing seemingly candid moments have now become staples of the arsenal of contemporary electioneering

(Lilleker, Tenscher, & Štětka, 2015). Technological affordances have also enabled completely novel forms of political self-expression among private citizens, from ballot selfies to Snapchat filters in support of specific causes (Gutterman, 2018).

In modern attention economies, visual forms of communication offer clear advantages over text; they are easier to process, elicit strong emotions and are effective at capturing viewers' attention and retention (Barry, 2005; Fahmy, Bock, & Wanta, 2014; Newhagen, 1998). In recent years, however, they have also emerged as popular catalysts of misinformation and disinformation across Europe and worldwide, both through inadvertent amplification and as part of larger social media manipulation campaigns (Bradshaw & Howard, 2019; Guy, 2017). The German Alternative für Deutschland (AfD) party, for example, has repeatedly used fake imagery involving immigrants relating to sexual abuse and violence to bolster anti-immigration sentiment during elections (Czuperski & Nimmo, 2017). This is particularly concerning given that key research on this topic suggests that most individuals struggle to distinguish between real and manipulated images (Nightingale, Wade, & Watson, 2017).

Despite their ubiquity in contemporary online political discourse, visuals have been somewhat neglected in the study of political communication, which overwhelmingly favors text-based approaches, especially in Europe (Weller, Bruns, Burgess, Mahrt, & Puschmann, 2014). To address this oversight in this study, we conduct a multilingual, cross-case comparative content and thematic analysis of Twitter images posted by users in six different European language spheres—English, French, German, Italian, Spanish and Swedish—during a two-week-long period leading up to the 2019 EU Parliamentary elections. Specifically, we investigate images on Twitter in the context of two conversations: One surrounding the EU elections in general, and one surrounding the more contentious issue of membership to the EU. Three main research questions drive our analysis:

> RQ1: What salient formats and modes of visual content were users in Europe sharing over Twitter during the 2019 EU Parliamentary election campaign?
>
> RQ2: How does this differ across two conversations with varying degrees of contention?
>
> RQ3: What were the most common themes embedded in different modes of visual communication?

In the following section, we begin by situating this research in the existing literature on visual media in political communication, their uses and effects. After detailing processes for data collection and sampling, we outline our coding scheme for analyzing the salient formats and modes of images shared over Twitter during the campaign, as well as their themes. We show that users shared substantial amounts of anti-European, populist and, to a lesser extent, extremist images, though these were largely disjointed from the mainstream public conversation. Our data also reveals political humor as a vessel for anti-establishment and Eurosceptic themes, especially in discussions critical of the European project. Furthermore, we find that while traditional visual formats dominate across both Twitter conversations, hybrid content in the form of memes, annotated screenshots and remixed media have also emerged as popular modes of visual communication. In the last section, we discuss these findings and their implications in the context of the 2019 European Parliamentary Elections and contemporary political communication more broadly.

2. Literature

2.1. Visual Political Communication in Europe

Digital modes of communication have been transformative for political communication. Through the continual expansion of networked infrastructures, social media platforms have emerged as arenas for the dissemination of political information to large audiences at marginal cost. Today, political actors rely on a wide array of such channels to engage with voters and other stakeholders year-round (Chadwick, 2013). Scholars underscore social media's vital role for advocacy (Karpf, 2012), e-governance (Margetts, John, Hale, & Yasseri, 2015) and democratization (Howard & Hussain, 2013). Non-party actors have also been shown to leverage social media to shape the political agenda, frame and amplify issues, and generate consensus among subsets of the electorate (Rohlinger, 2019).

Visual media has, in more ways than one, been at the heart of this transformation. While visual symbols themselves have long performed essential functions in political communication (Schill, 2012), the surging popularity of technologies whose affordances are specifically oriented towards the creation, dissemination and customization of images has awarded them newfound importance—especially times of heightened political tension such as electoral cycles. In Europe, both grassroots and mainstream political actors leverage visual formats as potent ways to communicate with their constituents and mobilize support. Cámara Castillo (2019, p. 49) demonstrates, for example, how European institutions carefully "advertise European identity" and foster civic interaction through immaculately curated Instagram feeds. In a departure from traditional forms of campaigning, 'satellite' and non-party intermediaries like Momentum, a British political organization, have also been credited with boosting support for Jeremy Corbyn's 2017 campaign through its clever use of memes, and short-form videos—tactics which were widely adopted by the political mainstream during the 2019 UK General Elections (Lyons, 2019).

Yet worryingly, digital platforms have also proven vulnerable to manipulation (Woolley & Howard, 2018)

and algorithmic gaming aimed at sowing discord in Europe and around the world (Jungherr, Posegga, & An, 2019; Marchal, Kollanyi, Howard, & Neudert, 2019). During recent elections in France, Germany, and Italy, for instance, hyper-partisan and conspiratorial junk news sites repeatedly outperformed professional news outlets on social media (Neudert, Howard, & Kollanyi, 2019). Research also shows that xenophobic, national-populist, and other extreme-right 'movement-parties' (Kitschelt, 2006) make strategic use of digital media to spread Eurosceptic messages across publics (Caiani & Pavan, 2017). A growing number of scholars note that hybrid visuals such as memes have become the format of choice to push political propaganda or troll social media users, often with the stated goal of driving division among key segments of the audience (Klein, 2019). Today, new digital visual media complement—if not altogether supplant—other, more traditional forms of political communication for party and non-party actors alike. To gain a better understanding of their formats and modes in the political context, in the next section we review existing literature on their uses on social media.

2.2. The Political Uses of Social Media Visuals

Visual media have diverse uses in digital political communication. With the widespread adoption of smartphones, social media users have been empowered to document and share real-time footage of political events in new forms of 'mobile witnessing' (Reading, 2011) and citizen journalism. Twitter, a micro-blogging platform favored by journalists, opinion leaders and politicians and geared towards opinion broadcasting (Marwick & boyd, 2011) has emerged as a prime arena for political image sharing. Images shared over Twitter, for instance, formed an important part of both the 2011 Egyptian revolution (Kharroub & Bas, 2015) and the 2012 Israel–Hamas conflict (Seo, 2014) as well as more recently during the 2020 Black Lives Matter protests.

Further, images are often shared as a means of interpersonal communication to express an opinion, persuade, or even manipulate. The adage that an image is worth a thousand words underscores the strong rhetorical impact that visuals have on those who view them (Birdsell & Groarke, 2007). Research suggests that audiences process images faster and more efficiently than text alone (Graber, 2012). Typically, images tap into a larger socio-political context (Schill, 2012) and are often used in conjunction with or in response to other images and text to highlight specific aspects of an issue (Blair, 2004; de Vreese, 2005).

Decontextualized, altered or altogether fake images have also become prolific on social media, where they bypass traditional gatekeepers and often elude content moderators (Gillepsie, 2018). In a hybrid media ecosystem, a large amount of viral visual content takes the form of "derivatives, responses, or copies of content generated by the mass-media producers" (Hemsley & Mason, 2013, p. 146) that can be re-worked to deceive audiences intentionally or make a political or satirical point (Hemsley & Snyder, 2018). As such, humorous memes, composite images and mixed media involving the use of irony have come to play an increasingly critical role in digital politics (Tay, 2015) as a way for users to express opinions, build community and mobilize action, as well as a tool for politicians to share policy ideas or to demean their opponents—for instance, during the 2016 Brexit referendum campaign (Dean, 2019; Segesten & Bossetta, 2017).

Beyond that, connected technological infrastructures have enabled citizens to engage more directly in democratic processes though 'tiny acts' of participation (Margetts et al., 2015). Today, citizens contribute to the public conversation about politics in more ephemeral and intangible ways than before: Broadcasting their support for a cause and seeking to influence others to do the same, through selfies with politicians, and pictures of themselves engaging in various political activities such as rallies and protests (Sorokowska et al., 2016).

Thus, while the literature on the use of visual content in digital communication is growing, it remains in its infancy and offers ample room for elaboration and empirical study. Notably few studies, if any, have explored how specific formats of visuals are mobilized by social media users during electoral campaigns and what political themes they express. Considering this, in this study, we take the 2019 EU Parliamentary elections as a case study to examine the types of images shared over Twitter among two issue publics: Twitter users discussing the elections themselves, and those discussing the more controversial issue of potential withdrawal from the EU.

3. Methods

3.1. Case Selection

The 2019 European Parliamentary Elections took place between 23–26 May 2019 and witnessed a turnout of 50.66% of more than 400 million eligible voters to the polls—making it the second largest democratic election in the world (European Parliament, 2019). Across member states, hundreds of candidates and dozens of parties and their supporters campaigned for months over social media, generating vast amounts of campaign material, media coverage and user-generated content. The elections took place against a backdrop of significant divisions in public attitudes towards the EU, with polls underscoring an erosion of trust in European institutions (Guerra & Serricchio, 2014). The establishment of extreme populist voices within the political mainstream over recent years had stoked experts' fears that Eurosceptic voices would make significant gains in 2019 European Parliamentary Elections. Ahead of the vote, Julian King, then European Commissioner for Security, noted that the dispersed nature and long duration of the European Parliament elections made them

a "tempting target for malicious actors" (Cerulus, 2019), while Vera Jourova, the then EU Justice Commissioner, warned against "external propagandist pressure" potentially playing out online (Stokel-Walker, 2019). These potential threats around social media interference, combined with the multi-campaigner, multi-issue nature of the mandate thus make the 2019 Parliamentary Elections a pertinent case study.

3.2. Data Collection and Sampling

Our data collection proceeded in four stages. We first identified a set of relevant hashtags in English, French, German, Italian, Polish, Spanish, and Swedish intended to capture Twitter traffic around these two separate conversations. While hashtag-based sampling has known limitations, it is a common technique to study 'ad-hoc publics' forming around discussion of specific topics, especially during key political moments (Burgess & Bruns, 2012; Larsson & Moe, 2012). The hashtags were compiled by a team of nine research assistants with expert knowledge of these countries' languages and political landscapes (full list of hashtags in the Supplementary File).

Using this set of 84 hashtags, our team then collected a total of 3,620,701 tweets in real time between 13 May and 26 May 2019 through Twitter's Streaming API. This method has knowns limitations: The API only collects 1% of the global public traffic related to a specific search query at any given time and the company's precise sampling method is unknown (Morstatter, Pfeffer, Liu, & Carley, 2013), but it remains the only legal way to collect Twitter data without violating terms of services (Freelon, 2018). From this initial dataset we extracted tweets that contained static visuals in their metadata fields. We included tweets if they satisfied one or more of the following criteria: (1) contained at least one of the relevant hashtags; (2) contained the hashtag in a URL shared, or the title of its webpage; (3) were a retweet of a message that contained a relevant hashtag or mention in the original message; or (4) were a quoted tweet referring to a tweet with a relevant hashtag or mention. We then rehydrated each tweet in our set to access the image files, identify tweets that had been removed or deleted since the initial time of posting, resulting in a final dataset of 307,951 tweets with visual content of which 256,204 related to the EU election and 3,164 related to EU exit. To make inferences about both populations, we determine the appropriate sample sizes based on a 95% confidence interval and a more or less 4% margin of error, resulting in a random sample of 599 tweets for what we henceforth refer to as the 'General' sample and 505 tweets for the 'Exit' sample. A very small sample of seven images was not accessible.

3.3. Comparative Content Analysis

Content analysis is a reliable method for the systemic classification and interpretation of visual representations (Bell, 2004, p. 20; Rose, 2012). We chose this technique over a more interpretive approach in the first instance, in order to allow for a crisp and objective classification and to avoid introducing cultural and personal biases in the analysis of more subtle nuances of messaging and meaning.

We take an inductive approach to codebook development, identifying units of meaning as they emerged from our data before grouping them into larger codes covering two separate aspects of the Twitter images: their format and mode. Here, format describes the type of media shared by users, based on their constitutive elements, while the notion of mode captures the way in which political information is being communicated, based on an image's manifest content and its apparent provenance. Two coders with extensive expertise in content analysis first identified emergent format and modal categories and coded a sub-sample of 100 images drawn from the 'General' and 'Exit' datasets. Any disagreements were discussed among the authors, and initial codes were later adjusted and integrated into broader ones in an iterative process, culminating in seven format and nine modal categories. Intercoder reliability was then determined using Krippendorf's alpha on two independent, non-overlapping sub-samples of 50 images randomly taken from each dataset, achieving high scores.

Visual format (α = 0.843) includes the following categories: 'Photograph' refers to pictures taken with a camera—including selfies, user-generated, official, and stock photos—that have not been visibly modified. 'Illustration' refers to drawings, sketches, cartoons, and computer-generated images. 'Screen capture' corresponds to images displaying the content of a phone, TV, or computer screen, including captures of webpages, newspaper articles, and screenshots of social media posts. 'Infographic' encompasses visual representations of information and data, including statistics, maps, and visual explainers. A 'Composite' is a visual that has been altered to combine different graphical elements (e.g., photo, text, and drawing), such as photo montages, memes, and GIFs. 'Quote' refers to images featuring a phrase attributed to an individual or plain text that has not been visibly altered. In the 'Poster' category, finally, we include promotional posters, campaign posters, leaflets, event announcements and party logos (Figure 1).

Visual mode (α = 0.865) categories comprise: 'Official campaign communication,' which applies to official campaign material, including political party programs, leaflet and event advertisements, and any communications from official candidate and party accounts. 'Campaign event' applies to images of campaign events, including pictures of rallies, candidate appearances on TV, and photo ops. 'Citizen political engagement' applies to images of private citizens engaging in political activities, such as photographs taken at demonstrations, and individual expressions of support for political causes. 'Political humor' applies to memes, humorous cartoons,

Figure 1. Examples of format categories. Note: From top left to bottom right, examples of an 'illustration,' 'composite,' 'photograph,' 'infographic,' 'quotation' and 'poster.'

satire, and other forms of humor directed at or derived from actors involved in the political process. 'News media reporting' represents images of news media reports, such as newspaper articles, but excludes composites of multiple media sources. 'Non-party and satellite campaigning' applies to campaigning material generated by non-party actors, such as satellite groups (Dommett & Temple, 2018), registered campaigners, and other "democratic intermediaries" (Edwards, 2006, pp. 8–9). This includes event announcements, unofficial campaign material, and get-out-the-vote initiatives. The category 'Voting day' describes visuals of the vote, such as pictures of ballot cards, and citizens or politicians engaging in the act of voting. 'Other political' applies to other images of political nature that do not specifically relate to the campaign. 'Miscellaneous,' finally, encompasses images unrelated to politics (Figure 2).

3.4. Thematic Analysis

Finally, to complement and enrich our systematic content classification, we perform a thematic analysis of visual materials in our samples. Although content and thematic analysis methods share similarities, content analysis lends itself to quantitative summarization of the coded variables, whereas thematic analysis is a more

Figure 2. Examples of modal categories. Note: From top left to bottom right, examples of 'official campaign communication,' 'other political,' 'citizen political activism,' 'political humor,' 'satellite campaigning' and 'voting day.'

interpretive approach that seeks to reveal patterns of meaning in data in context (Neuendorf, 2019). For this task we also followed an inductive process (Clarke, Braun, & Hayfield, 2015, p. 225), identifying recurrent themes and patterns of meaning as they emerged in the data through semantic and visual symbols without prior theoretical expectations. After familiarizing ourselves with the data, we devised a first round of descriptive codes, which we later grouped to form larger thematic categories based on salience and relevance. One such category pertained to references to policy issues addressed during the election campaign, such as 'economy,' 'security' or 'immigration.' Any references to the elections themselves were grouped under the 'General election' category. Our team treated images of individuals desecrating the European flag, admonishing the European project, or advocating for total disengagement from the EU as conveying a 'Eurosceptic' message, while treating positive references to European integration as 'pro-Europe.' Graphical violence and visual references to extreme ideology were grouped under 'extremism,' while those pushing an anti-elite/establishment rhetoric were assigned to 'populist and anti-elitist.' Up to two codes were assigned to a small fraction of images that touched on more than one theme. While thematic analysis is a qualitatively oriented approach (Braun & Clarke, 2006) we include frequency counts for each category to understand what topics different forms of political expressions touched on.

4. Findings

4.1. Format

Table 1 shows that photographs are by far the most prevalent format across both 'General' and 'Exit' samples, making up 38.7% and 23.8% of all images, respectively. Posters were shared slightly more frequently in the 'General' conversation, where they comprised 21.9% of all images compared to 20.4% for 'Exit.' Proportions of illustrations (6.2% in 'General,' 8% in 'Exit') and screen captures (12.9% in 'General,' 9% in 'Exit') were commensurate across both samples. Both datasets display stark differences when it comes to composites, however, with tweets focusing on EU withdrawal containing more than twice (22.4%) the number of composites than 'General' tweets (9.9%). Quotes and text only accounted for a fraction of all images in the 'General' dataset (2.7%) compared to the 'Exit' one (12.2%).

4.2. Mode

Table 1 reveals that official campaign material was the most shared mode in both 'General' and 'Exit'-related conversations, comprising 18.8% and 26.1% of images respectively. In the 'General' dataset, polls (10.6%), political humor (10.1%), images of voting day (10.1%) and unofficial campaign material (9.9%) made up almost equal proportions of content, while other political

Table 1. Frequency of visual formats and modes across both samples.

	'General' Sample		'Exit' Sample	
	N	%	N	%
Format Code				
Photograph	231	38.7	119	23.8
Illustration	37	6.2	40	8.0
Screen captures	77	12.9	45	9.0
Infographic	45	7.6	22	4.4
Composite	59	9.9	112	22.4
Quotes & Text	16	2.7	61	12.2
Poster	131	21.9	102	20.4
Total	596	100.0	501	100.0
Modal Code				
Official Campaign Material	112	18.8	131	26.1
Voting Day	60	10.1	24	4.8
Campaign Event	82	13.8	40	8.0
Citizen Political Activism	40	6.7	15	3.0
Polls	63	10.6	7	1.4
Political Humor	60	10.1	98	19.6
Satellite Campaign Material	59	9.9	60	12.0
Other Political	50	8.4	80	16.0
News Reporting	26	4.4	32	6.4
Miscellaneous/Spam	44	7.4	14	2.8
Total	596	100.0	501	100.0

images (8.4%), depictions of citizens engaging in political activity (6.7%) and news reports (4.4%) accounted for smaller proportions. Interestingly, politically humorous images were twice as present in 'Exit'-related tweets (19.6%), making them the second most popular mode of visual communication in this dataset, closely followed by other political images (16%). Here again, satellite and unofficial campaign material accounted for a substantial number of images with 12% of shares. Polls, finally, were mostly irrelevant to discussions around 'Exit' from the EU featuring in only 1.4% of tweets compared to 10.6% in the 'General' sample.

4.3. Themes

Table 2 displays the most salient themes in each sample. Comparing both datasets reveals some important differences. Most visual tweets in the 'General' dataset made reference to the 2019 EU Parliamentary Elections themselves (19%) or to multiple policy issues (10.6%). It is noteworthy, however, that the great majority of images shared were not attributed a thematic category. The salience of a policy issue in political discourse is a powerful indicator of its importance to the public. Yet, only a small proportion of visuals captured in our data made references to specific policies, such as security or immigration (less than 2% in both samples). In the 'Exit' sample, 41.9% of images propagated a Eurosceptic and anti-European message, making it the largest thematic category. In the 'General' sample, only 3.5% of images were classified as Eurosceptic, with double the share of images sharing pro-EU themes (6.7%). Finally, 9.4% of images shared in discussions of potential 'Exit' from the EU conveyed populist and anti-establishment sentiment, mainly consisting of derogatory or hateful messages vis-à-vis political elites.

4.4. Cross-Category Dependencies

Having identified the most salient content categories in each dataset, as a final step we investigate the relationships between format and modal categories on the one hand, and between modal and thematic categories on the other hand. Tables 3 and 4 reveal several interesting similarities and differences between samples. Across both 'General' and 'Exit' samples, photographs were mostly shared to depict campaign events, including party rallies (30% of photographs), with substantial proportions also alluding to voting day (24% in 'General,' 20% in 'Exit'). In 'Exit'-related tweets, photographs formed a core part of official campaign communication (20% of images in this mode, compared to 7% in the 'General' sample), where they often portrayed party volunteers canvassing. Illustrations, cartoons and drawings were the most common vehicle for political humor—the single largest category across samples for this format (62% in 'General,' 48% in 'Exit')—closely followed by composites and user-generated memes (42% in 'General,' 60% in 'Exit'). Across both samples, posters overwhelmingly corresponded to official campaign material, with 63% of posters in the 'General' and 85% of those in the 'Exit' sample taking the form of paper or digital campaign posters.

The relationships between modal and thematic categories are shown in Tables 5 and 6. In the conversation pertaining to the election at large, visual messages predictably revolved around the conduct of the election itself, often in the form of opinion and election polls (95% of images in this mode), as well as news reports (54%). Many also pushed a distinctly 'pro-EU' line, mostly through satellite campaign material from pro-EU groups and citizen-led initiatives to 'get-out-the-vote' for Europe (24%). Tweets shared with EU exit hashtags, on the other hand, overwhelmingly pushed Eurosceptic and populist/anti-establishment messages. Interestingly, these were mostly conveyed through satellite and unofficial campaigning material (87% of which carried a distinctly anti-EU message) and through political humor, where 50% and 33% of all images in this visual mode assumed either Eurosceptic or anti-elitist tones. In contrast to the 'General' dataset, polls shared in 'Exit'-related tweets were mostly shared to convey public

Table 2. Frequency of thematic categories across both samples.

	'General' Sample		'Exit' Sample	
Salient Theme	N	%	N	%
Security/Terrorism	3	0.5	7	1.4
Euroscepticism	21	3.5	210	41.9
Pro-EU	40	6.7	0	0.0
General Election	113	19.0	18	3.6
Extremism	7	1.2	21	4.2
Populist Anti-Elitism	11	1.8	47	9.4
Economy	8	1.3	5	1.0
Immigration	8	1.3	11	2.2
Multi-Issue	63	10.6	6	1.2
No Salient Theme	322	54.1	176	35.1
Total	596	100.0	501	100.0

Table 3. Cross tabulations of visual and modal categories in 'General' sample (N, %).

	Official Campaign	Voting Day	Campaign Event	Citizen Activism	Polls	Political Humor	Satellite Campaign	Other Political	News Reporting	Misc. & Spam	Total
Photograph	16 (7%)	56 (24%)	69 (30%)	26 (11%)	4 (2%)	6 (3%)	0 (0%)	17 (7%)	3 (1%)	34 (15%)	231
Illustrations	1 (3%)	1 (3%)	0 (0%)	0 (0%)	0 (0%)	23 (62%)	2 (5%)	8 (22%)	0 (0%)	2 (5%)	37
Screenshots	7 (9%)	0 (0%)	8 (10%)	8 (10%)	20 (26%)	1 (1%)	4 (5%)	4 (5%)	20 (26%)	5 (6%)	77
Infographic	0 (0%)	0 (0%)	0 (0%)	0 (0%)	39 (87%)	0 (0%)	1 (2%)	2 (4%)	1 (2%)	2 (4%)	45
Composite	3 (5%)	3 (5%)	4 (7%)	3 (5%)	0 (0%)	25 (42%)	15 (25%)	5 (8%)	1 (2%)	0 (0%)	59
Quotes	2 (13%)	0 (0%)	0 (0%)	1 (6%)	0 (0%)	1 (6%)	6 (38%)	6 (38%)	0 (0%)	0 (0%)	16
Poster	83 (63%)	0 (0%)	1 (1%)	2 (2%)	0 (0%)	4 (3%)	31 (24%)	8 (6%)	1 (1%)	1 (1%)	131
Total	112	60	82	40	63	60	59	50	26	44	

attitudes around membership in the EU or referred to various 2005 EU Constitution referenda. While depictions of citizen political activism were either pro-Europe or spoke to multiple policy issues in the 'General' sample (20% of images in each mode), in the 'Exit' sample they were almost exclusively mobilized to convey Eurosceptic visual symbols (87%).

5. Discussion

This research set out to identify and quantify the formats and modes of visual political communication mobilized by Twitter users in the lead up to 2019 European Parliamentary elections, and to determine if and how these varied in relation to contentiousness of discourse. Furthermore, our analysis sought to uncover the underlying themes conveyed through visuals. To this end, we developed a rigorous, multi-step scheme for categorizing visual content based on a multilingual, cross-case analysis of real-time Twitter data covering six European language spheres—English, French, German, Italian, Spanish and Swedish. Our findings underscore that visual media played a central role in the Twitter political discourse ahead of the 2019 European Parliamentary Elections, both as a conduit for official campaigning and candidate communications and for novel forms of political expression and user-generated political content. Three trends stand out from our analysis.

Table 4. Cross tabulations of visual and modal categories in 'Exit' sample (N, %).

	Official Campaign	Voting Day	Campaign Event	Citizen Activism	Polls	Political Humor	Satellite Campaign	Other Political	News Reporting	Misc. & Spam	Total
Photograph	24 (20%)	24 (20%)	36 (30%)	9 (8%)	0 (0%)	0 (0%)	1 (1%)	16 (13%)	1 (1%)	8 (7%)	119
Illustrations	1 (3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	19 (48%)	9 (23%)	8 (20%)	1 (3%)	2 (5%)	40
Screenshots	3 (7%)	0 (0%)	3 (7%)	3 (7%)	1 (2%)	0 (0%)	1 (2%)	6 (13%)	27 (60%)	1 (2%)	45
Infographic	0 (0%)	0 (0%)	1 (5%)	0 (0%)	5 (23%)	11 (50%)	2 (9%)	2 (9%)	1 (5%)	0 (0%)	22
Composite	2 (2%)	0 (0%)	0 (0%)	0 (0%)	1 (1%)	67 (60%)	30 (27%)	11 (10%)	0 (0%)	1 (1%)	112
Quotes	14 (23%)	0 (0%)	0 (0%)	3 (5%)	0 (0%)	1 (2%)	5 (8%)	35 (57%)	2 (3%)	1 (2%)	61
Poster	87 (85%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	12 (12%)	2 (2%)	0 (0%)	1 (1%)	102
Total	131	24	40	15	7	98	60	80	32	14	

Table 5. Cross tabulations of visuals modes and themes in 'General' sample.

	Pro-Europe	Eurosceptic	Populist Anti-Elitism	Extremism	General Election	Economy	Multi-Issue	Security/ Terrorism	Immigration
Official Campaigning	4 (4%)	13 (12%)	0 (0%)	1 (1%)	0 (0%)	4 (4%)	24 (21%)	0 (0%)	4 (4%)
Voting Day	1 (2%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Campaign Event	3 (4%)	0 (0%)	0 (0%)	0 (0%)	3 (4%)	0 (0%)	1 (1%)	0 (0%)	0 (0%)
Citizen Activism	8 (20%)	1 (3%)	1 (3%)	2 (5%)	0 (0%)	1 (3%)	8 (20%)	1 (3%)	0 (0%)
Polls	0 (0%)	1 (2%)	0 (0%)	0 (0%)	60 (95%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Political Humor	8 (13%)	5 (8%)	3 (5%)	2 (3%)	10 (17%)	1 (2%)	10 (17%)	2 (3%)	1 (2%)
Satellite Campaigning	14 (24%)	1 (2%)	2 (3%)	1 (2%)	0 (0%)	1 (2%)	11 (19%)	0 (0%)	2 (3%)
Other Political	2 (4%)	0 (0%)	3 (6%)	0 (0%)	0 (0%)	0 (0%)	3 (6%)	0 (0%)	0 (0%)
Reporting	0 (0%)	0 (0%)	2 (8%)	1 (4%)	14 (54%)	1 (4%)	6 (23%)	0 (0%)	1 (4%)
Misc./Spam	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)

Note: Number of images in each sample, followed by percentages of all images in the relevant mode.

Table 6. Cross tabulations of visuals modes and themes in 'Exit' sample.

	Pro-Europe	Eurosceptic	Populist Anti-Elitism	Extremism	General Election	Economy	Multi-Issue	Security/ Terrorism	Immigration
Official Campaigning	0 (0%)	42 (32%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Voting Day	0 (0%)	11 (46%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Campaign Event	0 (0%)	12 (30%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Citizen Political Activism	0 (0%)	13 (87%)	2 (13%)	1 (7%)	0 (0%)	0 (0%)	1 (7%)	0 (0%)	0 (0%)
Polls	0 (0%)	6 (86%)	0 (0%)	0 (0%)	1 (14%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Political Humor	0 (0%)	50 (51%)	32 (33%)	14 (14%)	0 (0%)	1 (1%)	0 (0%)	0 (0%)	1 (1%)
Satellite Campaigning	0 (0%)	52 (87%)	5 (8%)	3 (5%)	0 (0%)	0 (0%)	1 (2%)	0 (0%)	1 (2%)
Other Political	0 (0%)	22 (28%)	7 (9%)	3 (4%)	1 (1%)	1 (1%)	2 (3%)	7 (9%)	0 (0%)
Reporting	0 (0%)	2 (6%)	1 (3%)	0 (0%)	15 (47%)	7 (22%)	2 (6%)	0 (0%)	3 (9%)
Misc. & Spam	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)

Note: Number of images in each sample, followed by percentages of all images in the relevant mode.

First, anti-European visuals, populist anti-elite messages, and to a lesser extent extremist content around religion, were shared in substantial amounts ahead of the elections. However, this content was largely disjointed from the mainstream conversation about the election on social media and confined to critical discussions of the European project. Experts had expressed concerns about the spread of extremist propaganda and other forms of social media manipulation ahead of the vote, notably around questions of immigration (Dennison & Zerka, 2019). Contrary to these expectations, our data shows that extremist messaging predominately revolved around anti-Semitic tropes and pointed criticism of the purported Islamization of Europe.

