Troubling Matters: Examining the Spread of Misinformation and Disinformation on Social Media During Mass Disruption Events

Ahmer Arif

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Kate Starbird (Chair)

Jennifer Turns (Chair)

David M. Levy

Program Authorized to Offer Degree:

Human Centered Design & Engineering

University of Washington

Abstract

Troubling Matters: Examining the Spread of Misinformation and Disinformation on Social Media During Mass Disruption Events

Ahmer Arif

Chairs of the Supervisory Committee:

Associate Professor Kate Starbird
Department of Human Centered Design & Engineering

Professor Jennifer A. Turns
Department of Human Centered Design & Engineering

Most users want Twitter feeds, Facebook pages, and other information streams to be free of misleading content. Whether this misleading content was spread unintentionally (misinformation) or on purpose (disinformation), understanding and responding to its flows has never been more important. This is especially true during periods of collective stress and uncertainty — like large-scale emergencies, disasters, and political protests — where misleading information can have potentially broad-reaching societal consequences.

This dissertation aims to clarify some of the dynamics of online mis- and disinformation in such settings. It also seeks to provide insights to help researchers and designers formulate more

human-centered responses to the challenges posed by misleading information — i.e. responses that can support human skill and ingenuity rather than rendering them passive. It does this by asking the following questions:

> Research Question 1: How do well-intentioned members of the online crowd, like journalists and ordinary people, understand their own actions when they unwittingly circulate misleading information on social media while using it for sensemaking?
>
> Research Question 2: How do state-affiliated actors opportunistically exploit these sensemaking efforts to spread disinformation?
>
> Research Question 3: How do disinformation campaigns invite their audiences to make sense of the information landscape through a lens of suspicion?

To address the first two questions, I present three studies that give account of how different groups of social media users enacted the spread of misleading information as they participated in collective sensemaking, which is the process by which people build shared awareness around events that disrupt normal routines. Across the studies, I employ a combination of methods, including computational analyses of large Twitter datasets, interviews with social media users, and a qualitative analysis of alternative media websites. I address the third research question by integrating the findings from these separate studies and drawing on the literary theory of postcritique to unpack the interpretive gestures made by the disinformation campaigns examined in this research.

The results of this inquiry yield contributions along several dimensions. Some of the main results illuminate how: i) social media users engage in correcting misinformation and reason about their choices; ii) disinformation campaigns blend their activities with online activism, making their efforts participatory and difficult to isolate; and iii) these campaigns manipulate critical perspectives to foster doubt and division. While making these dynamics more legible, I also draw forward some implications and articulate a design direction to broaden our repertoire of ideas for how we might address mis- and disinformation.

In the larger picture, this inquiry makes empirical and theoretical contributions that can help us think about the appropriateness of how issues pertaining to misleading information are being framed, interventions are being shaped, and social and material possibilities are being constrained down the line.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

DEDICATION

بِسْمِ ٱللهِ ٱلرَّحْمٰنِ ٱلرَّحِيمِ

Bismillah hir-Rahman nir-Rahim
In the name of Allah, the Most Gracious, the Most Merciful.

ACKNOWLEDGMENTS

Any work that brushes against disinformation and conspiracy theories must inevitably confront the claim that "everything is connected." This is no more evident than in the writing of this dissertation, since it seems that every interaction I have had over the course of this work has had some hand in shaping its outcome. These interactions came in many forms, from the direct influences to the brief moments — a turn of phrase, a question, a blend of ideas, a tone. I am deeply grateful to all of them and to the institutions and spaces that made them possible.

The seeds of this dissertation were planted at the University of Washington's Department of Human-Centered Design & Engineering, a safe and warm home to study and do interdisciplinary research. So, first and foremost, thanks must go to my mentors who welcomed into this little corner of the world. They have shaped my work in profound ways and in the process become close friends. Kate Starbird's tireless efforts to support me and teach me concrete research practices were the bedrock I stood upon to do this work. Jennifer Turns' sagacious and generous guidance has been a major influence on my intellectual development, teaching me how to *have* challenging experiences during this project rather than be *had* by them. David Levy's intellectual and spiritual strength instilled me with the courage to do this work with a certain existential integrity — to find my way back when I was tired, when things would spin out of control. Simply put, these teachers did not just hone my ideas in this project, but have, each in their own way, welcomed me as family. They have continuously invested in me as a person and taught me how to be a more reflective, resilient and ethical lifelong learner. I am truly blessed to have had their mentorship.

By the same token, I am grateful to my larger network of teachers. Megan Finn for being my graduate student representative, for her invaluable feedback, and for helping to clarify the political stakes of my work. Ricardo Hidalgo for gently leading me to look more clearly and simply at the nature of experience, without any attempt to change it; this helped me write this dissertation with a greater acceptance of uncomfortable feelings. Naveed Arshad for giving me the opportunity to practice research when I was an undergraduate and for encouraging me to pursue this degree. Saeed Ghazi for spending countless evenings enthusiastically teaching me about the joys of literature, helping me explore what it is to be human; those conversations have exerted a persistent influence on this dissertation. Ashraf Iqbal for shaping me as an educator and showing me what it means to profess truths even when they are inconvenient. Yasser Hashmi for looking out for my friends and me, for his way of combining caring with a wry sense of humor, and for helping me understand that I didn't lack ambition as a student so much as have ones that are not easily recognized. Aasim Sajjad for his activism and contributing to my political outlook. Going much

further back, Naseerah aunty (for I know her no other way) for being my nursery schoolteacher, and for staying invested in my wellbeing to this day, exhorting me to keep on learning. It is a humbling experience to acknowledge the many teachers who have, mostly out of kindness, helped along the journey of my PhD.

Similarly, I am deeply grateful to my colleagues. Leo Stewart for his curiosity, dedication and wonderful warmth as a collaborator, which has significantly influenced not just the empirical research contained in this dissertation but also my motivation to become a better teacher and mentor to others. John Robinson for crafting and maintaining so much of the sociotechnical infrastructure that made any of this research possible. Dharma Dailey for our conversations and for enriching the emComp Lab in some very significant ways, making it feel welcoming of not only different peoples but different ways of knowing — which helped me embrace a broader range of theories and methodological techniques in this work. Tom Wilson for his fastidiousness and good humor, which made studying the disinformation surrounding the White Helmets in the Syrian Civil War not only more constructive but also less alienating. Himanshu Zaid for his excellent questions, which have pushed me to think more deeply about the implications of this work. The entire PJAM crew for their feedback and encouragement. Os Keyes for their deep commitment to human and social transformation, for our terrific conversations, and especially for recommending that I read Rita Felski's book, *The Limits of Critique* along with the Bardzells' work on Humanistic HCI.

More broadly, I owe a tremendous intellectual debt to the 75+ undergraduate and graduate students who came together from departments across the University of Washington to work with me in various research groups. I feel incredibly appreciative and blessed for having had the opportunity to work with fellow learners like Brian Nindo Illa, Charlie Yihao Huo, Elodie Fichet, Fang-Ju Chou, Feven Debela, Gordon Duncan, Jim Maddock, Katherine Van Koevering, Katya Yefimova, Kelly Shanahan, Kit Collins, Liz Crouse, Logan Walls, Nadir Tareen, Niat Emnetu, Ostin Kurniawan, Paul Townsend, Sam Spieth, Stephanie Stanek, Terri Lovins, Vera Liao, Will Sutherland, Yoanna Dosouto, Zena Worku and Ziyue Li. This work would be impoverished without all the questions and contributions made by these and my larger crowd of collaborators. I deeply appreciate all of them as well as the organizations that supported me in recruiting and working with such a rich array of wonderful people — such as the National Science Foundation, the Office of Naval Research, and the University of Washington.

I would like to give special thanks to the activists and volunteers supporting the Black Lives Matter movement and the Syrian Civil Defense Force for the work that they do and the hope they bring to me and so many others. This research would simply not exist without their dedicated. More generally, I wish to give thanks to all my participants (and especially the journalists). When I think of the information flows examined in this dissertation simply as the collective contributions of thousands of human beings, who carry in themselves certain power, consciousness and emotion, then the problems of mis- and disinformation fade. I am simply left with a profound sense of gratitude and the fundamental question of whether our societies are going to find a way to make themselves as healthy — and, if you will, as beautiful — as they potentially can be. From that perspective, I can even give some small thanks to the trolls — not to absolve them from the work that they have done but to recognize the humanity of the digital workers who have slipped into this underground industry due to their precarious conditions.

To my many friends and family, you should know that your support and encouragement was worth more than I can express on paper. To my friends, thank you for listening, making me laugh, and keeping me in good spirits through this entire process. Thanks to my UW and Seattle friends: Andrew Neang, Barbara Day, Brian Gunter, Chris Brummer, Cindy Atman, Cindy Bennet, Del Cross, Elena Agapie, Greg Nelson, Kathryn Shroyer, Kristen Dew, Lucas Colusso, Mania Orand, Melissa Marsh, Rafal Kocielnik, and Shana Hutchinson. The conversations, dinners, and game nights, and general help and friendship are all greatly appreciated. To my friends scattered around the world, thank you for your thoughts, well-wishes, phone calls, texts, visits, and being there whenever I needed a friend. A special thanks to Kit for his editing advice, encouragement, listening skills and willingness to be real. Eddie for all his help with proof-reading, wrangling of ideas, and for being the best college buddy and roommate I could have ever asked for. Noodle for mentally sparring with me over so many games (especially of *Twilight Struggle!)*, and for keeping me laughing and sane as I wrote this dissertation through a global pandemic, forest fires, an apartment flooding and more.

To my family, thank you for encouraging me in all of my pursuits and inspiring me to follow my dreams. I am grateful to my parents, who supported me and wanted the best for me. Thank you to my sister, Sana Arif for paving the way throughout my education and helping me remember how to travel and have fun. Thank you to my father, Arif Anwar for teaching me that my job in life was to learn, to be happy, and to know and understand myself; so that I might thrive and support others. Thank you to my mother, Shamsa Arif for guiding me as a person and for praying for me throughout this process (as well as encouraging myself and others to pray also). Thank you to my parents-in-law for being so generous and supportive, for being there for my parents so that I could rest easier being far away, and for raising Rida.

To Rida, a very special thank you for your love and constant support, for all those late nights and early mornings. Thank you for standing by me through all my travails, my absences, my bouts of anxiety and impatience. Thank you for being my muse, proof-reader, writing buddy, and sounding board. But most of all, thank you for being my best friend. You encouraged me to pursue my dreams, helped me apply to doctoral programs, gave me the courage to fly half-way across the world, built a life for us in the United States (and Canada!), handled every move, discussed my hare-brained ideas, wrote with me, and prevented countless wrong turns. I owe you everything.

Above all, thank you to Allah for granting me the spark of consciousness I used to carry out this work, and for being the source of all these connections. I can never count all of Allah’s blessings but I will strive to be ever grateful.

Our task is to make trouble, to stir up potent response to devastating events, as well as to settle troubled waters and rebuild quiet places.

Donna J. Haraway, *Staying with the Trouble*

Chapter 1. INTRODUCTION

1.1 THE 'PROBLEM' OF MISLEADING INFORMATION

In the past, social media technologies held out the promise of attaining progressive social goals, and of doing so effectively and without discrimination. Platforms like Twitter were positioned by some as "truth-machines" (Herrman 2012) and "self-cleaning ovens" (Frere-Jones 2012) that could harness the "wisdom of crowds" (Surowiecki 2005; see also Gao, Barbier, and Goolsby 2011, 10) at blistering speeds to circumvent the systemic problems of traditional news media. These utopian visions sprang from a persistent faith — etched deeply into the American ethos — that technologies exist outside the frailty of human politics. Since 2017, we can see the flipside of this coin: dystopic worries that map our social ills onto these tools. Concerns have erupted on a massive scale[1] about how these platforms are driving the spread of misleading information[2] and political polarization, both of which are also accentuating each other and simultaneously potentially undermining democratic quality (Tucker et al. 2018). Indeed, Tucker and colleagues (2018, 3) have noted how, within little over half a decade, the Journal of Democracy went from featuring an article on social media entitled "Liberation Technology" (Diamond 2010) to publishing one called "Can Democracy Survive the Internet?" (Persily 2017).

Many organizations are responding to these concerns by designing interventions to address the spread of misleading information. For example, universities like the University of Washington are moving to develop courses like 'Calling Bullshit' (Bergstrom and West 2020) to help students enhance their media literacy while technology companies are developing automated and crowdsourced systems to detect misleading information on larger scales. Several of these interventions have drawn sharp criticism. Verrit, a fact-checking site launched in 2017 and endorsed by Hillary Clinton, for instance, was roundly condemned by critics as "doomed to fail" (Paarlberg 2017) and quickly shut down (Marwick 2018, 475). Facebook's efforts to flag 'fake

[1] According to recent surveys, even though an increasing majority of American adults (67% in 2017, up from 62% in 2016) get their news from social media platforms on a regular basis, the majority (63% of respondents) do not trust the news coming from social media (Barthel and Mitchell 2017). Simultaneously, 64% of Americans say that fake news has left them with a great deal of confusion about current events, and 23% also admit to passing on misleading news stories to their social media contacts, either intentionally or unintentionally (Barthel, Mitchell, and Holcomb 2016; Shearer and Gottfried 2017).

[2] I use the term misleading information here to refer to misinformation, disinformation and other forms of networked manipulation, which can function not only to deceive and create divisions, but also to diminish trust in institutions such as journalism and science. Chapter 2 provides a more careful exploration of these concepts.

news' stories shared on the platform also met with unfavorable attention and the change was quickly retracted (Lyons 2019). Meanwhile, traditional media literacy efforts to 'spot fake news' have been critiqued for counterproductively exacerbating both existing epistemological differences and doubt in our information intermediaries (e.g. boyd[3] 2017a; Marwick 2018, 507; Tripodi 2018, 47).

The spirit of interventionism and debate surrounding misleading information points to a significant, complicated, and not yet entirely understood phenomenon at the intersection of social media, human behavior, journalism, and political propaganda. Considering the potential ramifications at a societal scale, there is a pressing need to stay vigilant and continue to understand how our sociotechnical systems are facilitating the spread of misleading information and what we — as researchers, educators, and designers — might do to address it.