Second, political humor emerged through our analysis as a popular vessel for Eurosceptic and anti-elite messaging in social media discourse, especially in the contentious conversation surrounding EU membership. Here, humorous visuals in the form memes, cartoons, and drawings, were predominantly mobilized to make ad populum arguments, attack political and economic elites, and to a lesser extent relay extremist viewpoints. These findings echo scholarship that flagged online political humor as a 'pipeline' to radicalization and extremism (Munn, 2019; Phillips & Milner, 2017). Several scholars also point out that memes, in-jokes, and political trolling successfully mobilize user and algorithmic attention (Marwick & Lewis, 2017; Wu, 2017)

Third, at odds with growing concerns surrounding the credibility and quality of political content circulating on social media, our data reveals that official campaign communication from candidates and political parties drove the largest proportion of visual traffic, both in the mainstream conversation and in the conversation specifically related to leaving the EU. This evidences the strong impact that traditional political actors continue to have on public conversations around elections and on the visual content that users encounter online. Traditional forms of political visuals like candidate posters and brochures, or official photographs from the campaign trail were widely shared on Twitter ahead of the vote. Likewise, material from non-party campaigners and 'satellite' (Dommett & Temple, 2018) issue groups prevailed among Twitter users, embracing novel forms of online political expression, such as annotated screenshots and remixed media, to campaign in support or opposition to the EU as a single issue.

Our study presents several limitations that highlight the need for further research. The first and most evident one is the focus on a single platform, Twitter. While Twitter remains a prime arena for political communication—favored by a wide range of political actors—electioneering typically takes place across several social media platforms. Future research should therefore investigate how the framework developed here applies to more visual-centric platforms and their unique affordances. Furthermore, by opting for topic-based sampling, we are necessarily restricted in the kind of claims we can make with respect to the political actors behind these visuals. It will therefore be valuable for future work to investigate how visuals are mobilized by party actors as compared to private citizens through actor-based sampling for instance. Lastly, by grounding our analysis in real-time social media data our findings are specific to both the temporal and socio-technical contexts in which they were collected. Studying visual information on social media is, finally, an inherently versatile exercise that must consider the multifaceted and changing nature of visuals as they develop over time. An interesting area for future work, in this respect, will be to further explore the link between the visual and textual elements of social media images as they are assembled and reworked through memetic practice, for example, to move towards more multi-modal understandings of platform vernaculars (Pearce et al., 2020). While there are many ways of analyzing visual content and our approach does not purport to be exhaustive, our analysis nonetheless provides a robust and situated look at visual political content shared across multiple language spheres. Rigorous classifications and thematic analyses of visual social media are not only critical for assessing the qualities and integrity of online political discourse, but also for bringing forward evidence-based policy and platform recommendations to effectively protect democratic freedoms online.

Acknowledgments

The authors gratefully acknowledge the support of the European Research Council for the project 'Computational Propaganda: Investigating the Impact of Algorithms and Bots on Political Discourse in Europe,' Proposal 648311, 2015–2020, Philip N. Howard, Principal Investigator. Project activities were approved by the University of Oxford's Research Ethics Committee, CUREC OII C1A 15–044. We are grateful to Civitates and the Omidyar Network for supporting our research in Europe. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the University of Oxford or our funders. We are also grateful to Samuelle Chinellato, Didac Fabregas, Freja Hedman, Tomasz Hollanek, Juan Lopez Martin, Karolina Partyga and Francesco Pierri for their contributions to this work.

Conflict of Interests

The author declares no conflict of interests.

Supplementary Material

Supplementary material for this article is available online in the format provided by the author.

References

Barry, A. M. (2005). Perception theory. In S. Josephson, J. Kelly, & K. Smith (Eds.), *Handbook of visual communication: Theory, methods, and media* (pp. 45–63). Abingdon: Routledge.

Bell, P. (2004). Content analysis of visual images. In T. Van Leeuwen & C. Jewitt (Eds.), *The handbook of visual analysis* (pp. 10–34). London: SAGE.

Birdsell, D. S., & Groarke, L. (2007). Outlines of a theory of visual argument. *Argumentation and Advocacy*, *43*(3/4), 103–113. https://doi.org/10.1080/00028533.2007.11821666

Blair, A. (2004). The rhetoric of visual arguments. In M. H. Helmers & C. Hill (Eds.), *Defining visual rhetorics* (pp. 41–61). Mahwah, NJ: Lawrence Erlbaum.

Bradshaw, S., & Howard, P. N. (2019). *The global disinformation order 2019 global inventory of organised social media manipulation* (Working Paper 2019.2). Oxford: Project on Computational Propaganda.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Burgess, J., & Bruns, A. (2012). Twitter archives and the challenges of "big social data" for media and communication research. *M/C Journal*, *15*(5). https://doi.org/10.5204/mcj.561

Caiani, M., & Pavan, E. (2017). "Inconvenient solidarities": Extreme-right online networks and the construction of a critical frame against Europe. In A. Grimmel & S. M. Giang (Eds.), *Solidarity in the European Union: A fundamental value in crisis* (pp. 145–160). Cham: Springer International Publishing.

Cámara Castillo, L. (2019). *Advertising the European identity: Instagram analysis on the visual self-presentation of the European Parliament and the European Commission* (Master thesis). University of Helsinki, Helsinki, Finland. Retrieved from https://helda.helsinki.fi/handle/10138/302878

Cerulus, L. (2019, January 16). Europe's most hackable election. *POLITICO*. Retrieved from https://www.politico.eu/article/europe-most-hackable-election-voter-security-catalonia-european-parliament-disinformation

Chadwick, A. (2013). *The hybrid media system: Politics and power*. Oxford: Oxford University Press.

Clarke, V., Braun, V., & Hayfield, N. (2015). Thematic analysis. In J. A. Smith (Ed.), *Qualitative psychology: A practical guide to research methods* (pp. 222–248). London: SAGE.

Czuperski, M., & Nimmo, B. (2017). #ElectionWatch: Germany's AfD utilizes fake imagery ahead of election. *DFRLab*. Retrieved from https://medium.com/dfrlab/electionwatch-germanys-afd-utilizes-fake-imagery-ahead-of-election-1fa1818ea82a

de Vreese, C. H. (2005). News framing: Theory and typology. *Information Design Journal + Document Design*, *13*(1), 51–62.

Dean, J. (2019). Sorted for memes and Gifs: Visual media and everyday digital politics. *Political Studies Review*, *17*(3), 255–266. https://doi.org/10.1177/1478929918807483

Dennison, S., & Zerka, P. (2019). The 2019 European election: How anti-Europeans plan to wreck Europe and what can be done to stop it. *European Council on Foreign Relations*. Retrieved from https://www.ecfr.eu/specials/scorecard/the_2019_European_election

Dommett, K., & Temple, L. (2018). Digital campaigning: The rise of Facebook and satellite campaigns. *Parliamentary Affairs*, *71*(1), 189–202. https://doi.org/10.1093/pa/gsx056

Edwards, A. (2006). ICT strategies of democratic intermediaries: A view on the political system in the digital age. *Information Polity*, *11*(2). https://dl.acm.org/doi/10.5555/1412569.1412572

European Parliament. (2019). The voting system. *European Parliament*. Retrieved from http://www.europarl.europa.eu/unitedkingdom/en/european-elections/european_elections/the_voting_system.html

Fahmy, S., Bock, M. A., & Wanta, W. (2014). *Visual communication theory and research*. New York, NY: Palgrave Macmillan.

Freelon, D. (2018). Computational research in the post-API age. *Political Communication*, *35*(4), 665–668. https://doi.org/10.1080/10584609.2018.1477506

Gillepsie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. New Haven, CT: Yale University Press.

Graber, D. A. (2012). *Processing politics: Learning from television in the Internet age*. Chicago, IL: University of Chicago Press.

Guerra, S., & Serricchio, F. (2014). Identity and economic rationality: Explaining attitudes towards the EIU in a time of crisis. In B. M. Stefanova (Ed.), *The European Union beyond the crisis: Evolving governance, contested policies, and disenchanted publics* (pp. 269–294). London: Lexington Books.

Gutterman, R. S. (2018). Ballot selfies: New political speech in search of first amendment protection in social media. *Wake Forest Journal of Law & Policy*, *8*(2), 211–258.

Guy, H. (2017, October 17). Why we need to understand misinformation through visuals. *First Draft News*. Retrieved from https://firstdraftnews.org:443/latest/understanding-visual-misinfo

Hemsley, J., & Mason, R. M. (2013). Knowledge and knowledge management in the social media age. *Journal of Organizational Computing and Electronic Commerce*, *23*(1/2), 138–167. https://doi.org/10.1080/10919392.2013.748614

Hemsley, J., & Snyder, J. (2018). Dimensions of visual misinformation in the emerging media landscape. In B. G. Southwell, E. A. Thorson, & L. Sheble (Eds.), *Misinfor-*

mation and mass audiences (pp. 91–109). Austin, TX: University of Texas Press.

Highfield, T., & Leaver, T. (2016). Instagrammatics and digital methods: Studying visual social media, from selfies and Gifs to memes and emoji. *Communication Research and Practice*, *2*(1), 47–62. https://doi.org/10.1080/22041451.2016.1155332

Howard, P. N., & Hussain, M. M. (2013). *Democracy's fourth wave? Digital media and the Arab Spring*. Oxford: Oxford University Press.

Jungherr, A., Posegga, O., & An, J. (2019). Discursive power in contemporary media systems: A comparative framework. *The International Journal of Press/Politics*, *24*(4), 404–425. https://doi.org/10.1177/1940161219841543

Karpf, D. (2012). *The moveon effect: The unexpected transformation of American political advocacy*. Oxford: Oxford University Press.

Kharroub, T., & Bas, O. (2015). Social media and protests: An examination of Twitter images of the 2011 Egyptian revolution. *New Media & Society*. https://doi.org/10.1177/1461444815571914

Kitschelt, H. (2006). Movement parties. In R. Katz & W. Crotty (Eds.), *Handbook of party politics* (pp. 278–290). London: SAGE.

Klein, O. (2019). *LOLitics: The content and impact of Dutch populist Internet memes*. Unpublished manuscript. Retrieved from https://ssrn.com/abstract=3371224

Larsson, A. O., & Moe, H. (2012). Studying political microblogging: Twitter users in the 2010 Swedish election campaign. *New Media & Society*, *14*(5), 729–747. https://doi.org/10.1177/1461444811422894

Lilleker, D. G., Tenscher, J., & Štětka, V. (2015). Towards hypermedia campaigning? Perceptions of new media's importance for campaigning by party strategists in comparative perspective. *Information, Communication & Society*, *18*(7), 747–765. https://doi.org/10.1080/1369118X.2014.993679

Lyons, K. (2019, October 30). Bring it on memes and "guy ropes of self-doubt": The first UK General Election ads. *The Guardian*. Retrieved from https://www.theguardian.com/politics/2019/oct/30/bring-it-on-memes-and-guy-ropes-of-self-doubt-the-first-uk-general-election-ads

Marchal, N., Kollanyi, B., Howard, P. N., & Neudert, L.-M. (2019). *Junk news during the 2019 EU parliamentary elections: Lessons from a seven-language study of Twitter and Facebook* (Data Memo 2019. 3). Oxford: Project on Computational Propaganda.

Margetts, H., John, P., Hale, S., & Yasseri, T. (2015). *Political turbulence: How social media shape collective action*. Princeton, NJ: Princeton University Press.

Marwick, A., & boyd, d. (2011). "I tweet honestly, I tweet passionately": Twitter users, context collapse, and the imagined audience. *New Media & Society*, *13*(1), 114–133. https://doi.org/10.1177/1461444810365313

Marwick, A., & Lewis, R. (2017). Media manipulation and disinformation online. *Data & Society*. Retrieved from https://datasociety.net/output/media-manipulation-and-disinfo-online

Meeker, M. (2019). Internet trends 2019. *Bond*. Retrieved from https://www.bondcap.com/report/itr19

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. *ArXiv*. Retrieved from http://arxiv.org/abs/1306.5204

Munn, L. (2019). Alt-right pipeline: Individual journeys to extremism online. *First Monday*. https://doi.org/10.5210/fm.v24i6.10108

Neudert, L.-M., Howard, P., & Kollanyi, B. (2019). Sourcing and automation of political news and information during three European elections. *Social Media + Society*, *5*(3). https://doi.org/10.1177/2056305119863147

Neuendorf, K. A. (2019). Content analysis and thematic analysis. In P. Brough (Ed.), *Research methods for applied psychologists: Design, analysis and reporting* (pp. 211–223). New York, NY: Routledge.

Newhagen, J. E. (1998). TV news images that induce anger, fear, and disgust: Effects on approach–avoidance and memory. *Journal of Broadcasting & Electronic Media*, *42*(2), 265–276. https://doi.org/10.1080/08838159809364448

Nightingale, S. J., Wade, K. A., & Watson, D. G. (2017). Can people identify original and manipulated photos of real-world scenes? *Cognitive Research: Principles and Implications*, *2*(1). https://doi.org/10.1186/s41235-017-0067-2

Pearce, W., Özkula, S. M., Greene, A. K., Teeling, L., Bansard, J. S., Omena, J. J., & Rabello, E. T. (2020). Visual cross-platform analysis: Digital methods to research social media images. *Information, Communication & Society*, *23*(2), 161–180. https://doi.org/10.1080/1369118X.2018.1486871

Phillips, W., & Milner, R. M. (2017). *The ambivalent Internet: Mischief, oddity, and antagonism online*. Cambridge: Polity.

Reading, A. (2011). The London bombings: Mobile witnessing, mortal bodies and globital time. *Memory Studies*. https://doi.org/10.1177/1750698011402672

Rohlinger, D. A. (2019). *New media and society*. New York, NY: New York University Press.

Rose, G. (2012). *Visual methodologies*. London: SAGE.

Schill, D. (2012). The visual image and the political image: A review of visual communication research in the field of political communication. *Review of Communication*, *12*(2), 118–142. https://doi.org/10.1080/15358593.2011.653504

Segesten, A., & Bossetta, M. (2017). Sharing is caring: Labour supporters use of social media #ge2017. *Election Analysis*. http://www.electionanalysis.uk/

uk-election-analysis-2017/section-5-the-digital-campaign/sharing-is-caring-labour-supporters-use-of-social-media-ge2017

Senft, T. M., & Baym, N. (2015). What does the selfie say? Investigating a global phenomenon: Introduction. *International Journal of Communication*, *9*, 1588–1606.

Seo, H. (2014). Visual propaganda in the age of social media: An empirical analysis of Twitter images during the 2012 Israeli–Hamas conflict. *Visual Communication Quarterly*, *21*(3), 150–161.

Sorokowska, A., Oleszkiewicz, A., Frackowiak, T., Pisanski, K., Chmiel, A., & Sorokowski, P. (2016). Selfies and personality: Who posts self-portrait photographs? *Personality and Individual Differences*, *90*, 119–123. https://doi.org/10.1016/j.paid.2015.10.037

Stokel-Walker, C. (2019, January 10). The EU doesn't really have a plan to stop its elections being hacked. *Wired UK*. Retrieved from https://www.wired.co.uk/article/eu-parliament-elections-hacking

Tay, G. (2015). Binders full of LOLitics: Political humour, internet memes, and play in the 2012 US Presidential Election (and beyond). *European Journal of Humour Research*, *2*(4). http://dx.doi.org/10.7592/EJHR2014.2.4.tay

Thomson, T. J., & Greenwood, K. (2020). Profile pictures across platforms. In S. Josephson, J. D. Kelly, & K. Smith (Eds.), *Handbook of visual communication* (1st ed., pp. 349–363). London: Routledge.

Weller, K., Bruns, A., Burgess, J., Mahrt, M., & Puschmann, C. (2014). *Twitter and society*. New York, NY: Peter Lang Publishing.

Woolley, S. C., & Howard, P. N. (Eds.). (2018). *Computational propaganda: Political parties, politicians, and political manipulation on social media*. Oxford: Oxford University Press.

Wu, T. (2017). *The attention merchants: The epic struggle to get inside our heads*. New York: Knopf.

About the Authors

Nahema Marchal is a Doctoral Candidate at the Oxford Internet Institute, University of Oxford, where her research focuses on the relationship between social media and political polarization. She is also a Researcher at the Computational Propaganda Project. Her other research interests include Internet regulation and governance, and the implications of digital technologies for public life.

Lisa-Maria Neudert is a Doctoral Candidate at the Oxford Internet Institute and a Core Researcher at the Computational Propaganda Project, where her work is located at the nexus of political communication, technology studies and governance. Her current research is looking into the public and private governance of policy issues surrounding disinformation through governments and social media platforms.

Bence Kollanyi is a Doctoral Candidate in Sociology at the Corvinus University of Budapest where his research focuses on the automation of social media accounts, including the development and the deployment of open-source Twitter bots. At the Computational Propaganda Project at the Oxford Internet Institute, Bence is responsible for collecting and analysing data from various social media platforms.

Philip N. Howard is a Statutory Professor of Internet Studies at the Oxford Internet Institute and a Professorial Fellow at Balliol College at the University of Oxford. He has published eight books and over 120 academic articles, book chapters, conference papers, and commentary essays on information technology, international affairs and public life. He is the author, most recently, of *Lie Machines: How to Save Democracy from Troll Armies*, *Deceitful Robots*, *Junk News Operations*, and *Political Operatives*.

Media and Communication (ISSN: 2183–2439)
2021, Volume 9, Issue 1, Pages 171–180
DOI: 10.17645/mac.v9i1.3399

Article

From Insult to Hate Speech: Mapping Offensive Language in German User Comments on Immigration

Sünje Paasch-Colberg *, Christian Strippel, Joachim Trebbe, and Martin Emmer

Institute for Media and Communication Studies, Freie Universität Berlin, 14195 Berlin, Germany;
E-Mails: s.colberg@fu-berlin.de (S.P.-C.), christian.strippel@fu-berlin.de (C.S.), joachim.trebbe@fu-berlin.de (J.T.), martin.emmer@fu-berlin.de (M.E.)

* Corresponding author

Submitted: 26 June 2020 | Accepted: 9 August 2020 | Published: 3 February 2021

Abstract
In recent debates on offensive language in participatory online spaces, the term 'hate speech' has become especially prominent. Originating from a legal context, the term usually refers to violent threats or expressions of prejudice against particular groups on the basis of race, religion, or sexual orientation. However, due to its explicit reference to the emotion of hate, it is also used more colloquially as a general label for any kind of negative expression. This ambiguity leads to misunderstandings in discussions about hate speech and challenges its identification. To meet this challenge, this article provides a modularized framework to differentiate various forms of hate speech and offensive language. On the basis of this framework, we present a text annotation study of 5,031 user comments on the topic of immigration and refuge posted in March 2019 on three German news sites, four Facebook pages, 13 YouTube channels, and one right-wing blog. An in-depth analysis of these comments identifies various types of hate speech and offensive language targeting immigrants and refugees. By exploring typical combinations of labeled attributes, we empirically map the variety of offensive language in the subject area ranging from insults to calls for hate crimes, going beyond the common 'hate/no-hate' dichotomy found in similar studies. The results are discussed with a focus on the grey area between hate speech and offensive language.

Keywords
comment sections; content analysis; Facebook; hate speech; refugees; text annotation; user comments; YouTube

Issue
This article is part of the issue "Dark Participation in Online Communication: The World of the Wicked Web" edited by Thorsten Quandt (University of Münster, Germany).

1. Introduction

In recent years, the use of offensive language in participatory online spaces has increasingly become the subject of public debate and scientific research in many countries (Keipi, Näsi, Oksanen, & Räsänen, 2017). Communication and media scholars analyze this phenomenon using various terms such as 'incivility' (e.g., Coe, Kenski, & Rains, 2014), 'flaming' (e.g., Cho & Kwon, 2015), or 'hate speech' (e.g., Erjavec & Kovačič, 2012). In particular, the term 'hate speech' receives much attention as it has a long tradition in a legal context where it is associated with hate crimes, genocide, and crimes against humanity (Bleich, 2011). In this context, the term refers to violent threats or expressions of prejudice against particular groups on the basis of race, religion, or sexual orientation.

However, due to its explicit reference to the emotion of hate (e.g., Brown, 2017a), 'hate speech' is also understood as a term referring to the expression of hatred (e.g., Post, 2009, p. 123). Accordingly, the term is often used as a general label for various kinds of negative expression by users, including insults and even harsh criticism. This ambiguity leads to fundamental misunderstandings in the discussion about hate speech and challenges its identification, for example, in online user comments (e.g., Davidson, Warmsley, Macy, &

Weber, 2017). Against this background, we formulate the following two research questions: How can we theoretically distinguish hate speech from neighboring concepts (RQ1)? And how can we empirically distinguish various forms of hate speech and offensive language using this theoretical framework (RQ2)? Answering these questions will allow for a more precise measurement of hate speech and offensive language, not only in academic research but also in practical content moderation and community management.

To this end, we introduce a modularized theoretical framework on the basis of which we can operationalize the defining features of hate speech and other forms of offensive language. We will first discuss challenges regarding the definition of hate speech and review how hate speech has been measured in content analyses so far. We then present a new approach to operationalize hate speech for the purpose of content analysis, combining qualitative text annotation and standardized labeling, in which hate speech is not directly identified by coders but rather results from the combination of different characteristics. This approach allows for quantitative description as well as for in-depth analysis of the material. In this article, we focus on the results of a qualitative content analysis of German user comments posted on the topic of immigration and refuge. The in-depth exploration of offensive user comments in the sample shows that our modularized approach allows us to go beyond the common 'hate/no-hate' dichotomy and empirically map the variety of hate speech and offensive language in the subject area.

2. Challenges in Defining Hate Speech

Hate speech is a complex phenomenon and defining it is challenging in several ways. According to Andrew Sellars, "any solution or methodology that purports to present an easy answer to what hate speech is and how it can be dealt with is simply not a product of careful thinking" (Sellars, 2016, p. 5). An important point of disagreement, for example, is the *group reference* of hate speech: On the one hand, some definitions tie the phenomenon to minority groups or list specific group characteristics such as race, religion, gender, or sexual orientation (e.g., Waltman & Mattheis, 2017). On the other hand, some authors stress that hate speech can target every possible group (e.g., Parekh, 2006).

From a theoretical perspective, there are three main approaches in defining hate speech that each *emphasize different aspects*: approaches that (1) refer to the *intentions* behind hate speech; (2) address the *perception* and possible damage of hate speech; and (3) focus on the *content* level and attempt to define hate speech by certain content characteristics (Sellars, 2016, pp. 14–18). For the purpose of content analysis, content-based definitions seem to be most appropriate. For example, Saleem, Dillon, Benesch, and Ruths (2017) focus on speech containing an expression of hatred and use the term 'hateful speech' to emphasize the nuance. Bhikhu Parekh also argues in favor of a content-based understanding and defines 'hate speech' as speech that singles out individuals or groups on the basis of certain characteristics, stigmatizes them and places them outside of society; as such, hate speech "implies hostility, rejection, a wish to harm or destroy, a desire to get the target group out of one's way" (Parekh, 2006, p. 214).

Many scholars approach the heterogeneity of hate speech with rather *broad framework*s: For example, Alexander Brown argues "that the term 'hate speech' is equivocal, that it denotes a family of meanings, for which there is no one overarching precise definition available" (Brown, 2017b, p. 562). He proposes a family resemblances' concept to address hate speech, that is, a "network of similarities overlapping and criss-crossing" (Brown, 2017b, p. 596). Using speech act theory and its basic distinction between locutionary, illocutionary, and perlocutionary speech acts, Sponholz (2017) differentiates hateful speech, hate-fomenting speech, and dangerous speech. The characteristic features of these types are the content, the intent of the speaker, and the context-dependent impact. However, they also differ in respect to language: While hateful speech is typically emotional and uses derogatory language (such as insults or slurs), hate-fomenting speech tends to follow the principles of rationality and reasoning (Sponholz, 2017, pp. 3–5). Nevertheless, empirical studies are challenged with in-between forms of these types, in which the emotional and the rational side of hate speech "coexist to varying degrees" (Keipi et al., 2017, p. 54).

Several authors stress that the emotion or attitude of hatred is not necessarily an essential part of hate speech. Moreover, hate speech can also be rooted, for example, in (religious) beliefs, power relations, boredom, attention-seeking, or negligence (Brown, 2017a). That is why spontaneous and unconsidered forms of hate speech can be expected particularly in participatory online spaces (Brown, 2018, pp. 304–306).

Another problem with hate speech identification is its *overlap with neighboring concepts*. Obviously, hate speech is not the same as dislike or disapproval (Parekh, 2006). However, it is a challenge to consistently identify hate speech and distinguish it from other forms of negative evaluation, since our understanding of hate speech is shaped by changing societal norms, context, and interpretation (Post, 2009; Saleem et al., 2017). This issue becomes evident in low reliability scores for hate speech identification reported in some studies, for example by Ross et al. (2016).

Against this theoretical background, our framework is based on a content-related understanding of hate speech, which seems most appropriate for the purpose of content analysis. In order to avoid assumptions about the intentions of the speaker or possible consequences of a given statement, we argue in favor of the legal origins of the term 'hate speech' and focus on discriminatory content and references to violence within a given statement.

3. Operationalizing Hate Speech for Content Analysis

In order to measure hate speech, the theoretical definitions and dimensions have to be transferred into empirically operable instructions for identifying and categorizing occurrences of the concept. To answer RQ1, in this section we first review existing approaches of measuring hate speech, before we develop our theoretical framework to identify hate speech content in public communication in a multi-dimensional way.

3.1. Existing Approaches of Measuring Hate Speech

Contradicting the elaborate theoretical discussion of hate speech, many empirical studies follow a 'hate/no-hate' dichotomy when categorizing communication content (e.g., Lingiardi et al., 2020). This applies also to non-scientific classification, e.g., in the context of the 'Network Enforcement Law' in Germany, which forces platform companies to identify and block "apparently unlawful" content, including hate speech. Here, it is solely the criterion of 'unlawfulness' that differentiates hate speech from non-hate speech. As a result of this approach, the number of identified (and blocked) items is rather small, given the far-reaching guaranties of free speech in western countries (Facebook, 2020) and does not allow for many insights into the content dimension of hate speech. Thus, many studies in the field avoid a formal law-based, narrow operationalization of hate speech and operate with broader concepts such as 'incivility' (Coe et al., 2014) or 'negative speech' (e.g., Ben-David & Matamoros-Fernández, 2016). A common approach to operationalize such general categories for content analyses is the use of dictionaries that provide pre-categorizations of search terms (e.g., Cho & Kwon, 2015) and that allow for more advanced manual and automated coding of hateful content (e.g., Davidson et al., 2017).

A more differentiated categorization of hate speech can be provided by qualitative approaches (e.g., Ernst et al., 2017), which have the capacity to identify multiple aspects of hate speech and relate them to theoretical dimensions in detail. However, qualitative analyses usually focus on in-depth analysis of specific cases and cannot handle large bodies of text. An example of a multi-dimensional approach to identifying and categorizing different levels of incivility and hate speech on a larger scale following a quantitative approach is presented by Bahador and Kerchner (2019). They applied a computer-aided manual categorization model that ranked the intensity of hate speech on a six-point scale, allowing both for systematic analysis of larger amounts of text and a differentiated recording of aspects of hate speech.

3.2. Introducing a New Approach

Following our theoretical argument, we developed a flexible labeling scheme that measures three key elements of hate speech in text: First, the negative evaluation of a group as a whole, i.e., *negative stereotyping*, is one common element of many hate speech definitions (e.g., Parekh, 2006). For the purpose of our coding scheme, we define negative stereotyping as the attribution of negatively connotated characteristics, roles, or behaviors to the whole group or to individuals on the basis of their group membership (see also, Trebbe, Paasch-Colberg, Greyer, & Fehr, 2017).

Secondly, *dehumanization* is often singled out as one element of hate speech (e.g., Bahador & Kerchner, 2019). On the basis of this literature, we define statements as dehumanization that equate or compare humans with inanimate things (e.g., "scum" or "pack"), animals (e.g., "rats") or inhuman beings (e.g., "demons," "vampires") or characterize humans as savage or animalistic (see also, Maynard & Benesch, 2016). As such, dehumanization is a form of negative stereotyping. However, we agree with Bahador and Kerchner who argue that "dehumanization is a particularly extreme type of negative characterization…and a well-established tool for justifying political violence, and thus merits its own category" (Bahador & Kerchner, 2019, p. 6).

Third, the *expression of violence, harm, or killing* is another important element of hate speech (e.g., Bahador & Kerchner, 2019; Parekh, 2006). Our approach, therefore, defines all statements as hate speech that justify, incite, or threaten physical violence against an individual or a group or that justify, incite, or threaten the killing of individuals or members of a group.

These three elements are measured independently of each other in the sense that they can, but do not have to, apply simultaneously to a comment in order to qualify as hate speech. Thus, our approach allows us to distinguish between forms of hate speech using various combinations of these three elements. In this respect, our approach differs from the scale developed by Bahador and Kerchner (2019), which conceptualizes negative actions, negative characterization, demonizing/dehumanization, violence, and death as different points on a hate speech intensity scale. However, with the help of the hate speech elements of our framework and their various combinations, different types and intensities of hate speech can be identified in the empirical analysis.

For such an analysis, the use of offensive language below the level of hate speech needs to be included, too. Therefore, our coding scheme accounts for three different forms of offensive language that are measured independently of the three hate speech elements: insults and slurs, degrading metaphors, and degrading wordplays.

4. Method

In order to both test this approach empirically and answer our research questions, we conducted a structured text annotation of user comments on news about immigration and refuge to Germany posted in March

2019 in the comment sections of three German news sites (*Compact Magazin*, *Epoch Times*, *Focus Online*), one right-wing blog (*PI news*), four Facebook pages (*FOCUS Online*, *The Epoch Times*, *WELT, Zeit Online*) and 13 YouTube channels (*ARTEde*, *BILD*, *COMPACTTV*, *DW Deutsch*, *Epoch Times Deutsch*, *euronews (deutsch)*, *KenFM*, *Laut Gedacht*, *MrMarxismo*, *Oliver Flesch*, *RT Deutsch*, *tagesschau*, *Tagesschlau*). These sources were selected on the basis of a preliminary study in August 2018, which considered a much broader variety of sources (8 news sites, 3 right-wing blogs, 7 Facebook pages, 31 YouTube channels, and 1 Q&A platform) chosen on the basis of their high reach, their relevance to the public debate of immigration (indicated by the number of user comments), and the variation in their discourse architectures (i.e., comment section, discussion forum, social media, Q&A platform). Following the results of this preliminary study, we selected those sources that contained most hate speech against refugees and immigrants in order to collect as much material as possible for the following analysis. Accordingly, the sample is not designed for a systematic comparison of hate speech in different types of sources.