1.2 Misleading Information and Mass Disruption Events

Addressing the spread of misleading information is especially important in the context of mass-disruption events. Mass-disruption events have been defined as "events affecting large numbers of people that cause disruption to normal social routines" (Starbird 2012, 1). These include political protests, natural disasters like earthquakes, and man-made crises such as wars and terrorist attacks. Misleading information is more dangerous in such settings because it can lead people to make harmful decisions.

Numerous research studies attest to the widespread adoption of social media for information sharing and organizing efforts during these events (e.g. Bruns et al. 2012; Lotan et al. 2011; Palen et al. 2010; Starbird et al. 2010). Unsurprisingly, the spread of misleading information has become a major (or at least a much remarked upon) feature of social media activity during such events. Several researchers (e.g. Friggeri et al. 2014; Starbird et al. 2014; Oh, Agrawal, and Rao 2013) have begun to explore this phenomenon in terms of misinformation (i.e. information that is unintentionally misleading). At the same time, the online discourse surrounding these events has also become increasingly politicized and a point of access for a range of actors to spread disinformation (i.e. information that is deliberately misleading). This is reflected, for instance, in reports about the increasing visibility of 'crisis actor' narratives on social media platforms, which claim that various mass shootings and other public tragedies are fabricated events staged by media elites (e.g. Barnes 2018).

[3] No, I did not forget to capitalize danah boyd's name. The lack of capitalization accurately reflects the name of the author in question (boyd n.d.).

Complicating these matters further is how online misleading information (both mis- and disinformation) can actually help manufacture faux mass disruption events. For example, on September 11, 2014, reports of an alleged explosion at a chemical plant in Centerville, St. Mary Parish, Louisiana caused by the militant group ISIS were sent to local residents via text messages and spread through various social media (Kirgan 2014). No explosion had taken place. The texts, tweets, news websites, and eyewitness footage were all part of what journalist Adrian Chen (2015) reportedly traced to a "highly coordinated disinformation campaign," being run by a Russian organization known as the Internet Research Agency. To date, few studies have examined the intersection of misleading information with faux mass disruption events that essentially never took place. Given the potential consequences, comprehensive and evaluative research on how to address misleading information during mass disruption events is increasingly important.

1.3 AIMS AND STRUCTURE OF THIS DISSERTATION

In this dissertation, I examine the phenomenon of online mis- and disinformation in the context of mass-disruption events. I conceptualize the online activity taking place during these events using the concept of sensemaking, which names the work people do to build shared awareness during times of collective stress and uncertainty. I aim to answer several questions as directly as possible.

> **RQ1: How do well-intentioned members of the online crowd, like journalists and ordinary people, understand their own actions when they unwittingly circulate misleading information on social media while using it for sensemaking?** How do they choose to (or not to) correct the information once they realize it is misleading? How do they think that their context (e.g. their situation, social media, the event itself) mediates their ability to recognize and correct misleading information?
>
> **RQ2: How do state-affiliated actors opportunistically exploit these sensemaking efforts to spread disinformation?** What kinds of content do these actors promote? How do they interact with other members of the online crowd?
>
> **RQ3: How do disinformation campaigns invite their audiences to make sense of the information landscape through a lens of suspicion?** What are some interpretive practices that these campaigns call upon to do this work? What are some alternative practices not rooted in suspicion that we might promote in turn to disrupt this work?

To answer RQ1 and RQ2, I provide three studies that give account of how mis- and disinformation spread and engaged social media audiences on Twitter during some significant mass-disruption events. These events include the 2015 Paris Attacks, a faux plane-hijacking, the

#BlackLivesMatter movement[4], and the Syrian Civil War. Across the studies, I employ a combination of methods, including computational analyses of large Twitter datasets, interviews with social media users, and a qualitative analysis of alternative media websites. Integrating the findings from these separate studies, I draw on postcritical theory to theorize about how disinformation campaigns undermine collective sensemaking and to discuss a positive intellectual orientation for resisting these campaigns. I do this theorizing in the form of a humanistic essay that tracks my thoughts as I ask RQ3.

An important aim for asking these three research questions is to explore potentially fresh perspectives on designing for this domain. In focusing on the identification of fresh perspectives, I will be paying deep attention to people's experiences of engagement with information and evaluating it to understand what skills designers could support to help address the spread of misleading information. This emphasis on human skill and learning is a core theme that I will keep returning to throughout this work. For example, I use my descriptive findings about people's emergent practices for correcting misinformation in chapter 4 to discuss what kinds of new literacies we might want to consider designing for. Similarly, my studies of disinformation campaigns in chapters 5 and 6 bring me to a more philosophically oriented analysis of what may be involved in evaluating information (specifically in the experiences of critiquing it) in the interest of generating a new design direction.

This dissertation therefore aims to make two levels or forms of knowledge contributions — empirical and theoretical. The empirical contributions are the findings based on observation and data gathering contained within the three naturalistic studies of how different groups of social media users enacted the spread of misleading information during mass-disruption events. Collectively, these studies help make the structure and dynamics of misleading information more legible as it manifests and spreads through online domains and across social media. This has value because phenomena like disinformation campaigns are of increasing interest to researchers, journalists, formal emergency responders, humanitarian agencies, and the public at large. The information gleaned from these contributions can, for example, provide support for external validity for other investigations and interventions in this area.

The theoretical contributions come through the ways in which the empirical data is organized, juxtaposed, interpreted, and presented to make an argument that reveals something about the dynamics of misleading information. These contributions are tied to using collective sensemaking

[4] Specifically, I examine online discourse about the #BlackLivesMatter movement and police-related shootings in the U.S. during 2016.

and postcritique as a lens through which to see misleading information in online settings. This is useful because it can provide us with new ways of imagining the relationships between misleading information, people, and technology, which in turn can inspire new kinds of research and design interventions around this topic. Of course, observations are always theory-laden, so a rigid separation between the two forms of contributions is impossible. But it can be useful to distinguish between these two ways in which this inquiry makes a contribution — in terms of "this is what happens" (Dourish 2006, 547), and in terms of the ideas it offers for thinking about misleading information.

The rest of this dissertation is organized as follows. The remaining parts of chapter 1 will provide some additional prefatory material that is necessary to understand this inquiry. This material includes descriptions of 1) my positionality; 2) my scholarly motivations for asking the questions that I have formulated above; and 3) the disciplinary perspectives I deploy in this inquiry.

Chapter 2 integrates existing research from human computer interaction, computational propaganda as well as the sociology of disasters to provide background on the key concepts that are used in this research. These concepts include online rumoring, disinformation, social media and collective sensemaking.

Chapter 3 describes the epistemological and methodological underpinnings of this research to minimize redundancy in later chapters and address topics like the role of extant theory in this research. I will also reflect on some limitations of this research in this chapter.

The three studies are then presented in separate chapters (i.e. Studies 1-3 will be Chapters 4-6). Each chapter includes background, methods, and findings for that study, as well as short discussions that point to the larger implications of the research.

Chapter 4 (Study 1 - *A Closer Look at the Self-Correcting Crowd*) describes different patterns of online rumoring behaviors and interviews with 15 Twitter users that exhibited those behaviors. This provides two viewpoints on the dynamics of online misinformation: an etic understanding of how unverified and false information spreads at the aggregate level; and a more emic one regarding how responsibility for verifying and correcting information can be positioned during breaking news events.

Chapter 5 (Study 2 - *Acting the Part*) describes how accounts affiliated with the Internet Research Agency, a known Russian 'troll farm' utilized Twitter and other platforms to influence a highly charged conversation that linked police violence in America during 2016 to the #BlackLivesMatter movement. This work shows how these disinformation actors cultivated personas that were all but indistinguishable from other participants in these spaces to appeal to the values of both right-leaning anti-#BlackLivesMatter voices and politically left-leaning pro-#BlackLivesMatter groups.

Chapter 6 (Study 3 - *Ecosystem or Echo System*) describes how a diverse network of alternative media websites propagate narratives regarding the White Helmets, a volunteer humanitarian group working in the Syrian conflict zone. By combining network and content analyses, this work illuminates how these websites — which are integrated with Russian government-funded media — draw up a shared narrative that the White Helmets are terrorists and a propaganda construct propped up by Western imperialists.

Chapter 7 builds on some of the key findings of this research. Specifically, it discusses some of the significant results from studies 2 and 3 to elaborate on how disinformation campaigns invite readers to engage with information through a lens of suspicion. By bringing their rhetoric and interpretive gestures into focus, the chapter describes how disinformation campaigns work to disrupt established understandings and entice audiences into their sphere of influence.

Chapter 8 will summarize the conclusions of this dissertation.

1.4 DISPELLING THE GOD-TRICK: INTRODUCING YOUR NARRATOR

What follows is a scholarly analysis that draws on several different disciplines, which I use to construct a certain view of online mis- and disinformation, and then provide some contributions guided by that view. Conventions of Western academic writing makes me present some of this in the third person, invoking an illusion that feminist philosopher Donna Haraway (1988, 581) has called a “gaze from nowhere” and the “god trick”. The illusion lies in how a third-person academic voice rhetorically suppresses writers’ contingent identities and prevents readers from assessing the role researchers play as narrators of their work. One way[5] to dispel this illusion is to disclose myself and my positioning as speaker, so that you can understand my claims in relation to my position, rather than as the voice of some disembodied being.

Here are some relevant details about my positionality and the implications I anticipate they will have on my dissertation. I am an interdisciplinary scholar with a background in computer science and English literature who came to the United States in the mid-2010s from my home country of

[5] I plucked the idea to write my positionality statement in this form from the Bardzells’ book on Humanistic HCI (2015, 8). Their words and ideas have been very helpful to me in general for shaping the outlook and tone of this work.

Pakistan to pursue a doctorate in human centered design & engineering. I initially began studying the spread of misinformation on social media during crisis situations in early 2015 as a member of Kate Starbird's Emerging Capacities of Mass Participation Lab at the University of Washington. I entered this territory because I sensed it was a way for me to draw together different disciplines that I was trained in. This intuition was based on my understanding of crisis informatics, which strikes me as a field that is relatively open to blending computational techniques with the interpretive approaches that are a paradigm contribution of the humanities (e.g. close reading, reader-response theory, deconstruction). An implication of this is that part of my agenda here is to create some fruitful dialog between the humanities and information sciences in the service of understanding mis- and disinformation.

My relationship with the humanities itself has been influenced, and perhaps biased by my study of literature. Had I been, say, a student of law, I might have used different conceptual systems and framings. This influence is reflected, among other ways, through my extensive use of long quotations. I see my use of long quotations as a reflection of my humanistic commitment to allowing sources to speak for themselves and for you to hear the voices I cite. My engagements with literature also inform this research in other ways including my orientation to human growth, my epistemological position (which I will discuss in chapter 3), my attempts to avoid disembodied rationalism, and my attempt to maintain a critical and reflexive stance with the 'big data' I use in my studies.

Another influence or bias is that although I have spent the past 5 years in an interdisciplinary environment, that environment is still firmly situated in a college of engineering and not the humanities. This bias has a key advantage and disadvantage. The disadvantage is that digital humanities were simply not on my radar when I was in Pakistan studying traditional literary figures like Shakespeare and Dostoevsky along with theorists like Aristotle and Harold Bloom. Being less familiar with the works of contemporary thinkers, I cannot properly leverage parallel conversations about misleading information in that area. The advantage is that I am now in a better position to situate my study of misleading information within the fields of human-computer interaction, computer-supported cooperative work, and crisis informatics; hopefully I can build on this to improve the bridges these fields have with the humanities.

My training in computer science also plays into my positionality. For example, without this training I would not have been able to analyze large amounts of sociotechnical system log data to provide insights into human behaviors during mass-disruption events, as mediated by online platforms at scale. This training has also given me the tremendous privilege of working with several large organizations like The World Bank, the UN, Yahoo! and Facebook. These experiences have given me a window into how neoliberal, anti-humanistic philosophies influence many of the people who drive today's emerging technological developments, and who, in turn, shape our ethical lives on a global scale in ways that often bypass our understanding or conscious choice. Among other things, this has created a bias in my research against the idea that social media companies can reliably address misleading information by themselves.

Finally, my life in Pakistan shapes my position as a narrator in profound ways. Twenty-eight years in Pakistan has taught me, in some very existential ways, how liberating ourselves from oppressive societal forces foremost requires people who possess a certain consciousness and ability to see the humanity in others. This has led me towards being very concerned about how mis- and disinformation can be used to stoke the fires of hate and suspicion, and that has influenced how I interpret my empirical data. Simultaneously, the legacy of colonialism and the War on Terror has sensitized me to counter-hegemonic views that are suspicious of western actors like media outlets and non-profits. This is a powerful prophylactic against the notion that 'misleading' information is something that can be determined using binary frames, and it has lent strands to some of the mental knots I will discuss in chapter 7.

Knowing all of this, I have tried to account for my positionality and biases during the research and writing of this dissertation. Much of the scientific research was done in collaboration with multiple authors who participated in the overall interpretive process. I have read up on areas where I know I am weak, and when I import these extant ideas into my research, I strive to respect the traditions they come from. But I am also committed to the humanistic stance that the value of my contributions lies partly in the perspective I bring to the work — and so I draw on that perspective where it seems most fruitful.

1.5 Integration of Prior Work

This dissertation incorporates research that has previously been published across three conference papers. Each publication has multiple co-authors and I am the first author for two of these and second author for one of them. For that study (*Ecosystem or Echo System*), I have requested and been granted permission by the first author, Kate Starbird, to include the work in my dissertation. I helped drive the writing, conceptualization and analysis for each of these projects.

In the interest of honoring the voices of multiple authors, as well as being clear, consistent and transparent, I have chosen to incorporate the original work as-is, changing the font to indicate that you are reading the original work. This change in font will also function to signify the change in narrative tone (that comes from the 'I' becoming a 'we', writing for a CSCW audience, my evolving perspective and abilities as a writer etc.). In the chapters that contain these studies, I will add some additional interpretations (e.g. through prefatory remarks) that pertain to the larger themes and findings of this research. Each of these chapters will cite the original work and in case of any deviations (intentional or otherwise), the previously published work ought to be considered the original and this version a derivative.

1.6 MOTIVATION AND BACKGROUND ON RESEARCH QUESTIONS

There are many reasons that moved me to open this inquiry and formulate its three research questions as I have described them above. I will explain some of the more important developments that led to the identification of these research questions and shaped how I interpret the significance of the three studies that comprise this inquiry.

1.6.1 *Investigating Emergent Social Practices for Correcting Misleading Information*

One of the questions that this inquiry asks is about how well-intentioned people, like journalists and ordinary people, understand their own actions when they accidentally circulate misleading information on social media during mass-disruption events. Asking this question is important in the present moment because lacking the perspectives of the people passing along inaccurate information risks overlooking their agency. This can feed the easy temptation to lay the blame for misleading information solely at the feet of new technologies, as if their impact is independent of our ability to make judgments based on some notion of right and wrong. The reality is that technologies like social media or twitter bots neither actively determine nor passively reveal our ability to judge information, they *mediate* it by conditioning and being conditioned by our abilities and habits. Developing fuller understandings of such processes of mediation means leveraging the complementarity of etic and emic viewpoints — to study not only ‘what things do’ (Verbeek 2005) but also how humans give meaning to these mediations.

A diverse range of scholars are working to understand how human agency and technical affordances mutually shape the spread of misleading information (e.g. Lukito et al. 2020; Marwick 2018; Wardle and Derakhshan 2017). This dissertation contributes to these efforts by focusing on the context of mass-disruption events and particularly ‘corrective behaviors’ — i.e. when people correct misleading information that they encountered or even shared themselves. These behaviors are important because people can make mistakes and change their minds about the credibility of information, especially during fast-moving situations. However, the resulting ‘corrections’ or ‘self-corrections’ that manifest in online settings have remained an understudied topic, which is unfortunate because this overlooks human skills like humility and self-reflection that might be supported and nurtured. Scholars such as Shannon Vallor (2016, 186) and Nick Couldry (2013, 13) have argued — and I agree — that studies of new media practices need to focus not just on people who exhibit patterns of vice but of virtue so that we can learn how *they* might design or modify our sociotechnical systems to be more conducive to civic flourishing than they are at present.

There is another reason for studying and foregrounding 'self-correcting' behaviors as a first step in this inquiry. I believe that opening digital research up to more emic perspectives and appreciating the stories of social media users who have struggled to recognize and correct misleading information can help create some much-needed sympathy. It can be tempting to lay the blame for misleading information squarely on the shoulders of social media users and away from the powerful situational factors that could be influencing their behaviors. This line of thinking can lead us down the path to seeing people who pass along misleading information like rumors or conspiracy theories as people making poor decisions. Too often, the question seems to become "Why did the person decide to pass along misleading information?" which can quickly slide towards "Why did the person make this wrong decision?"

From this angle, inquiring into how people make sense of their own actions matters because doing so helps us attend to cultural differences when approaching 'misleading information'. Researchers are increasingly highlighting the importance of considering cultural differences as part of the complex politics involved in labeling certain kinds of online information as one thing or another. For example, Francesca Tripodi (2018) describes how American conservatives in her study rely on media literacy practices that they were taught in church to critique mainstream media as 'fake news'. Outside the west, Pohjonen and Udupa (2017) point toward the situatedness of online practices in the cultural and political milieus of India and Ethiopia to reject the use of top-down, binary frames that separate online messages into speech that is acceptable and speech that is not. Examining self-corrections, as I do in *A Closer Look at the Self-Correcting Crowd* (study 1) doesn't directly answer questions about cultural differences around misleading information, but it helps reveal something about how 'misleading information' is a fluid category, even for people passing it along. This can support researchers and designers in being more constructive, self-aware, and generous as they engage with the murky phenomenon of misleading information.

1.6.2 *Examining the Entanglements Between Disinformation Campaigns and Other Actors*

One of the aims of this research is to better understand the complicated relationships (or entanglements) that emerge between disinformation campaigns and other members of the online crowd during mass-disruptions events. My motivation to understand these entanglements initially arose out of a desire to embrace more complexity within my own thinking and research. Between 2014 and 2016, I was researching the spread of online rumors during mass-disruption events as a member of the University of Washington's Emerging Capacities of Mass Participation Lab. I encountered problematic content that seemed to be linked to conspiracy theorists and even state-sponsored disinformation campaigns while doing this research. However, when I would analyze this data, I would bracket off such content by consigning it to my 'residual categories' (Star and Bowker 2007) — the bucket of messy things that were not my object of study. With hindsight, I know I was able to do this because I was organizing my thoughts using a binary opposition that my colleagues and I came to call organic/orchestrated. This organic/orchestrated binary refers to the act of perceiving a clear dichotomy between the 'organic' rumoring activity of ordinary citizens, journalists and others vs. the 'orchestrated' activity of media manipulation efforts. This binary provided me a sense of illusory order, and it was my way of naively simplifying the complexity and ambiguity of today's misleading information.

Reality eventually made these simplifications untenable. As my research unfolded, more and more of the datasets I studied to understand the dynamics of 'organic' rumoring activity also ended up revealing 'orchestrated' activity linked to state-affiliated disinformation campaigns. Guided by my teachers, I allowed myself to become more curious about this connectedness. I eventually came to feel that enhancing the integrity of my research requires me to shift away from synthesizing orchestrated and organic activity in opposition and instead, move towards examining their interplay.

A return to history shows that this interplay is not new. For example, Ladislav Bittman ([1972] 1981), a former practitioner of disinformation who defected to the U.S. in 1968 and later became an academic, explained in his memoirs that disinformation campaigns often employ 'unwitting agents' to advance their objectives. Similarly, in World War II Nazi Germany, orchestrated propaganda efforts cultivated a network of neighborhood volunteer propagandists called "The Ring" (Bytwerk 2010; see also Starbird, Arif, and Wilson 2019). Members of the Ring didn't wear a party badge but were friends and neighbors who used their credibility to support a system of "gentle persuasion" that worked to quell dissent and reduce the spread of troubling rumors

(Gellately 2002). The Ring served as a useful component in Nazi Germany's propaganda system by adjusting national narratives to fit local circumstances and by serving as an interpersonal channel for spreading enthusiastic assent. Bytwerk (2010, 113) captures some of the value of this approach by remarking how "Germans knew that party members were obligated to say the right things. It was something else when a shopkeeper or teacher with no obvious party connection made the point". Ultimately these ordinary citizens were implicated in the propaganda system, but the system itself initially emerged to help them feel empowered by allowing them to "participate in the great events of the day" (Goebbels 1933). Looking even further back, Sawyer (1990) makes similar observations in the context of pamphlet propaganda in France during the seventeenth century.

We can learn from these insights in the context of social media. A great deal of research published in the Association for Computing Machinery's digital library thus far tends to focus on isolated techniques or phenomena, such as fake news, trolls, or botnets (Wanless and Berk 2017; Starbird, Arif, and Wilson 2019). This research is invaluable in my estimation because it provides new insights into how human-computer interactions are shaping modern techniques for misinforming others. However, it has become increasingly clear that we also need to understand these techniques in ways that consider social media audiences not only as passive objects of influence, but as active subjects as well. For example, Caroline Jack (2017) points out the need for new typologies and ways of thinking that can appreciate the interrelated concerns raised by different kinds of problematically misleading information:

> The words we have are not perfect. Sometimes they don't fully capture the complexity of current events, in which new media platforms afford new strategies for a range of actors, from individuals to companies to governments, for using information to manipulate, control, or profit from audiences under the guise of informing them...In today's information environment, we may need to modify and qualify the terms we have, or find new metaphors and models that acknowledge the complexity and ambiguity of today's problematic information. (Jack 2017, 13)

Similarly, Claire Wardle and Hossein Derakshan (2017) have put forward a model and comprehensive agenda that acknowledges the need for closer inquiry into how misleading information is moved through online spaces by different agents of "information disorder." Ong and Cabañes (2019) build on this agenda to argue for the value of viewing misleading information as a culture of production. They suggest stepping back from the headline-grabbing politicians, bots and influencers to whom we attribute misleading content, and moving towards understanding their position in relation to the network of paid workers and ordinary people participating in the circulation of misleading information. Parallel arguments are evident in the insightful works of Alice Marwick and Rebecca Lewis (2017); Alicia Wanless and Michael Berk (2017); Gregory Asmolov (2018); and Joan Donovan and Brian Friedberg (2019), who all apply approaches from their diverse backgrounds of media fan studies, cybersecurity, social anthropology, and science and technology studies to consider the participatory dynamics of mis- and disinformation.

I view the contributions offered in this inquiry as adding to and further enriching this growing set of perspectives. For example, by considering the spread of misleading information during mass-disruption events, and within the context of activism, *Acting the Part* and *Ecosystem or Echo System* (studies 2 and 3) allow us to appreciate the connectedness of organic and orchestrated social media from a new frame of reference. To say more will require describing the perspectives I draw on for my research. For now, I wish to turn to the motivations that inform the third aim of this inquiry.

1.6.3 *Describing how Disinformation Campaigns Promote Suspicion*

My third research question aims to help clarify, both theoretically and empirically, how disinformation campaigns promote suspicion[6]. This goal is motivated by my recognition that disinformation is not simply false information. As we will see in later chapters, disinformation can also be viewed as the spread of a cognition of mistrust, one that — to pluck a phrase from Rita Felski (2015, 45) — is promiscuous rather than partisan, capable of attaching itself to a broad spectrum of views. For example, in chapter 6, I will show how a group of websites that appear to promote different ideologies (e.g. anti-imperialist left, alt-right) fused together with Russian government-funded media to promote skepticism about the White Helmets, a humanitarian rescue group operating in Syria. These websites sought to discredit the White Helmets not only by spreading false narratives, but also by encouraging their readers to look behind the group—for its hidden causes, determining conditions, and noxious motives.

Many writers and thinkers have been going head-to-head with these dynamics — consider, for example, the rich interdisciplinary literature on conspiracy theories. However, I am approaching these dynamics obliquely in the interest of teasing out one specific thread and exploring a path that might help us design more human-centered responses to disinformation campaigns. Specifically,

[6] To give a concise and 'just enough' explanation of what I mean by suspicion here, we can turn to a helpful quote from Rita Felski (2015), who summarized an essay written by the psychologist Alexander Shand (1922) on the phenomenology of suspicion. Suspicion can be understood as "an elusive and complex attitude, a secondary emotion composed out of basic affects such as fear, anger, curiosity, and repugnance. It is a sensibility that is oriented toward the bad rather than the good, encouraging us to presume the worst about the motives of others—with or without good cause" (Felski 2015, 37).

I am interested in exploring the idea that disinformation campaigns can abuse the critical sensibilities and perspectives promoted within and outside of academia to disrupt established understandings, and energetically contest the credentials of those who get to decide 'the facts'.

Several journalists and scholars have made observations in support of this idea. For example, Pomerantsev coined the memorable phrase "the menace of unreality" (Pomerantsev and Weiss 2014) to capture how the Kremlin abuses arguments regarding the social construction of knowledge to promote the view that everything is a farce and a sham. Similarly, Rid (2020, 433) briefly alludes to this aspect of disinformation by describing it as a "postmodern intelligence practice", while boyd (2018) empathically argues that critical thinking has become weaponized to "dismantle the very foundations of elite epistemological structures that are so deeply rooted in fact and evidence". That these observations come from diverse lines of work testifies to the many-sidedness of this emerging problem of how suspicion gets mobilized. This research seeks to clarify some of the implications of these observations for designers and to help us understand the appeal and consequences of this suspicion.

The invigorating works of Bruno Latour (and building on Latour, Rita Felski) have been a fruitful resource for me in theorizing about this suspicion by linking it to the interpretive techniques of criticism. Latour (2004) invites us to consider what has become of the critical spirit when French villagers are convinced that the September 11 attacks were a conspiracy and when there exists a whole industry devoted to disproving the moon landings. When arguments about the social construction of truth are used for plausible deniability and to challenge hard won evidence for political reasons. To quote:

> "Maybe I am taking conspiracy theories too seriously, but it worries me to detect, in those mad mixtures of knee-jerk disbelief, punctilious demands for proofs, and free use of powerful explanation from the social neverland many of the weapons of social critique. Of course conspiracy theories are an absurd deformation of our own arguments, but, like weapons smuggled through a fuzzy border to the wrong party, these are our weapons nonetheless." (Latour 2004, 230)

In such circumstances, is developing more and better tools to help users critique information really all we need? Questions like this — raised by thoughtful scholars like danah boyd (2017a) and Mike Caulfield (2018) — have motivated me to examine how disinformation can function to promote suspicion and explore some of the limitations of critiquing information as a response. To be clear, there is no question of giving up such approaches entirely (an impossible scenario in any case). Rather, my hope is to help my readers arrive at more well-considered positions.

Theorizing about this aspect of disinformation in an essay (chapter 7) is also my way of making the doubts and challenges of this research more generative (Locke, Golden-Biddle, and Feldman 2008). For example, to connect disinformation with promoting suspicion risks looking for the hidden causes, determining conditions, and noxious motives of actors who commonly prescribe looking for hidden causes, determining conditions, and noxious motives — and that should give one pause. And sometimes suspicion is warranted in the face of larger structures of economic and political injustice that sustain them. Left unacknowledged, such tensions can become an impediment to reflexive research in this emerging area. By provoking us to think about some of these tensions, I hope to support myself and other researchers in conducting more socially responsible research and design work in this space.

1.7 KEY DISCIPLINARY PERSPECTIVES THIS INQUIRY DRAWS ON

This inquiry integrates insights from several different lines of research. I will now briefly summarize the key perspectives that my work is extending from.

1.7.1 *CSCW*

This research is empirically anchored in three studies that are situated in the field of Computer Supported Cooperative Work (CSCW). CSCW can be viewed as a sub-field of Human-Computer Interaction that explores "the technical, social, material, and theoretical challenges of designing technology to support collaborative work and life activities" (CSCW 2021, n.d.). CSCW orients this research towards approaching online misleading information as a 'social-technical gap' that must be explored and understood before it can be ameliorated. Ackerman (2000) offers the metaphor of the social- technical gap — the gap essentially between what we know we must support socially and what we can support technically — as the central intellectual preoccupation of CSCW. While many other technical and design-oriented disciplines seek to narrow the gap or to bridge it, established perspectives within CSCW suggest that the gap is "where all the interesting stuff happens, a natural consequence of human experience." (Dourish 2006, 436). From this orientation, misleading information in online settings is both a problem to be solved, and a phenomenon to be understood — one that emerges out of collective activity.