Using topic related search terms, these sources were screened for articles and posts referring to the topic of immigration and refuge capturing all related user comments. We then randomly selected 178 articles and posts with a total of 6,645 related user comments (for each initial article or post the first up to 50 user comments) for the subsequent analysis. This material was annotated using the *BRAT rapid annotation tool*, a browser-based software for structured text annotation (Stenetorp et al., 2012).

The method of structured text annotation includes that each text is examined for relevant words, sentences, or sections ('entities'), which are then selected and labeled with predefined categories ('entity attributes'). Thus, this method is basically a combination of the inductive identification of relevant text segments as we know it from computer-assisted qualitative text analysis and the assignment of codes to these text segments as we know it from standardized content analysis. As such, it is particularly helpful for content analysis as the classification is explicitly related to specific parts of a text, which are at the same time recorded for subsequent analysis. This allows us to conduct both a standardized and a qualitative content analysis of the annotated user comments.

Both the methodological approach and our focus on immigration and refuge to Germany were chosen due to the broader research context of this study, which aims at the automatization of detecting hate speech against refugees and immigrants in German user comments. For this reason, we will first take a closer look at the situation in Germany (see Section 4.1). While our methodological approach is suitable for analyzing hate speech against various groups (see Section 4.2), the results presented in this article are limited to hate speech against refugees and immigrants. This is not to say that only refugees and immigrants are recently affected by hate speech in Germany; anti-Semitic hate speech, for example, has dramatically increased again as well (Hänel, 2020; Schwarz-Friesel, 2019). Nevertheless, we consider a focus on a specific target group to be helpful in order to distinguish hate speech from neighboring concepts.

4.1. Immigration and Refuge to Germany

The topic of immigration and refuge was chosen for our case study as it has been heavily discussed in public since 2015, when the German Chancellor Angela Merkel decided to keep the state's borders open and the number of refugees entering Germany rose sharply. Even though the number of asylum applications dropped drastically in 2017 and continues to decrease (Bundesamt für Migration und Flüchtlinge, 2020), questions of immigration have repeatedly triggered heated political debates in Germany in the following years and have long been high on the media agenda (e.g., Krüger & Zapf-Schramm, 2019). The public opinion was increasingly divided on the issue, and dissatisfaction with political institutions and the processes that deal with it is widespread (Arlt, Schumann, & Wolling, 2020).

This social division has become apparent, for example, in anti-immigration protests (Bennhold, 2018), the rise of the populist extreme right-wing party 'Alternative für Deutschland' (Bennhold, 2018) and a growing mistrust regarding the accuracy of media coverage on refugees (Arlt & Wolling, 2016). However, this issue is of particular relevance as the growing online hate speech against refugees and immigrants has been accompanied by an increase in racist hate crimes against these groups in Germany in recent years (Eddy, 2020; Hille, 2020). Examples include the attacks in Hanau (February 2020), Halle (October 2019) and Munich (June 2016) as well as the murder of the Hessian politician Walter Lübcke, who publicly supported liberal refugee politics (June 2019). There is reason to believe that such hate crimes are verbally prepared, socially backed, and ideologically legitimized by hate speech on far-right websites and forums, but also on social media and in comment sections of news websites (see e.g., Scholz, 2020).

4.2. Annotation Rules and Coding Scheme

On the basis of a detailed theory-based manual, three trained coders annotated the user comments in our sample following three steps: First, the initial article or post was read and checked for thematic relevance. The net sample contains 135 relevant articles or posts with 5,031 corresponding user comments.

In the second step, all judgments of individuals or groups within these comments were identified and annotated as 'entities' on a sentence level. Thereby, a judgment is defined as a statement expressing an opinion or an evaluation of the person/group by ascribing negative characteristics, roles, or behavior to it. Such judgments

can be recognized by attributions of adjectives, judgmental subjectivizations, the attribution of behavior that violates social standards, or by association with certain consequences (e.g., damage). In addition to such explicit judgments, the coders were also instructed to identify and annotate implicit judgments, expressed by rhetorical questions, ironic statements, or historical references. In order to capture such implicit forms as validly as possible, the manual includes dimensions and examples taken from a qualitative expert survey of German community managers (Paasch-Colberg, Strippel, Laugwitz, Emmer, & Trebbe, 2020), qualitative pre-coding and literature.

In the third step, all annotated judgments were further qualified by attributing predefined labels to them, such as the targets of judgment (e.g., politicians, journalists/media, German citizens, right-wing groups, Muslims, and refugees/immigrants) and the subject of judgment (e.g., culture, sexuality or character/behavior). Judgments that were attached to a specific group membership (i.e., ethnicity, nationality, religion, gender, profession) and are thus stereotyping were labeled accordingly. It was further specified whether a judgment includes a dehumanization (as defined in Section 3.2) or a response to the target group. Possible responses range from non-violent forms (i.e., rejection) to violent forms (i.e., the legitimization, threat or call for physical violence or killing of the person/group).

Finally, the manual contains three attributes to specify different forms of offensive language, i.e., insults and slurs, derogatory metaphors and comparisons as well as derogatory wordplays. The manual includes examples for these forms of offensive language used in German user comments, which were drawn primarily from the aforementioned expert survey.

The context unit of the annotation of judgments in a user comment were the news item or social media posting and the preceding comments, i.e., the coders were instructed to use textual references within this context to identify judgments.

4.3. Data Analysis

A qualitative content analysis was conducted for all user comments in the sample that contains at least one element of hate speech or offensive language according to our framework. The hate speech and offensive language elements described in Section 3.2 were thus used as deductive categories to pre-structure the material. In the first step, the comments in these pre-set categories were close-read in order to describe and exemplify the categories as such. In the second step, the analysis was focused on those comments that target refugees or other immigrants and segmented into (1) comments that qualify as hate speech and (2) comments that qualify as offensive language but not as hate speech. These comments were then further explored using the technique of structuring qualitative content analysis according to Mayring (2015). This form of qualitative content analysis focuses on patterns and co-occurrences of selected characteristics in the material and aims at the description of different types in the material (Mayring, 2015, pp. 103–106; Schreier, 2014). To assure consistency, the material was close-read by two researchers independently and inconsistencies were resolved in discussing.

5. Results

In our sample of 5,031 user comments, 2,602 negative judgments were identified. Hate speech was identified in 25% of the judgments (n = 701) and, since a comment can contain more than one judgment, in 11% of the comments (n = 538). With regard to the three hate speech elements, negative stereotyping is by far the most frequent element. Every fifth judgment in our sample (n = 539) uses negative stereotypes while only 155 judgments (6%) dehumanize the target. Calls for violence or death were identified even less frequently (n = 56 and n = 57). The majority of judgments with hate speech are targeting the group of refugees and immigrants.

Offensive language is more frequent in our sample than hate speech. And if offensive language is used in a comment, it is often used more than once: Offensive language was identified in 16% of the comments (n = 796) and 38% of the judgments (n = 1,070). About 60% of these judgments use offensive language without qualifying as hate speech according to our framework.

5.1. Describing Hate Speech in German User Comments on Immigration and Refuge

The following sections present examples of user comments that contain potentially offensive and upsetting terms, particularly racist and islamophobic. They are solely used as examples to illustrate the results of this research and do not reflect the views of the authors in any way. The user comments were translated, the German originals can be found in the supplementary document.

In a first step, we close-read the user comments to illustrate in more depth how the three hate speech elements defined in Section 3.2 can be identified. This aims at describing the main categories of our framework, illustrates them with examples and thus makes them applicable for further analysis in the field. As Table 1 shows, hate speech in our sample is expressed through different kinds of rhetoric and can be identified by different indicators. At the least extreme level, groups are negatively stereotyped by referring to the average or majority of its members or by calling a behavior or role typical for the group. Another form of stereotyping is to criticize the behavior of a group as a negative deviation from supposedly normal behavior.

Dehumanizing hate speech refers to humans as things, animals, or other inhuman beings, considered inferior, disgusting, or dangerous.

Table 1. Description of hate speech elements.

Hate speech element	Example (English translation)
Negative stereotyping	
Referring to everybody, most people, or the average or typical person	"These newcomers are all potential killers, they pull out their knives on every little thing"
Social groups, religious groups, professional roles, or nationalities are generalized	"Muslim and Black African, the recipe for murder and manslaughter"
Critique is tied to the deviation of 'normality'	"Nowhere else in the world do criminal asylum seekers get so much support and so many murderers can run free like they do here"
Dehumanization	
Humans are equated as or compared to inanimate things	"Whoever takes the stuff out has to be well paid. What kind of sewer man digs in shit without proper pay?"
Humans are equated as or compared to animals or inhuman beings	"Unfortunately, the money is not enough to get rid of even a small portion of these parasites"
Violence and killing	
Fantasies of violence/killing	"Let the cops beat him until he's crippled! Then fly him across the desert and throw him out"
Violence/killing as only effective means or remedy	"The only thing that helps is violence"
Violence/killing as a right/appropriate solution	"It would have been faster, cheaper and more sustainable to just shoot him"
Specific calls for violence/killing	"When all subjects are still in deep sleep, let's blow up the asylum center!"

The user comments in the category 'violence and killing' address a broad spectrum of violence, ranging from general physical violence and more specific forms such as sexual violence, violence in law enforcement, extreme punishment (i. e., forced labor, torture), or (civil) war to murder, suicide, deadly revenge, or death penalty. Furthermore, the category includes violent fantasies, rhetoric describing violence or killing as the only effective means or the appropriate solution, and specific calls for violent action or killing.

5.2. Mapping the Variety of Hate Speech and Offensive Language towards Immigrants and Refugees

To answer RQ2, we then used our multi-dimensional annotations to identify patterns by grouping the user comments in our sample to general types. To derive the types, the common occurrence of the labeled characteristics (including the three hate speech elements and forms of offensive language as defined in Section 3.2) was examined. For those user comments that target immigrants or refugees, five types of hate speech emerged which partly build on each other, so that their borders tend to be blurry; also, individual user comments may recur to more than one type at once.

Racist othering: Key characteristics of this type are an 'us against them'-rhetoric and a sharp devaluation of the designated out-group. At least implicitly, this type is the basic motive of hate speech. The element of negative stereotyping applies to all comments of this type: Immigrants and refugees are negatively stereotyped (e.g., as lazy, stupid, rude), and framed as a burden and imposition to the ingroup, as this example shows: "Anyone who comes here to participate in what our forefathers built, and their ancestors did not contribute anything at all, is unwanted because he is only scrounging, no matter what else he says or does." The devaluation is often associated with descriptions of an allegedly abnormal sexual life, as in this example: "Illiterate Afros, Arabs, and Afghanis have no access to women of their ethnicity and consequently suffer a hormonal emergency."

Racist criminalization: This type is a special form of negative stereotyping, which focuses on the description of immigrants and refugees as a threat. Crime is culturized and associated particularly with the male gender. In this context, it is striking that the knife is coined as the central tool of crime, shaping the image of an uncivilized wild: "We live in hard times in which one must constantly count on getting a knife from foreigners, who were raised differently." Forms of self-victimization are also identified, whereby the sexual motif reappears as the narrative of the threatened German woman: "These murderers, rapists, and thieves from Morocco, Algeria, or Mauritania cause the most damage to the population and are therefore Merkel's darlings."

Dehumanization: This type builds on the previous types, but is characterized by an additional dehumanization of the target group; in other words, comments of this type are distinguished by the common presence of the elements of negative stereotyping and dehumaniza-

tion. Immigrants and refugees are compared or referred to as non-human things or beings that are connoted as inferior, disgusting or even dangerous, as this example shows: "The scum from which the whole world protects itself is integrated into the social systems here." The second example is a hybrid of racist criminalization, expressed through a play on the words criminal and migrant, and dehumanization: "These Crimigrants are predators. They lurk and choose their victims."

Raging hate: User comments of this type are distinguished by the element of violence and killing. Physical violence against immigrants and refugees or even their death is legitimized or demanded; other comments imply fantasies of violence and killing as acts of revenge. Some comments of this type also contain the element of dehumanization, as if to justify (lethal) violence. Further, this type is characterized by the use of offensive language, i. e., insults, which imply malice, cynicism, disgust, and aggression: "That filthy mutt is still alive?"

Call for hate crimes: The main characteristic of this type is the occurrence of the element of violence and killing. However, in contrast to the type of raging hate, this is done without the use of offensive language, but in a distanced and calm form, as this example shows: "The attacker should be shot, stabbed, or beaten to death immediately. Another language is not understood by Muslim Africans. Otherwise they understand: Keep up the good work." Calls for hate crimes often use negative stereotyping (e. g., by criminalizing immigrants and refugees) and dehumanization as a justifying rhetoric: "They should not be stoned, but fed to the lions. Something like that must not live."

We further analyzed the use of offensive language in the user comments, to assess its role for the five types of hate speech as well as the grey area that exists in the demarcation of hate speech and offensive language. The in-depth analysis showed that comments targeting immigrants and refugees use different forms of offensive language.

First, the target group is described with common racial slurs and insults. User comments that contain racial insults but none of the hate speech elements described in Section 3.2 do not qualify as hate speech according to our framework. However, they would do so on the basis of other definitions in the literature (e.g., Saleem et al., 2017). Thus, they are clearly sitting in a grey area between hate speech and offensive language.

In addition, derogatory group labels are identified that either use neologisms or wordplays. The distinction between this form and common racial insults is temporary and fluent, as such, these labels can also be considered as a grey area. However, they are difficult to capture in standardized approaches and require special knowledge. The same holds for ironic group labels (e.g., "gold pieces") that are highly context-sensitive.

Another form of offensive language can be referred to as distancing, as it denies refugees their legal status, e.g., by using quotation marks ("so-called 'refugees'"), adjectives such as "alleged" or neologisms such as "refugee actors." Distancing can be understood as a preliminary stage to racist othering. Finally, user comments referring to immigrants and refugees also use common insults (e.g., "wanker") against them without referring to the group of refugees as a whole. Therefore, this form qualifies as incivility (in the sense of impoliteness) but clearly not as hate speech.

Offensive language was found to be used in all hate speech types, however, the type 'call for hate crimes' seems to be an exception to that.

6. Conclusions

In this article we developed a new approach to hate speech definition and identification that aims at solving some of the described challenges in the field of research and goes beyond the common 'hate/no-hate' dichotomy. To add more depth to the concept of hate speech and answering RQ1, our theoretical approach first developed a multi-dimensional understanding of the term based on the dimensions of discriminatory content and references to violence, which in the second step was measured using a set of independent labels. In contrast to most existing studies in the field, hate speech thus could be measured indirectly and in a multi-dimensional way.

In a structuring content analysis of user comments targeting immigrants and refugees, we showed how this approach allows an in-depth analysis of the character of hate speech statements in a content analysis as well as, in a second step, the development of distinct types of hate speech that form a dark spectrum of discrimination and violence-related statements. Answering RQ2, our approach captures recurring patterns of hate speech, as identified and described in other qualitative studies, and enables their standardized measurement: The types of racist othering, racist criminalization, and dehumanization correspond largely to some of the hate myths identified by Waltman and Mattheis (2017) in hate novels of US-American white supremacists. Dehumanization and racist criminalization resemble closely some of the justificatory hate speech mechanisms identified by Maynard and Benesch (2016, pp. 80–82) in the context of mass atrocities.

The results further show that two of the hate speech types are characterized by a special relationship to language and thus deepens our knowledge on the role of offensive language for hate speech: While the use of offensive language is constitutive for 'raging hate,' the type 'call for hate crimes' is characterized by a quite rational language. Hence, the empirical analysis supports our argument that a deeper theoretical conceptualization of hate speech and offensive language as two distinct dimensions allows for much more detailed insights into the nature of this phenomenon.

Our case study is limited in several ways. Firstly, our analysis addresses hate speech in user comments. While this is a relevant perspective because most hate content

emerges in this sphere, it is only one facet of the problem of offensive language in participatory online discussions. In order to better understand the dynamics of escalating discussions, future studies should therefore consider the broader context and, for example, analyze discriminatory speech in the discussed news pieces and social media posts themselves.

Secondly, the analysis is not based on a representative set of sources, but biased by the right-wing news sites and the right-wing blog selected for analysis. Therefore, our typology can only be preliminary and must be validated and quantified in further studies. Such further empirical applications of our framework should in particular consider the differences between different types of sources systematically.

Thirdly, we limited our analysis to hate speech targeting immigrants and refugees, as this seems to be particularly relevant against the background of recent hate crimes in Germany (see Section 4.1). Nevertheless, the question of what forms of hate speech are used to target other social groups should definitely be answered in future studies.

Finally, capturing implicit forms of hate speech is quite difficult. In order to prevent corresponding user comments from being deleted directly, hate speech is sometimes strategically disguised (e.g., Warner & Hirschberg, 2012). Another challenge with regard to right-wing blogs and websites in specific is the strategy of right-wing extremists to use their websites for image control and to avoid open racisms and calls for violence (e.g., Gerstenfeld, Grant, & Chiang, 2003). Through a previous expert survey, we were able to supplement our manual with many current examples of implicit hate speech. However, this form of hate speech can change significantly over time, which is why our manual at this point is more of a snapshot that needs updating for and through future research. Moreover, our framework focuses on text and does not include forms of hate speech expressed by non-textual communication, such as memes for example.

Nevertheless, we argue that our framework provides a sensitive tool to describe the prevalence of hate speech in more detail than existing approaches, while also considering borderline cases and rhetoric that prepare hate speech. This extended perspective on the phenomenon of hate speech is promising to better understand escalating dynamics in participatory online spaces and to empirically test different counter-measures, for example. This is of particular importance for practical social media community and content management. When integrated into existing (semi-)automated content management systems, such a tool that distinguishes between several types and intensities of incivility and hate speech may contribute to more adequate strategies of dealing with disturbing content than many of the existing keyword-based and binary 'hate/no-hate' systems. This is even more important as simple deletion of 'hate'-labeled postings often raises concerns of censorship, particularly when measurement is blurry and mistakenly covers also non-hate speech content.

Finally, with reference to the various hate speech definitions in the literature, we want to point out the flexibility of our approach: It can be adapted to answer specific research questions and make different or broader hate speech definitions operational for content analysis, e.g., definitions of 'hateful speech' that would include racial insults but exclude the element of negative stereotyping. By combining it with surveys or experiments, the content-related perspective of our approach can also be related to other perspectives on hate speech in order to provide additional insights, for example, into the interplay of text characteristics and their perception by different population groups.

Acknowledgments

This research is part of the project "NOHATE—Overcoming crises in public communication about refugees, migration, foreigners," funded by the German Federal Ministry of Education and Research (grant number 01UG1735AX). The authors would like to thank Laura Laugwitz for her support of the study and the coding students for their work.

Conflict of Interests

The authors declare no conflict of interests.

Supplementary Material

Supplementary material for this article is available online in the format provided by the author (unedited).

References

Arlt, D., Schumann, C., & Wolling, J. (2020). Upset with the refugee policy: Exploring the relations between policy malaise, media use, trust in news media, and issue fatigue. *Communications*. Advance online publication. https://doi.org/10.1515/commun-2019-0110

Arlt, D., & Wolling, J. (2016). The refugees: Threatening or beneficial? Exploring the effects of positive and negative attitudes and communication on hostile media perceptions. *Global Media Journal German Edition*, *6*(1), 1–21.

Bahador, B., & Kerchner, D. (2019). *Monitoring hate speech in the US media* (Working Paper). Washington, DC: The George Washington University.

Ben-David, A., & Matamoros-Fernández, A. (2016). Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain. *International Journal of Communication*, *10*, 1167–1193.

Bennhold, K. (2018, August 13). Chemnitz protests show new strength of Germany's far right. *The New*

York Times. Retrieved from https://www.nytimes.com/2018/08/30/world/europe/germany-neo-nazi-protests-chemnitz.html

Bleich, E. (2011). The rise of hate speech and hate crime laws in liberal democracies. *Journal of Ethnic and Migration Studies*, *37*(6), 917–934.

Brown, A. (2017a). What is hate speech? Part 1: The myth of hate. *Law and Philosophy*, *36*(4), 419–468.

Brown, A. (2017b). What is hate speech? Part 2: Family resemblances. *Law and Philosophy*, *36*(5), 561–613.

Brown, A. (2018). What is so special about online (as compared to offline) hate speech? *Ethnicities*, *18*(3), 297–326.

Bundesamt für Migration und Flüchtlinge. (2020). *Asyl und Flüchtlingsschutz. Aktuelle Zahlen* [Asylum and refugee protection. Current figures] (04/2020). Berlin: Bundesamt für Migration und Flüchtlinge.

Cho, D., & Kwon, K. H. (2015). The impacts of identity verification and disclosure of social cues on flaming in online user comments. *Computers in Human Behavior*, *51*, 363–372.

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, *64*(4), 658–679.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). *Automated hate speech detection and the problem of offensive language.* Paper presented at the Eleventh International Conference on Web and Social Media, Montreal, Canada.

Eddy, M. (2020, February 21). Far-right terrorism is no. 1 threat, Germany is told after attack. *The New York Times*. Retrieved from https://www.nytimes.com/2020/02/21/world/europe/germany-shooting-terrorism.html

Erjavec, K., & Kovačič, M. P. (2012). "You don't understand, this is a new war!" Analysis of hate speech in news web sites' comments. *Mass Communication and Society*, *15*(6), 899–920.

Ernst, J., Schmitt, J. B., Rieger, D., Beier, A. K., Vorderer, P., Bente, G., & Roth, H.-J. (2017). Hate beneath the counter speech? A qualitative content analysis of user comments on YouTube related to counter speech videos. *Journal for Deradicalization*, *2017*(10), 1–49.

Facebook. (2020). *NetzDG Transparenzbericht* [Network enforcement act: Transparency report]. Menlo Park, CA: Facebook. Retrieved from https://about.fb.com/wp-content/uploads/2020/01/facebook_netzdg_Januar_2020_German.pdf

Gerstenfeld, P. B., Grant, D. R., & Chiang, C.-P. (2003). Hate online: A content analysis of extremist Internet sites. *Analyses of Social Issues and Public Policy*, *3*(1), 29–44.

Hänel, L. (2020, May 7). Germany: Anti-semitism despite remembrance culture. *Deutsche Welle*. Retrieved from https://www.dw.com/en/germany-anti-semitism-despite-remembrance-culture/a-53360634

Hille, P. (2020, February 20). Right-wing terror in Germany: A timeline. *Deutsche Welle*. Retrieved from https://www.dw.com/en/right-wing-terror-in-germany-a-timeline/a-52451976

Keipi, T., Näsi, M., Oksanen, A., & Räsänen, P. (2017). *Online hate and harmful content: Cross-national perspectives*. London: Routledge.

Krüger, U. M., & Zapf-Schramm, T. (2019). InfoMonitor 2018: GroKo und Migrationsdebatte prägen die Fernsehnachrichten. Analyse der Nachrichtensendungen von Das Erste, ZDF, RTL und Sat.1 [InfoMonitor 2018: Grand coalition and migration debate dominate the television news—Analysis of the news programs of Das Erste, ZDF, RTL and Sat.1]. *Media Perspektiven*, *2019*(2), 44–73.

Maynard, J., & Benesch, S. (2016). Dangerous speech and dangerous ideology: An integrated model for monitoring and prevention. *Genocide Studies and Prevention*, *9*(3), 70–95.

Lingiardi, V., Carone, N., Semeraro, G., Musto, C., D'Amico, M., & Brena, S. (2020). Mapping twitter hate speech towards social and sexual minorities: A lexicon-based approach to semantic content analysis. *Behaviour & InformationTechnology*, *39*(7), 711–721.

Mayring, P. (2015). *Qualitative Inhaltsanalyse. Grundlagen und Techniken* [Qualitative content analysis: Basics and techniques] (12th ed.). Weinheim and Basel: Beltz.

Paasch-Colberg, S., Strippel, C., Laugwitz, L., Emmer, M., & Trebbe, J. (2020). Moderationsfaktoren: Ein Ansatz zur Analyse von Selektionsentscheidungen im Community Management [Moderation factors: A framework to the analysis of selection decisions in community management]. In V. Gehrau, A. Waldherr, & A. Scholl (Eds.), *Integration durch Kommunikation. Jahrbuch der Publizistik- und Kommunikationswissenschaft 2019* [Integration through communication. Yearbook of Journalism and Communication Studies 2019] (pp. 109–119). Muenster: DGPuK.

Parekh, B. (2006). Hate speech. Is there a case for banning? *Public Policy Research*, *12*(4), 213–223.

Post, R. (2009). Hate speech. In I. Hare & J. Weinstein (Eds.), *Extreme speech and democracy* (pp. 123–138). Oxford: Oxford University Press.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2016). *Measuring the reliability of hate speech annotations: The case of the European refugee crisis.* Paper presented at the 3rd Workshop on Natural Language Processing for Computer-Mediated Communication, Bochum, Germany.

Saleem, H. M., Dillon, K. P., Benesch, S., & Ruths, D. (2017). *A web of hate: Tackling hateful speech in online social spaces*. Paper presented at the first Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS), Portorož, Slovenia.

Scholz, K.-A. (2020, February 21). How the inter-

net fosters far-right radicalization. *Deutsche Welle*. Retrieved from https://www.dw.com/en/how-the-internet-fosters-far-right-radicalization/a-52471852

Schreier, M. (2014). Qualitative content analysis. In U. Flick (Ed.), *The Sage handbook of qualitative data analysis* (pp. 1–19). London: Sage.

Schwarz-Friesel, M. (2019). "Antisemitism 2.0": The spreading of Jew-hatred on the World Wide Web. In A. Lange, K. Mayerhofer, D. Porat, & L. H. Schiffman (Eds.), *Comprehending and confronting antisemitism: A multi-faceted approach* (pp. 311–338). Boston, MA: De Gruyter.

Sellars, A. F. (2016). *Defining hate speech* (Research publication No. 2016–20). Cambridge, MA: Berkman Klein Center.

Sponholz, L. (2017). *Tackling hate speech with counter speech? Practices of contradiction ans their effects*. Paper presented at the International Conference Worlds of Contradiction, Bremen, Germany.

Stenetorp, P., Pyysalo, S., Topic, G., Ohta, T., Ananiadou, S., & Tsujii, J. I. (2012). *BRAT: A web-based tool for NLP-assisted text annotation*. Paper presented at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France.

Trebbe, J., Paasch-Colberg, S., Greyer, J., & Fehr, A. (2017). Media representation: Racial and ethnic stereotypes. In P. Rössler (Ed.), *The international encyclopedia of media effects*. Hoboken, NJ: John Wiley & Sons.

Waltman, M. S., & Mattheis, A. A. (2017). Understanding hate speech. In *Oxford Research Encyclopedia of Communication*. New York, NY: Oxford University Press. Retrieved from http://communication.oxfordre.com/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-422

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web. In S. Owsley Sood, M. Nagarajan, & M. Gamon (Eds.), *Proceedings of the 2012 Workshop on Language in Social Media* (pp. 19–26). Montreal: Association for Computational Linguistics.

About the Authors

Sünje Paasch-Colberg is a Postdoc Researcher at the Institute for Media and Communication Studies at Freie Universität Berlin, Germany. She currently works in the research project "NOHATE—Overcoming crises in public communication about refugees, migration, foreigners." Her research interests are digital communication, migration, integration and media, and content analysis. In her dissertation she worked on media effects in election campaigns.

Christian Strippel is a Research Assistant and PhD candidate at the Institute for Media and Communication Studies at Freie Universität Berlin, Germany. He currently works in the research project "NOHATE—Overcoming crises in public communication about refugees, migration, foreigners." His research interests include digital communication, media use, public sphere theory, and sociology of scientific knowledge.

Joachim Trebbe is Professor for media analysis and research methods at the Institute for Media and Communication Studies at Freie Universität Berlin, Germany. His research interests include the media use of migrants in Germany, the media representations of ethnic minorities, and research methods for television content analyses.

Martin Emmer is Professor for media use research at the Institute for Media and Communication Studies at Freie Universität Berlin and Principle Investigator of the research group "Digital Citizenship" at the Weizenbaum Institute, Berlin, Germany. His research focusses on digital communication and digital publics, political communication and participation, and methods of empirical communication research.

COGITATIO

Media and Communication (ISSN: 2183–2439)
2021, Volume 9, Issue 1, Pages 181–194
DOI: 10.17645/mac.v9i1.3469

Article

Constructive Aggression? Multiple Roles of Aggressive Content in Political Discourse on Russian YouTube

Svetlana S. Bodrunova [1], Anna Litvinenko [2,*], Ivan Blekanov [3,4] and Dmitry Nepiyushchikh [4]

[1] School of Journalism and Mass Communications, St. Petersburg State University, 199004 St. Petersburg, Russia; E-Mail: s.bodrunova@spbu.ru
[2] Institute for Media and Communication Studies, Freie Universität Berlin, 14195 Berlin, Germany; E-Mail: anna.litvinenko@fu-berlin.de
[3] School of Mathematics and computer Science, Yan'an University, Yan'an Shanxi, China;
[4] Faculty of Applied Mathematics and Control Processes, St. Petersburg State University, 199004 St. Petersburg, Russia; E-Mails: i.blekanov@spbu.ru (I.B.), st075408@student.spbu.ru (D.N.)

* Corresponding author

Submitted: 13 July 2020 | Accepted: 19 August 2020 | Published: 3 February 2021

Abstract
Today, aggressive verbal behavior is generally perceived as a threat to integrity and democratic quality of public discussions, including those online. However, we argue that, in more restrictive political regimes, communicative aggression may play constructive roles in both discussion dynamics and empowerment of political groups. This might be especially true for restrictive political and legal environments like Russia, where obscene speech is prohibited by law in registered media and the political environment does not give much space for voicing discontent. Taking Russian YouTube as an example, we explore the roles of two under-researched types of communicative aggression—obscene speech and politically motivated hate speech—within the publics of video commenters. For that, we use the case of the Moscow protests of 2019 against non-admission of independent and oppositional candidates to run for the Moscow city parliament. The sample of over 77,000 comments for 13 videos of more than 100,000 views has undergone pre-processing and vocabulary-based detection of aggression. To assess the impact of hate speech upon the dynamics of the discussions, we have used Granger tests and assessment of discussion histograms; we have also assessed the selected groups of posts in an exploratory manner. Our findings demonstrate that communicative aggression helps to express immediate support and solidarity. It also contextualizes the criticism towards both the authorities and regime challengers, as well as demarcates the counter-public.

Keywords
communicative aggression; hate speech; networked discussions; obscene speech; political protest; Russia; verbal aggression; YouTube

Issue
This article is part of the issue "Dark Participation in Online Communication: The World of the Wicked Web" edited by Thorsten Quandt (University of Münster, Germany).

1. Introduction

In contemporary networked discussions, aggressive verbal behavior is a widespread phenomenon. The initial optimism about the democratic potential of online communicative milieus as new deliberative spaces (Diakopoulos & Naaman, 2011) was replaced in the 2010s by a pessimistic perception of them as dominated by trivia (Fuchs, 2017) as well as incivility, false information, and hate speech, conceptualized as 'dark participation' (Quandt, 2018). Aggressive speech has, for years, been almost exclusively considered a negative phenomenon worth detecting and filtering out, as aggression challenges the argumentative integrity of online

discussions (Vollhardt, Coutin, Staub, Weiss, & Deflander, 2007) and their normatively understood democratic quality (Cortese, 2006). In this capacity, it is simultaneously a digital threat of a non-political nature (Salter, Kuehn, Berentson-Shaw, & Elliott, 2019) and a threat to a rational and politically relevant public sphere (Miller & Vaccari, 2020; Pfetsch, 2018).

So far, this view upon aggressive speech has been challenged from several viewpoints. Thus, many works have addressed the dilemma of 'free speech vs. hate speech' (Hare & Weinstein, 2010; Howard, 2017; Weinstein, 2017). This well-known debate, including how hate speech and its being banned from public use relates to equality, autonomy, and legitimacy, has been reconstructed by Massaro (1990) and Waldron (2012). As a rule, bans on hate speech may be found in laws that prevent group hate and promote inter-group tolerance (Waldron, 2012, p. 8). However, this line of debate does not, in effect, challenge the understanding of aggressive speech as a threat, and only a threat, to democratic discussion; here, only the boundaries of what may be banned are debated. Also, it does not address the issue of politically motivated offensive language, as, in democracies, political groups are not considered disadvantaged minorities.

Only a few studies have so far tackled the issue of aggressive content as a form of individual/group empowerment or discussion fuel, while linguistic literature on offensive speech points to its positive functions for the speaker, such as release from tension or marking group belonging. Burns (2008, p. 61) has stated that "this type of linguistic behavior reflects and supports both the successful functioning of societies and individuals." This might be especially true for restrictive political and legal environments where various types of public offense are prohibited by law and the political environment does not give much space for voicing discontent. In these contexts, dark participation could be a way to voice political dissent and rebel against the hegemonic discourses of the public sphere.

Our article aims at exploring the roles of aggressive language in political discussions in a so-far heavily underresearched context of countries with no sustainable democratic tradition. For such a study, today's Russia represents a nearly perfect case. First, Russian society and Russian public communication of the 2010s have been fundamentally fragmented and increasingly polarized (Bodrunova & Litvinenko, 2015). Scholars describe a post-perestroika values-based division of the nation into a dominant traditionalist majority and an outlier minority of a mostly liberal-oppositional stance (Berezuev & Zvonareva, 2019), with a high level of mutual hostility. This allows for the exploration of the phenomenon of politically motivated hate speech. Second, the recent tightening of the political regime has, inter alia, brought along new bans on extremist speech, public swearing, and offending civil servants, thus giving offensive language the new connotation of an act of political disobedience. Third, in addition to hate lexicons, Russian has a highly developed obscene sublanguage (*mat*), the functions of which go far beyond just expression of aggression, and which is even considered an "exceptionally rich" "linguistic system in its own right" and "a special genre of folk-art" (Dreizin & Priestly, 1982, p. 233–234).

Fourth, the Russian-speaking segment of Internet, or Runet, grew intensely and remained relatively unregulated for quite a long time before the mid-2010s (Vendil Pallin, 2017). By 2020, Internet penetration in Russia has reached 79% and was expected to exceed 100 million citizens by the end of 2020 (data by Krivoshapko, 2020), with 70.5% of people using mobile Internet (Elagina, 2020). On the one hand, since the mid-2000s, habits of discussions free from any bans have formed in Runet (Bodrunova, in press). To this day, Runet talk remains largely unbound by legal limitations, and the aforementioned restrictions introduced in the 2010s started to have their impact only very recently. For over two decades, Runet has served as a constellation of arenas in the online public sphere 'parallel' to the offline media landscape dominated by pro-state and loyal actors (Kiriya, 2014), including arenas with alternative agendas (Bodrunova & Litvinenko, 2015) and large publics critical to leadership (Toepfl, 2020). On the other hand, Runet is known for its platform-wide echo chambering where, for example, Facebook is recognized as a liberal-oppositional filter bubble (Bodrunova & Litvinenko, 2015), while, in some Twitter discussions, nationalist discourses are dominant (Bodrunova, Blekanov, Smoliarova, & Litvinenko, 2019). Scholars also documented radicalization of Russian-language online speech quite early (Salimovsky & Ermakova, 2011).

Most studies of aggressive verbal behavior on Runet have focused on how to conceptualize, detect, and filter it out from ongoing or past online discussions (Koltsova, 2019). However, politicization of public swearing combined with wide de-tabooing of *mat* by cultural communities, the rural populace, and younger generations calls for reassessment of the possible roles of aggressive speech in political discussions. In particular, we ask what roles different types of aggressive language play in Runet political discussions.

For our enquiry, we use the case of Moscow protests against non-admission of oppositional and independent candidates to the elections of the Moscow city parliament of July to September 2019. We have analyzed 77,847 comments under the most viewed YouTube videos on this topic. We have assessed the roles of various types of aggressive content in the dynamics of the discussion by quantitative and qualitative instruments, including Granger tests and interpretation of discussion histograms.

The remainder of the article is structured as follows. In Section 2, we reconstruct the relevant theories and contextual knowledge on YouTube discussions in Russia. In Section 3, we pose the research questions and hypotheses. In Section 4, we describe data collection, the sample, and research steps. Section 5 provides the results and discussion, including the discovered roles

of aggressive content. In the final section, we place the results in the wider context of the dark participation and communicative aggression studies.

2. Aggressive Speech in Runet Discussions: Theory and Context

2.1. Communicative Aggression in Networked Discussions: Only a Threat?

Scholars, lawyers, and public institutions have produced a myriad of definitions for verbal aggression, hate speech, and other illegitimate ways of expression (Brown, 2017a, 2017b). Many legal definitions, though, are narrowed down by listing particular social groups vulnerable to verbal hatred or aggressive beliefs, like in case of anti-Semitism (see, for example, Council of Europe, 1997, p. 107). As mentioned above, they do not include groups defined by political views. However, the rise of authoritarianism and polarization in many countries demands an extension of the understanding of aggressive speech online.

In our study, we use a wide definition of aggressive speech by Cohen-Almagor (2011, p. 1): It is "bias-motivated, hostile, malicious speech aimed at a person or a group of people because of some of their actual or perceived innate characteristics," which might include political groups. Here, we use the terms '[acts of] verbal aggression' and 'aggressive speech' interchangeably, although we well realize that aggressive speech is a manifestation of aggression. In order to tailor this definition to the focus of our research, we use two other approaches that frame our understanding of aggressive speech.

First, for a formal definition of types of aggressive communication, we use the concept of 'communicative aggression,' as elaborated by Sidorov (2018). This umbrella concept allows for systematizing various forms of aggression in mediated communication. Following Sidorov's logic, we argue that distinguishable pragmatic types of communicative aggression, such as cyberbullying, virtual racism, political hate speech, or swearing, link a certain speaker's goal to a certain lexicon. Below, we operationalize the types of communicative aggression relevant for conflictual political discussions. Second, Parekh (2012), as cited and commented upon in Howard (2017), provides criteria for differentiating communicative aggression from political opinion, often also sharp and provocative. Parekh (2012, p. 41) has noted that hate speech "stigmatizes the target group by implicitly or explicitly ascribing to it qualities widely regarded as undesirable" and objectifies this group as a legitimate object of hostility. This approach allows us to select, within types of aggression, the lexical conglomerates linked to undesirable objects, events, or features. This approach also allows for distinction between verbal aggression and expressions of anger and other negative emotions. Undoubtedly, acts of verbal aggression are often used to express anger and hatred, but not always: For instance, an obscene lexicon that substitutes normal speech in affective circumstances could express nearly any emotion, from disappointment to puzzlement to even joy. Our focus, thus, is not on emotions but on the speech used for humiliation and offence.

The effects of hate speech on political discussions have, so far, been almost exclusively assessed negatively (Van Aken, Risch, Krestel, & Löser, 2018). However, in certain contexts, aggressive content might help minorities voice political dissent not heard otherwise. Before the era of social networks, Delgado and Stefancic (1995) claimed that, among students, expressions of hatred may spur substantial on-campus debates on social discrimination. Davidson, Warmsley, Macy, and Weber (2017) have pointed to the possible use of potentially offensive language in a positive sense in groups that face discrimination, such as the LGBTQ community.

With the rise of social networking platforms, the issue of communicative aggression within them has become sharp to an extent unprecedented offline, where speech is not usually anonymized and detached from the speaker. The freedom of 'dark participation' has become part of a wider growth of dissonant public spheres (Pfetsch, 2018) where users neither seek consensus nor limit themselves by norms of public speech. The structure of such discussions has already been conceptualized as affective (Papacharissi, 2015; that is, hardly rational or reflexive, mostly highly emotional and quick to react) and ad hoc (Bruns & Burgess, 2011), which may lead to quick dissipation after the trigger event is over and does not allow the commenting evolve into a meaningful discussion. However, it is exactly this state of the online discourse that calls for rethinking of the roles of aggressive content within it. As stated above, it may allow users to shape a wide variety of thoughts and feelings expressed in rational discussions in another manner.

Aggressive content might also, presumably, influence discussion dynamics. Thus, Platonov and Svetlov (2020) have found that negative posts on the social network VK.com provoke a larger number of comments than neutral and positive ones. Our earlier works have shown that anger and aggression in tweets might be related to discussion intensity and pivotal turns in its topicality (Bodrunova et al., 2020; Smoliarova, Bodrunova, Yakunin, Blekanov, & Maksimov, 2019).

Given this, we might expect that aggressive speech may play multiple (positive and negative) roles in online discussions, for both its dynamics and content.

2.2. Aggressive Speech on Russian YouTube and the Moscow Protests of 2019

2.2.1. Aggressive Language on Runet in the 2010s

Reassessment of aggression online might be especially relevant for political cultures like Russia's. As stated above, relatively liberal regulation of online communication in the 2000s (Vendil Pallin, 2017) left room for both

free political discussion and its radicalization. In Russia, the rise of online radical, extremist, and nationalist speech became a scholarly concern much earlier than in most democratic countries (Etling et al., 2010; Kronhaus, 2012; Salimovsky & Ermakova, 2011). Widespread de-tabooing of obscene language contributed to this process. After the protests of 2011–2012 and the Ukrainian crisis, a range of legal restrictions were introduced (Litvinenko & Toepfl, 2019), and hundreds of legal cases against online posting opened (Gabdulhakov, 2020). Thus, in 2014, swear words were banned from use in media and the arts (Federal Law, 2013; enacted in 2014). In 2016, the so-called 'Yarovaya law package' provided formal grounds for recognition of online speech as extremist (Federal Law, 2016a, 2016b). In 2019, two other laws expanded the ban of 'disrespectful' statements about representatives of state power to include online space (Federal Law, 2019a, 2019b). These laws scaled up the range of instruments that could be used by the elites to curb dissent online (Litvinenko & Toepfl, 2019). Despite this, the online discourse has largely remained free of taboos.

In the 2010s, most English-language studies of aggressive speech on Runet were focused on inter-ethnic hostility (Bodrunova, Koltsova, Koltcov, & Nikolenko, 2017; Koltsova, 2019) and online stigmatization practices (Dudina, Judina, & Platonov, 2019). Several critical studies have linked aggressive speech to activity of pro-governmental trolls (Zvereva, 2020). In rare Russian-language studies of verbal aggression online, authors have mostly raised the issue of "degraded online talk" (Salimovsky & Ermakova, 2011, p. 74), without empirically testing its possible functions in online discussions. The linguo-pragmatic functions of obscene language have been explored by linguists (up to 27 functions; Havryliv, 2017; Zhel'vis, 1997). However, in none of these studies has communicative aggression been linked to discussion dynamics or freedom of expression in a restrictive media environment.

2.2.2. *Mat* in Russian Culture and in Online Speech

In this context, the Russian swearing lexicon deserves special attention: "Russia has an incredibly rich and versatile swear sublanguage, called *mat* [emphasis added], which is based on four key stems" (Pilkington, 2014; see also Pluzer-Sarno, 2000), with several more stems equally tabooed in public speech. Due to its inflective nature, Russian provides these stems diverse opportunities to perform the functions of nouns, verbs, adjectives, adverbs, and even interjections via the use of prefixes and suffixes. As *mat* has undergone the strictest tabooing within the Russian lexicon, we can expect that it bears the highest level of aggression and, if used, changes the discussion fabric. Similar to other languages, Russian *mat* had experienced de-tabooing long before social networks were in place. In today's oral and online speech, swearing performs a variety of constructive pragmatic functions, such as an increase of emotionality; release from psychological tension, demonstration of relaxedness and independence of the speaker; demonstration of disregard to restrictions; marking in-group belonging, etc. (Kosov, 2011, p. 37).

In today's online speech, *mat* unites with politically motivated hate speech, another under-explored type of aggressive language. Their combination produces hybrid pejorative neologisms directed at both political camps: For example, towards the liberal-oppositional camp, *майданутый* (*maidanuty*, or, in English, 'f***ed in head by the Ukrainian Maidan revolution') and *либерast* (*liberast*, that is, 'liberal pederast') are used, while *кремлядь* (*kremlyad*, or 'a prostitute from Kremlin') and *пропагандон* (*propagandon*, or 'propagandistic condom') are used against the authorities and loyalist media. These very neologisms are a sign of a politicization of the obscene lexicon and mark the importance of both obscene language and politically motivated hate speech for online political discussions.

2.2.3. YouTube as an Alternative to TV: A Crossroads of Polarized Opinions

We have chosen YouTube for our research because, in the recent years, YouTube has moved to the front of political communication in Russia. Since 2017, It has been the third most popular Runet website, according to Google Russia, with a monthly reach of 26% of Russians (Polyakova, 2017). For Russians, YouTube has gradually become an alternative to the state-dominated TV channels (Litvinenko, in press), with politics being one of the popular topics.

A recent study of political discourse on Russian YouTube during the 2018 presidential campaign has shown that the 'Popular' section, featuring top Russian YouTube videos, was dominated by oppositional actors (Litvinenko, in press). The leading oppositional channels, like Navalny's Alexey Navalny (with 3,94 million subscribers) and Navalny LIVE (1,89 million subscribers); pro-liberal independent media/journalist channels, like VDud' by Yury Dud' (7,8 million), Alexey Pivovarov's Redaktsia (1,4 million), and TV Rain (1,37 million); human-interest news vloggers like Roman Usachev (2,36 million); and critical political vloggers like Anatoly Shariy (2,38 million) and kamikazedead (an estimated 1,5 million of subscribers, as the actual figures are hidden) are comparable in viewership to the main state-affiliated federal channels, like NTV (9,56 million), Pervy kanal (4,95 million), and Rossiya 1 (3,62 million) and outperform many channels of national entertainment TV. Navalny's anti-corruption investigation published on YouTube in 2017 'Don't call him Dimon' (Navalny, 2017) gathered more than 37,5 million views (as of January 2021) and triggered nationwide street protests (Gorbachev, 2017).

Another advantage of YouTube is that, unlike Facebook or VK.com, this platform has not yet been the major focus of legal action against political activists. Also,

Russian YouTube culture has been influenced by highly popular rap battles where obscene and offensive speech are core. As a result, Russian YouTube of the late 2010s has grown into a freely speaking opinion crossroads with a predomination of oppositional agenda and audience; this constituted a suitable case for our research.

For our analysis, we have taken the case of the Moscow protests of summer 2019, a vivid example of polarization of Russian public communication. Non-admission of oppositional and independent candidates on ballots for the Moscow city parliament elections triggered peaceful demonstrations in July 2019. They were suppressed by police, which led to an escalation of the conflict throughout August and September 2019. During those months, videos featuring the riots were widely circulated on YouTube and attracted a large number of comments, politically split and manifestly aggressive.

3. Research Questions

In this article, we ask whether various types of communicative aggression play active roles in both the dynamics and content of online political discussions. Here, in general, our approach is of an exploratory nature; we orient to the multiple roles of obscene language but explore whether other types of aggressive content may play similar roles. We follow the argument by Thelwall (2018), who has argued for the use of a mixed-method and exploration-oriented approach for studies of YouTube comments.

In the following research question, we ask whether various types of communicative aggression can spur discussions, what are the patterns of their appearance in the course of discussions, and whether users tend to radicalize individually and in discussion micro-clusters:

> RQ1: Does communicative aggression affect discussion dynamics?

Moreover, in the next research question, we ask whether communicative aggression may be linked to democratic functions of the public sphere, including fostering both cross-group dialogue and counter-publicity. We also have in mind the above-mentioned functions of aggressive speech, where one can distinguish between psychological functions (like individual release of tension), social-psychological functions (like marking belonging), and political functions (like struggle for privilege or power):

> RQ2: What roles do various types of communicative aggression play in political discussions online?

4. Method and Sampling

4.1. Data Collection

For our analysis, we have chosen the comments under the most popular YouTube videos about the Moscow protests of 2019. To form a sample, we tried three strategies: assessing the 'Popular' section, examining the popular political accounts, and searching by protest-related keywords. As a result, we decided to focus not on accounts but on individual videos, and have selected the 15 most popular videos in the search results for 'Moscow protests' that reached over 100,000 views uploaded to YouTube between July 27 (the start of the active phase of the street protests) and September 8 (the election day). Of those, two videos did not allow for comment upload: *Radio Liberty* disabled comments, and comments for the video by the oppositional news outlet *Current Time* were automatically blocked for download. Thus, we crawled and downloaded the comments for 13 videos from ten accounts, which resulted in 77,847 comments altogether, the number of comments per video ranging from 538 to more than 42,400. This collection allowed us to gather the most intense discussion fragments on the protests within Russian social media and trace its dynamics over time.

4.2. Methods of Data Processing and Research Steps

4.2.1. (Re)Conceptualizing Aggressive Speech for Conflictual Political Discussions

Following what was stated above, we first needed to conceptualize the types of aggression and the lexical conglomerates within them for automated analysis. A preliminary reading of over 3,000 posts in the dataset led us to rejecting the usual types of aggression, including cyberbullying, inter-ethnic hostility, and homophobia, as they were not the focus of the discussion. We have formulated four types of communicative aggression, all politically relevant and tabooed/prohibited by law: 1) Humiliation, including politically motivated hate speech and discriminative expressions—this type is partly directed at authorities and police and, thus, might fall under the law on offence of civil servants; 2) radical political claims (similar to 'Carthage must be destroyed'), including calls for aggressive action against individuals or groups—this is prohibited, as it may be considered extremism or a call to overthrow the existing political system; and 3) obscene language (*mat* and equally tabooed lexemes), which is tabooed and prohibited by the law against public swearing.

Then, we have further narrowed down our research via assessing the possibilities for automated analysis. We have found that, in over the 3,000 posts of the preliminary study, the number of radical calls for action amounted to only several dozen and were very diverse in lexical and grammatical terms; this was not enough even for deep learning and was only suitable for manual analysis. Thus, politically motivated hate speech and obscene language became our major focus.

4.2.2. Selection of the Lexicon for Automated Analysis

Out of the preliminary analysis, the following lexical clusters were manually detected in obscene speech:

mat with case endings; *mat* with flexions (nouns, pronouns, verbs, adjectives, adverbs, etc.); lexemes tabooed equally to *mat*, with flexions; *mat* altered not to break the law (using *, @, '.,' '$,' etc.); euphemisms of *mat*; and a lexicon of the lower body and defecation, which is less tabooed.

The following lexical clusters were manually detected in politically motivated hate speech inside anti-state discourse, pointing to: the police, in comparisons to garbage, lapdogs of the regime, cosmonauts (due to the uniforms), and the KGB; institutions, mostly the 'United Russia' party and the State Duma; the regime, by comparing its rule to that of the Nazis, to the Tsarist gendarmerie, and to organized crime (e.g., mafia); the state-affiliated media as 'purchasable' or bot-like; and individual politicians and their visual appearance.

The following lexical clusters were manually detected in politically motivated hate speech inside anti-opposition discourse: pejoratives of 'liberal' and 'democracy' made via flexions and stem combinations; 'fifth column,' the insider betrayers; diminutives and pejoratives formed out of the names of oppositional leaders Alexey Navalny and Luybov Sobol; linking the opposition to the USA, its grant funding, and philanthropists like George Soros; the Ukrainian pro-European Maidan revolution of 2013; LGBTQ and *gayropa* ('gay-favoring Europe'); and slacktivism.

We also detected offensive lexicon addressed to both political camps: markers of stupidity and ugliness, and prison slang; and to other groups: ethnic pejoratives.

Pre-tests on detection of aggression showed that the two latter groups were minor and did not play any significant role in the fabric of the discussion, while the three former groups mattered. Thus, the last two groups were excluded from further analysis.

The overall number of stems submitted to the stem detector amounted to 286. Also, 216 of them potentially contained flexions and case endings; thus, for the three types of aggressive speech, the overall expected number of lexical units was over 500.

4.2.3. Addressing the RQs

To assess the role of aggression in discussion dynamics, we created the scripts to automatically detect the selected lexicons in the dataset and marked the comments as neutral, obscene, anti-state, or anti-opposition. The cases where the lexicon originally directed against the state or the opposition was used to form the opposite meaning (e.g., sarcastic) were fewer than 2% for all the videos; the comments where both anti-state and anti-opposition lexicons were found were even fewer, which shows that these lexicons are, indeed, markers of commenters' distinct political positioning.

Then, we conducted Granger tests juxtaposing the overall number of posts and the number of neutral posts with the number of the following types of posts: aggressive, obscene, politically aggressive, anti-state, and anti-opposition, where 'politically aggressive' was anti-state and anti-opposition combined. As the discussion was intense for around four days in all the cases, three-day and four-day periods were selected for testing. We tested one-hour, two-hour, and six-hour increments; in addition, 15-minute and 30-minute increments were tested for the most commented-on videos. Altogether, 34 tests were performed. For the Granger tests, the comments that contained both anti-state and anti-opposition lexicons were excluded from analysis.

To assess individual vs. micro-group patterns of presence of aggressive speech, we decided to go for linear and interval histograms. While, for the Granger tests, the posts were singled out independently, for the histograms, we first singled out the posts with obscene language, then those with anti-state claims, and then those with anti-opposition claims; if the marking overlapped, the post was marked preferentially red, then yellow, then green. The number of overlapping posts, again, never grew over 3% of all aggressive posts. Hourly steps were introduced to demonstrate the discussion dynamics. Axis *x* ranged the users by time of their first post; axis *y* ranged the discussion in time. This allowed for assessing individual and micro-group user speech in time.

RQ2 was addressed in a grounded-theory manner. We did not pre-suppose any roles in advance, but were orienting to the lists of possible functions of obscene and aggressive speech mentioned above and our results from RQ1.

5. Results

5.1. RQ1

The videos that happened to be in our collection belonged to three types: foreign media in Russian (*BBC News*, *Euronews*, *DW*, and a Ukrainian correspondent in Moscow), Russian oppositional public political figures also active in media/online (Alexey Navalny, Maxim Shevchenko), and activist channels focused mainly on independent news production (*My Protest*, *Mordor Channel*, and *Superpower News*). As stated above, we have looked at whether aggressive posts fueled the subsequent discussion—or whether aggression just grew when the discussion itself grew, for which we used Granger tests. The results are presented in Figures 1 and 2 (two-hour increment). They show that, in 12 cases out of 13, the dynamics of the discussion was spurred by aggressive speech, and in 8 cases at least to the medium level. In at least five cases out of 13, this was not reciprocated—that is, if aggression grew or fell, the number of subsequent posts grew or fell accordingly. The videos showed different patterns of interaction between aggression and the overall discussion, but, in any case, we can make several conclusions.

Thus, both obscene and political aggression can play the role of discussion fuel; but, when one does it, the other is rarely involved. How aggression works seems

Videos	The overall number of posts in the discussion vs. communicative aggression									
	Direct effect					Reverse effect				
	All / aggressive	All / obscene	All / political	All / anti-state	All / anti-opposition	Aggressive / all	Obscene / all	Political / all	Anti-state / all	Anti-opposition / all
video_1	V STRONG	V STRONG	V STRONG	V STRONG	No	STRONG	No	MEDIUM	MEDIUM	No
video_2	No	No	No	No	No	MEDIUM	STRONG	STRONG	STRONG	V STRONG
video_3	No	WEAK	No	No	No	No	STRONG	No	No	No
video_4	No	No	No	WEAK	n/a	No	WEAK	No	No	n/a
video_5	No	No	No	No	n/a	No	WEAK	No	No	n/a
video_6	WEAK	No	No	No	No	V STRONG	MEDIUM	No	No	STRONG
video_7	WEAK	No	MEDIUM	MEDIUM	n/a	No	No	No	No	n/a
video_8	MEDIUM	MEDIUM	No	No	n/a	No	No	No	WEAK	n/a
video_9	No	No	No	STRONG	No	WEAK	WEAK	MEDIUM	WEAK	WEAK
video_10	No	No	No	No	No	MEDIUM	STRONG	STRONG	STRONG	STRONG
video_11	No	WEAK	No	n/a	n/a	No	No	STRONG	n/a	n/a
video_12	No	STRONG	No	No	n/a	No	No	WEAK	WEAK	n/a
video_13	MEDIUM	No	WEAK	n/a	n/a	V STRONG	No	V STRONG	n/a	n/a

Videos	The number of non-aggressive posts in the discussion vs. communicative aggression									
	Direct effect					Reverse effect				
	All / aggressive	All / obscene	All / political	All / anti-state	All / anti-opposition	Aggressive / all	Obscene / all	Political / all	Anti-state / all	Anti-opposition / all
video_1	V STRONG	V STRONG	V STRONG	V STRONG	No	STRONG	No	MEDIUM	MEDIUM	No
video_2	No	No	No	No	No	MEDIUM	STRONG	STRONG	STRONG	V STRONG
video_3	No	WEAK	No	No	No	No	STRONG	No	No	No
video_4	WEAK	No	No	WEAK	n/a	No	WEAK	No	No	n/a
video_5	No	No	WEAK	WEAK	n/a	No	WEAK	No	No	n/a
video_6	No	No	No	No	No	WEAK	MEDIUM	No	No	STRONG
video_7	WEAK	No	MEDIUM	STRONG	n/a	No	No	No	No	n/a
video_8	MEDIUM	MEDIUM	No	No	n/a	No	No	No	WEAK	n/a
video_9	WEAK	No	WEAK	STRONG	No	WEAK	WEAK	MEDIUM	WEAK	WEAK
video_10	No	No	No	No	No	MEDIUM	STRONG	STRONG	STRONG	STRONG
video_11	No	WEAK	No	n/a	n/a	No	No	No	n/a	n/a
video_12	No	MEDIUM	No	WEAK	n/a	No	No	WEAK	WEAK	n/a
video_13	WEAK	No	No	n/a	n/a	No	No	V STRONG	n/a	n/a

Figure 1. The results of the Granger test (two-hour increment). Notes: 'Direct effect' appears when the number of aggressive posts grows due to the overall growth of a discussion; 'reverse effect' appears when a discussion grows due to growth of the number of aggressive posts. 1) Blue: direct effect; orange: reverse effect; grey: data too scarce for the test. 2) Weak: $F < 10$, $0.01 < p \leq 0.05$; medium: $p = 0.001 \leq p \leq 0.01$; strong: $F > 10$, $0.0001 \leq p < 0.001$; very strong: $F > 10$, $1e\text{-}12 < p < 0.0001$.

to depend on two dimensions: author type and genre. Thus, aggression clearly fuels discussion under commentaries by political activists (even if the size of the discussion under Navalny's video makes the trend reciprocal), while under commentaries by foreign news media both effects appear in weak to medium state, and, under news and the interview, the effects are clearly much less sound. For political celebrities, discussion fuels aggression (again, reciprocated under Navalvy's video)—that is, the volume of the discussion matters. The smaller-sample cases show that obscene language might be provocative and bring on politically harsh speech, while, in bigger discussions, this effect is overcome. Due to scarcity, anti-liberal talk did not have an impact upon the discussions, but, overall, aggression played a significant role in how half of the discussions developed.

However, somewhat surprisingly, it is doubtful that one can call the comment conglomerates 'discussions.'