To study this collective activity, I draw on some methodological sensibilities that have been established within CSCW. The field currently enjoys a 30-year relationship with various traditions of ethnographic and ethnomethodological inquiry which have been an essential resource in furthering the development of concepts such as situated action, situated awareness, articulation work and invisible work (Blomberg and Karasti 2013). This research integrates some of the basic principles of these methods. These principles include providing a descriptive understanding of phenomena in their natural settings, taking a holistic view, and practicing reflexivity (Blomberg, Burrell, and Guest 2009).

As I mentioned earlier, the three studies that comprise this research are descriptive and interpretive in their methodology (as opposed to, say, correlational, quasi-experimental or experimental). They are foremost concerned with portraying how the work of misinforming occurs in online settings, without attempting to evaluate the efficacy of people's practices. These descriptive understandings however enable the possibility of more interventionist agendas. This research also takes another principle of ethnographic research formed in offline environments — an emphasis on gathering information in the settings in which the activities of interest 'naturally' take place — and applies it to the study of misleading information in large-scale virtual environments, with the necessary adjustments. Related to the emphasis on natural settings in ethnography is the view that activities must be understood within the larger context in which they occur (Blomberg, Burrell, and Guest 2009, 967). This is sometimes referred to as holism (Blomberg and Karasti 2013, 374) which holds that studying an activity in isolation, without reference to the other activities with which it is connected in time and space, can result in limited or even problematic understandings of that activity.

The three studies follow the principle of holism in two ways. First, they attempt to attain a more holistic understanding of the online discourses that they examine by diligently taking the patterns, themes and categories surfaced through qualitative research and situating them within a broader picture of the research environment by using complimentary quantitative techniques such as social network analysis, log analysis and visualizations. Second, the studies draw on structuration theory — as adopted and evolved by CSCW — to better account for how misleading information in online settings is mutually shaped by technological affordances, social structures, and human agency (Giddens 1984; see also Orlikowski and Robey 1991). Within this context, technological affordances are essentially the ways in which social media platforms enable and constrain the actions of their users through things like their interfaces (e.g. the ability to 'follow' another user or 'retweet' their message) and the algorithms that determine what content gets displayed to whom. Social structures comprise both the collective norms that guide our behaviors within these systems, and the virtual communities that come to participate in online discourse. Human agency (the actions we choose to take) is enabled and constrained by these social and technical structures, but these structures are also recursively shaped by the results of human action (e.g. through the development of new features or norms in response to people adopting and adapting technologies in new ways). This view of the co-evolution of technological and social structures, and human action informs how this research situates misleading information within broader ecologies and histories.

Finally, this research draws inspiration from the reflexive character of ethnographic analysis. This reflexive character means that this inquiry is not only ‘about’ misleading information, but equally, implicitly or explicitly, ‘about’ the cultural perspective from which this analysis is written and presented. The question of positionality is particularly important in this inquiry because there is an element of uncontrolled subjectivity surrounding ‘misleading information’. Misleading information cannot be unproblematically extracted from a setting, since ‘misleading’ is a pejorative judgement that is generated by me, the researcher, who is situated as an instrument in that setting. So, in places, this research depends critically on understanding misleading information from the perspective of the people studied, and elsewhere it strives to be explicitly reflective about the nature of my engagement with misleading information.

1.7.2 *Crisis Informatics*

This inquiry builds upon a line of research that connects CSCW to the field of crisis informatics. Crisis informatics can be viewed as a “multidisciplinary field combining computing and social science knowledge of disasters; its central tenet is that people use personal information and communication technology to respond to disaster in creative ways to cope with uncertainty” (Palen and Anderson 2016, 224). The expansion of crisis informatics, like the expansion of social media, has been rapid and diverse (Palen and Hughes 2018, 499). Within the context of CSCW, crisis informatics scholars have contributed new understandings about how technologies mediate collaborative activity among individuals, organizations and society as a whole during mass-disruption events including emergencies (e.g. Qu et al. 2011), large protests (Fichet et al. 2016), and political uprisings (Wulf et al. 2013; Starbird and Palen 2012)[7]. In the course of formulating some of these contributions, crisis informatics scholars have cultivated a rigorous methodology to examine the ‘mass participation’ that takes place on social media in response to arising events. I will elaborate on how my research connects to these methodological innovations in chapter 3.

Crisis informatics provides this inquiry an important frame of reference for examining the information activities that occur after disasters and other breaking news events, and by extension, the activities that give oxygen to misleading information in these contexts. It does this by substantively drawing on disaster sociology research (e.g. Fritz and Mathewson 1957; Dynes 1970; Kendra and Wachtendorf 2003) using an information science perspective to understand how social

[7] Reuter and Kaufhold (2018) have recently contributed an excellent meta-review to recapitulate some of the achievements and future potentials of the field.

media communications serve important tactical, community-building and emotional functions. Relevant here, crisis informatics scholars have built on a nearly eight-decade-long line of research that has repeatedly found that people generate misinformation like rumors to facilitate collective sensemaking, whereby a group of individuals attempt to interpret ambiguous, uncertain, or confusing situations (Prasad 1935; Allport and Postman 1947; Rosnow 1980; Schneider 1987; Rosnow, Esposito, and Gibney 1988; DiFonzo, Bordia, and Rosnow 1994; Bordia and DiFonzo 2004; Starbird et al. 2016). This sensemaking perspective draws attention to how misleading information is joined with our collective desire to reduce stress and uncertainty during mass-disruption events. It suggests that we turn to tools like our phones and social media to hear, pass along, and speculate around misleading information because doing so helps restore our feelings of being prepared through having an understanding of our environment after it has been disrupted.

This perspective, along with the concept of sensemaking that it rests upon, are both central to this inquiry, and I will provide further background on them in chapter 2. They describe the overriding viewpoint I use to contextualize social media activity that spreads misleading information during mass-disruption events. I believe that alternative ways of theorizing this activity might better emphasize important facets of misleading information like current economic and political arrangements (e.g. Penney 2017; Ong and Cabañes 2019). But the perspective of sensemaking that I have inherited through crisis informatics is closely aligned with the goals I outlined earlier in this chapter.

Sensemaking provides this inquiry an initial vantage point from which to conceptualize the complicated relationships that arise between social media, its users, and disinformation campaigns during mass-disruption events. From a sensemaking perspective, social media platforms can be seen as sites of collaborative work where people self-organize to generate and process a flood of data into useful information during mass disruption events. These people, if they share misleading information, can be seen as "good people struggling to make sense," rather than as "bad ones making poor decisions" (Snook 2002, 206-207) because sensemaking shifts emphasis away from decision-making and towards the processes that constitute the meaning of decisions that are enacted in behavior (Weick 1988). Finally, we can view disinformation campaigns as a third set of actors that opportunistically tap into this sensemaking process to further their own interests.

This conceptualization foregrounds the breadth and importance of information work taking place through social media systems that is particular to mass-disruption events. In emphasizing the inherent goodness of people, it also better equips this research to engage with how the technical affordances of social media and people's information sharing and verification practices around these mediums can interact in complex ways to fuel misleading information. It also helps us appreciate the conditions in which the ‘orchestrated’ activity of disinformation campaigns intermingles with ‘organic’ crowds, underscoring the participatory nature of these campaigns.

1.7.3 *Humanistic HCI*

The three studies that comprise this dissertation came out of CSCW and crisis informatics, but in the process of doing this research, a direction for further study emerged that I am exploring using a style or mode of scholarship that’s been characterized as humanistic HCI. Humanistic HCI can be understood as “any HCI research or practice that deploys humanistic epistemologies and methodologies in service of HCI processes, theories, methods, agenda-setting, and practices” (Bardzell and Bardzell 2015, 3). Jeffery and Shaowen Bardzell (2015) have recently used the term to reintroduce the humanities to a general HCI readership, and to create support for a more inclusive and broad reach for humanistic thought within the field. Because of certain strands of humanistic thought employed within this investigation, as well my background and broader motivations, this dissertation inherits some key commitments and practices of humanistic HCI, and it extends them to the study of misleading information.

Humanistic HCI guides my approach to theorizing about how disinformation campaigns sow doubt and division. It has, for example, informed my choice to do this theorizing in the form of an essay rather than a traditional scientific report. In contrast to a scientific report (like my three studies), a humanistic essay “tracks a person’s thoughts as he or she tries to work out some mental knot, however various its strands. An essay is a search to find out what one thinks about something...enacting the struggle for truth in full view” (Lopate 1998, 281; quoted in Bardzell and Bardzell 2015, 67).

The specific strand of humanistic thought that I will deploy in this essay comes from Rita Felski's book *The Limits of Critique* (2015), which sketches out a turn to the "postcritical," set in motion, Felski writes, by Paul Ricouer (1965), Eve Kosofsky Sedgwick's (2003) call to "reparative reading", Bruno Latour’s (2004) influential essay “Has Critique Run Out of Steam?”, and Heather Love's (2010) argument for "close but not deep" reading. Building upon these sources, Felski (2015) argues for the postcritical as a way of reassessing the disaffected stances of both a suspicious hermeneutics ("digging down") and a prevaricating postmodernism ("standing back"). These ideas form the theoretical substrates of my position as I describe the ‘suspicious’ quality of disinformation. I will give a fuller explanation of postcritique in chapter 7 after the reader has understood the results of the three studies in this dissertation.

To summarize this chapter: I briefly introduced the phenomenon of misleading information within the context of mass-disruption events. I then explained the aims and structure of this dissertation, followed by a statement about my positionality. After this, I elaborated on some of the developments that led to the identification of these aims and supplied some reasons for why these aims are important. Finally, I explained how this research draws on CSCW, crisis informatics and humanistic HCI. Let us now move on to develop some of the key-concepts that will be used in this inquiry.

Chapter 2. BACKGROUND: MISLEADING INFORMATION AND COLLECTIVE SENSEMAKING

To answer the questions asked in this dissertation, I first take a step backward to consider the question, how might we understand the processes associated with misleading information on social media? I have several reasons for beginning with this question. Above all, I want to direct readers' understanding of misleading information into a relation with sensemaking. To do so, I will provide some background on social media, sensemaking, and mis- and disinformation. I will explicate the relations between these things to pull forward a set of themes I will call up throughout the inquiry.

2.1 Social Media

The scope of this dissertation is limited in its precision and transferability by the constantly shifting forms and boundaries of the information technologies that help comprise its focus. These technologies, alongside mobile media devices and the encompassing architecture of Web 2.0, have facilitated both the rise of modern digital culture, and the widespread dissemination of misleading information within it. In the early 2000s, Web 2.0 standards helped shift internet content from a collection of relatively static webpages to a more interactive multimedia environment.

Social media sites like Facebook and Twitter dominate this new landscape alongside other corporate giants like Google, Apple and Amazon. Yet the distinctive feature of Web 2.0 is that it is individual users who author and share the vast majority of the videos, photos, stories, hashtags, memes and other forms of content that comprise digital media culture (Kaplan and Haenlein 2010). Equally significant is the collective nature of this information activity. It is these exchanges, between the immense number of content creators and sharers, that imbue these internet-based applications with their present character as global social networks and generate some of the emergent effects that this dissertation grapples with.

The global nature of these networks makes the theorizing I do in this dissertation of a specific kind. It tries to be explicit about its local origins. Thus, the studies presented in this dissertation can be seen as snap-shot stories: capturing patterns of activity in a large-scale online environment (Twitter, for it positions itself as a source for real-time news) at particular moments before zooming in on the doings of a handful of (Western) Twitter accounts around highly situated discourses. In attempting this, I am trying to move my theorizing away from formats that carry universalistic pretensions yet hide the locality to which they pertain.

A related term that I employ in this research is *online crowds*. Other than adding the qualifier of being internet-connected, my use of the term aligns with the Oxford English Dictionary's definition of "A mass of spectators; an audience" (s.v. "crowd, n.3," accessed October 21, 2020). The term also reflects some relationship to Surowiecki's (2005) perspective on 'The Wisdom of Crowds,', where people can be mobilized to work intentionally and unintentionally on some information processing problem. It is also worth noting that the crowd in this work is not necessarily one of millions of participants; in fact, some of the activity studied here occurs in relatively small groups of people. Rather, I tend to use the term online crowds to capture the heterogeneity of actors: including ordinary people, trolls, bots, fake-news websites, politicians, highly partisan media outlets, the mainstream media, and foreign governments, all of which play overlapping—and sometimes competing—roles in producing and amplifying misleading information on social media.

2.2 SENSEMAKING

The literature on sensemaking can offer a useful perspective for examining the information activities that manifest over social media during mass disruption events. The term has been used to describe how individuals collectively make sense of surprising, confusing, or ambiguous situations (Sandberg and Tsoukas 2015). Its use can be traced back to the beginning of the twentieth century (Dewey 1922; James 1890), and since then it has been conceptually explored and integrated within research from disparate fields, including the social psychology of rumor (Schneider 1987), organizational studies (e.g. Choo 1996; Drazin, Glynn, and Kazanjian 1999), communications (Dervin 1983) and human-centered computing (e.g. Klein, Moon, and Hoffman 2006; Ntuen, Park, and Gwang-Myung 2010).

This inquiry particularly draws on Karl Weick's (1995) conception of sensemaking, who studied it during time and safety critical situations. For Weick, sensemaking is the process of social construction that occurs when discrepant cues interrupt individuals' ongoing activity, and involves the retrospective development of plausible meanings that rationalize what people are doing (Weick 1995; Weick, Sutcliffe, and Obstfeld 2005). This process "is (1) grounded in identity construction, (2) retrospective, (3) enactive of sensible environments, (4) social, (5) ongoing, (6) focused on and by extracted cues, (7) driven by plausibility rather than accuracy" (Weick 1995, 17). Let's take stock of some important ideas embedded within this perspective. Inspired by Maitlis and Christianson (2014), I've organized this around some questions about how sensemaking unfolds in practice. With each answer, I explore a core aspect of sensemaking and connect it to the present inquiry.