Video Metadata									The overall number of posts					Non-aggressive posts				
#	Video # in the dataset	Days assessed	Upload date	Overall Nposts	Channel type	Channel name	Content type	Topic	Aggressive	Obscene	Political	Anti-state	Anti-opposition	Aggressive	Obscene	Political	Anti-state	Anti-opposition
1	video_7	4	August 3	5037	Personal account of a public politician	Maksim Shevchenko	Comment	The Moscow protests. A leftist view										
2	video_8	4	July 28	592	Foreign media	Euronews in Russian	News	Moscow after the protests										
3	video_11	3	August 3	538	Foreign media: Ukrainian journalist	Roman Tsymbalyuk	Comment	Streaming on the Moscow protests										
4	video_12	3	August 10	929	Foreign media	DW in Russian	News	Reporting on protests and arrests										
5	video_1	4	July 30	42511	Personal account of a public politician	Navalny LIVE	Comment	The Moscow protests. What next?										
6	video_13	3	July 30	1099	Foreign media	BBC News Russian	Interview	Interview with victims hurt at the protests										
8	video_3	4	July 27	2028	Foreign media	BBC News Russian	News	Mass arrests at the Moscow protests										
7	video_6	4	August 5	7312	Activist oppositional / 'News&entertainment'	Moy protest / My protest	Comment	Commenting on police action										
9	video_2	4	August 31	8901	Activist oppositional / 'Independent news'	Kanal Mordor / Mordor Channel	Comment	Call for violent action against police										
10	video_10	4	July 28	2756	Activist oppositional / Road police watchdog	Dvizhenie / Traffic	Comment	Criticism on the Moscow authorities										
11	video_9	4	August 3	1078	Foreign media	BBC News Russian	Comment	The Moscow protests. What next?										
12	video_4	4	July 29	1455	Foreign media	BBC News Russian	Comment	Consequences of the Moscow protest										
13	video_5	4	August 21	4161	Activist oppositional / 'Independent news'	Novosti Sverhderzhavy / Superpower News	News	Putin compares the protests to jillets jauns										

Figure 2. The mixed effects of communicative aggression on the dynamics of the overall number of posts and the number of neutral posts (two-hour increment), accompanied by video metadata. Note: Blue—direct effect; green—reciprocal (both direct and reverse) effect; orange—reverse effect; grey—data too scarce for the test.

In the aforementioned histograms (Figures 3 and 4), we reproduced the commenting patterns by showing each user's posts in time, to see if they commented multiple times, and the intensity of the discussion by hourly intervals. The histograms clearly point to the affective and ad hoc nature of commenting, as, for all the discussions except Navalny's (Figure 3), repeated commenting is very rare (Figure 4). And, despite that, the fabric of the discussions is imbued with aggression, regardless of legal prohibitions; moreover, it rarely looks like an exchange of offensive comments, but rather like expressive 'shouts to the air' that condemn one of the political sides. We have also spotted micro-outbursts of aggression in smaller cases (Figure 4, circled); in big discussions, though, when individual users post multiple times (Figure 3, circled), they are likely to be aggressive, but when several users post simultaneously in a dynamic micro-cluster (Figure 3, squared), they, against expectations, abandon aggressive discourse. Thus, users mostly come to express to the author their aggressive support and solidarity in condemnation, and they rarely address their anger to fellow commenters; if engaged in further talking, they talk non-aggressively.

5.2. RQ2

Beyond immediate expression of support and solidarity, aggressive speech performs an array of functions that tie together the affective nature of the discussions and the outer context. The most evident function is, by nature, sociological, as use of politicized obscene speech and pejoratives fosters political identities of the polarized camps sharply demarcating 'us' from 'them' (Van Dijk, 1993). This is already evident from the anti-state and anti-opposition lexical conglomerates described above, where an obscene lexicon intertwines with political pejoratives to the extent of neologisms. Additionally, a 'normal' group is put in open opposition to a group marked as undesirable, in agreement with Parekh (2012):

> Cop mutts, they go against the people, shameful!!!
>
> Every normal person disrespects rogues, thieves, and organized criminal groups!
>
> Police are the best. One must nip the white-stripe coup d'état in the bud. Mercenary liberast beasts should be jailed!

Destructive to potential consensus between political antagonists, anti-state hate speech becomes a constructive means of counter-public consolidation. Interestingly, predominance of statements that favor oppositional discourse makes pro-state commenters, whose opinions overwise dominate in the country, take a defensive tone:

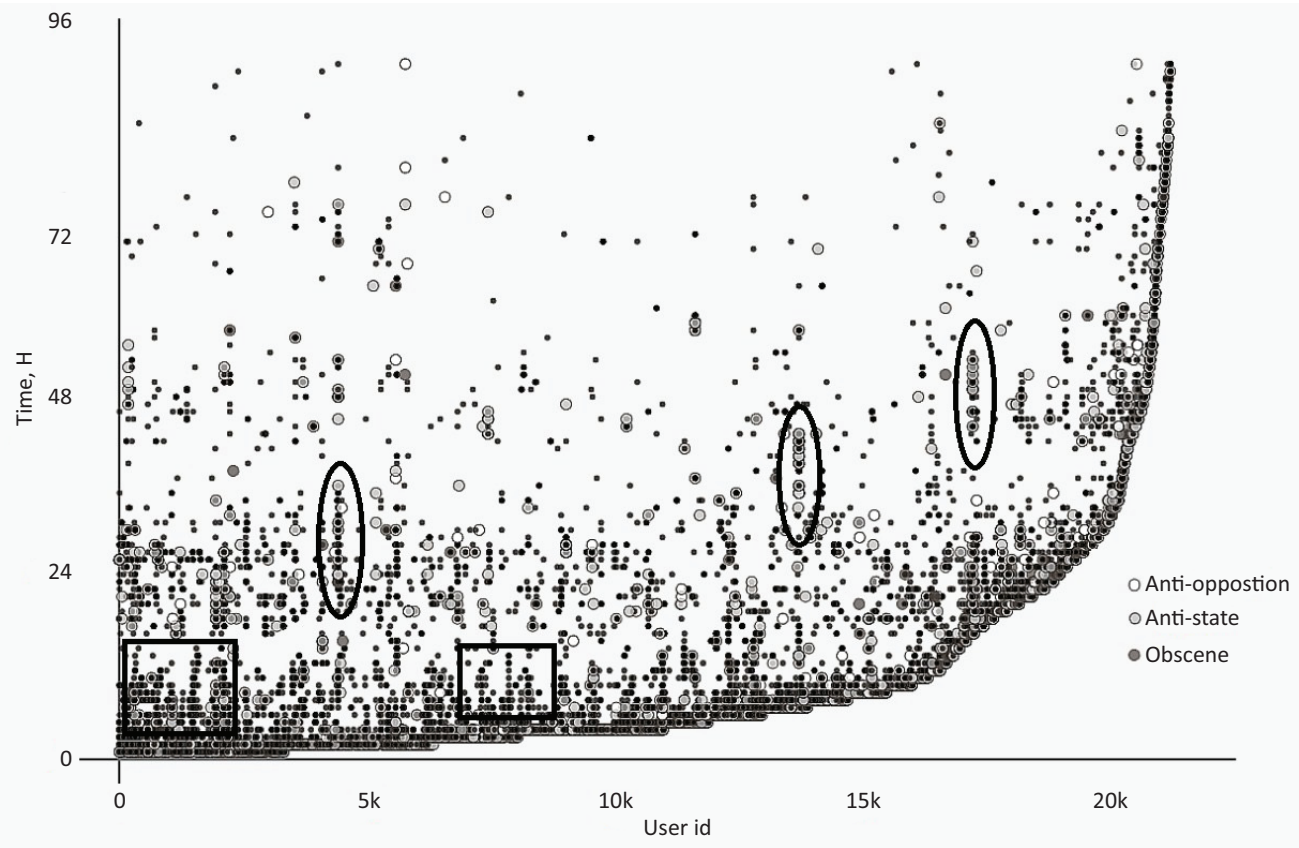


Figure 3. Interval histogram of the discussion, video 1: Individual aggressive vs. interactive neutral speech.

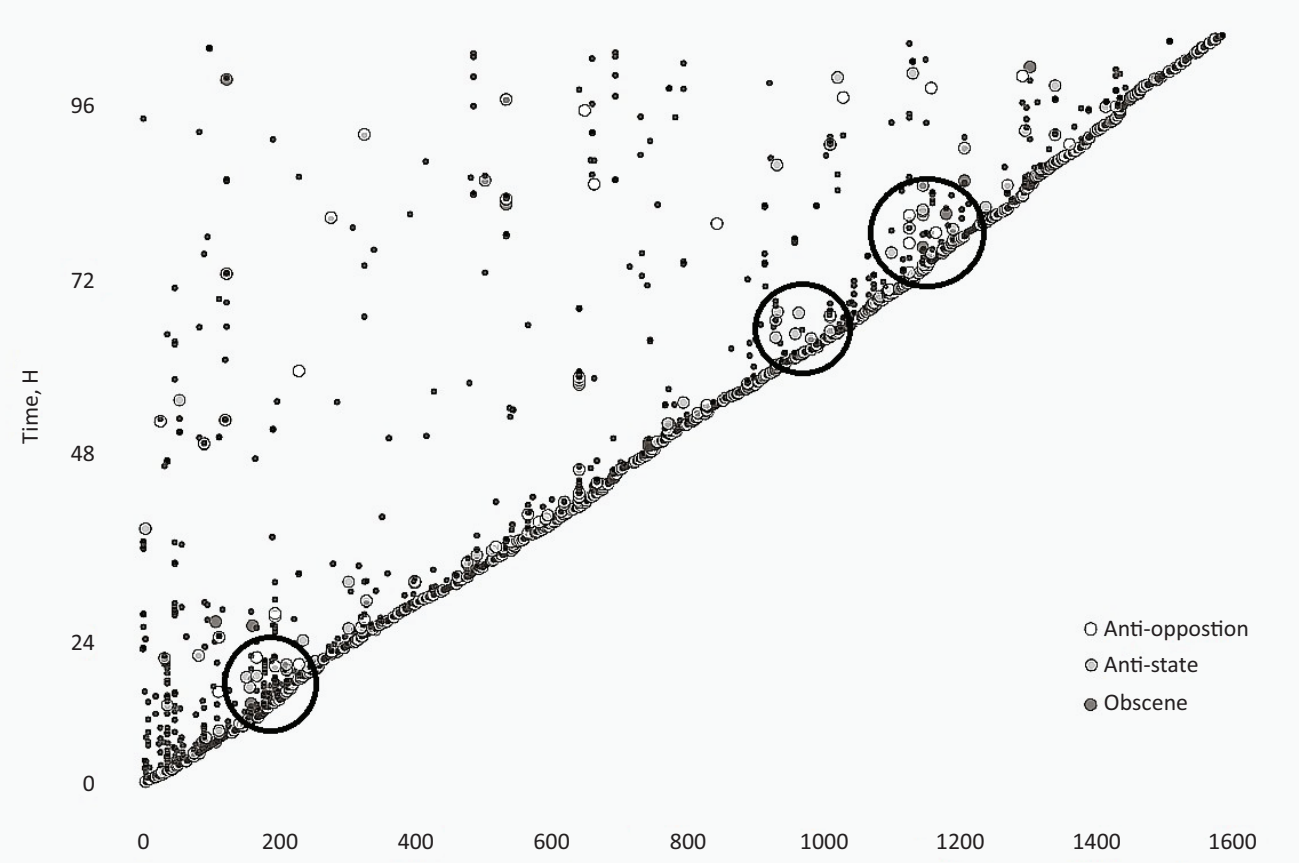


Figure 4. Linear histogram, video 13: Micro-clusters of aggression.

> [To Lyubov Sobol whose hunger strike is depicted in the video] You think you put on glasses and got smart? Stupid. Realize one thing: Russia will not lie under your curators, even if you crap your pants together with the [US] State Dept.! As for your rallies…in the same Europe people like you are killed with fly swatters….Who do you serve, evil forces?
>
> I am not going to prove anything. If you're idiot enough to be buying this whole show with Sobol's hunger strike, swollen Navalny, etc., it's up to you, but don't change flags after your idols leave for other countries.

In contrast, for the pro-opposition speakers, the very use of prohibited language may delineate a restriction-free realm. In political terms, aggressive speech challenges the hegemonic discourse (otherwise unchallengeable) and seeks to protect a way of thinking. And this, to an extent, is true for both political camps.

Via the use of particular metaphors in harsh lexicon, aggressive critique puts the current political regime in a row of national traumas from the 20th century, especially with World War II and the uncontrolled capitalism of the 'frantic 1990s,' united by the underlying reference to a dysfunctionality of the political scene captured by force:

> Occupation, Siege, Nazism.
>
> A 100% OCG [organized criminal group] runs the show. There are tops, and there are 'thugs.' Just the classic nineties but on a national scale!

Another metaphorical line points to the issue of freedom vs. excessive control:

> In a concentration camp, there are no rights.
>
> Democracy and freedom are the least of it. Russia has been captured by Nazis!!!
>
> Bandits in uniform. They do not protect the citizens; they enslave them.

Anti-liberal speech, in its turn, radically addresses fears of yet another system revolt: post-Cold War cleavages in values, deep suspicion towards American and European intentions, and post-imperial resentment. Here, post-Maidan Ukraine is often posed as an example of a country destroyed by a 'color revolution,' and, likewise, the 1990s are a symbol of poverty and instability:

> Go on rallying! [The USA] State Dept. will be grateful…1991 and 1993 forgotten? Moscow b****es! Trying to sell Russia again? That time [it was] for jeans and salami, and now what for?

> For a second Maidan! For the return to the 1990s! For hunger and destitution! Hurraaaaaaaa!!!!

While obscene language marks the emotion intensity, it is exactly the aggressive political talk that is responsible for contextualization—and, thus, for re-appropriation of the right of political interpretation for both camps. Thus, two types of aggression (obscene and political) support each other in user posts. Moreover, in the absence of sound public debate and chances for peaceful dissent, aggressive 'speaking out what is suppressed' serves as a way to vent anger. This, inter alia, can be beneficial for the regime, as users who vent their anger online may refrain from offline protest, as described by Toepfl (2020) in his conceptualization of risks and benefits of critical publics for autocracies.

Last but not least, we observed a high level of creativity in political use of tabooed language, from stem fusion to 'bleeping' of prohibited words. An opposition to the sterile and deliberate official language, it marks the grassroots, ingenuous, and censorship-free character of political discussions online.

6. Discussion and Conclusion

In this article, we have challenged the view of communicative aggression online as a necessarily negative means of disempowerment and de-legitimation. We have shown that under-studied types of communicative aggression, such as obscene language and politically motivated hate speech, may not only escalate conflict in public debate but, by contrast, also play constructive democratic roles for individuals, groups, and discursive processes in a restrictive political atmosphere. This includes expression of support and solidarity, counter-public consolidation, re-appropriation of rights to interpret historical context, and manifestation of creativity and wit that opposes restrictions and official discourses. Some of these functions are, of course, double-edged swords: Thus, consolidation of a counter-public leads to widening gaps between it and the wider public, and interpretation of context available for both pro- and counter-establishment groups may potentially lead to abuse of historical memory. However, such elements of public deliberation are in any case more characteristic of democratic contexts than of autocratic ones. This is why we claim that one can speak of 'constructive aggression' and differentiate it from destructive hate speech and an obscene lexicon if it plays constructive roles in public discussions; similarly, misinformation may be shared online with constructive motivations (Metzger, Flanagin, Mena, Jiang, & Wilson, 2021).

Our results have supported the claim that communicative aggression not only thrives within YouTube reactions to oppositional videos but also fuels user talk. We have spotted micro-outbursts of both aggression (by newcomers) and neutral talk (by discussants), while, in general, aggression was spread throughout the affective conglomerates of posts.

By analyzing the roles of aggressive content, we have seen that it may play a significant role in the formation of ad hoc counter-publics around a politically polarizing issue in online discussions, even when the patterns of user interaction reveal a disrupted public. It is politically motivated hate speech that, at least partly, forms the fabric of collective expression in comments. We have explored how the personal-level functions of aggressive speech, including *mat*, manifest themselves on the group level and gain political relevance. In the Russian case, communicative aggression is linked to giving voice to political opposition, which is overwise excluded from the mainstream discourse, and may foster counter-publics and offline action, as was the case during the Russian protests of 2011–2012 (Bodrunova & Litvinenko, 2015). This function of aggressive speech is in line with the strand of research on agonistic public spheres that emphasizes the importance of political conflict and political voices 'from the margins' for public deliberation (Dahlberg, 2007, p. 128). If, as in Russia, offensive language is officially banned in the media, using this kind of language per se might become a way to challenge the hegemonic official discourses.

The limitations of our research come from its exploratory nature, as well as from the limited number of videos around which the core discussion evolved. But even more, they stem from the nature of the discovered publics, as the lack of interaction between commenters prevents the use of, for example, social network analytics. Our findings support the idea of affective and dissonant publics, but partly re-interpret dissonance and 'dark participation' as democratically functional. They also point out to cumulative effects in online communication.

Our research shows that fighting aggressive speech on global social media platforms can, inter alia, give autocrats a tool to curb political dissent online. It can also mean depriving marginalized groups of the opportunity to vent their anger, which may lead, instead of to the expected harmonization of political communication, to an escalation of violence, both online and offline. One may ask how true this might be for democratic contexts; this would be intriguing to explore in future research.

Acknowledgments

This research has been supported in full by Russian Science Foundation, grant 16-18-10125-P.

Conflict of Interests

The authors declare no conflict of interests.

References

Berezuev, E. A., & Zvonareva, L. V. (2019). Quasi-modernization in today's Russia as a confrontation of values of modernism and traditionalism. *Manuscript*, *12*(1), 93–98.

Bodrunova, S. S. (in press). Information disorder practices in/by contemporary Russia. In H. Tumber & S. Waisbord (Eds.), *The Routledge companion to media misinformation and populism* (pp. 279–289). London: Routledge.

Bodrunova, S. S., Blekanov, I., Smoliarova, A., & Litvinenko, A. (2019). Beyond left and right: Real-world political polarization in Twitter discussions on inter-ethnic conflicts. *Media and Communication*, *7*(3), 119–132.

Bodrunova, S. S., Koltsova, O., Koltcov, S., & Nikolenko, S. (2017). Who's bad? Attitudes toward resettlers from the post-Soviet south versus other nations in the Russian blogosphere. *International Journal of Communication*, *11*, 3242–3264.

Bodrunova, S. S., & Litvinenko, A. A. (2015). Four Russias in communication: Fragmentation of the Russian public sphere in the 2010s. In B. Dobek-Ostrowska & M. Glowacki (Eds.), *Democracy and media in Central and Eastern Europe* (pp. 63–79). Bern: Peter Lang.

Bodrunova, S. S., Nigmatullina, K., Blekanov, I. S., Smoliarova, A., Zhuravleva, N., & Danilova, Y. (2020). When emotions grow: Cross-cultural differences in the role of emotions in the dynamics of conflictual discussions on social media. In G. Meiselwitz (Ed.), *Proceedings of the 12th international conference SCSM 2020* (Vol. 12194, pp. 433–441). Cham: Springer.

Brown, A. (2017a). What is hate speech? Part 1: The myth of hate. *Law and Philosophy*, *36*(4), 419–468.

Brown, A. (2017b). What is hate speech? Part 2: Family resemblances. *Law and Philosophy*, *36*(5), 561–613.

Bruns, A., & Burgess, J. E. (2011). *The use of Twitter hashtags in the formation of* ad hoc *publics.* Paper presented at the European Consortium for Political Research (ECPR) General Conference 2011, Reykjavík, Iceland.

Burns, M. C. (2008). Why we swear: The functions of offensive language. *Monash University Linguistics Papers*, *6*(1), 61–69.

Cohen-Almagor, R. (2011). Fighting hate and bigotry on the Internet. *Policy & Internet*, *3*(3), 1–26.

Cortese, A. J. P. (2006). *Opposing hate speech.* Westport, CT: Greenwood Publishing Group.

Council of Europe. (1997). *Recommendation of The Committee of Ministers to member states on 'hate speech'* (No. R (97) 20). London: Councli of Europe. Retrieved from https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680505d5b

Dahlberg, L. (2007). The Internet and discursive exclusion: From deliberative to agonistic public sphere theory. In L. Dahlberg & E. Siapera (Eds.), *Radical democracy and the Internet* (pp. 128–147). London: Palgrave Macmillan.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In O. Varol, E. Ferrara, C. A. Davis, F. Menczer, & A. Flammini (Eds.), *Proceedings of the eleventh international AAAI conference on web and social media (ICWSM 2017)* (pp. 512–515). New York, NY: AAAI.

Delgado, R., & Stefancic, J. (1995). Ten arguments against hate-speech regulation: How valid. *Law Review*, *23*. https://heinonline.org/HOL/LandingPage?handle=hein.journals/nkenlr23&div=32&id=&page=

Diakopoulos, N., & Naaman, M. (2011). Towards quality discourse in online news comments. In P. Hinds, J. Tang, J. E. Bardram, J. Wang, & N. Ducheneaut (Eds.), *Proceedings of the ACM 2011 conference on computer supported cooperative work* (pp. 133–142). New York, NY: ACM.

Dreizin, F., & Priestly, T. (1982). A systematic approach to Russian obscene language. *Russian Linguistics*, *6*, 233–249.

Dudina, V., Judina, D., & Platonov, K. (2019). Personal illness experience in Russian social media: Between willingness to share and stigmatization. In S. El Yacoubi, F. Bagnoli, & G. Pacini (Eds.), *Proceedings of the 6th international conference on Internet science (INSCI 2019)* (Vol. 11938, pp. 47–58). Cham: Springer.

Elagina, D. (2020, July 20). Internet penetration rate in Russia from September 2019 to February 2020, by type of device. *Statista*. Retrieved from https://www.statista.com/statistics/1051507/russia-internet-penetration-by-medium/#statisticContainer

Etling, B., Alexanyan, K., Kelly, J., Faris, R., Palfrey, J. G., & Gasser, U. (2010). *Public discourse in the Russian blogosphere: Mapping RuNet politics and mobilization* (Research Publication 2010–2011). Cambridge, MA: Harvard Berkman Center. Retrieved from https://dash.harvard.edu/bitstream/handle/1/8789613/Public_Discourse_in_the_Russian_Blogosphere_2010.pdf?sequence=1&sa=U&ei=qGVQU8_NAoSTyAS57oFg&ved=0CEAQFjAG&usg=AFQjCNEMhqk9I61iT2bGDO2ITuGgV_hOTw

Federal Law. (2013). *On Amendments to Clause 4 of the Law of the Russian Federation 'On Mass Media' and Clause 13.21 of the Code of Administrative Offences of the Russian Federation: Federal law #34-FZ of April 5, 2013*. Moscow: Federal Law.

Federal Law. (2016a). *On Amendments to the Federal Law 'On Countering Terrorism' and Certain Legislative Acts of the Russian Federation regarding the establishment of additional measures to counter terrorism and ensure public safety: Federal law #374-FZ of July 6, 2016*. Moscow: Federal Law.

Federal Law. (2016b). *On Amendments to the Criminal Code of the Russian Federation and the Code of Criminal Procedure of the Russian Federation regarding the establishment of additional measures to counter terrorism and ensure public safety: Federal law #375-FZ of July 6, 2016*. Moscow: Federal Law.

Federal Law. (2019a). *On Amendments to the Code of Administrative Offences of the Russian Federation:*

Federal law #28-FZ of March 18, 2019. Moscow: Federal Law.

Federal Law. (2019b). *On Amendments to the Federal Law 'On Information, Information Technologies and Information Protection': Federal law #30-FZ of March 18, 2019*. Moscow: Federal Law.

Fuchs, C. (2017). *Social media: A critical introduction*. London: Sage.

Gabdulhakov, R. (2020). (Con)trolling the Web: Social media user arrests, state-supported vigilantism and citizen counter-forces in Russia. *Global Crime*. https://doi.org/10.1080/17440572.2020.1719836

Gorbachev, A. (2017, March 24). Pochemy na akzii protesta vyshlo stolko podrostkov? [Why have so many teenagers attended the protests?]. *Meduza*. Retrieved from https://meduza.io/feature/2017/03/27/pochemu-na-aktsii-protesta-vyshlo-stolko-podrostkov-i-chego-oni-hotyat

Hare, I., & Weinstein, J. (Eds.). (2010). *Extreme speech and democracy*. Oxford: Oxford University Press.

Havryliv, O. (2017). Verbale Aggression: das Spektrum der Funktionen [Verbal aggression: The spectrum of functions]. *Linguistik*, *82*, 27–47.

Howard, J. W. (2017). Free speech and hate speech. *Annual Review of Political Science*, *22*, 93–109.

Kiriya, I. (2014). Social media as a tool of political isolation in the Russian public sphere. *Journal of Print and Media Technology Research*, *3*(2), 131–138.

Koltsova, O. (2019). Methodological challenges for detecting interethnic hostility on social media. In S. S. Bodrunova, O. Koltsova, A. Følstad, H. Halpin, P. Kolozaridi, L. Yuldashev, . . . H. Niedermayer (Eds.), *Proceedings of international the 5th international conference on Internet science (INSCI 2018)* (Vol. 11551, pp. 7–18). Cham: Springer.

Kosov, A. V. (2011). Mythological conscience and invective vocab. *Novosibirsk State University Bulletin*, *5*(2), 35–40.

Krivoshapko, Y. (2020, April 15). Runet audience is to break through to 100 million users in 2020. *Rossiyskaya Gazeta*. Retrieved from https://rg.ru/2020/04/15/auditoriia-runeta-v-2020-godu-probet-planku-v-100-mln-polzovatelej.html

Kronhaus, M. (2012). *The Russian language on the verge of a nervous breakdown*. Moscow: Astrel.

Litvinenko, A. A. (in press). YouTube as alternative television in Russia: Political videos during the presidential election campaign 2018. *Social Media + Society*.

Litvinenko, A. A., & Toepfl, F. (2019). The 'gardening' of an authoritarian public at large: How Russia's ruling elites transformed the country's media landscape after the 2011/12 protests 'For fair elections.' *Publizistik*, *64*(2), 225–240.

Massaro, T. M. (1990). Equality and freedom of expression: The hate speech dilemma. *William & Mary Law Review*, *32*. https://scholarship.law.wm.edu/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=1923&context=wmlr

Metzger, M. J., Flanagin, A. J., Mena, P., Jiang, S., & Wilson, C. (2021). From dark to light: The many shades of sharing misinformation online. *Media and Communication*, *9*(1), 134–143.

Miller, M. L., & Vaccari, C. (2020). Digital threats to democracy: Comparative lessons and possible remedies. *The International Journal of Press/Politics*. https://doi.org/10.1177/1940161220922323

Navalny, A. (2017, March 2). *Don't call him Dimon* [Video file]. Retrieved from https://www.youtube.com/watch?v=qrwlk7_GF9g&ab_channel=%D0%90%D0%BB%D0%B5%D0%BA%D1%81%D0%B5%D0%B9%D0%9D%D0%B0%D0%B2%D0%B0%D0%BB%D1%8C%D0%BD%D1%8B%D0%B9

Papacharissi, Z. (2015). *Affective publics: Sentiment, technology, and politics*. Oxford: Oxford University Press.

Parekh, B. (2012). Is there a case for banning hate speech? In M. Herz & P. Molnar (Eds.), *The content and context of hate speech: Rethinking regulation and responses* (pp. 37–56). Cambridge: Cambridge University Press.

Pfetsch, B. (2018). Dissonant and disconnected public spheres as challenge for political communication research. *Javnost-The Public*, *25*(1/2), 59–65.

Pilkington, A. (2014, May 8). The rich, sweary sublanguage that will protect Russia from Putin's latest crackdown. *The Conversation*. Retrieved from https://theconversation.com/the-rich-sweary-sub-language-that-will-protect-russia-from-putins-latest-crackdown-26362

Platonov, K., & Svetlov, K. (2020). *Involvement in discussions on Khachaturian sisters' case on SNS Vkontakte: Agenda polarization online*. Paper presented at Networks in the Global World (NetGloW'2020) conference, St. Petersburg, Russia.

Pluzer-Sarno, A. (2000). *Mat vocabulary as a phenomenon of Russian culture*. Moscow: Novaya russkaya kniga.

Polyakova, V. (2017, December 6). YouTube zanyal tretje mesto po ohvatu auditorii v RuNete [YouTube takes third place in audience outreach on RuNet]. *Seonews.Ru*. Retrieved from https://www.seonews.ru/events/YouTube-zanyal-trete-mesto-po-okhvatu-auditorii-v-runete

Quandt, T. (2018). Dark participation. *Media and Communication*, *6*(4), 36–48.

Salimovsky, V. A., & Ermakova, L. M. (2011). Extremist discourse in Runet mass communication. *Rossiyskaya i zarubezhnaya filologia*, *3*(15), 71–80.

Salter, L. A., Kuehn, K., Berentson-Shaw, J., & Elliott, M. (2019). Digital threats to democracy: Literature review part 1: Threats and opportunities. *Luminate Workshop*. Retrieved from https://mro.massey.ac.nz/bitstream/handle/10179/14949/3.DD-background-paper-lit-review-1-WEB.pdf?sequence=1

Sidorov, V. A. (2018). Communicative aggressions of the 21st century: Definition and analysis of the prerequisites. *Vestnik*, *15*(2), 300–311.

Smoliarova, A. S., Bodrunova, S. S., Yakunin, A. V., Blekanov, I., & Maksimov, A. (2019). Detecting pivotal points in social conflicts via topic modeling of Twitter content. In S. S. Bodrunova, O. Koltsova, A. Følstad, H. Halpin, P. Kolozaridi, L. Yuldashev, . . . H. Niedermayer (Eds.), *Proceedings of the 5th international conference on Internet science (INSCI 2018)* (Vol. 11551, pp. 61–71). Cham: Springer.

Thelwall, M. (2018). Social media analytics for YouTube comments: Potential and limitations. *International Journal of Social Research Methodology*, *21*(3), 303–316.

Toepfl, F. (2020). Comparing authoritarian publics: The benefits and risks of three types of publics for autocrats. *Communication Theory*, *30*(2), 105–125.

Van Aken, B., Risch, J., Krestel, R., & Löser, A. (2018). *Challenges for toxic comment classification: An in-depth error analysis*. Unpublished manuscript.

Van Dijk, T. A. (1993). *Elite discourse and racism*. Newbury Park, CA: Sage Publications.

Vendil Pallin, C. (2017). Internet control through ownership: The case of Russia. *Post-Soviet Affairs*, *33*(1), 16–33.

Vollhardt, J., Coutin, M., Staub, E., Weiss, G., & Deflander, J. (2007). Deconstructing hate speech in the DRC: A psychological media sensitization campaign. *Journal of Hate Studies*, *5*(15), 15–35.

Waldron, J. (2012). *The harm in hate speech*. Cambridge, MA: Harvard University Press.