2.2.1 *How do events become triggers for sensemaking?*

Sensemaking is triggered by cues—such as issues, events, or situations—for which the meaning is ambiguous and/or outcomes uncertain. Such occurrences, when noticed, interrupt people's routines, disrupting their understanding of the world and creating uncertainty about how to act. This uncertainty or confusion is usually caused by a violation of expectations, which can range between feelings "that something is not quite right, but you can't put your finger on it" (Weick and Sutcliffe [2001] 2011, 31), to "cosmology episodes" that occur "when people suddenly and deeply feel that the universe is no longer a rational, orderly system" (Weick 1993, 633). Researchers have outlined how this experience of a violation, or discrepancy is subjective and influenced by a variety of factors like its impact on individual or social identity, and personal goals (Corley and Gioia 2004; Maitlis, Vogus, and Lawrence 2013). If events grab enough attention, they can also trigger 'environmental sensemaking' (Nigam and Ocasio 2010), a process in which actors make sense not only of a triggering event (e.g. a public health emergency), but also of larger socioeconomic systems and their ability to cope with disruptive events (e.g U.S. healthcare).

From this view, social media can be (and have been) interpreted as sites where users gather for collective sensemaking (e.g. Hagar and Haythornthwaite 2005; Palen et al. 2010; Heverin and Zach 2012) during and after mass-disruption events. Qu, Wu, and Wang (2009) report that members of the public now converge via social media to carry out many of the same informational and sensemaking activities that sociologists of disaster like Fritz and Matthewson (1957), and Dynes (1970) documented in the pre-digital world. These include, for example, attempts to contact family or friends, inquiring about the status of places, and making offers of assistance. Social media also increases the scale of these sensemaking activities, allowing more people to emotionally connect with events (Huang et al. 2015) and join in on relief activities. Further, social media can afford new kinds of sensemaking activities by supporting the work of emerging actors like 'voluntweeters' (Starbird and Palen 2011) and 'citizen journalists' (Gillmor 2004) who can provide additional perspectives from the ground. Part of the value of sensemaking lies in sensitizing us to these information activities as part of the complex terrain in which misleading information is produced and circulated.

2.2.2 *What is the role of action in sensemaking?*

Taking actions and seeing what happens next is an integral part of sensemaking. As Weick asserts, "Cognition lies in the path of action. Action precedes cognition and focuses cognition" (Weick 1988, 307). People can take actions as fodder for new sensemaking and to test provisional understandings generated through prior sensemaking. But these actions can alter what people encounter and shape the very situations that provoked sensemaking in the first place. This mutual co-shaping between action and environment during sensemaking is known as enactment, and it is one of the aspects that differentiates sensemaking from more passive (or purely cognitive) acts of interpretation (Maitlis and Christianson 2014, 84).

The idea of enactments— individuals simultaneously shaping and reacting to the environments they face — informs how this inquiry conceptualizes the relationship[8] between social media users and their activity around misleading information. For example, enactments guard this inquiry against presuming that people—social media audiences—are passive news consumers; that misleading information affects all people the same way; and that it has the impact that its creators intended. Rather, it suggests that it is not enough to see how many users of social media were exposed to a misleading tweet or video; we must understand what these users do with it.

The co-shaping of people's actions and environments during sensemaking also directs this inquiry's attention away from individual decision makers to "where context and individual action overlap" (Snook [2000] 2002, 207). The context this inquiry attends to can be characterized as sociotechnical in nature, which includes both the technical affordances of social media and how they amplify or stifle certain types of activities (e.g. self-correcting misleading information), and the hand people have in creating the environments that then constrain them. For example, we know that people are much more likely to be committed to the meanings they have constructed to justify actions they have taken when their actions are public, volitional, and irrevocable (Festinger 1957; Salancik 1977). Many of the ways in which action constrain future sensemaking are heightened during times of mass-disruption. During disasters, for instance, actions can become much more public and irrevocable, narrowing sensemaking precisely when flexibility and improvisation are required (Maitlis and Christianson 2014, 86).

2.2.3 *How is meaning collectively constructed?*

Sensemaking is inherently social because even individuals making sense on their own are embedded in sociomaterial contexts where their thoughts and actions are shaped by the "actual,

[8] Beyond sensemaking, this relationship can be theorized using various models of media effects from communication and media studies. In the interest of not straying too far from sensemaking, I will summarily disclose that my perspective on media effects most closely aligns with the sociotechnical model sketched out by Alice Marwick (2018). This model consists of three parts: first, that people make meaning from information based on their identity, social positioning, and skill set; second, that media messaging is often structured in particular ways to further a variety of agendas—like furthering a political viewpoint; and third, that the materiality of media consumption have particular technical affordances (e.g. the algorithms that drive social media) that affect both meaning-making and messaging.

imagined, or implied presence of others" (Allport 1985, 3; quoted in Weick 1995, 39). Weick (1998) holds up jazz orchestras as a classic example of how people build understandings together: members must listen closely to each other, take turns leading and following, and improvise together in real-time to novel or unexpected performance. In more complex settings, meaning can be highly contested and negotiated among a wide range of actors with different backgrounds and interests.

To understand this plurivocality, a significant stream of research has highlighted the importance of narratives as sensemaking resources. For example, Kaplan and Haenlein (2010) have noted how narrative "framing contests" can develop between peers as they attempt to persuade each other to adopt their perspective. Some scholars have gone further and equated sensemaking with the contestation of narratives, describing narratives as "the primary form by which human experience is made meaningful" (Polkinghorne 1988, 1; quoted in Maitlis and Christianson 2014, 81) and "the preferred sensemaking currency" (Boje 1998, 106). This emphasis on narratives alerts us to another important property of sensemaking:

> In an equivocal, postmodern world, infused with the politics of interpretation and conflicting interests and inhabited by people with multiple shifting identities, an obsession with accuracy seems fruitless, and not of much practical help, either. Of much more help are the symbolic trappings of sensemaking, trappings such as myths, metaphors, platitudes, fables, epics, and paradigms. Each of these resources contains a good story...They explain. And they energize. And those are two important properties of sensemaking that we remain attentive to when we look for plausibility instead of accuracy. (Weick 1995, 61)

This stance on accuracy — born from the recognition that people look for plausibility instead of accuracy in accounts of events and contexts — informs my work in several ways. For example, it influenced[9] my decision to treat narratives and memes as units of analysis in my empirical studies of disinformation (*Acting the Part* and *Ecosystem or Echo System)*. Weick's remark (1995, 61) on obsession with accuracy also informs how I approach definitions of mis- and disinformation, which I will describe further below.

The contestation of narratives has also helped two sensemaking related constructs gain traction. The first of these is *sensegiving*, "the process of attempting to influence the sensemaking and

[9] At the time, the actual decision to do so emerged inductively through conversation with my data, but I must acknowledge the role that extant concepts like sensemaking played in my making, and being comfortable with, such analytic moves.

meaning construction of others toward a preferred redefinition of organizational reality" (Gioia and Chittipeddi 1991, 442). Sensegiving is not simply a one-way process of influence since people can adopt, alter, resist, or reject the sense they have been given. The second construct is *sensebreaking*, defined as "the destruction or breaking down of meaning" (Pratt 2000, 464). While there is less research on sensebreaking, it captures an important part of processes involving sensemaking and sensegiving (Maitlis and Christianson 2014). Sensebreaking can move people to reevaluate the sense that they have already made, to question or doubt themselves, and to reconsider their course of action. It is often a prelude to sensegiving, in which others move to fill the meaning gap left by sensebreaking with new interpretations (Pratt 2000). Sensegiving and sensebreaking can be useful concepts when thinking about the dynamics of disinformation campaigns and their instrumental use of narratives.

2.2.4 *How can sensemaking be organized to reduce mistakes and bounce back from them?*

Efficient, reliable sensemaking is associated with cognitive processes that help people acknowledge and learn to cope with complexity in group settings[10] (Westrum 1997; Thordsen and Klein 1989; Sharma 2008; Weick and Sutcliffe 2006). Westrum, for instance, vividly illustrates this in alluding to "generative" organizations where information is actively pursued, new ideas are welcomed, and failures inspire inquiry, a pattern which he described as a "license to think" (1993, 405) and a "protective envelope of human thought" (1997, 237).

To provide a clearer specification of these mechanisms, Weick and Sutcliffe (2001) observed high reliability organizations (like hospital emergency departments) using the concept of mindfulness (jointly informed by Eastern and Western thinking), seeing it as:

> The combination of ongoing scrutiny of existing expectations, continuous refinement and differentiation of expectations based on newer experiences, willingness and capability to invent new expectations that make sense of unprecedented events, a more nuanced appreciation of context and ways to deal with it, and identification of new dimensions of context that improve foresight and current functioning. (Weick and Sutcliffe [2001] 2011, 42)

[10] This is in line with the proposition that the "order or confusion of society corresponds to and follows, the order or confusion of individual minds" (Thera 1996, 22; quoted in Weick and Putnam 2006, 75).

Weick and Putnam (2006) note how this description builds extensively on Langer's more compact formulation of mindfulness as "a flexible state of mind in which we are actively engaged in the present, noticing new things and sensitive to context" (Langer 2000, 220; quoted in Weick and Putnam 2006, 280). The additional specificity is intended to help extend Langer's (2000) conception of mindfulness to the group level, and to selectively draw on both Eastern and Western views of mindfulness as they converge on organizational issues (Weick and Putnam 2006; Bodhi 2000; Kabat-Zinn 2003).

For Weick and Putnam (2006, 280), mindlessness is characterized by a "reliance on past categories, acting on 'automatic pilot,' and a fixation on a single perspective without awareness that things could be otherwise". These things predispose people to misunderstand and incorrectly specify matters in the course of sensemaking. Mindfulness on the other hand, stabilizes attention, weakens the tendency to simplify events into familiar ones, and helps people stay attuned to unfolding events for longer time intervals which increases the likelihood that they will be able to comprehend puzzling interactions or errors.

This perspective adds awareness of the mind itself as a skill for better sensemaking. In this research, I argue that supporting this skill could be a human-centered pathway for reducing the spread of misleading information. Notably, mindfulness here can be understood as not only an individual practice, but also as a collective process that is a complex mix of human alertness, skill, collaboration, and infrastructure. One implication here is that this process can be supported. For example, Weick and Sutcliffe (2001) propose that mindful infrastructures can help formal organizations be more reliable and adaptive in the face of unexpected events. Such infrastructures can include elements like technology, policy and norms. If we accept that small moments on any scale can cumulate, enlarge, and have disproportionately large consequences as complexity theorists keep telling us, then there might be some value in exploring and accounting for these infrastructures not just for the big 'W' work that occurs within formal organizations, but also the little 'w' work that emerges within distributed, online crowds (e.g. Starbird and Palen 2011; Nagar 2012).

Moving on from sensemaking, we are now more prepared, with the concept's help, to unpack misleading information as a class of things connected to ongoing processes. That is to say, sensemaking pushes us to think of misleading information not only in terms of nouns (mis- and disinformation), but also verbs (mis- and disinforming).

2.3 MISLEADING INFORMATION

This dissertation uses the term ‘misleading information’ as a shorthand for two other terms: misinformation and disinformation. Throughout the dissertation, I will use these two terms separately where appropriate.

2.3.1 *Misinformation*

In accounts provided by philosophers and information studies scholars, misinformation refers to information that is unintentionally misleading (i.e. likely to lead an agent to hold a false belief) (Fox 1983; Dretske 2008; Karlova and Lee 2011; Fallis 2015; Søe 2016). This information can be true, false, unverified or somewhere in between. To understand this, consider the following example drawn from Karlova and Lee (2011):

> After an earthquake hit Japan in March 2011, social media such as Twitter and Facebook, were the only functioning communication tools, and many utilized these tools to tell their friends and family they were safe. On March 14th, a Twitter user tweeted that his friend was waiting to be rescued in the mountain area of Sendai, and asked people to retweet the message as much as possible. The message spread quickly as people retweeted it, and although the friend was rescued the next day, people continued to retweet the message. (Karlova and Lee 2011, 7)

This example highlights both the complexity of contemporary media practices and the limitations of defining misinformation as ‘inaccurate information.’ The original message in this case was true and propagated by people who likely held no intent to mislead anyone. However, due to changes in context, the message became misinformation.

Misinformation from a sensemaking perspective

While there are many species of public misinformation, such as misprints, misinterpretations by journalists and gossip, this dissertation specifically focuses on rumors as misinformation. A nearly eight-decade-long line of rumoring research has repeatedly found that people generate rumors as a resource for sensemaking (Prasad 1935; Allport and Postman 1947; Shibutani 1966; Rosnow, Esposito, and Gibney 1988; DiFonzo, Bordia, and Rosnow 1994). When connected to the concept of sensemaking, a rumor is not necessarily a false claim, but instead represents “an unverified proposition for belief that bears topical relevance for persons actively involved in its dissemination” (Rosnow and Kimmel 2000, 122). Focusing on the process of rumoring, rather than the content of rumors, Shibutani (1966) described rumor as improvised news. He proposed that when information is not available from formal channels, people compensate by informally interpreting the situation. Explicating this point of view, DiFonzo and colleagues explained that rumors arise in circumstances that are ambiguous or cognitively unclear and when explanations from trusted or official sources such as news media or government agencies are not readily available (DiFonzo, Bordia, and Rosnow 1994).

From a sensemaking perspective, rumoring is not inherently a negative activity. Early work by Prasad (1935) detailing rumors surrounding the aftermath of a devastating earthquake in Northern India explained that rumors restored a sense of meaning and a semblance of control within the context of an unfamiliar and anxiety-provoking situation. Several subsequent studies have positioned rumoring as a strategy by which members of a group can act to reduce the loss of control inherent in many undesirable events (Walker and Blaine 1991; Bordia and DiFonzo 2004). Walker and Blaine (1991) suggest that hearing, passing along, and speculating around rumors helps restore "interpretive control" that ties to one's feelings of being prepared through having an understanding of their environment. Aligned with these claims, several researchers have suggested that the act of rumoring also functions to relieve or reduce anxiety (Prasad 1935; Rosnow 1991). Less understood is how social media mediates, amplifies or otherwise alters these dynamics, which is part of this inquiry's motivation, especially for study 1, *A Closer Look at the Self-Correcting Crowd*.

2.3.2 *Disinformation*

Disinformation is information that is deliberately misleading. Again, philosophers of information have argued that disinformation need not be inaccurate or false. While Walczyk et al. (2008) have noted that disinformation allows us to accomplish goals both malevolent and benevolent (such as lying about a surprise party), in practice there are often strong associations between disinformation and malicious intent. One cultural reason for this might be the term's etymology: The Oxford English Dictionary (s.v. "disinformation, n" accessed October 21, 2020) states that the term disinformation bears some relation to what in the Soviet Union of the 1950s was called *dezinformatsiya* —a type of population-scale information campaign based on strategies that go beyond simple deception (Bittman [1983] 1985; Jack 2017). These strategies involved the use of active measures, techniques with hostile intent that include:

> Spreading disinformation, especially with the goals of widening existing rifts; stoking existing tensions; and destabilizing other states' relations with their publics and one another. It also included various types of subversive action, such as, for example, establishing 'front' organizations or financing opposition political movements. (Jack 2017, 9)

Of course, such acts of professional deception were not limited to the Soviet Union. In his invigorating work on the history of disinformation, Rid (2020) notes that, after World War II, American intelligence agencies professionalized covert truthful revelations, forgeries, and outright subversions under the sprawling label of "political warfare":

> Political warfare was deadliest in 1950s Berlin, just before the Wall went up. The Eastern bloc, by contrast, then preferred the more honest and precise name 'disinformation'. Whatever the phrase, the goals were the same: to exacerbate existing tensions and contradictions within the adversary's body politic, by leveraging facts, fakes, and ideally a disorienting mix of both...But just when the CIA had honed its political warfare skills in Berlin, U.S. intelligence retreated from the disinformation battlefield almost completely. When the Berlin Wall went up in 1961, it did more than block physical movement between the West and the East; it also came to symbolize an ever-sharper division: the West deescalated as the East escalated. (Rid 2020, 9)

Disinformation therefore points not only to information that is deliberately misleading, but a style of information campaigning that is associated with tactics developed and deployed by intelligence agencies. It can label both deliberate lies (like those spontaneously told by politicians) and the methodical output of organizations. In this dissertation, I will use the term disinformation campaign, and occasionally, *information operations*[11] to disambiguate between these two meanings where appropriate.

Seen as the methodical output of organizations, disinformation does not have to be false information— at least, not necessarily. Disinformation campaigns can also leverage information based on reality to inflict harm. For example, in 1960, the KGB produced a pamphlet detailing actual lynchings and other acts of racial violence against African Americans in areas like Tennessee and Texas; the agency then translated and distributed these pamphlets in more than a dozen African countries, using a fake African American activist group as a front-organization (Rid 2020). Such examples can illustrate why "an obsession with accuracy seems fruitless, and not of much practical help, either" (Weick 1995, 61). It can be more helpful to attend to the processes underpinning disinformation.

[11] I provide background on this term in both chapters 5 and 6 where it becomes relevant.

Disinformation from a sensemaking perspective

While the concept of sensemaking has often been connected to the development and propagation of misinformation, I have not found literature that makes similar connections to study disinformation. In this research, I move to change this by examining online discourse in *Acting the Part* and *Ecosystem or Echo System* as a process of collective sensemaking, in which participants (including ‘bad actors’) identify, share, question, and discuss information related to police-related shootings and the White Helmets respectively. Through this lens, I bring the participatory dynamics of disinformation campaigns into focus, contributing a more nuanced understanding of the entanglements that exist between these orchestrated, explicitly coordinated campaigns and the emergent, organic behaviors of online crowds.

Sensemaking positions the work of disinforming as a collaborative activity[12]. That is, it foregrounds how disinformation campaigns interact with, and take advantage of, other members of the online crowd as they attempt to gather information and theorize about unfolding events. Disinformation campaigns come to be seen as being less about the one-way transmission of problematic content, and more about the opportunistic exploitation of people’s efforts to enact order into chaos. This improvised ‘exploitation’ can take various forms, like strategically amplifying unintentional misinformation, reinforcing antagonistic narratives, and persuading audiences to become ‘citizen marketers’ (Penney 2017) that advance the campaign’s messages at a grassroots level.

There is another advantage to understanding the work of disinforming in this way. I believe it helps us avoid overstating the effects of disinformation campaigns and highlights the difficulty of measuring those effects using straightforward cause-and-effect models. When seen as the exploitation of sensemaking, disinformation becomes a social activity in which narratives are constructed and shared. Sensemaking suggests that the audiences of those narratives include the speakers themselves[13] and that the narratives are "both individual and shared...an evolving product

[12] I have helped frame disinformation as a collaborative activity before with my colleagues in (Starbird, Arif, and Wilson 2019). The concept of sensemaking was not explicitly brought up in that work, but I wish to flag that research as one more entry-point into some of these ideas.

[13] This raises the question of how practitioners of disinformation are shaped by the work they do. From a historical perspective, Thomas Rid (2020, 455) remarks that not only did disinformation actors often believe their own lies; but that driven by bureaucratic logic, they tended to overstate the value of their own disinformation work. On the contemporary side, Ong and Cabañes (2018) have noted the sense of empowerment that can be sometimes felt by digital workers employed to spread disinformation. Undermining these views is another reason to avoid overstating the potential of disinformation campaigns.

of conversations with ourselves and with others" (Currie and Brown 2003, 565). The raw material of disinformation is made of existing narratives and existing divisions, so causal effects are very difficult to establish. Moreover, these narratives can be resisted or taken forward by people in unexpected ways as they engage in sensemaking. So, it can be useful to think of the architects of disinformation campaigns less as cunning masterminds, and more as improvisational jazz performers (to recall Karl Weick's [1998] compelling example). By highlighting not only the potential[14] of disinformation, but also its limitations, sensemaking can help us avoid expanding and escalating that very threat and its potential.

Finally, sensemaking, along with the sub-concepts of sensegiving and sensebreaking, highlights how disinformation can be less about accurate/inaccurate information and more about influencing the processes we use to *interpret* information. That is, disinformation is not just a matter of content (knowing that something is the case) but also a matter of method, and orientation (knowing how to interpret information or pursue a line of reasoning). This becomes important in Chapter 7, as we consider some of the findings in *Acting the Part* and *Ecosystem or Echo System* to try to imagine what interpretive practices we might promote to disrupt the work of disinformation campaigns.

2.3.3 *A coda about truth and intent*

The definitions of mis- and disinformation offered above are ideal categories and struggle to reflect how reality is often much more complex. For instance, the boundaries between mis- and disinformation as they have been conceptualized hinge on the intent of the agent sharing it. If the intent is viewed to be benign, then the content is misinformation. If the intent is viewed to be malicious, then the content is disinformation. The intent of an agent however, is often unknowable and subject to change — particularly on the internet (Phillips and Milner 2018), where remixing and replication make it extremely difficult to pinpoint where a particular idea or piece of content originated, much less determine the intent of its author.

Similarly, a reason for employing sensemaking and selecting definitions that de-emphasize truth or falsity in favor of context and intent is to guard against the notion that reality is objective, independent, and not socially mediated. Even so, what counts and doesn't count as 'misleading' also largely comes down to one's values and processes for determining what representations of reality are 'accurate'.

[14] Of course, we must also not understate the risks posed by disinformation to influence what is politically real and what is not; to establish certain political agendas for social attention and to contain, channel and exclude others; and to shape our images of political 'others'.

Media historian Caroline Jack (2017) highlights these limitations to persuasively argue that the term *problematic information* works best for the current information ecosystem. This term has the benefits of foregrounding the writer’s positionality and the ambiguous nature of today’s information environment. This dissertation adopts this mindset but employs the term misleading information instead of problematic information to better fit with its specific focus on rumors (which are not always problematic but do mislead) and Russian disinformation campaigns. Beyond acknowledging the subjectivity surrounding these concepts, I will explore some of the highlighted issues and their implications in the next section and the coming chapters.

Chapter 3. METHODOLOGY

In this chapter I give an overview of the methodological and epistemological aspects of this research. I also highlight some of this inquiry's limitations to contextualize the knowledge it constructs. My main purpose here is to orient readers so I will keep most of my comments at a relatively high-level. More detailed information on specific research activities are disclosed in each of the studies where relevant.

3.1 EPISTEMOLOGICAL ASPECTS OF THIS INVESTIGATION

To clarify the scientific, intellectual and moral reasoning undergirding this inquiry, it is useful to sketch out a piece of the epistemological terrain that it operates in. Not only can this serve as a vantage point for evaluating this research, but having a more explicit articulation on the table can also provide some transparency around the tensions present in this work. I embrace a variety of tensions in this research: from remaining disciplined while holding an interdisciplinary stance towards methods; to maintaining a critical and interpretivist engagement with big data while leveraging it; to approaching 'misleading information' without assuming that what is correct and what is incorrect are objective truths.

My internal compass for navigating these tensions is best[15] summarized using feminist standpoint theory[16]. This theory emphasizes that knowledge is situated and produced through socio-material practices. In other words, as Donna Haraway (1988) argues, what we call 'facts' are actually 'artifacts' of scientific inquiry that are enmeshed in a complex web of human politics, needs and values. This investigation is particularly driven by Sandra Harding's (1987) standpoint theory, which carefully lays down the researcher's perspective as one that's enmeshed in a physical place, shaped by interests like concern for one's personal advancement and reinforced through discourses that are created by governments, media, universities etc. By implicating researchers in power dynamics, standpoint theory places normative injunctions to hold researchers intellectually and morally accountable for their practices (Bardzell and Bardzell 2011, 680). At a high-level, intellectual accountability is achieved by avoiding disembodied rationalism, abstractions and claims to neutrality in favor of emphasizing the particulars of a given context, understanding the

[15] Ultimately, I am not aiming to provide a full accounting of my position so much as trying to be reflexive and self-disclosing. What I want to hold up in the end is a kind of 'system of systems' thinking that I use to instantiate my research methodology and negotiate between my various competing commitments.

[16] Like other postmodern positions, it is important to recognize that there are many different standpoint theories. My purpose for invoking the label here is to draw upon the central themes of the science-oriented standpoint theories to characterize my own position.

experiences of marginalized groups and maintaining reflexivity. Similarly, working towards moral accountability involves trying to produce knowledge that is less about controlling or managing people and more about nurturing them (Sprague 2005).

This humanistic and less traditionally scientific view of knowledge production helps me cope with several of the tensions I have named above. For one, it helps me see a multiplicity of methods not as idiosyncratic, but as cumulative. Since feminist standpoint theory views methods as a way of embedding values into the research process and it prioritizes the inclusion of marginalized perspectives, it has historically encouraged ‘a multiplicity of methods’ (Bardzell and Bardzell 2011, 681). It is useful to be more pragmatic and less dogmatic around methodological traditions for this investigation because the challenges presented by large-scale online environments call more traditional strategies for doing research into question (Rotman et al. 2012); and this is especially true in the context of mass-disruption events (Palen and Anderson 2016) and, one could argue, misleading information.

Standpoint theory also orients this investigation to maintain a critical and interpretivist engagement with big data. This matters because I tap into large volumes of sociotechnical system log data across all three studies. As boyd and Crawford (2012) note, doing such work can tempt us to make inaccurate claims towards ‘objective truth’, ignoring the social context behind the numbers, and leave us generally struggling to say something meaningful and ethical with it. By foregrounding the researcher as a person and their subjective filter, standpoint theory helps this investigation maintain an interpretive and critical basis for analyzing and communicating around big data. This same focus on situated, embodied forms of knowledge also helps this research be more politically engaged. Whereas positivistic ideals of objectivity would regard the political as an impediment to good science, feminists such as Harding (1987) have argued that engaging in politics *directs* research and offers new avenues of insight (Bardzell and Bardzell 2011). I rely on this argument extensively throughout this inquiry, and especially when I approach contexts like the Syrian Civil War and the #BlackLivesMatter movement.

Most significantly, feminist versions of objectivity support the integrity of this research by helping me negotiate the tensions that arise from studying ‘misleading information’, while also striving to acknowledge a diversity of unique human perspectives, none of which can claim absolute knowledge authority. The regulative ideal of objectivity is important for this inquiry because making all truth into a matter of contingencies calls up a powerful resistance in me when I study state-sponsored disinformation campaigns. Haraway (1997) eloquently captures something of this resistance while writing about the importance of being able to “talk about reality”:

> So much for those of us who would still like to talk about reality with more confidence than we allow the Christian Right when they discuss the Second Coming and their being raptured out of the final destruction of the world. We would like to think our appeals to real worlds are more than a desperate lurch away from cynicism and an act of faith like any other cult's, no matter how much space we generously give to all the rich and always historically specific mediations through which we and everybody else must know the world. But the further I get in describing the radical social constructionist program and a particular version of postmodernism, the more nervous I get. (Haraway [1997] 2018, 577; quoted in Love 2017, 59-60)

In acknowledging her need to continue to talk about real worlds, Haraway positions herself as a person who has "tried to stay sane in these disassembled and disassembling times by holding out for a feminist version of objectivity" ([1997] 2018, 578; quoted in Love 2017, 60). Her formulation of feminist objectivity provides this research a way[17] to avoid an endless questioning of reality followed by a complete lapse into relativism or strong essentialism. For Haraway, "feminist objectivity means quite simply situated knowledges" ([1997] 2018, 581; quoted in Love 2017, 61); it is about admitting bias as inevitable, acknowledging the limits of reason, and cultivating an "embodied, therefore accountable, objectivity" ([1997] 2018, 588; quoted in Love 2017, 61). In the context of this research, it also means speaking not just of multiplicity (as postmodernists have done for years), but also of how we collectively navigate through this multiplicity, to something stable (which requires trust).