Weinstein, J. (2017). Hate speech bans, democracy, and political legitimacy. *Constitutional Commentary*, *32*, 527–583.

Zhel'vis, V. (1997). *Pole brani: Skvernoslovie kak sotsial'naya problema v yazykah i kul'turah mira* [Field of word battle: Swearing as a social problem in world languages and cultures]. Bryansk: Ladomir.

Zvereva, V. (2020). 11 State propaganda and popular culture in the Russian-speaking Internet. In M. Wijermars & K. Lehtisaari (Eds.), *Freedom of expression in Russia's new mediasphere*. London: Routledge.

About the Authors

Svetlana S. Bodrunova (Dr. Hab.) is Professor at School of Journalism and Mass Communications, St. Petersburg State University, Russia. She has led six research projects and published two books and over 80 research papers in Russian and English on Russian and European journalism, media and politics, social media, and ethnicity in communication. She leads the Center for International Media Research in her university.

Anna Litvinenko (PhD) is Researcher at Digitalization and Participation department of Institute for Media and Communication Studies, Freie Universität Berlin, Germany. After receiving her PhD in 2007, she worked as Associate Professor at the Department of International Journalism of St. Petersburg State University, Russia. Her research focusses on the interrelation of media and politics in the digital age and on the role of social media platforms in various socio-political contexts.

Ivan Blekanov has PhD in System Analysis, Control and Data Processing. He is Assistant Professor and Head of Department of Programming Technology at St. Petersburg State University, Russia. He has published over 40 academic publications in Russian and English and has lead data science teams within 5 interdisciplinary projects. His research interests include information retrieval and data mining, webometrics and web analytics, social network analysis and computer forensics. He also works as IT industry consultant.

Dmitry Nepiyushchikh is a Graduate Student at St. Petersburg State University, Russia. He has been involved into research projects that dealt with social media data collection and social network analysis. His research interests include crawling of web and social media, user text classification, and visualization of the results of data processing.

Media and Communication (ISSN: 2183–2439)
2021, Volume 9, Issue 1, Pages 195–208
DOI: 10.17645/mac.v9i1.3360

Article

Roots of Incivility: How Personality, Media Use, and Online Experiences Shape Uncivil Participation

Lena Frischlich [1,2,*], Tim Schatto-Eckrodt [1], Svenja Boberg [1] and Florian Wintterlin [1]

[1] Institute for Communication, University of Münster, 44843 Münster, Germany;
E-Mails: lena.frischlich@uni-muenster.de (L.F.), tim.schatto-eckrodt@uni-muenster.de (T.S.E.),
svenja.boberg@uni-muenster.de (S.B.), florian.wintterlin@uni-muenster.de (F.W.)
[2] Department of Media and Communication, University of Munich, 80538 Munich, Germany

* Corresponding author

Submitted: 15 June 2020 | Accepted: 11 September 2020 | Published: 3 February 2021

Abstract
Online media offer unprecedented access to digital public spheres, largely enhancing users' opportunities for participation and providing new means for strengthening democratic discourse. At the same time, the last decades have demonstrated that online discourses are often characterised by so-called 'dark participation' the spreading of lies and incivility. Using 'problematic behaviour theory' as framework and focusing on incivility as a specific form of dark participation, this article investigates the role of users' personal characteristics, media use, and online experiences in relation to offensive and hateful online behaviour. Using a random-quota survey of the German population, we explored how dark personality traits, political attitudes and emotions, the frequency and spaces of online-media use, and users' experiences with both civil and uncivil online discourses predicted participants own uncivil behaviour, such as posting, sharing, or liking uncivil content. We found that 46% of the participants who had witnessed incivility in the last three months also engaged in uncivil participation. A hierarchical logistic regression analysis showed that incivility was associated with manipulative personality traits as measured by the dark triad, right-wing populist voting intentions, and frequent social-media use. Experiences with both civil comments and hate speech predicted higher levels of uncivil participation. The strongest predictor was participants' personal experiences with online victimisation. Overall, the results confirmed that dark participation in the sense of uncivil engagement results from the interplay of personality traits, an online environment that allows for deviant engagement, and, most importantly, participants' experiences in said environment.

Keywords
dark participation; dark triad; hate speech; incivility; offensive speech; personality; political anger; problematic behaviour theory; social media; victimisation

Issue
This article is part of the issue "Dark Participation in Online Communication: The World of the Wicked Web" edited by Thorsten Quandt (University of Münster, Germany).

1. Introduction

Online media provide unprecedented access to digital public spheres. While the eased access to digital public spheres offers new possibilities for deliberative participation (Shane, 2004), it also provides new opportunities for so-called 'dark participation' (Quandt, 2018, p. 36), norm transgressing forms of online engagement, which includes acts such as spreading disinformation and/or uncivil and hateful content. Most citizens in Western democracies report experiences of online incivility (Geschke, Klaßen, Quent, & Richter, 2019). Witnessing online hate contributes to political polarisation (Hwang, Kim, & Huh, 2014), jars social trust (Näsi, Räsänen, Hawdon, Holkeri, & Oksanen, 2015), fuels discrimination (Hsueh, Yogeeswaran, & Malinen, 2015),

and reduces pro-social activities addressing minorities (Ziegele, Koehler, & Weber, 2018). So far, comparably little is known about the factors motivating individuals' uncivil online behaviour. Our study aims to fill this gap. Using Germany, the first country in the world with a legal framework for dealing with deviant online engagement (the so-called network enforcement act), as context and data from a random-quota survey ($N = 5000$), we aimed to answer the following overarching question: Which factors motivate uncivil participation?

2. Dark Participation and Incivility

Dark participation describes an online setting whereby (a) wicked actors (individuals, groups, and state actors), driven by (b) sinister strategical, tactical, or "pure evil" (Quandt, 2018, p. 41) motives, attack (c) despised objects/targets either directly or indirectly with the aim of (d) manipulating different audience(s). In broader terms, dark participation can be understood as norm-transgressing participation that violates either the norms of civil discourse (i.e., by name-calling or using racial slurs) or honesty (i.e., by spreading falsehoods). Our study focuses on the first type of dark participation: incivility.

Incivility is a "notoriously difficult term to define" (Coe, Kenski, & Rains, 2014, p. 660). While there is a consensus that incivility can be understood as norm-transgressing communication (Kenski, Coe, & Rains, 2020; Mutz, 2015; Papacharissi, 2004), it is less clear which norms are being transgressed. At least two types of norms need to be distinguished: norms related to interpersonal communication and norms related to intergroup communication. Muddiman (2017) refers to these two types as transgressions of personal norms—"Communication that violates the norms of politeness" (Mutz, 2015, p. 6)—and transgressions of public norms—messages that "threaten a collective founded on democratic norms" (Papacharissi, 2004, p. 271). Gagliardone et al. (2016, p. 19) suggested labelling the first type as 'offensive speech,' often studied under labels like 'flaming' (e.g., O'Sullivan & Flanagin, 2003), 'trolling' (Buckels, Trapnell, & Paulhus, 2014), or 'cyber-bullying' (e.g., Festl, 2016; van Geel, Goemans, Toprak, & Vedder, 2017), and the second type as 'hate speech' (Silva, Mondal, Correa, Benevenuto, & Weber, 2016) intersecting with phenomena such as discrimination, racism, or 'group-focussed enmity' and sometimes characterized as 'harmful' or 'dangerous speech' (for an overview, see Kümpel & Rieger, 2019), a distinction we will also use throughout this article.

Complicating the definition of incivility further, online norms are context-dependent (Coe et al., 2014). What is considered acceptable on message boards like 4Chan (Hine et al., 2016) might be inappropriate in a public Facebook group. Plus, as O'Sullivan and Flanagin (2003) argue, there is a sender, a receiver, and an observer perspective involved in detecting norm violations. Messages can be intentionally norm-transgressing or not and can be classified as such by an observer or the target. Consequently, "Incivility is very much in the eye of the beholder" (Herbst, 2010, p. 3). Finally, incivility uses multiple channels. Although much research has focused on text-based incivility (e.g., Stroud, 2010), uncivil communication can also use images as well as audio or video material (Kümpel & Rieger, 2019).

With this context in mind, the current article focuses on uncivil participation that (a) entails offensive speech and hate speech, (b) in online media channels using different forms (e.g., video or text), and (c) is perceived as mocking or attacking the target by the perpetrator (though might not be intended to harm).

3. Problematic Behaviour Theory and Uncivil Participation

Jessor's problematic behaviour theory (1991; Jessor & Jessor, 1977) argues that norm-transgressing behaviour, such as uncivil participation, results from the interplay between a person's characteristics, his or her environment, and, most importantly, that person's perceptions of said environment. Traditionally, problematic behaviour theory distinguishes the following three systems: the 'personality system' (including beliefs and attitudes); the 'environmental system' (including the 'social system,' such as one's peers or parents); the 'perceived environment' (i.e., the norms within that system); which all influence a forth system, namely the 'behaviour system,' which involves a covariation of different norm-transgressing problematic behaviours. The theory has been developed and employed mostly in the context of (adolescent) norm-deviances (e.g., alcohol abuse as examined in Hays, Stacy, & di Matteo, 1987; for a review of the theories' application, see Jessor, 2017), but there is also scholarship arguing that problematic behaviour theory provides a useful perspective on online behaviour (De Leo & Wulfert, 2013; Lee, Kim, Hong, & Marsack-Topolewski, 2019).

Most of the scholarship on problematic behaviour theory employed a developmental perspective, underlining the role of family values and peers. However, scholarship on online incivility suggests that personality and the perceived environmental system are valuable organising structures for adults' online behaviour as well. In a recent overview, Kümpel and Rieger (2019) have identified two main drivers of uncivil online discourses: the characteristics of the sender (e.g., their personality, attitudes, and emotions) and the online environment (e.g., the attention-driven social-media logic). In addition, theories of computer-mediated communication, such as the 'social identity deindividuation' framework (Postmes, Spears, & Lea, 1998), have argued that perceptions of the online realm matter for behaviour in that realm. Accordingly, we will review prior research on incivility by using an adapted version of the problematic behaviour theory and distinguishing between variables

related to the personality system, participants' social online environment, and their experiences in and perceptions of this environment.

3.1. The Personality System

3.1.1. Dark Personality

Norm-transgressing behaviour has been frequently associated with the so-called 'dark triad.' The dark triad describes three sub-clinical forms of offensive personalities: 'narcissism,' 'Machiavellianism,' and 'psychopathy' (Paulhus & Williams, 2002). Although all three covary, they are not superimposable. Narcissists are characterised by grandiosity perceptions, a belief in their own superiority (Paulhus & Williams, 2002), social manipulativeness and a lack of empathy (Raskin & Hall, 1981). Machiavellianism involves manipulative and cold behaviour and psychopathy impulsive and thrill-seeking behaviour by individuals showing reduced levels of empathy and anxiety (Paulhus & Williams, 2002). The dark triad is associated with uncivil forms of behaviour (Kurek, Jose, & Stuart, 2019), such as trolling (Buckels et al., 2014) and bullying (van Geel et al., 2017). However, Koban Stein, Eckhardt, and Ohler (2018) found no statistically significant association between the dark triad and intentions to comment in an uncivil manner. We thus formulated our first research question openly:

RQ1. Does the dark triad predict uncivil participation?

3.1.2. Political Attitudes and Political Emotions

In a series of interviews with users who produce hate speech, Erjavec and Kovačič (2012) identified ideological motivations, like defending one's ingroup against a perceived enemy, as the core characteristics of a certain type of uncivil actors, the so-called "believers." For the current study, we focused on the following three aspects of ideological motivations: 'political ideology' and political frustration as indicated by participants' 'political anger' and their feelings of 'political inefficacy' (i.e., the feeling that they are unable to influence political decision-making processes using norm-consistent forms of participation). Both anger and inefficacy have been linked to norm-deviant behaviour in collective-action research (Becker & Tausch, 2015).

For political ideology, we looked at the extremity of political leanings and voting intentions. Extreme political leanings correlate strongly with partisan identification and reflect a polarisation of attitudes (Jost, 2017). Polarisation, in turn, is correlated with uncivil participation (Suhay, Blackwell, Roche, & Bruggeman, 2015). We thus formulated the following hypothesis:

H1. The extremity of political attitudes predicts higher levels of uncivil participation.

In addition to extreme political leanings, right-leaning audiences could be particularly prone to uncivil discourses. US data show that conservatives generally evaluate hateful posts as being less disturbing than do the democrats (Costello, Hawdon, Bernatzky, & Mendes, 2019). Although this lack of sensitivity might be related to white male republicans being less seldomly attacked online, conservatives are also more single-minded and have grown more extreme over time, arguing for underlying ideological asymmetries (Jost, 2017). In Germany's multi-party system, incivility has been linked to followers of the right-wing party 'Alternative for Germany,' AfD (Kreißel, Ebner, Urban, & Guhl, 2018) and right-leaning politicians (Jaki & Smedt, 2019), although there are also isolated incidents of conservative politicians contributing to uncivil discourses—for instance, the current state president of Bavaria, Markus Söder, derogated refugees as "asylum tourists" in 2018 (for a media report, see dpa, 2018). We thus formulated the following hypothesis:

H2. Voters of right-wing populists and conservative parties are more likely to engage in uncivil participation.

Expressions of incivility (e.g., the use of slurs, see Coe et al., 2014) tend to mirror expressions of anger. According to the cognitive-functional model of emotions, anger arises from demeaning offenses or goal-blockage, creating a sense of injustice and motivating retributive action (cf. Nabi, 2002). It is thus not surprising that anger fuels users' incivility (Gervais, 2016). Although the relationship between incivility and anger is most obvious in offensive speech, Fischer, Halperin, Canetti, and Jasini (2018) have argued that anger is also functionally related to hate. Like anger, hate emerges when a situation is perceived as being unjust and powerful figures are perceived as responsible for the anger-evoking state. We thus expected that anger, more precisely anger against the government, might be associated with uncivil participation more generally:

H3. Political anger predicts higher levels of uncivil participation.

Related to this assumption, we expected that perceived political inefficacy also plays a role in fuelling incivilities. People who perceive themselves as inefficient in a political system are less likely to engage in normative political behaviours, such as voting (Finkel, 1985), and more likely to engage in norm-deviant forms of participation (Becker & Tausch, 2015). We thus expected that people who felt unable to influence political conditions would be more likely to express themselves in an uncivil manner:

H4. Perceived political inefficacy predicts higher levels of uncivil participation.

3.2. The Social Environment

As our research focused on online incivility, we first and foremost focused on the online realm as the relevant 'social-environmental' system (Jessor, 2014). More time spent online and the use of specific social media formats is associated with a higher likelihood of encountering uncivil content (Barnidge, Kim, Sherrill, Luknar, & Zhang., 2019; Koban et al., 2018; Oksanen, Hawdon, Holkeri, Näsi, & Räsänen 2014), making it plausible that this is also the case with uncivil behaviour. Hence, we formulated the following hypothesis:

> H5. Frequent use of online media, particularly of social media, predicts higher levels of uncivil participation.

Besides the mere time spent online, concrete virtual spaces are also likely shape uncivil online behaviour. In their examination of adolescents' contact points with specifically harsh forms of incivility, extremism, Reinemann, Ninierza, Fawzi, Riesmeyer, and Neumann (2019) found that social-networking sites are the largest contact point. However, little is known about the specific social-networking sites that people use to act uncivilly. Therefore, we formulated the following research question:

> RQ2. Which social-networking sites are associated with uncivil participation?

3.3. The Perceived Environment

Problematic behaviour theory assumes that environmental norms guide behaviour. As the social-identity de-individuation model argues, this can be particularly true for online media (Postmes et al., 1998). We thus expected that observing civil online behaviour would be associated with less uncivil participation and observing incivility with more uncivil participation. Observing civil interactions might even reduce the impact of incivility by breaking 'hate norms.' Evidence supporting this expectation comes from research showing that counter-speech can re-civilise online discourses (Garland, Ghazi-Zahedi, Young, Hébert-Dufresne, & Galesic, 2020; Ziegele, Jost, Frieß, & Naab, 2019). Overall, we formulated the following two hypotheses:

> H6. Exposure to civil speech predicts lower levels of uncivil participation.

> H7. Exposure to uncivil speech predicts higher levels of uncivil participation.

Research using the problematic behaviour theory to explain (cyber-)bullying showed that being a victim is a strong predictor for future aggression. Although one might argue that being victimized can lead to a variety of outcomes, research on media violence has shown that victimisation (e.g., through parental aggression or abuse) is a crucial factor in predicting violent behaviour (e.g., Ferguson et al., 2008). Victimised and bullied children are likely to become aggressors themselves, although some do escape the spiral of aggression (Davis, Ingram, Merrin, & Espelage, 2020). We thus predicted the following:

> H8. Personal victimisation predicts higher levels of uncivil participation.

Figure 1 summarises our assumptions.

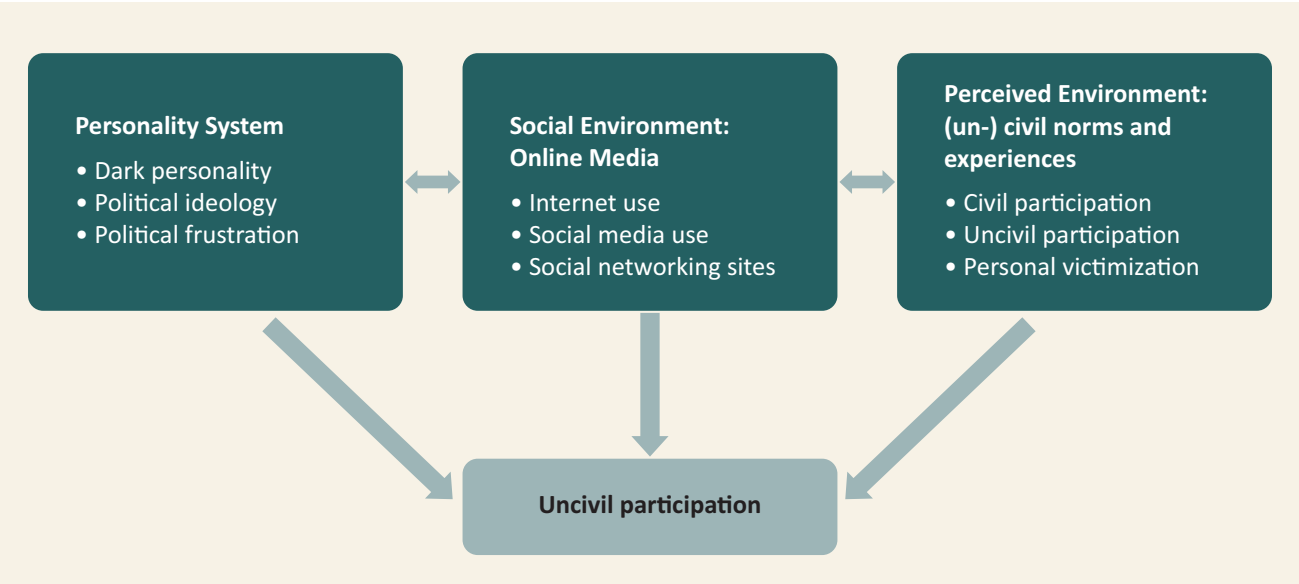


Figure 1. Predictors of uncivil online behaviour.

4. Methods and Measurements

The data for this study were collected by Kantar Emnid via a random-quota web-based survey during the first two weeks of September 2017. The questions of interest were embedded in a larger survey covering sociodemographics, media use, personality traits, experiences with different types of online content, political attitudes, and voting intentions. Our analysis focuses on the variables theoretically hypothesised to motivate uncivil participation.

4.1. Sample

A total of 5,000 individuals finished the survey, representing the German population in terms of age, gender, and region. As it is typical for online surveys, the sample population was slightly more educated than the general population, corresponding to the overall larger online activity by those with a higher formal education (Statistisches Bundesamt, 2020). As we were interested in participants' behavioural responses to uncivil material, only the participants who had witnessed at least one type of uncivil user-generated content (i.e., hate speech or offensive speech) were included in the main analysis (n = 2,968, 59% of the entire sample). This selection was necessary as we asked participants whether they liked, shared, and created uncivil online content that had encountered (see section 4.2.4). As only 0.05% of the participants stated that they had created uncivil content without witnessing it, we are confident that our selection did, indeed, narrow down the sample of interest. The analysed subsample was largely comparable to the overall sample, although the participants who had witnessed uncivility reported higher levels of political anger, were less likely to have a low educational status or income, and reported higher levels of online media use, matching prior research about factors leading to hate speech exposure (e.g., Costello,Barret-Fox, Bernatzky, Hawdon, & Mendes, 2020).

4.2. Measurements

4.2.1. Personality System Variables

We measured the dark triad using the 9-item German short-scale by Küfner, Dufner, and Back (2014). The scale is a psychometrically optimised and translated version of the international Dirty Dozen scale (Webster & Jonason, 2013), which has been demonstrated to have good structure, internal consistency, and stability (Küfner et al., 2014). Three items measured everyday psychopathy (e.g., "I tend to not feel remorse"), narcissism (e.g., "I tend to want to be admired by others") and Marchiavellism ("I deceived and lied to get my way"), respectively. To ensure consistency across the survey, all items were answered using 5-point scales (1 = "absolutely does not apply," 5 = "absolutely applies").

We measured political ideology by combining participants' attitudes and extremity of political leanings. Political attitudes were measured via participants' voting intentions for either of the large parties running for government at that time. Extreme political leanings were measured by coding the distance to the scale mean of the political ideology scale, resulting in a new 5-point scale (0 = "non-extreme," encompassing former values 5 and 6, 4 = "extreme," encompassing former values 1 and 10).

We measured political anger using three items modelled after Stürmer and Simon (2009), using the following questions "How angry/furious/irritated are you about the politics of the current government?" We measured political inefficacy using two items from Zick, Küppers, and Hövermann (2011): "Politicians do not care about people like me" and "People like me have no control over what the government does." All items were answered on a 5-point scale.

We further controlled for gender, age, education, income, and—due to large political differences between the former Eastern and Western parts of Germany where the study took place—region.

4.2.2. Social-Environment Variables

We measured participants' engagement with the online environment by asking participants about how much time they spent using the Internet in a normal week (hours and minutes) and their subjective social media usage (1 = "never," 5 = "always"). Furthermore, we provided the participants with a list of the most prominent social-networking sites in Germany (Schröder, 2016)—Facebook, Odnoklassniki, Twitter, Instagram, VKontakte, Pinterest, Tumblr, Reddit, LinkedIn, and Xing—and asked whether they had a social media account that they used at least occasionally (dummy coded: 1 = active user, 0 = non-user). Tables 1 and 2 summarise the descriptives.

4.2.3. Perceived-Environment Variables

Before asking the participants about their experiences with civil and uncivil participation, we explained our definition of user-generated content using the following statement:

> The next step is about your experiences with user-generated content on the Internet in the last three months. User-generated content is everything that private individuals publish on the Internet. This can include text, images, or videos, for example in social media, on websites, or in the comment sections of online newspapers.

We then asked the participants about their experiences with different forms of participation, always using a 5-point scale (1 = "never," 5 = "at least once per day").

Table 1. Descriptives and reliabilities for parametric variables.

	α	*M*	*SD*
Age	—	42.57	15.9
Machiavellianism	.77	1.98	0.92
Psychopathy	.82	1.91	0.93
Narcissism	.87	2.28	1
Extreme political leanings	—	1.05	1.24
Political anger	.95	3.27	1.11
Political inefficacy	.79	3.51	1.07
Internet use per day (in hrs)	—	2.4	2.43
Social-media use	—	3.97	1.33

Civil participation was measured with a single item:

> Some user-generated content discusses political and social issues in an objective and helpful way. How often have you seen such contributions, pictures, or videos on political or social topics on the Internet in the last three months?

Incivility as offensive speech (explained as "user-generated content [that] mocks or attacks someone personally, insults, or abuses him or her") and as hate speech ("user-generated content [that] attacks someone because he or she belongs to a certain group") were measured using multiple items. Offensive speech was divided into personal victimisation and observed attacks on other users, journalists, or politicians. Hate speech was measured via attacks on people based on their gender, sexual orientation, skin colour, religion or nationality, political attitude, relationship with refugees, relationship with nature and animals (e.g., their diet), and fandom. A polychoric factor analysis identified three underlying factors, jointly explaining 83% of the variance. All items measuring hate speech loaded onto the first factor (Cronbach's $\alpha = .97$), and all items measuring

Table 2. Frequencies for dummy-coded categorical variables.

	Yes	%
Male	1,622	51.4
East German	392	12.4
Low education	1,399	44
High education	756	24
Low income	1,399	44.3
High income	769	25.1
Facebook	2,217	77.9
Odnaklassniki	113	3.97
Twitter	600	19
VKontakte	108	3.79
Instagram	816	28.7
Pinterest	631	20
Tumblr	191	6.71
Reddit	129	4.53
LinkedIn	296	10.4
Xing	517	18.2
Alternative for Germany (AFD) voter	349	11.7
Christian Democratic Union (CDU)	517	17.3
Christian Social Union (CSU) voter	123	4.12
Social Democratic Party (SPD) voter	519	16.35
Green voter	194	6.5
Left voter	337	11.3
Free Democratic Party (FDP) voter	216	5.36
Witnessed civic participation	2,753	87.3
Witnessed offensive speech	2,692	85.3
Witnessed hate speech	2,902	100
Personally victimised	887	28.1

observed offensive speech loaded onto the second factor (Cronbach's $\alpha = .94$). Personal victimisation was the only item loading onto the third factor, although it also loaded onto the second one. As this pattern was compatible with our theoretical assumptions that observing incivility has a different impact on the individual compared to experiencing victimisation oneself (see also Geschke et al., 2019), we used personal victimisation as a single item for the subsequent analyses. As the distribution for all exposure items was u-shaped, we formed dummy-coded variables representing exposure.

4.2.4. Uncivil Participation

Based on research that recommends behaviour-based measures for norm-transgressing online behaviour to reduce social-desirability bias (Festl, 2016), we asked the participants, "If you take all the comments together that mock, insult, abuse, or threaten someone, how often in the last three months did you" followed by eights statements representing positive responses to uncivil content ("evaluate such posts, images, or videos positively [e.g., via "likes" or positive comments]?) and negative reactions (e.g., "reported such a post, image, or video"). A polychoric factor analysis identified four underlying factors, explaining 81% of the variance.

Our study focuses on uncivil participation, such as liking, sharing, or producing uncivil content. All items representing this kind of behaviour loaded onto the same factor (44% variance explanation, Cronbach's α = .96). The other three factors were countering (two items, "disliking" and "commenting," 19% variance explanation, Cronbach's $\alpha = .90$), ignoring (one item, 14%), and avoiding (one item, 5%). Two items ("reporting" and "consuming") showed cross-loadings and were thus excluded from the scale construction. Although our article focuses only on factors associated with uncivil participation, the inclusion of negative responses allowed us to provide a more natural image of the behavioral options in the current media environment.

As the distribution for the single uncivil behaviours and the sum score of the factor was highly positively skewed and mean aggregation did misrepresent such distributions, we dummy-coded whether participants had engaged in uncivil behaviour for the following analyses. Overall, 46% of those who had seen uncivil content admitted amplification (corresponding to roughly one quarter of the overall population when those who had never seen any kind of incivility and were accordingly not able to "like" or "share" this kind of material were also considered.

5. Results

5.1. Analytical Approach

We used R (Version 3.5.1; R Core Team, 2018) for all our analyses. 7% of the variables had at least one missing value. Based on preliminary inspections, missing values were treated as missing at random and imputed using the MissForest algorithm. MissForest is an iterative imputation procedure based on the random forest algorithm (Stekhoven & Buehlmann, 2012). All analyses were based on this imputed data set. To account for the different scale levels, parametric variables were mean-centred before the main analysis.

5.2. Preliminary Analyses

A preliminary inspection of the zero-order correlations found no statistically significant associations between uncivil participation and the German region, low income, or low education (see Table 3), and we removed these variables from the subsequent analyses. As income and education were perfectly correlated (see Supplementary File, Table A), we included only high education in the next step to avoid multi-collinearity.

5.3. Main Analysis

We tested our hypotheses using a logistic hierarchical regression analysis. Parametric variables were mean-centred. The variables of the personality system were entered first. Block 1 included the socio-demographics and Block 2 the dark triad. Block 3 included political ideology (i.e., voting intentions and extreme political leanings) and Block 4 political frustration (anger and inefficacy). The variables related to the online environment were entered in Block 5, and the variables related to the perceived environment in Block 6.

As all blocks reached statistical significance, we focused on the last block to evaluate our hypotheses. This allowed us to examine all predictor variables in concert, thus reflecting the logic underlying problematic behaviour theory (Jessor, 1991). Block 6 explained 34% of the variance. All socio-demographic control variables failed to reach statistical significance. Answering RQ1, the dark personality traits Machiavellianism and psychopathy, though not narcissism, were associated with a higher likelihood of uncivil participation. Extreme political leanings did not predict uncivil participation (H1) but, partially confirming H2, intentions to vote for the AfD did. Voting for conservative parties (CDU and CSU) did not predict uncivil participation, and neither did political anger (H3) or political inefficacy (H4). Although general Internet use was not predictive of uncivil participation, the subjective frequency of social media use and using VKontakte were (H5, RQ2). Surprisingly, observing civic participation was positively related to uncivil participation (H6). Observing offensive speech was not related to uncivil participation, but experiences of both hate speech (H7) and personal victimisation increased the likelihood of uncivil participation (H8). Table 4 shows the results of this last block. The full table is provided in the Supplementary File (see Table B in the Supplementary File).

Table 3. Correlates of uncivil participation.

		Pearson correlation with uncivil participation
Socio-demographics	Age	−0.20**
	Male	0.07**
	Low formal education	−0.01
	High formal education	−0.05*
	Low income	−0.01
	High income	−0.04*
	East Germany	−0.01
Personality system	Machiavellianism	0.33**
	Psychopathy	0.33**
	Narcissism	0.24**
	AFD	0.08**
	CDU	−0.02
	CSU	−0.01
	SPD	0.02
	Green	−0.01
	Left	−0.01
	FDP	−0.01
	Extreme political leanings	0.01
	Political anger	0.11**
	Political inefficacy	0.01
Social-environment variables	Internet use per day	−0.03
	Social-media use	0.12**
	Facebook	0.09**
	Odnoklassniki	0.14**
	Twitter	0.12**
	VKontakte	0.17**
	Instagram	0.17**
	Pinterest	0.08**
	Tumblr	0.14**
	Reddit	0.16**
	Xing	0.11**
	Linkedin	0.11**
Perceived environment	Civil speech	0.15**
	Offensive speech	0.05**
	Hate speech	0.28**
	Victimisation	0.39**

Notes: *** $p < 0.001$, ** $p \leq 0.01$, * $p < 0.05$.

6. Discussion

Our study examined uncivil participation from a problem-behaviour perspective. In line with problematic behaviour theory (Jessor, 1991; Jessor & Jessor, 1977), uncivil participation was predicted by a combination of the participants' 'personality system,' the online 'social system,' and participants' experiences in and perceptions of this online realm (see Figure 1).

First, our data showed that uncivil participation, when measured by behavioural indicators, is much more frequent than direct questions regarding hate-spreading suggest (Isenberg, 2019). Nearly half of those who had witnessed incivility had contributed to its spread, corresponding to roughly one quarter of the German online users. As the data for this study were collected in 2017, the concrete percentages must be interpreted with care. Nevertheless, long-term comparisons show that Germans witnessed more hate speech in 2020 than three years earlier (Landesanstalt für Medien NRW, 2020). In light of the observed relationship between witnessing hate speech and uncivil participation, our results can thus be considered a conservative estimate for uncivil participation in 2020. Although our measure of uncivil behaviours included both mild and more severe forms of attacks against others—and the most extreme attacks

Table 4. Hierarchical logistic regression analysis, Block 6.

	B	*SE*	OR	LL	UL
Intercept	−2.00	0.23 ***	0.14	0.09	0.21
Age	−0.01	0.00	0.99	0.99	1.00
Male	0.18	0.10	1.19	0.98	1.45
High education	−0.17	0.11	0.84	0.68	1.04
Machiavellianism	0.30	0.07 ***	1.35	1.17	1.55
Psychopathy	0.29	0.07 ***	1.33	1.17	1.52
Narcissism	0.04	0.06	1.04	0.93	1.17
AFD	0.41	0.16 **	1.51	1.10	2.06
CDU	0.04	0.15	1.04	0.78	1.38
CSU	−0.04	0.24	0.96	0.60	1.53
SPD	0.20	0.14	1.22	0.92	1.61
Green	0.05	0.19	1.05	0.72	1.52
Left	0.09	0.17	1.09	0.79	1.53
FDP	−0.07	0.19	0.94	0.65	1.34
Extreme political leanings	−0.02	0.04	0.98	0.91	1.06
Political anger	0.10	0.05	1.11	1.00	1.23
Political inefficacy	0.04	0.05	1.04	0.94	1.14
Internet use (hrs/d)	−0.02	0.02	0.98	0.95	1.02
Social-media use	0.13	0.05 ***	1.14	1.04	1.25
Facebook	0.01	0.12	1.01	0.80	1.29
Odnoklassniki	0.29	0.34	1.34	0.69	2.64
Twitter	−0.08	0.12	0.92	0.72	1.17
Vkontakte	1.04	0.40 **	2.83	1.33	6.57
Instagram	0.19	0.12	1.21	0.96	1.52
Pinterest	0.05	0.12	1.05	0.83	1.32
Tumblr	−0.09	0.24	0.91	0.58	1.45
Reddit	0.35	0.33	1.42	0.76	2.76
Xing	0.23	0.14	1.26	0.96	1.65
LinkedIn	−0.07	0.19	0.93	0.64	1.35
Civic speech	0.46	0.15 ***	1.59	1.18	2.16
Offensive speech	0.02	0.14	1.02	0.77	1.35
Hate speech	1.04	0.12 ***	2.83	2.26	3.55
Victimisation	1.23	0.10 ***	3.42	2.78	4.20
Nagelkerke's Pseudo-R^2			0.34***		

Notes. *** $p < 0.001$, ** $p \leq 0.01$, * $p < 0.05$.

are usually being driven by very few users (Kreißel et al., 2018)—our results show the larger context in which incivility flourishes.

The association between users' personal characteristics and uncivil participation was overall relatively weak. In line with prior research about dark personalities' uncivil behaviour (Kurek et al., 2019), we found that psychopathy and Machiavellianism increased self-reported uncivil participation (RQ1), whereas narcissism did not. When participants scored one scale point above average on the dark triad, they were roughly 20% more likely to engage in uncivil participation. This modest effect size could also explain why other studies with fewer participants failed to find a link between the dark triad and uncivil participation intentions (Koban et al., 2018).

In line with prior research linking the spread of incivility and hate in Germany to the AfD (Kreißel et al., 2018), the participants intending to vote for AfD were 50% more likely to report uncivil participation (H2). We did not find any statistically significant associations between uncivil participation and the intention to vote for any of the other major parties. In contrast to prior correlational studies (Suhay et al., 2015), extreme political leanings were not linked to uncivil participation (H1), suggesting that incivility, at least in Germany, is asymmetrically more compatible with right-wing as compared to (extreme) left-wing ideologies.

We did not observe a statistically significant link between participants' feelings of political anger (H3) or political inefficacy (H4) and uncivil participation. Uncivil participation, at least in our sample, cannot be understood as a participation driven by feelings of anger and inefficacy towards the political system. Notably, prior research found that incivility by an opposing

party does create anger, consequently fuelling incivility (Gervais, 2016, 2019); therefore, a more nuanced understanding of the roots of political anger and its effects on uncivil online discourses would be a valuable avenue for future research.

In line with prior research (Costello et al., 2019; Koban et al., 2018; Oksanen et al., 2014), we found that social-media use (H5) was associated with uncivil participation, whereas general Internet use was not. Particularly users of the Russian-based network VKontakte, which has gained public attention for having lax moderation rules and hosting ultra-right-wing content (Udupa et al., 2020), reported uncivil participation in our study, suggesting that such platforms might be an attractive environment for those engaging in uncivil participation (RQ2).

Only partially confirming our expectations, experiences with both uncivil (H7) but also civil speech (H6) predicted higher levels of uncivil participation, although the effect for hate speech was substantially stronger. Participants who witnessed civil speech were 1.5 times more likely to report uncivil participation, but those who had noted hate speech were nearly three times as likely to report engaging in uncivil behaviour online. In line with prior research on bullying (Lee et al., 2019) and studies on the toxic effects of having been victimised (Davis et al., 2020), the participants who had been personally victimised were 3.4 times as likely to report uncivil participation (H8) compared to those without such experiences.

Taken together, our data add to the theoretical understanding of uncivil participation in numerous ways. First of all, our study confirmed the central assumption of problematic behaviour theory for norm-deviant online behaviour, showing that the 'personality system,' the online environment, and the experiences within this environment jointly contribute to uncivil participation. Regarding the 'personality system,' our data did not find meaningful associations between uncivil participation and extreme political leanings, political anger, or political inefficacy. Overall, uncivil participation in the German population was not driven by these motives. We did, however, find uncivil participation to be rooted in sinister personality traits and to be prevalent among right-leaning voters.

With regard to the digital 'social system,' general Internet use did not increase the likelihood of uncivil participation, while social media was associated with uncivil participation. We did not find a statistically significant association between observing offensive speech and uncivil participation. Instead, the toxic effects of incivility were mostly related to hate speech—that is, to violations of 'public norms' that are foundational to a democratic society as Muddiman (2017) has summarised the matter. Although offensive speech can trigger nasty replies (Ziegele, Jost, Bohrmann, & Heinbach, 2018), our data show that hate speech is the kind of discourse that is likely to erode civil discussion norms (see Papacharissi, 2004, for a similar argument) and might need counterstrategies. Noteworthy, counter-measures such as removing content thereby must be carefully balanced against values of free speech (Masullo Chen, Muddiman, Wilner, Pariser, & Stroud, 2019).

When it comes to concrete steps to counter hate speech, our results suggest that interventions need to account for both users' personalities and their online environments. At the level of the 'personality system,' our results are compatible with the argument that empathy might prevent the spread of incivility (Bilewicz & Soral, 2020). Diminished empathy is the defining characteristic of the dark-triad personalities (Paulhus & Williams, 2002; Raskin & Hall, 1981), and fostering empathy can improve prejudiced intergroup attitudes (Batson & Ahmad, 2009). Fostering empathy thus seems to be a promising approach to fighting the roots of incivility at the level of the 'personality system.'

At the same time, the strongest predictors for uncivil participation were related to the 'social system' and the participants' experiences therein. Using social media sites known to tolerate hateful rhetoric and experiencing hate speech and personal attacks substantially increased the likelihood of reporting uncivil participation. Therefore, our data underline the need for proprietors of online spaces and platforms to care for the spaces they provide. Research has shown that community management upholding civil norms can be a valuable strategy to accomplish this (Ziegele, Jost, et al., 2018; Ziegele, Jost, Frieß, & Naab, 2019).

6.1. Limitations and Further Research Directions

Our study had several limitations that must be considered. First, we focused on Germany, a country where harsh forms of incivility are legally sanctioned. The generalisation of our findings to other cultural contexts is a question for future research. Second, we used a cross-sectional design. Although this allowed us to collect a sample large enough to detect uncivil participation, the reported associations cannot be interpreted causally. Even when we think that it is most plausible that, for instance, personality predicts behaviour, long-term measurements are needed to disentangle the direction of the relationships reported in our study. Furthermore, our sample was slightly more educated than the general population. Although this reflects a typical online public and is thus suitable to study online incivility, future research on the roots of incivility amongst those with lower levels of formal education would be worthwhile. Finally, we focused on self-reported behaviours. Although our prevalence rates by far exceed prior work using definition-based approaches, supporting the notion that behaviour-based approaches might be less susceptible to social-desirability biases (Festl, 2016), it would be beneficial if future research were to combine our findings with observational data.

6.2. Conclusion

Overall, our study confirmed the central assumption of problematic behaviour theory for uncivil participation, showing that the 'personality system,' the online environment, and the experiences therein jointly contribute to our understanding of norm-transgressing dark participation. As such, we have provided unique empirical evidence for the ongoing debate about addressing the downsides of participatory online media by highlighting the factors that contribute to the spread of incivility.

Acknowledgments

Data collection for this study was supported by the Federal Ministry of Education and Research (BMBF) via the grant number 16KIS0496, "PropStop: Detection, Proof, and Combating of Hidden Propaganda Attacks via Online Media." Kantar-Emnid Germany conducted the data collection. The first and second author are supported by the Digital Society research program funded by the Ministry of Culture and Science of the German State of North Rhine-Westphalia via the grant number 005–1709-001. We further thank the anonymous editor and the two reviewers for their constructive feedback and support throughout the review process.

Conflict of Interests

The authors declare no conflict of interest.

Supplementary Material

Supplementary material for this article is available online in the format provided by the author (unedited).

References

Barnidge, M., Kim, B., Sherrill, L. A., Luknar, Ž., & Zhang, J. (2019). Perceived exposure to and avoidance of hate speech in various communication settings. *Telematics and Informatics*, *44*, 1–13. https://doi.org/10/gghcnr

Batson, C. D., & Ahmad, N. Y. (2009). Using empathy to improve intergroup attitudes and relations. *Social Issues and Policy Review*, *3*(1), 141–177. https://doi.org/10.1111/j.1751-2409.2009.01013.x

Becker, J. C., & Tausch, N. (2015). A dynamic model of engagement in normative and non-normative collective action: Psychological antecedents, consequences, and barriers. *European Review of Social Psychology*, *26*(1), 43–92. https://doi.org/10/gf3gvc

Bilewicz, M., & Soral, W. (2020). Hate speech epidemic: The dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, *41*(1). https://doi.org/10/gg4whj

Buckels, E. E., Trapnell, P. D., & Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and Individual Differences*, *67*, 97–102. https://doi.org/10/f58bzw

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, *64*(4), 658–679. https://doi.org/10/f6dxrx

Costello, M., Barrett-Fox, R., Bernatzky, C., Hawdon, J., & Mendes, K. (2020). Predictors of viewing online extremism among America's youth. *Youth and Society*, *52*(5), 710–727. https://doi.org/10/gf3gq3

Costello, M., Hawdon, J., Bernatzky, C., & Mendes, K. (2019). Social group identity and perceptions of online hate. *Sociological Inquiry*, *89*(3), 427–452. https://doi.org/10/gghcnc

Davis, J. P., Ingram, K. M., Merrin, G. J., & Espelage, D. L. (2020). Exposure to parental and community violence and the relationship to bullying perpetration and victimization among early adolescents: A parallel process growth mixture latent transition analysis. *Scandinavian Journal of Psychology*, *61*(1), 77–89. https://doi.org/10/ggs53d

De Leo, J. A., & Wulfert, E. (2013). Problematic Internet use and other risky behaviors in college students: An application of problem-behavior theory. *Psychology of Addictive Behaviors*, *27*(1), 133–141. https://doi.org/10.1037/a0030823

dpa. (2018, August 7.). «Asyltourismus»: Söder verteidigt Begriff ["Asylum tourist": Söder defends the term]. *Welt Online*. Retrieved from https://www.welt.de/regionales/bayern/article178969630/Asyltourismus-Soeder-verteidigt-Begriff.html

Erjavec, K., & Kovačič, M. P. (2012). "You don't understand, this is a new war!" Analysis of hate speech in news web sites' comments. *Mass Communication and Society*, *15*(6), 899–920. https://doi.org/10/gfgnmm

Ferguson, C. J., Cruz, A. M., Martinez, D., Rueda, S. M., Ferguson, D. E., & Negy, C. (2008). Personality, parental, and media influences on aggressive personality and violent crime in young adults. *Journal of Aggression, Maltreatment & Trauma*, *17*(4), 395–414. https://doi.org/10/bbc4pg

Festl, R. (2016). Perpetrators on the Internet: Analyzing individual and structural explanation factors of cyberbullying in school context. *Computers in Human Behavior*, *59*, 237–248. https://doi.org/10/f8hw28

Finkel, S. E. (1985). Reciprocal effects of participation and political efficacy: A panel analysis. *American Journal of Political Science*, *29*(4), 891–913. https://doi.org/10/fg9xv5

Fischer, A. H., Halperin, E., Canetti, D., & Jasini, A. (2018). Why we hate. *Emotion Review*, *10*(4), 309–320. https://doi.org/doi/10.1177/1754073917751229

Gagliardone, I., Pohjonen, M., Zerai, A., Beyene, Z., Aynekulu, G., Bright, J., . . . & Teferra, Z. M. (2016). *Mechachal: Online debates and elections in Ethiopia—From hate speech to engagement in social media*. Oxford: University of Oxford.

Garland, J., Ghazi-Zahedi, K., Young, J.-G., Hébert-Dufresne, L., & Galesic, M. (2020). Countering hate on social media: Large scale classification of hate and counter speech. *arXiv.org*. Retrieved from http://arxiv.org/abs/2006.01974

Gervais, B. T. (2016). More than mimicry? The role of anger in uncivil reactions to elite political incivility. *International Journal of Public Opinion Research*, *29*(3), 384–405. https://doi.org/10/gftpws

Gervais, B. T. (2019). Rousing the partisan combatant: Elite incivility, anger, and antideliberative attitudes. *Political Psychology*, *40*(3), 637–655. https://doi.org/10/gghcns

Geschke, D., Klaßen, A., Quent, M., & Richter, C. (2019). *Hass im Netz: Der schleichende Angriff auf unsere Demokratie* [Hate on the net: The creeping attack on our democracy]. Jena: Institut für Demokratie und Zivilgesellschaft.

Hays, R. D., Stacy, A. W., & di Matteo, M. R. (1987). Problem behavior theory and adolescent alcohol use. *Addictive Behaviors*, *12*(2), 189–193. https://doi.org/10.1016/0306-*4603*(87)90026-8

Herbst, S. (2010). *Rude democracy: Civility and incivility in American politics*. Philadelphia, PA: Temple University Press.

Hine, G. E., Onaolapo, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Samaras, R., . . . & Blackburn, J. (2016). Kek, cucks, and god emperor Trump: A measurement study of 4chan's politically incorrect forum and its effects on the Web. *arXiv.org*. Retrieved from http://arxiv.org/abs/1610.03452

Hsueh, M., Yogeeswaran, K., & Malinen, S. (2015). "Leave your comment below": Can biased online comments influence our own prejudicial attitudes and behaviors? *Human Communication Research*, *41*(4), 557–576. https://doi.org/10.1111/hcre.12059

Hwang, H., Kim, Y., & Huh, C. U. (2014). Seeing is believing: Effects of uncivil online debate on political polarization and expectations of deliberation. *Journal of Broadcasting & Electronic Media*, *58*(4), 621–633. https://doi.org/10/gf3g62

Isenberg, M. (2019). Hate speech und Diskussionsbeteiligung im Netz [Hate speech and involvement in discussions on the net]. *Landesanstalt für Medien NRW*. Retrieved from https://www.medienanstalt-nrw.de/fileadmin/user_upload/lfm-nrw/Service/Veranstaltungen_und_Preise/Ergebnisbericht_Hate_Speech_Sonderstudie_LFMNRW.pdf

Jaki, S., & Smedt, T. D. (2019). Right-wing German hate speech on Twitter: Analysis and automatic detection. *arXiv.org*. Retrieved from http://arxiv.org/abs/1910.07518

Jessor, R. (1991). Risk behaviour in adolescents: A psychosocial framework for understanding and action. *Journal of Adolescent Health*, *12*, 597–695. https://doi.org/10.1016/1054-139X(91)90007-K

Jessor, R. (2014). Problem behavior theory: A half-century of research on adolescent behavior and development. In R. M. Lerner, A. C. Petersen, R. K. Silbereisen, & J. Brooks-Gunn (Eds.), *The developmental science of adolescence: History through autobiography* (pp. 239–256). New York, NY: Psychology Press.

Jessor, R. (2017). Problem drinking and psychosocial development in adolescence. In R. Jessor (Ed.), *Problem behavior theory and adolescent health: The collected works of Richard Jessor* (Vol. 2, pp. 105–121). Wiesbaden: Springer. https://doi.org/10.1007/978-3-319-51349-2_6

Jessor, R., & Jessor, S. (1977). *Problem behavior and psychosocial development: A longitudinal study of youth*. Cambridge: Academic Press. Retrieved from https://www.scienceopen.com/document?vid=75f5b901-e506-4382-8d2d-307b63a5b851

Jost, J. T. (2017). Asymmetries abound: Ideological differences in emotion, partisanship, motivated reasoning, social network structure, and political trust. *Journal of Consumer Psychology*, *27*(4), 546–553. https://doi.org/10/gdh72z

Kenski, K., Coe, K., & Rains, S. A. (2020). Perceptions of uncivil discourse online: An examination of types and predictors. *Communication Research*, *47*(6), 795–814. https://doi.org/10/gghcnf

Koban, K., Stein, J. P., Eckhardt, V., & Ohler, P. (2018). Quid pro quo in Web 2.0: Connecting personality traits and Facebook usage intensity to uncivil commenting intentions in public online discussions. *Computers in Human Behavior*, *79*, 9–18. https://doi.org/10/gf3gv4

Kreißel, P., Ebner, J., Urban, A., & Guhl, J. (2018). Hass auf Knopfdruck: Rechtsextreme Trollfabriken und das Ökosystem koordinierter Hasskampagnen im Netz [Hate when the button is pressed: Right-wing extremist troll fabrics and the ecosystem of coordinated hate campaigns in the Internet]. London: Institute for Strategic Dialogue.

Küfner, A. C. P., Dufner, M., & Back, M. D. (2014). Das dreckige Dutzend und die niederträchtigen Neun: Kurzskalen zur erfassung von Narzissmus, Machiavellismus und Psychopathie [The dirty dozen and the naughty nine: Short scales for measuring narcissicm, Marchiavellism, and psychopathy]. *Diagnostica*, *61*(2), 76–91. https://doi.org/10/gf3gzv

Kümpel, A. S., & Rieger, D. (2019). *Wandel der Sprach: und Debattenkultur in sozialen Online Medien. Ein Literaturüberblick zu Ursachen und Wirkungen von inziviler Kommunikation* [Changing debate cultures in social online media: A literature review on the causes and consequences of uncivil communication]. Berlin: Konrad-Adenauer Stiftung.

Kurek, A., Jose, P. E., & Stuart, J. (2019). 'I did it for the LULZ': How the dark personality predicts online disinhibition and aggressive online behavior in adolescence. *Computers in Human Behavior*, *98*, 31–40. https://doi.org/10/ggft9d

Landesanstalt für Medien NRW. (2020). *forsa-Befragung zu: Hate Speech 2020* [forsa survey on: Hate

speech 2020]. Düsseldorf: Landesanstalt für Medien NRW. Retrieved from https://www.medienanstalt-nrw.de/fileadmin/user_upload/NeueWebsite_0120/Themen/Hass/forsa_LFMNRW_Hassrede2020_Ergebnisbericht.pdf

Lee, J. M., Kim, J., Hong, J. S., & Marsack-Topolewski, C. N. (2019). From bully victimization to aggressive behavior: Applying the problem behavior theory, theory of stress and coping, and general strain theory to explore potential pathways. *Journal of Interpersonal Violence*. Advance online publication. https://doi.org/10.1177/0886260519884679

Masullo Chen, G., Muddiman, A., Wilner, T., Pariser, E., & Stroud, N. J. (2019). We should not get rid of incivility online. *Social Media + Society*, *5*(3). https://doi.org/10/gghcnh

Muddiman, A. (2017). Personal and public levels of political incivility. *International Journal of Communication*, *11*, 3182–3202.

Mutz, D. C. (2015). *In-your-face politics: The consequences of uncivil media*. New Jersey, NY: Princeton University Press.

Nabi, R. L. (2002). Anger, fear, uncertainty, and attitudes: A test of the cognitive-functional model. *Communication Monographs*, *69*(3), 204–216.

Näsi, M., Räsänen, P., Hawdon, J., Holkeri, E., & Oksanen, A. (2015). Exposure to online hate material and social trust among Finnish youth. *Information Technology & People*, *28*(3), 607–622. https://doi.org/10.1108/ITP-09-2014-0198

O'Sullivan, P. B., & Flanagin, A. J. (2003). Reconceptualizing "flaming" and other problematic messages. *New Media & Society*, *5*(2), 69–94. https://doi.org/10/b3txz4

Oksanen, A., Hawdon, J., Holkeri, E., Näsi, M., & Räsänen, P. (2014). Exposure to online hate among young social media users. In N. M. Warehime (Ed.), *Soul of society: A focus on the lives of children & youth* (pp. 253–274). Oklahoma City, OK: Emerald Books.

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, *6*(2), 259–283. https://doi.org/10/dz4rp6

Paulhus, D. L., & Williams, K. M. (2002). The dark triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality*, *36*(6), 556–563. https://doi.org/10/d2jxm9

Postmes, T., Spears, R., & Lea, M. (1998). Breaching or building social boundaries? SIDE-effects of computer-mediated communication. *Communication Research*, *25*(6), 689–715. https://doi.org/10/ffsbdn

Quandt, T. (2018). Dark participation: Manipulative user engagement in the news making process. *Media and Communication*, *6*(4), 36–48. http://dx.doi.org/10.17645/mac.v6i4.1519

Raskin, R., & Hall, C. S. (1981). The narcissistic personality inventory: Alternate form reliability and further evidence of construct validity. *Journal of Personality Assessment*, *45*(2), 159–162. https://doi.org/10/ch4b6b

Reinemann, C., Ninierza, A., Fawzi, N., Riesmeyer, C., & Neumann, K. (2019). *Jugend—Medien—Extremismus* [Youth—Media—Extremism]. Wiesbaden: Springer Fachmedien VS.

Schröder, J. (2016, December 1). Die 20 populärsten sozialen Netzwerke in Deutschland [The 20 most popular social network sites in Germny]. *Meedia.de*. Retrieved from https://meedia.de/2016/12/01/die-20-populaersten-sozialen-netzwerke-in-deutschland-facebook-klar-vorn-instagram-snapchat-und-musical-ly-boomen

Shane, P. M. (2004). *Democracy online: The prospects for political renewal through the Internet*. New York, NY: Psychology Press.

Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). Analyzing the targets of hate in online social media. *arXiv.org*. Retrieved from http://arxiv.org/abs/1603.07709

Statistisches Bundesamt. (2020). *Private Haushalte in der Informationsgesellschaft: Nutzung von Informations- und Kommunikationstechnologien* [Private households in the information society: Use of information and communication technologies] (No. 15-4). Wiesbaden: Statistisches Bundesamt.

Stekhoven, D. J., & Buehlmann, P. (2012). MissForest: Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118.

Stroud, N. J. (2010). Polarization and partisan selective exposure. *Journal of Communication*, *60*(3), 556–576. https://doi.org/10/d6p3kx

Stürmer, S., & Simon, B. (2009). Pathways to collective protest: Calculation, identification, or emotion? A critical analysis of the role of group-based anger in social movement participation. *Journal of Social Issues*, *65*(4), 681–705. https://doi.org/10/d3nfv3

Suhay, E., Blackwell, A., Roche, C., & Bruggeman, L. (2015). Forging bonds and burning bridges: Polarization and incivility in blog discussions about occupy wall street. *American Politics Research*, *43*(4), 643–679. https://doi.org/10.1177/1532673X14553834

van Geel, M., Goemans, A., Toprak, F., & Vedder, P. (2017). Which personality traits are related to traditional bullying and cyberbullying? A study with the Big Five, Dark Triad and sadism. *Personality and Individual Differences*, *106*, 231–235. https://doi.org/10/f9g9n6

Webster, G., & Jonason, P. (2013). Putting the "IRT" in "Dirty": Item response theory analyses of the Dark Triad Dirty Dozen: An efficient measure of narcissism, psychopathy, and Machiavellianism. *Personality and Individual Differences*, *54*(2), 302–306. https://doi.org/10/gg8hm2

Zick, A., Küpper, B., & Hövermann, A. (2011). *Die Abwertung der Anderen. Eine europäische Zustandsbeschreibung zu Intoleranz , Vorurteilen und Diskri-*

minierung [The derogation of others: A European description of the state of intolerance, prejudice, and discrimination]. Berlin: Friedrich Ebert Foundation.

Ziegele, M., Jost, P., Bormann, M., & Heinbach, D. (2018). Journalistic counter-voices in comment sections: Patterns, determinants, and potential consequences of interactive moderation of uncivil user comments. *Studies in Communication and Media*, *7*(4), 525–554. https://doi.org/10/gf3gk2

Ziegele, M., Jost, P., Frieß, D., & Naab, T. (2019). *Aufräumen im Trollhaus: Zum Einfluss von Community-Managern und Aktionsgruppen in Kommentarspalten* [Tidy up the troll house: On the influence of community managers and activist groups in the comment sections] Düsseldorf: Institute for Internet and Democracy.

Ziegele, M., Koehler, C., & Weber, M. (2018). Socially destructive? Effects of negative and hateful user comments on readers' donation behavior toward refugees and homeless persons. *Journal of Broadcasting & Electronic Media*, *62*(4), 636–653. https://doi.org/10/gf8pn4

About the Authors

Lena Frischlich (PhD) is currently an Interim Professor at the LMU Munich, Department of Communication and Media Studies and the Principal Investigator of the junior research group 'Democratic Resilience in Times of Online Propaganda, Fake News, Fear and HateSpeech (DemoRESILdigital)' at the University of Münster, Institute for Communication Science. Her research interests include the staging and effects of online-propaganda and related phenomena in a changing media environment and digital communication more broadly.

Tim Schatto-Eckrodt is a Research Associate at the Department of Communication at the University of Münster, Germany. He holds an MA degree in Communication Studies from the same department. He is currently working on his PhD project about online conspiracy theories and is part of the junior research group 'Democratic Resilience in Times of Online Propaganda, Fake News, Fear and HateSpeech (DemoRESILdigital).' His further research interests include computational methods and online propaganda.

Svenja Boberg is a Communication Scientist at the University of Münster, who researches the dynamics of online debates and social media hypes using computational methods. In her dissertation she studies the spread of outrage in social media networks. In 2016 she joined the BMBF-funded project 'Detection, Proof and Combating of Covert Propaganda Attacks via Online Media.' Beforehand, she completed her MA in Communication Studies at the University of Münster addressing news consumption on Facebook.

Florian Wintterlin (PhD) is a Research Associate at the University of Münster. After finishing his PhD in trust and journalism research, he worked on disinformation in the BMBF-funded project 'Detection, Proof and Combating of Covert Propaganda Attacks via Online Media.' Currently, his research includes questions of science communication as well as political communication with a focus on digital phenomena.

COGITATIO

Media and Communication (ISSN: 2183–2439)
2021, Volume 9, Issue 1, Pages 209–214
DOI: 10.17645/mac.v9i1.1770

Commentary

Advancing Research into Dark Participation

Oscar Westlund [1,2,3]

[1] Oslo Metropolitan University, 0167 Oslo, Norway; E-Mail: oscarw@oslomet.no
[2] Volda University College, 6101 Volda, Norway
[3] University of Gothenburg, 405 30 Gothenburg, Sweden

Submitted: 21 October 2020 | Accepted: 3 December 2020 | Published: 3 February 2021

Abstract
Dark participation is and should be an essential concept for scholars, students and beyond, considering how widespread disinformation, online harassment, hate speech, media manipulation etc. has become in contemporary society. This commentary engages with the contributions to this timely thematic issue, which advance scholarship into dark participation associated with news and misinformation as well as hate in a worthwhile way. The commentary closes with a call for further research into four main areas: 1) the motivations that drive dark participation behaviors by individuals and coordinated groups; 2) how these individuals and groups exploit platforms and technologies for diverse forms of dark participation; 3) how news publishers, journalists, fact-checkers, platform companies and authorities are dealing with dark participation; and 4) how the public can advance their media literacy for digital media in order to better deal with dark participation. Authorities must advance and broaden their approaches focused on schools and libraries, and may also use emerging technologies in doing so.