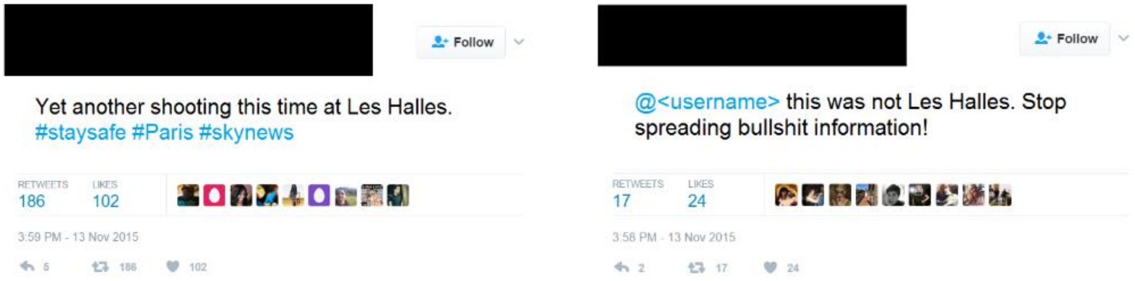
3.2 EMPIRICAL ASPECTS OF THIS INVESTIGATION

In this research, I develop empirically derived understandings of misleading information by collecting social media data based on the 'mass participation' that unfolds on Twitter in response to arising events. I process this data using both computational and interpretive techniques. In doing so, this research expands upon methodological innovation from crisis informatics, which has evolved rigorous techniques to facilitate the rapid collection, storage and network analysis of social media data. I also draw on more canonical qualitative inquiry, specifically by interviewing individuals to understand their interpretations of situations and coding texts. I bring these analyses together through a grounded theory orientation, specifically a constructivist grounded theory approach drawing on the invitation to bring not just textual accounts but also other forms of evidence into the grounded theory process (Charmaz [2006] 2014; see also Glaser 1998). By including quantitative and visual representations as both artifacts for interpretative analysis and

[17] The larger philosophical issues of realism vs. relativism is beyond the scope of this humble dissertation. My engagement with these issues must remain pragmatic because I see my research as a practical activity for now.

methods for stressing and refining my emerging theories, I construct rich grounded theories of the particular situations that I study. Such configurations of work have required me to draw on large teams of individuals, to be able to clearly communicate expectations for specific activities, and to find ways for others to participate in the overall theorizing process.

Let me illustrate how these elements come together by way of example. In *A Closer Look at the Self-Correcting Crowd* (study1, chapter 4), I led a team of 8 student-researchers to construct a grounded theory about how members of the online crowd shaped and corrected rumors after the 2015 Paris Attacks and a possible plane hijacking. I began my investigation from an etic perspective by leading these students to identify particular rumors and specific subsets of tweets related to them from within 10 million+ tweets collected around these larger events. We manually coded 55,011 unique tweets related to two rumors to determine whether they affirmed or denied these rumors (see Figure 3.1 below). I computationally analyzed this coded data to sequentially summarize each recorded user's involvement in these rumors, also identifying tweets they had deleted. I then switched to a more emic perspective by interviewing 15 of these users about their participation, intentions, and reflections on the information space at the time.


[redacted]

Yet another shooting this time at Les Halles. #staysafe #Paris #skynews

RETWEETS 186 LIKES 102

3:59 PM - 13 Nov 2015

[redacted]

@<username> this was not Les Halles. Stop spreading bullshit information!

RETWEETS 17 LIKES 24

3:58 PM - 13 Nov 2015

Figure 3.1**:** Two anonymized tweets — one affirming, and the other denying — a rumored shooting during the 2015 Paris attacks. Such tweets were qualitatively analyzed by me and a group of student-researchers that I supervised (Arif et al. 2017).

A more complex example of my approach is visible in *Ecosystem or Echo System* (study 3, chapter 6). For this research, I led a team of 7 students to analyze 135,827 tweets, leading to the creation of a grounded theory about how a content sharing network of websites micro targeted anti-White Helmets content to different audiences. We started with a collection of Twitter data related to the Syrian conflict, scoped to tweets that contain an explicit reference to the White Helmets. After analyzing these tweets from both high level views (e.g. visualizing network ties and temporal patterns of content production) and close up qualitative engagement with the content (open coding for competing narratives, investigating prominent accounts etc.), we looked beyond Twitter at the surrounding information ecosystems contributing to the White Helmets discourse. To do this, we examined the links embedded in these tweets and extracted 1680 news articles from 270 websites using a tool I built. Again, we blended quantitative and qualitative analyses, constructing a network graph to see larger patterns of content sharing across domains, and then using it for a closer examination of the influential domains within this ecosystem. Throughout this, I engaged my fellow researchers with activities to increase our knowledge of the context we were studying, and reflect on our how own positionalities were shaping the research and how it was affecting us (e.g. I had us interview each other about our reflections and discuss a reading about secondary trauma).

These examples make plain how I have developed empirical insights about misleading information, as mediated by online platforms at scale. They also highlight how I enrich and triangulate these insights by drawing on data like in-depth interviews, and content from alternative news websites, which also help me bound what is and is not visible in the digital record. More broadly, they demonstrate how I repeatedly switch perspectives to do the empirical work of this dissertation. I develop high-level views using visualizations and descriptive statistics to isolate patterns and anomalies and bound my scope and 'unit of analysis.' After making my 'big data' smaller, content analysis helps me understand the nature of those patterns and anomalies, so as to determine whether the decisions were reasonable. Systematic content analysis also lets me generate new hypotheses about underlying patterns in the data and devise new ways of surveying those patterns on larger scales (e.g. by creating a network graph with an unexplored edge property).

In furnishing this explanation of my methodology, I have striven to be direct and to provide the right level of detail—neither becoming pedantic nor, at the other extreme, becoming piecemeal. To further guard against the latter, I will now provide some additional details on how the methodology of this inquiry is influenced by crisis informatics and constructivist grounded theory.

3.2.1 *Crisis informatics*

I inherited the above mixed methods approach from crisis informatics, and specifically from my advisor, Kate Starbird who showed me how this approach can be adapted for studying online rumors and misinformation (Starbird et al. 2014), and then later for studying intentional disinformation (Starbird 2017). This approach is closely aligned with network ethnography (Howard 2002) and trace ethnography (Geiger and Ribes 2011), which articulate similar techniques for following human activities through and across large-scale online environments. Compared to these other traditions, crisis informatics evolved its techniques to closely track the evolution of social media use by the public and emergency responders during mass-disruption events. These roots have influenced my methodology in at least two other ways that I haven't discussed so far.

Events as objects of study: Crisis informatics has influenced how I scope my research by the focus it places on time-delimited events in specific places as the object of study. Crawford and Finn (2015) have traced this focus back to the view of disasters as moments "concentrated in time and space, in which a society, or a relatively self-sufficient subdivision of a society, undergoes severe danger" (Fritz 1961, 655; quoted in Tierney 2007, 505). This view has limitations that I will describe below, but it has helped shape how my social media datasets were operationalized and how I filter analytical considerations down to those most salient to particular periods.

Avoiding the fetishization of social media: As the field of crisis informatics was taking shape, Palen and her colleagues (2010) recognized that social media data can make it possible to study phenomena related to mass-disruption events after the fact and at scale, but they also highlighted the pitfalls of fetishizing this data. They suggested that multidisciplinary, multi-method approaches would allow for more robust uses of big social data. This perspective has influenced my research "to approach big data qualitatively and even ethnographically" (Palen and Anderson 2016, 225), particularly by inspiring me to draw on constructivist grounded theory. The field has also guided how I harness the structural analysis furnished by quantitative methods like social network analysis, log analysis and visualizations to address some of the challenges for qualitative research in large-scale online environments.

3.2.2 *Grounded theory*

Methodologically, this dissertation has made use of a subset of practices from constructivist grounded theory. At a glance, constructivist grounded theory (CGT) has provided this inquiry with a flexible but systematic set of guidelines on how to recursively collect and analyze data from different data sources to describe the dynamics of misleading information. CGT is different from Glaser and Strauss's (1967) original, objectivist version of grounded theory in that it joins standpoint theory in assuming that the world, and grounded theories, are not separate from the observer, but rather constructed by them. Rather than attempt to summarize and recapitulate CGT in the abstract, I will simply highlight some of the specific ways that this methodology has informed this inquiry.

Methodological eclecticism when it comes to data gathering: CGT holds that data collection methods flow from the research question (Charmaz [2006] 2014, 27), and that everything from elicited to extant documents can be seen as a socially constructed source of data. This logic actively encourages creation and innovation around data collection methods at any point in the research process so long as we can advance our understanding around emergent ideas (Charmaz [2006] 2014, 29). This has freed this research to move across different categories of qualitative and quantitative data.

Flexibility in terms of following leads during data gathering: As Muller (2014) notes, emotions like surprise, doubt, curiosity, and passion are reframed as cognitive tools in CGT through a strategy called theoretical sampling. Theoretical sampling engages Charles Peirce's (1960) idea of abductive reasoning and a disciplined focus on engaging in constant comparisons between different pieces of data and theory to give a way to direct further data collection and expound upon emergent ideas. This notion of theoretical sampling has influenced how this inquiry's questions and data collection bent towards disinformation campaigns as my understanding of the phenomenon shifted. Similarly, Charmaz's (2014, 113) conception of theoretical saturation as that point when "fresh data no longer sparks new theoretical insights, nor reveals new properties of these core theoretical categories" partially informed when my data collection and analysis came to an end in my studies.

Qualitative analytical process: The primary bones of my qualitative analytic process derive from CGT's interpretivist method of initial open coding, focused coding and informal memo writing. To be brief, open coding of my collected data focuses on teasing out subtle insights and playing with different meanings to generate ideas that I could cluster and check against the larger picture via focused coding.

Working with extant theories: CGT guides me on how to make proper use of previous knowledge in ways that help me navigate around the tension between 'contaminating' my analysis with external literature and a sort of naïve 'theoretical agnosticism' (Henwood and Pidgeon 2003, 138). Specifically, the social constructivist perspective embedded within CGT suggests that I should guard myself against using existing theory to paint accounts of how participants thought and felt, but that I cannot (and should not) pretend to erase my preconceived conceptions. Much of this goes back to an underlying epistemological position that rejects positivistic conceptions of theory (that separate facts from values) in favor of interpretivist theories that "allow for indeterminacy rather than seeking causality" (Charmaz 2014, 230). Instead, CGT suggests that one can use existing literature and previous experiences as a starting point for working with data. From that perspective, I can use theories like sensemaking and postcritique to help develop sensitizing concepts, contextualize and situate this research.

3.3 Limitations of this investigation

This research confronts several issues and challenges that were not fully overcome. I've highlighted some of these in the studies themselves when relevant, but I believe that disclosing and reflecting upon them up-front can help generate new research ideas and improve how the rest of this work is interpreted. I've organized this non-exhaustive list of issues into three categories: ontological, epistemic and ethical.

3.3.1 *Ontological issues*

The way phenomena are defined and framed influences how data is put to use, and the definitions and metaphors I apply in this research come with some limitations in that regard. For example, I've explained why this research uses the term 'misleading information' to foreground how the distinctions between misinformation and disinformation get drawn up are a subjective enterprise in practice. To make my analyses tractable, I sometimes draw such distinctions without having complete knowledge of the situations I study (e.g. knowing the intentions of people). When I do this, I strive to disclose my methods and the context with enough detail to enable readers to determine whether my process is credible.

A similar issue lies in the use of terms like information operations, disinformation campaigns and dezinformatsiya. These terms are unsatisfying for their connection to militaristic metaphors like 'information war' which risk legitimizing nationalist or nativist sentiments. But they must suffice as the uncertain terrain for my task of broadening our consideration of how population-scale information campaigns use strategies other than simple deception. I encourage readers to reflect upon how such terms highlight particular actors and purposes over others, inviting us to overlook—or flatten through simplistic equivalencies—the complexity and ambiguity inherent to other closely related activities, like advertising and public diplomacy campaigns (Starbird, Arif, and Wilson 2019).

There is also a need to critically reflect on the notion of using 'mass-disruption events' to scope my analyses. 'Mass-disruption' conveniently combines a range of different phenomena — including protests, disasters and wars — and this can make it easier to overlook the long-term reasons for and implications of these things. A number of researchers have also highlighted this risk in seeing physical losses and social disruptions as 'events' bound in space and time (e.g. Burns 2015; Crawford and Finn 2015). For example, Crawford and Finn (2015, 493) point out that "the analysis of social media during and after a disaster can resemble traditional media coverage, which has been often accused of paying attention to only the most sensational stories in a truncated timeframe”. Indeed, I believe the time-delimited[18] event framework I have been influenced by has rendered my work relatively silent on both the structural causes and aftermaths of the affairs I study when compared to research published in related fields like critical disaster studies. I encourage readers to take a longer view of the information work after disasters by looking to the work of scholars like Finn (2018) and Dailey (2020).

3.3.2 *Epistemic issues*

My research relies on sources of knowledge that have constraints. For one, I extensively use public information provided by organizations like social media companies and news agencies to identify the online components of modern disinformation campaigns. For example, I have used Twitter’s reports on information operations in *Acting the Part* (study 2; chapter 5), and I cannot fully account for the methodology they used to identify accounts affiliated with Russia’s Internet Research Agency. Similarly, I do not have direct knowledge that contradicts narratives about the White Helmets that I have judged to be disinformation. Instead, I rely on indirect knowledge such as the fact that no reputable body has ever found the White Helmets involved in, say, chemical incidents in Syria in any capacity other than as first responders to attacks. The use of indirect knowledge is hardly a significant or novel limitation, but it points to assumptions I am making plain in the interest of transparency.

The Twitter datasets I use are also imperfect sources of knowledge that present many[19] epistemological challenges to my work. Addressing these exhaustively is not possible, so I will focus on explaining four significant examples.

- **Polysemy:** In today’s information environment, people can use ‘polysemy’ or coded communication to simultaneously appeal to different, even oppositional audiences (Fiske 1986; Marwick and boyd 2011). A story like ‘Israel Evacuates White Helmets' Members From Southern Syria’ can be interpreted as a report on the precarious conditions the group faces, or it could be used to position the group as an operation conducted by Western intelligence agencies. This polysemy can potentially increase the audience for news stories, in that people with many different political leanings might be motivated to share them. It

[18] It has also been argued that time-delimited views do not provide the context necessary to understand the meaning of, say, a spike in social media activity in a particular community (e.g. Tufekci 2014). However, I believe that my mixed-methods approach and the length of time I’ve spent acquiring contextual knowledge for each of my studies helps address this potential limitation.