Keywords
dark participation; disinformation; journalism; misinformation; platforms; platform exploitation

Issue
This commentary is part of the issue "Dark Participation in Online Communication: The World of the Wicked Web" edited by Thorsten Quandt (University of Münster, Germany).

1. Introduction

The moment a person decides to engage with digital or social media, and billions have globally, they enter an online world filled with information and opportunities and also various forms of dark participation. Individuals as well as more coordinated groups and organizations use social media platforms for various forms of *dark participation*, such as media manipulation, misinformation, hate speech and online harassment. These in turn connect with several important aspects of dark participation, such as actors, reasons, objects/targets, audiences and processes.

In his original article featured in a 2018 *Media and Communication* thematic issue, Quandt (2018) problematized the intersections of journalism and publication participation via digital technologies and advanced the concept dark participation. This is a valuable concept that can guide empirical research and makes for an important and timely theme for a thematic issue entering the 2020's. This thematic issue makes a substantial advancement of knowledge into some more specific areas of dark participation. Altogether, the thematic issue consists of 10 original articles, authored or co-authored by highly respected scholars. The call for papers sought for diverse contributions from all corners of the world, including efforts to engage the Global South. Notwithstanding such efforts, all submissions must go through rigorous peer-review, and in the end the articles meeting quality standards for this thematic issue turn out to be authored mostly by scholars from Europe and the United States.

This commentary will engage with the thematic issue, highlighting some of its key contributions. The article contributions span diverse forms of dark participation, such as a study of visual forms of political communication on Twitter and its association with social media manipulation. This study advances our understanding

of euro-sceptic imagery and anti-systemic communication in the salient case of the European Union (Marchal, Neudert, Kollanyi, & Howard, 2021). Ultimately, I have chosen to sort most of the contributions into two main thematic areas for discussion. The first thematic area is *news and misinformation*. Many scholars may immediately come to think of traditional journalists and normatively important news about public affairs and politics, but here it includes a broader spectrum of actors, such as "alternative news media." The second thematic area is *hate* and similarly includes several thematic issue contributions.

2. News and Misinformation

What makes journalism, and who is a journalist, has been a recurring issue of discussion, debate and boundary work in journalism studies literature for decades. This is not only a form of normative academic exercise but can be closely linked to the practice of policy. What is news is crucially important for assessments of financial subsidies (such as in Scandinavia), and authoritarian regimes enforce rules for defining who has the right to work as a journalist, and in such ways controlling who gets to scrutinize the authorities (Badr, 2020). Journalism studies scholars have advanced diverse conceptualizations of emerging social actors associated with journalism, such as so-called interlopers (Eldridge, 2017), in-betweeners (Ahva, 2017) and peripheral actors (Belair-Gagnon, Holton, & Westlund, 2019). In this context, and for this thematic issue, von Nordheim and Kleinen-von Königslöw (2021) have added "parasites," discussed here as a subsystem that act in ways relating to and even threatening journalism as the primary system. They have co-authored a theoretically oriented article titled "Uninvited Dinner Guests: A Theoretical Perspective on the Antagonists of Journalism Based on Serres' Parasite." Von Nordheim and Kleinen-von Königslöw take their point of departure in how journalism, in the traditional sense with legacy news media and its journalists, have become confronted with emerging actors of various kinds. Their article highlights how so-called "parasites" have increasingly entered into the journalistic system, albeit operate in a much different way and with other norms and logics. The authors explicitly seek to theorize the parasites in the role of the antagonists, and how such parasites threaten the well-established journalistic system. Thus, this article builds on a normative perspective that the historically established journalistic system is something being harmed by emerging parasites, whom take advantage of its resources and affect output and values. Parasites as a concept clearly is associated with actors having a negative effect. Von Nordheim and Kleinen-von Königslöw clarify four key characteristics, including but not limited to them acting from within the system with journalistic resources and thus difficult to eliminate without affecting the system itself. The authors do not discuss concrete examples who such parasites are, with the exception of intermediary platform companies, for which there fortunately is a discussion on dissolvement of the host vs. parasite distinction.

Well established institution(s) of journalism are indeed associated with epistemic and journalistic authority, and for producing different forms of knowledge relevant for citizens (Ekström & Westlund, 2019a). It is common that publishers and journalism studies scholars, policymakers and pundits take their departure in some sort of established journalistic system, positioning this as the center and everything else as peripheral, alternative or parasitic. However, such positioning has normative underpinnings, and the fundamental idea that journalism is placed at some sort of center is problematic (Steensen & Westlund, 2020; Tandoc, 2019). There are ongoing processes of dislocation of news journalism, including but not limited to how platform companies as intermediaries have shifted the dynamics for how news is shared, consumed and engaged with, moving it further away from journalistic actors and their epistemic practices (Ekström & Westlund, 2019b). A recent article has developed and introduced the concept of *local political information infrastructure* for a study of platform civics and Facebook. The concept suggest we should broaden our perspectives when it comes to actors producing local news and local information beyond traditional news media, and account for the role of networked media logics (Thorson et al., 2020).

Multiple actors can benefit from being like or associated with, journalistic news. On the one hand, producers of "fake news" and disinformation repeatedly imitate the form and style of journalistic news, presumed to increase likelihood to deceive. On the other hand, alternative news media oftentimes imitate the form and style of journalistic news, while potentially being explicit in their offering of an alternative voice and their intentions towards narrowing their focus, the scope and plurality of voices. The relational aspects are central to scholarly conceptualizations of alternative news media, and how such intentionally produce and publish "news" that legacy news media do not bring forth (Holt, Ustad Figenschou, & Frischlich, 2019). In Scandinavia, some members of the public have discontinued turning to the public service broadcasters for news, questioning their credibility and instead turning to alternative news media. In this thematic issue, Schwarzenegger (2021) presents an interview-based study focusing on how users of diverse alternative media connect with their media and its community. He identifies different nuances of grey, concluding that these communities, or audiences, experience ambivalence in relation to aspects such as alternative sources, experiences of community and comfort, and anti-systemness (see also the article on anti-systemic communication in this thematic issue by Marchal et. al., 2021).

As this issue proposes, journalistic authority has been challenged by diverse stakeholders and actors, including

powerful politicians in various countries, repeatedly questioning and delegitimizing legacy news media by using "fake news" as a label. Additionally, "fake news" is a genre of producing intentionally false news (e.g., Egelhofer & Lecheler, 2019), an area that Tandoc, Thomas, and Bishop (2021) thematic issue's article tackles. The authors contend that much is known about motivations for "fake news," and the different kinds of deceptions emerge, but less about how this genre imitates "traditional journalism." Following a content analysis of fake news materials, they found overall similarities with traditional journalism in terms of inverted pyramid format storytelling, timeliness, negativity and prominence. The authors discuss that the main difference comes down to fake news articles oftentimes featuring an opinion by its author. Notwithstanding this, also news journalism comes with many choices of different kinds, following epistemic values and practices, genre conventions etc., making it difficult to neutralize opinion even though only resorting to voicing opinion of sources etc. Wahl-Jorgensen (2020), for example, discusses that emotional labor linked to news production often has been made invisible when portraying journalists as detached observers.

To continue, we may ask what effect corrective messages have for citizens that have exposed themselves to online misinformation and disinformation? Martel, Mosleh, and Rand (2021) discuss that existing research witnesses mixed results in terms of what approaches are effective. Their experiment-based study is focusing on how correction style may affect the efficacy of corrections. Martel et al.'s study finds that analytic thinking and active open-minded thinking are most clearly associated with citizens absorbing corrective messages in a way that results in updating their beliefs. To continue, another thematic issue article advancing knowledge about online sharing of misinformation comes from Metzger, Flanagin, Mena, Jiang, and Wilson (2021). The authors stress the importance of studying the motivations people have for sharing news and misinformation online, and the beliefs associated with the misinformation they are exposed to. They analyze a large dataset of comments online, leading to the conclusion that misinformation being spread on social media oftentimes is disbelieved. The next thematic issue article, authored by Chang, Haider, and Ferrara (2021), focuses on the intersection of citizen's online political participation and misinformation, in the salient case of Taiwan and its 2020 presidential election. The authors studied online participation in discussions across three social platforms, and reveal that some topics are selectively discussed, and others are largely avoided. In studying misinformation, the authors argue the importance of acknowledging clashes associated with practices, ideologies and cultural history. Ultimately, a red thread for the articles discussed concerns a sort of interrelationship between journalism, publishers and news on the one hand, and misinformation on the other. While it is problematic to juxtapose these in relation to each other, yet there remains a strong dynamic between the two.

3. Hate

While the world wide web initially was associated with visions regarding access, participation etc., a growing body of literature witness to the prevalence of what Quandt (2018) refers to as dark participation, in its diverse forms. Publishers have been struggling to deal with participatory journalism, and much of the participation with news has been displaced to platforms non-proprietary to the publishers (Westlund & Ekström, 2018). Some publishers have maintained participatory features such as comment fields but have had to develop their content moderation strategies to cope with hate speech, disinformation and other forms of dark participation (Wintterlin, Schatto-Eckrodt, Frischlich, Boberg, & Quandt, 2020). Similarly, platform companies are wrestling with both human—and technology-led approaches towards content moderation, for which disinformation has become a central concern (Napoli, 2020), not least during political elections such as the United States' 2020 presidential election where platform companies such as Twitter have flagged misinformation coming from various actors, including but not limited to the president himself.

This thematic issue features articles on aggressive behaviors, hate speech and uncivility, and hate is a common denominator across these studies although incivility extends to also anger and fear. A basic definition of hate suggests it has to do with people's feelings of hostility towards a person or a group. Hatred towards others as an inner feeling kept to oneself is problematic in itself but may in such instances actually do no harm. However, harm will likely take place when people enact or communicate their hatred. Social media platforms affordances have most certainly enabled people to take advantage of platforms for such purposes. Social media platforms have lowered the threshold for individuals to express hate, and for coordinated groups and organizations to give expression for intentional and systemic hate towards someone or something.

In their thematic issue article, Paasch-Colberg, Strippel, Trebbe, and Emmer (2021) have focused on hate speech, which they discuss in the broader sense in terms of forwarding expressions about one's emotion of hate vis-a-vis the more specific legal understanding of the concept as referring to prejudice or violent expressions towards specific groups in society. The article advances a more nuanced understanding and framework for different forms of hate speech as well as offensive language, which is used to analyze the materials collected through a rich mixed-method study focusing on migration and refugees, conducted in Germany and across news publishers, a blog, Facebook and YouTube. Moreover, Bodrunova, Litvinenko, Blekanov, and Nepiyushchikh (2021) focuses on obscene speech

and politically motivated hate speech and aggressive content in the salient case of Russian YouTube. The authors have carefully selected 13 videos that altogether have generated a large amount of comments and views. Their study of obscene and hate speech reveals a link to expressing solidarity and support, shapes dynamics of public discussions, and helps place criticism towards authorities and regimes in context. And Frischlich, Schatto-Eckrodt, Boberg, and Wintterlin (2021) further expands on hate, its roots and uncivility. Their article "Roots of Incivility: How Personality, Media Use, and Online Experiences Shape Uncivil Participation" takes its point of departure in a situation where dark participation is salient in society and offers a survey-based study from Germany focusing on how personality, media use and online experiences influence incivility. The article finds that a relatively high proportion of the citizens exposed to uncivil actions have themselves engaged in uncivil participation, and those who are exposed to both hate speech and civil comments are most likely to engage in incivility themselves.

In extension of the research advanced in this thematic issue, scholars may explore the ways in which social actors (humans) and technological actants (machines) have agency in the development or pursuit of emotions (Lewis & Westlund, 2015), if there are systematic reactions to individual feelings and emotions, and what the causes are? Research associated with the so-called emotional turn in journalism has brought forward how social media platform affordances have impacted the space for emotion (e.g., Wahl-Jorgensen, 2020).

4. Concluding Reflection

This thematic issue advances research into several critically important aspects of dark participation, such as hate (including hate speech and incivility) and the tensions between journalists and other news—and misinformation producing actors that may or may not be harmful. This thematic issue provides worthwhile insights for scholars, students, policymakers and practitioners in diverse fields. Notwithstanding the above, this does not mean that the findings can easily be turned into actions that substantially reduce dark participation. We thus must reconcile our perspectives on the mediascape and dark participation, acknowledging that such is and will remain to be a core component. Lending from the 4 A's framework (Lewis & Westlund, 2015), the prevalence of dark participation can be seen as a complicated interplay of activities between diverse actors (such as publishers, fact-checkers, policymakers, platform companies, fake news producers, alternative news media etc.), distinct technological actants (such as platform algorithms, software for editing phots and videos, artificial intelligence etc.) and audiences (e.g., citizens and their media—and information literacy skills).

These four areas—actors, actants, audiences and activities—are generally important to consider when advancing research. First, scholars should study motivations further to understand better the "roots" to the various emerging forms of dark participation, offering much more granular understanding of political and financial motivations, and also seeking to identify potential other motivations.

Second, researchers should ask how do such motivations intersect with the current mediascape, and the opportunities for enacting different forms of dark participation. Emerging technologies in combination with the social architecture of the Web, offering low thresholds for "produsage" (Bruns, 2012) and participation, has enabled laymen as well as coordinated groups to achieve high impact with their dark participation. Science and technology studies (STS) has consistently shown how the uses of technologies are not determined beforehand (i.e., technological determinism), but rather can be seen as socially constructed. This means that whatever good purposes platform and tech companies may have in terms of building platforms that are safe and useful and marked by positive participation that can be associated with civil conversations and informed citizens there will be motivated people and groups taking advantage of the very same tech and platforms for purposes of dark participation. Ultimately, people and groups exploit platform affordances for their own interest and motives, fueling dark participation and causing substantial concerns for societies and democracy such as through hate speech and incivility etc.

Third, scholars should inquire how do news publishers, journalists, fact-checkers, platform companies and authorities deal with dark participation. Researchers need to recognize that there is a broad spectrum of actors (and actants and audiences) engaging in different activities, some of their own and some in collaboration with others, to combat dark participation. While some stakeholders complain that Facebook and other platform companies are unpredictable in changing their algorithms, such changes may well be necessary to undermine systematic exploitation of their platforms for dark participation. While the accessibility to data via some platforms (most notably Twitter) has enabled research, other platforms have enforced significant restrictions to their data sharing (most notably Facebook). Numerous scholars have called for an improved collaboration with platforms, involving social media platforms sharing relevant data for research (e.g., Pasquetto et. al., 2020). Importantly, social media platforms are used for questioning journalism, publishers and journalists, through various forms of digital press criticism (Carlson, Robinson, & Lewis, 2020). Moreover, journalists have become targets of online harassment (Lewis, Zamith, & Coddington, 2020) and mob censorship (Waisbord, 2020). Ultimately, various actors are taking advantage of platforms and their affordances for dark purposes, such as to destabilize journalism as an institution, and the journalists carrying out newswork. It is important that scholars study such behaviors and actors further, for instance

by applying the lens of parasites (c.f. von Nordheim & Kleinen-von Königslöw, 2021) for more concrete empirical work.

Fourth, an important question for scholars relates to what can be done in terms of the public advancing their media literacy for digital media in order to better deal with dark participation. Media—and information literacy should not be approached as an explicit form of knowledge that one can develop theoretically by reading or watching instructions, but must also be approached as a form of tacit knowledge that is developed through experiences, ideally together with someone having tacit knowledge that can supervise (such as a school teachers, alternatively an interactive program designed to simulate situations and offer feedback). Scholars should explore and study ways that authorities, NGO's and other stakeholders (with the public's interest in mind) *can* and possibly *should* take advantage of emerging technologies and platforms for purposes of countering dark participation. For example, how can schools and libraries develop or appropriate AR/VR technologies into instructional role play games that allow individuals to embody others (e.g., age, gender, race etc.) and get such first-hand and emotional experiences in the interaction with others? A prerequisite may well be to conduct and integrate basic science with applied science, enrolling key stakeholders such as funding bodies for research and innovation, commercial companies, together with authorities and governmental institutions such as schools, libraries and media oversight institutions.

Acknowledgments

First, let me extend my thanks to Valerie Belair-Gagnon, Avery Holton as well as Thorsten Quandt for their constructive comments on earlier versions of this commentary. Avery offered really useful advice to the first draft, Thorsten helped clarify many arguments and reflections, and Valerie has played an important role for the final polishing of the article. Second, it is a great honor to me that Thorsten Quandt invited me to author this commentary. Mats Ekström and I were intellectually stimulated when processing the original article on dark participation by Quandt for the thematic issue we guest edited for *Media and Communication* in 2018. As I have said numerous times elsewhere, this is an exceptional article with a much-needed critical perspective and exciting multi-level rhetoric, and it should and surely will impact research and higher education for many years to come. In extension of this, I envision this thematic issue will be well received by scholars. Third, I wrote this commentary as a member of Source Criticism and Mediated Disinformation, a research project funded by the Norwegian Research Council.

Conflict of Interests

The author declares no conflict of interests.

References

Ahva, L. (2017). How is participation practiced by "in-betweeners" of journalism? *Journalism Practice*, *11*(2/3), 142–159. https://doi.org/10.1080/17512786.2016.1209084

Badr, H. (2020). The Egyptian syndicate and (digital) journalism's unresolved boundary struggle. *Digital Journalism*. Advance online publication. https://doi.org/10.1080/21670811.2020.1799424

Belair-Gagnon, V., Holton, A. E., & Westlund, O. (2019). Space for the liminal. *Media and Communication*, *7*(4), 1–7. https://doi.org/10.17645/mac.v7i4.2666

Bodrunova, S. S., Litvinenko, A., Blekanov, I., & Nepiyushchikh, D. (2021). Constructive aggression? Multiple roles of aggressive content in political discourse on Russian YouTube. *Media and Communication*, *9*(1), 181–194.

Bruns, A. (2012). Reconciling community and commerce? *Information, Communication & Society*, *15*(6), 815–835. https://doi.org/10.1080/1369118X.680482

Carlson, M., Robinson, S., & Lewis, S. C. (2020). Digital press criticism: The symbolic dimensions of Donald Trump's assault on U.S. journalists as the "enemy of the people." *Digital Journalism*. Advance online publication. https://doi.org/10.1080/21670811.2020.1836981

Chang, H.-C. H., Haider, S., & Ferrara, E. (2021). Digital civic participation and misinformation during the 2020 Taiwanese presidential election. *Media and Communication*, *9*(1), 144–157.

Egelhofer, J. L., & Lecheler, S. (2019). Fake news as a two-dimensional phenomenon: A framework and research agenda. *Annals of the International Communication Association*, *43*(2), 97–116. https://doi.org/10.1080/23808985.2019.1602782

Ekström, M., & Westlund, O. (2019a). Journalism and epistemology. In *Oxford Encyclopedia of Journalism Studies*. Oxford: Oxford University Press. https://doi.org/10.1093/acrefore/9780190228613.013.806

Ekström, M., & Westlund, O. (2019b). The dislocation of news journalism: A conceptual framework for the study of epistemologies of digital journalism. *Media and Communication*, *7*(1), 259–270. https://doi.org/10.17645/mac.v7i1.1763

Eldridge, S. A. (2017). *Online journalism from the periphery: Interloper media and the journalistic field*. London: Routledge.

Frischlich, L., Schatto-Eckrodt, T., Boberg, S., & Wintterlin, F. (2021). Roots of incivility: How personality, media use, and online experiences shape uncivil participation. *Media and Communication*, *9*(1), 195–208.

Holt, K., Ustad Figenschou, T., & Frischlich, L. (2019). Key dimensions of alternative news media. *Digital Journalism*, *7*(7), 860–869. https://doi.org/10.1080/21670811.2019.1625715

Lewis, S. C., & Westlund, O. (2015). Actors, actants, audiences, and activities in cross-media sews work: A matrix and a research agenda. *Digital Journalism*, *3*(1), 19–37. https://doi.org/10.1080/21670811.2014.927986

Lewis, S. C., Zamith, R., & Coddington, M. (2020). Online harassment and its implications for the journalist–audience relationship. *Digital Journalism*, *8*(8), 1047–1067. https://doi.org/10.1080/21670811.2020.1811743

Marchal, N., Neudert, L.-S., Kollanyi, B., & Howard, P. N. (2021). Investigating visual content shared over Twitter during the 2019 EU parliamentary election campaign. *Media and Communication*, *9*(1), 158–170.

Martel, C., Mosleh, M., & Rand, D. G. (2021). You're definitely wrong, maybe: Correction style has minimal effect on corrections of misinformation online. *Media and Communication*, *9*(1), 120–133.

Metzger, M. J., Flanagin, A. J., Mena, P., Jiang, S., & Wilson, C. (2021). From dark to light: The many shades of sharing misinformation online. *Media and Communication*, *9*(1), 134–143.

Napoli, P. M. (2020). Connecting journalism and public policy: New concerns and continuing challenges. *Digital Journalism*, *8*(6), 691–703. https://doi.org/10.1080/21670811.2020.1775104

Paasch-Colberg, S., Strippel, C., Trebbe, J., & Emmer, M. (2021). From insult to hate speech: Mapping offensive language in German user comments on immigration. *Media and Communication*, *9*(1), 171–180.

Pasquetto, I., Swire-Thompson, B., Amazeen, M. A., Benevenuto, F., Brashier, N. M., Bond, R. M., . . . Yang, K. C. (2020). Tackling misinformation: What researchers could do with social media data. *The Harvard Kennedy School Misinformation Review*, *1*(8). https://doi.org/10.37016/mr-2020-49

Quandt, T. (2018). Dark participation. *Media and Communication*, *6*(4), 36-48. https://doi.org/10.17645/mac.v6i4.1519

Schwarzenegger, C. (2021). Communities of darkness? Users and uses of anti-system alternative media between audience and community. *Media and Communication*, *9*(1), 99–109.

Steensen, S., & Westlund, O. (2020). *What is digital journalism studies?* London: Routledge.

Tandoc, E. C. (2019). Journalism at the periphery. *Media and Communication*, *7*(4), 138–143. https://doi.org/10.17645/mac.v7i4.2626

Tandoc, E. C., Jr., Thomas, R. J., & Bishop, L. (2021). What is (fake) news? Analyzing news values (and more) in fake stories. *Media and Communication*, *9*(1), 110–119.

Thorson, K., Medeiros, M., Cotter, K., Chen, Y., Rodgers, K., Bae, A., & Baykaldi, S. (2020). Platform civics: Facebook in the local political information infrastructure. *Digital Journalism*, *8*(10). 1231–1257.

von Nordheim, G., & Kleinen-von Königslöw, K. (2021). Uninvited dinner guests: A theoretical perspective on the antagonists of journalism based on Serres' parasite. *Media and Communication*, *9*(1), 88–98.

Wahl-Jorgensen, K. (2020). An emotional turn in journalism studies? *Digital Journalism*, *8*(2), 175–194. https://doi.org/10.1080/21670811.2019.1697626

Waisbord, S. (2020). Mob censorship: Online harassment of US journalists in times of digital hate and populism. *Digital Journalism*, *8*(8), 1030–1046. https://doi.org/10.1080/21670811.2020.1818111

Westlund, O., & Ekström, M. (2018). News and participation through and beyond proprietary platforms in an age of social media. *Media and Communication*, *6*(4), 1–10. https://doi.org/10.17645/mac.v6i4.1775

Wintterlin, F., Schatto-Eckrodt, T., Frischlich, L., Boberg, S., & Quandt, T. (2020). How to cope with dark participation: Moderation practices in German newsrooms. *Digital Journalism*, *8*(7), 904–924. https://doi.org/10.1080/21670811.2020.1797519

About the Author

Oscar Westlund (PhD) is Professor at Oslo Metropolitan University, where he co-leads the OsloMet Digital Journalism Research Group. He holds secondary positions at Volda University College and University of Gothenburg. Westlund is the Editor-in-Chief of *Digital Journalism*. Westlund has published widely on digital journalism, media management, mobile news and epistemology. His four most recent books were all co-authored/co-edited and published with Routledge in 2020: *What is Digital Journalism Studies?*, *Critical Incidents in Journalism, Definitions of Digital Journalism (Studies)*, and *Mobile News*. Westlund wrote this commentary as a project member of Source Criticism and Mediated Disinformation, a project funded by the Norwegian Research Council.

COGITATIO

Media and Communication (ISSN: 2183–2439)
2021, Volume 9, Issue 1, Pages 215–216
DOI: 10.17645/mac.v9i1.3815

Commentary

Beyond the Darkness: Research on Participation in Online Media and Discourse

Claes de Vreese

Amsterdam School of Communication Research, University of Amsterdam, The Netherlands; E-Mail: c.h.devreese@uva.nl

Submitted: 2 November 2020 | Accepted: 11 November 2020 | Published: 3 February 2021

Abstract
This commentary reflects on the notion of 'dark participation' which is central in this thematic issue. It asks whether there are patches of light and whether our research is becoming too obsessed with the darkness?

Keywords
democratic backsliding; misinformation; participation; polarization; trust

Issue
This commentary is part of the issue "Dark Participation in Online Communication: The World of the Wicked Web" edited by Thorsten Quandt (University of Münster, Germany).

In an era of polarization, democratic backsliding, and decreasing media freedom, it is important for the scholarly community to occasionally take a step back. It is upon our community to provide clear analyses of these developments and to provide systematic empirical evidence about the nature, scope, and conditionalities of them. The current thematic issue of *Media and Communication* is welcome because it does just that.

The topics are not trivial: a) Spread of mis- and disinformation; b) polarization of and by political elites, in the media, in media use; c) increasing distrust of journalism and news avoidance; and d) changing (read 'worsening') civil discourse, especially online. Welcome to the research agenda of a communication scholar! Given the potential far reaching consequences for democratic processes and our societies at large, evidence that reaches beyond single cases or single countries is most welcome.

This thematic issue offers insights from Germany, Taiwan, the European Union, Russia, and the United States. Some research looks at the *content* of political discourse, in 'fake news,' in comments sections, and in visuals. The studies include social media platforms like Twitter and YouTube. Other research focuses on the *user*, and how personality, motivation, and gratifications affect both information usage and processing. The studies span from interviews to content analyses and survey data. In this respect the thematic issue delivers on important dimensions: Evidence from different countries, content, and user perspectives; and multiple methods being deployed.

The thematic issue takes its departing point in the concept of 'dark participation.' This concept has received the most comprehensive treatment by Quandt (2018). It is a broad concept, trying to capture "negative, selfish, or even deeply sinister contributions" (Quandt, 2018, p. 40). Quandt offers a generic introduction to the concept, distinguishing how dark participation can come from different *actors*, with different *reasons*, focusing on different *objects*, reaching different *audiences*, through different *processes*. This is a useful overview, but also one that leaves the concept fully open for further delineation.

The dark participation concepts can leave a scholar or citizen in a depressed mood. As Westlund and Ekström (2018) point out we came from a period where there was a focus on 'positive forms of participatory journalism.' In their view, this research did not signal or problematize dark participation. Quandt (2018) paints the dark picture. He makes intentional reference to Hobbes but an attentive reader will also have found this important sentence: "If you now believe that the future is all doom and gloom, then you have stepped into a trap that I intentionally set"

(p. 44). This (for scholarly articles) somewhat unconventional tool is important, beyond the stylistic effect.

In recent years there has been a turn in the societal discussions and in the research agenda, and the 'doom and gloom' perspective now seems pervasive. This not only holds true for the perspective of citizen participation in (online) (news) media, but also for the broader field of political journalism and communication. Van Aelst et al. (2017), including myself, indeed focus on six *concerns* when describing the changes in contemporary media ecologies: 1) Declining supply of political information; 2) declining quality of news; 3) increasing media concentration and declining diversity; 4) increasing fragmentation and polarization; 5) increasing relativism; and 6) increasing inequality in political knowledge. Across these concerns, a core challenge is 'epistemic relativism,' where all information is treated equal, whether provided by journalists or citizens, whether positive or negative, whether distributed through traditional or online, automated, digital channels. In the review of these concerns it is pointed out that far from all empirical evidence supports an unequivocal legitimacy of the concerns.

This raises a bigger question: In the midst of worries about, and research into trolling, incivility, conspiracy, mis- and disinformation, automated pollution of the information environment, populism, and democratic backsliding, is there also space for optimism and a positive research agenda? Whether that is work driven by an 'always look on the bright side of life' or '*post tenebras lux*,' light after darkness philosophy, can remain open. But it seems important to balance our fascination with 'darkness' with questions about positive engagements with media.

Whether it be instances of increased media trust, the possible upsides for journalism during the Covid-19 pandemic, examples of constructive news, the still positive correlates between political interest and news media usage, or the focus on engagement in media and politics which is also evidenced in recent elections. I am not advocating a return to past decades. It is an invitation for us to think about the conditions and mechanisms for *positive* contributions to a healthy public debate. And to think about how we can make research contributions constructive and actionable. There are tangible examples like the Center for Media Engagement (https://mediaengagement.org) or the Media for Democracy (https://mediafordemocracy.org) initiative around the 2020 United States' elections, offering advice for both media and citizens. The bottom line is, that in the era of darkness, it will *also* be a task of scholars to provide guidance on the upsides.

Conflict of Interests

The author declares no conflict of interests.

References

Quandt, T. (2018). Dark participation. *Media and Communication*, *6*(4), 36–48.

Van Aelst, P., Strömbäck, J., Aalberg, T., Esser, F., de Vreese, C. H., Matthes, J., . . . Stanyer, J. (2017). Political communication in a high-choice media environment: A challenge for democracy? *Annals of the International Communication Association*, *41*(1), 3–27.

Westlund, O., & Ekström, M. (2018). News and participation through and beyond proprietary platforms in an age of social media. *Media and Communication*, *6*(4), 1–10.

About the Author

Claes de Vreese is Faculty Professor of AI & Democracy and Professor of Political Communication, ASCoR, University of Amsterdam. He is President (2020-2021) of the International Communication Association, ICA. Twitter: @claesdevreese.

Media and Communication (ISSN: 2183-2439)

Media and Communication is an international open access journal dedicated to a wide variety of basic and applied research in communication and its related fields. It aims at providing a research forum on the social and cultural relevance of media and communication processes.

www.cogitatiopress.com/mediaandcommunication

COGITATIO