[19] The work of boyd and Crawford (2012); Kitchin (2014); and Crawford and Finn (2015) provide some useful starting points to learn about the myriad epistemic issues at play.

can also make the qualitative stories or tweets difficult to analyze without seeing the context in which they are shared. My methodology strives to overcome this limitation by paying attention to more than just content. For example, I try to reconstruct information flows using social network analysis to understand how particular messages travelled over time.

- **Intent:** Related to the issue of multiple meanings is the issue of multiple contexts. To wit, people tweet in cultural contexts that can be particularly difficult for geographically distant researchers to parse. As I've noted earlier, this can make it challenging to infer intent. I've tried to account for this limitation by selecting theories and definitions that de-emphasize intent and by conducting research activities like interviews.

- **Bots vs. Humans:** Automated actors like bots constitute a part of my collected data and they further complicate issues of interpretation because I strive to focus on 'human' activity. This is another ongoing, non-trivial technical challenge, but I have tried my best to address it by manually analyzing individual social media accounts that seemed significant to my analysis. For example, in another study (Stewart et al. 2017) that directly informs *Acting the Part*, my close analysis of accounts helped reveal a set of information activists that occasionally scheduled their activity via a website—which resulted in many of their accounts appearing to be 'cyborgs,' at least partially operated by machines. Such observations have sensitized me to the problems of drawing up simple distinctions between human and automated activity.

- **Demographic skew:** The social media data that I use is partial and incomplete due to factors such as my ability to only analyze messages in English; API rate limits; and the terms used to scope collection results. Even if these limitations were absent, it is important to remember that the platform's users skew towards younger, more urban demographic groups, even in wealthy western nations like the US. Consequently, this means older, and more vulnerable communities are less likely to be self-representing on the platform. This highlights the need to avoid drawing universalistic conclusions from my studies. It also highlights a direction of future research that draws on other data channels to triangulate this inquiry's findings.

These issues notwithstanding, social media data can provide useful insights into phenomena that have historically been very challenging to study — like rumors and disinformation campaigns. That is why I use Twitter data and try to pay care and attention to what is being represented in that data, and how it carries multiple contexts and meanings.

3.3.3 *Ethical issues*

This research manages some specific risks that require me to explicitly reason[20] about the rights and wrongs of my conduct. These risks were assessed to be acceptable at the procedure-level by the Human Research Protection Program's Institutional Review Board at the University of Washington. In the interest of promoting ethical mindfulness, I will highlight two issues that warrant special attention.

First, this research has tried to balance principles of individual privacy against the security and well-being of communities. Respecting privacy is an important part of minimizing harm in this research because passing along misleading information is a socially stigmatized activity. Even though my 'participants' are subjects who have publicly posted information, not all of them have necessarily consented to be a part of my research project. Naming them can cause damage to reputations or create unwanted feelings. This is why I have anonymized participant names for the vast majority of this research. I have also taken additional steps in *A Closer Look at the Self-Correcting Crowd* like altering demographic data, and not including any actual tweets made by my participants. That said, I believe it is important[21] not to absolve those who have intentionally spread misleading information from the work that they have done. So in my studies of disinformation, in a few cases, I have chosen to publish real names for authors who self-identify as academics and journalists.

Second, my colleagues and I have had to pay attention to our psychological and physical safety to do parts of this work. In the same way that oils splatter on the painter's shirt or dirt gets under the gardener's nails, research on disinformation and disasters has an impact. I've come to appreciate that when the sources of anxiety go unrecognized, the anxiety cannot be managed. When that is the case, a range of possible emotions, ideas, and behaviors can show up to indicate that the work is taking a toll. To mitigate these harms, I've tried to make this research collaborative, to promote reflexivity (so that we can pay attention to how our work affected us) and learn from research on vicarious trauma. On a physical register, I have learned that studying disinformation can attract harassment and unwanted attention. I've tried to maintain vigilance to mitigate this potential harm, and this too is a collaborative activity since it requires looking out for, and relying upon other

[20] The ethical framework that most visibly informs my research is Shannon Vallor's (2016) articulation of technomoral virtues. I also draw on Peter-Paul Verbeek's (2011) book, *Moralizing Technology*.

[21] Jack (2017, 4) has made an interesting note of how the professional cautiousness of researchers can be exploited to create an imbalance of power: "actors who distribute deceptive or misleading content can do so without facing major threats to their own credibility, while posing potential legal and reputational threats to those who report on or critique them".

members of my research team. I especially relied a great deal on my advisor, Kate Starbird, who stepped up to be first-author on the *Ecosystem or Echo-System* study in order to strategically take on some key burdens and risks in this regard.

These are certainly not the only ethical issues embedded within this inquiry. For example, another issue that consistently concerns me is how to guard this research against participating in the kind of expropriation and dispossession through data that has been described as Data Colonialism (Thatcher, O'Sullivan, and Mahmoudi 2016; Couldry and Mejias 2019) — i.e. the logics of universal data extraction and management of human beings through data. Helping people reflect on their own social media data was a small way for me to push back against this larger systemic issue. But I cannot get into the details of every ethical issue here. We must now turn to the empirical studies that comprise this inquiry.

Chapter 4. STUDY 1: A CLOSER LOOK AT THE SELF-CORRECTING CROWD

Addressing Research Question 1, this chapter offers *A Closer Look at the Self-Correcting Crowd*[22], a study that examines some emergent practices for correcting misinformation on social media. The preliminary direction for this study was set in 2015, at a point in time when online misinformation did not occupy the same position in public consciousness as it does today. Although emergency officials and members of the press had called out the threat of misinformation on social media (e.g. Madrigal 2013; Hilt, Kushma, and Plotnick 2014; Zeier and Perez 2016), there was still some optimism that social media platforms might function like “truth machines” (Herrman 2012), where the collective intelligence of large-scale online crowds could be harnessed by algorithms to weed out falsehoods. In research on disaster-related communication, this optimism was reflected in, for instance, the hypothesis that rumors could be automatically detected based on social media crowds’ potential to self-correct non-factual content (Mendoza, Poblete, and Castillo 2010).

The original motivation for this project included critically examining[23] this vision of self-correcting crowds. The study highlights that conventional understandings of self-correcting crowds, which build on theories of computational collective intelligence, prioritize classifying and measuring behaviors at scale — to extract value at the crowd level. This perspective is not relevant to the larger focus of this research, especially as trends in the diffusion of misinformation have become clearer. Rather, the study argues that this focus on measuring how much crowds are or are not self-correcting risks obscuring important aspects of this context, like the work people *are* doing to correct misleading information within these crowds, and how that work can be supported.

From this human-centered perspective, there is still some value in remembering the idea of self-correcting crowds. It helps us resist the natural tendency to focus upon those for whom social

[22] This study is previously published work. To cite material from this Chapter, please cite this original work as well as the dissertation:

Arif, Ahmer, John J. Robinson, Stephanie A. Stanek, Elodie S Fichet, Paul Townsend, Zena Worku, and Kate Starbird. 2017. "A Closer Look at the Self-Correcting Crowd: Examining Corrections in Online Rumors". In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 155–68. https://doi.org/10.1145/2998181.2998294.

[23] Prior research carried out by my colleagues and I (Starbird et al. 2014; Arif et al. 2016) had found evidence of crowd-correction for various rumors. But in most of our case studies, we observed that the proportion of messages that corrected or denied rumors compared to the volume that affirmed or supported them was relatively small. So I was motivated to identify, question, and assess assumptions about the ‘wisdom of crowds’ in this context.

media appears to be a barrier to civic flourishing, freeing us to focus instead on the relatively few who are striving to overcome this barrier. Let us consider an example (in a similar hue to Vallor [2016, 185]). Among those Twitter users who choose to affirm or deny a rumor about the present Covid-19 pandemic, we can assume a significant number did so only because they anticipated robust agreement from a like-minded audience. Let us set them aside. Who are the minority who ventured to share information, even though they knew they might be challenged? Assume that some were insincere actors like trolls. Set them aside too. Certainly, some non-negligible subset of the original group remains that is trying to put the information space on a path towards health. How did *they* come by the capacities that help them address misleading information on social media? What habits and practices, on or offline, do these people find helpful in their work?

This chapter offers a closer look at the activity of such individuals in the context of correcting misinformation during mass-disruption events. It interprets this activity as one of sensemaking rather than decision making to foreground that these individuals were working at the edge of codified knowledge. The study first analyzes online rumoring activity at scale to map and describe patterns of behavior. It then describes my interviews with fifteen individuals who both passed along and corrected these rumors. In the interviews, I asked people questions about their motivations for participating in the online discourse, their own social media activity, and what they might do differently etc. The study then moves on to describe features of corrective behavior and constructs a preliminary model to discuss how people’s understanding of how social media functions as a sociotechnical system and their sense of responsibility mediates those corrective behaviors. After the study, I reflect on some of its key insights to pull forward some implications that relate to this dissertation’s aim of informing —and perhaps broadening— our perspectives on how we might address misleading information.

A Closer Look at the Self-Correcting Crowd: Examining Corrections in Online Rumors

Published in Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work

Authors: Ahmer Arif, John J. Robinson, Stephanie A. Stanek, Elodie Fichet, Paul Townsend, Zena Worku & Kate Starbird

4.1 ABSTRACT

This paper examines how users of social media correct online rumors during crisis events. Focusing on Twitter, we identify different patterns of information correcting behaviors and describe the actions, motivations, rationalizations and experiences of people who exhibited them. To do this, we analyze digital traces across two separate crisis events and interviews of fifteen individuals who generated some of those traces. Salient themes ensuing from this work help us describe: 1) different mechanisms of corrective action with respect to who gets corrected and how; 2) how responsibility is positioned for verifying and correcting information; and 3) how users' imagined audience influences their corrective strategy. We synthesize these three components into a preliminary model and explore the role of imagined audiences—both who those audiences are and how they react to and interact with shared information—in shaping users' decisions about whether and how to correct rumors.

4.1.1 *Author Keywords*

Social media; social computing; Twitter; rumoring; crisis informatics; imagined audience; folk theories

4.1.2 *ACM Classification Keywords*

H.5.3 [Information Interfaces & Presentation]: Groups & Organization Interfaces - Collaborative computing, Computer-supported cooperative work; K.4.2 Social Issues

4.2 INTRODUCTION

Finding and disseminating information rapidly can be crucial to building situational awareness during periods of collective stress and uncertainty. This is particularly true for crisis events, where activities such as offering support, learning from eye-witness accounts, and checking in with loved ones can help people both physically and mentally (Corvey et al. 2010; Palen et al. 2010). Consequently, social platforms such as Twitter that allow people to quickly communicate with a wide audience have come to be seen as an important medium for collective sensemaking activities during crises.

However, the conditions of disasters can also give rise to another closely related human response—the propagation of rumors. In these cases, the affordances of social media platforms serve to rapidly spread unverified information or even misinformation. This can have negative consequences for the efforts of emergency responders and the general public. Reflecting these concerns, both mainstream media and emergency officials have called out the threat of misinformation as limiting the utility of social media as a source of actionable information (Hiltz, Kushma, and Plotnick 2014; Madrigal 2013).

A growing body of research on the dynamics of online rumoring and detecting rumors on social media attests to the continued interest of disaster scholars in addressing this challenge. One aspect of online rumors that has drawn attention pertains to how these rumors are ultimately 'self-corrected' by the online crowd. For example, Mendoza et al. find evidence suggesting that the online crowd posts more questions or challenges when presented with a false rumor (Mendoza, Poblete, and Castillo 2010). Zhao et al. build upon this idea by examining how rumors might be identified by seeking out and clustering 'enquiry' tweets that express skepticism about a story (Zhao, Resnick, and Mei 2015). Examining this further, Starbird et al. (2014) measured the volume of corrections across the lifespan of several rumors and found them to constitute a relatively small portion of a rumor's overall propagation.

As such, digital traces of the crowd that question, deny or otherwise 'correct' rumors have been a focus of study for some time now. However, most of this research concentrates on leveraging these traces to detect online rumors. Consequently, very few studies have moved past the analysis of digital traces to interact directly with the users who were responsible for creating them.

We argue that this has led to an important gap in the literature when it comes to developing a deeper understanding of how members of the crowd choose to correct online rumors, and the different strategies they use to do this work. To address this, we examine two false rumors from two different crisis events through combined analysis of trace data (tweets) and interviews with 15 individuals who left some of these traces—specifically those who engaged in rumoring and correcting behaviors during these events. In doing so, we are interested in developing a deeper understanding of the drivers of and barriers to different rumor-correction behaviors.

4.3 BACKGROUND AND RELATED WORK

4.3.1 *Social Media use during Crisis Events*

A growing body of research attests to the widespread adoption of social media and the increasingly critical role they are playing in facilitating information-sharing during crisis events—from natural disasters like earthquakes (Acar and Muraki 2011; Gao, Barbier, and Goolsby 2011; Starbird and Palen 2011) to man-made crises such as acts of terrorism (Cassa et al. 2013). Summarizing the breadth of this literature, in the wake of disaster events, people are repeatedly turning to these platforms with several purposes: to share information about their own circumstances, to seek actionable information for their own response actions and about impacts to the people and places they care about, to request and offer assistance, and to seek and provide emotional support to others. Increasingly, to be part of this conversation, emergency responders are integrating social media into their formal work practices as well. Though the capacity of social media to facilitate information-sharing practices has been described, in general, with a great deal of optimism, these platforms also bring new challenges. One commonly noted weakness of social media use in the crisis context is the vulnerability of these platforms to the spread of rumors and misinformation (Hiltz, Kushma, and Plotnick 2014; Hughes and Palen 2012; Starbird et al. 2014).

4.3.2 *Online Rumoring during Crisis Events*

The dangerous potential of online rumors, especially in the context of crisis events, is a popular refrain (Madrigal 2013). Emergency managers cite the fear of misinformation—and the difficulty in discriminating between true and false information—as a barrier to adopting and utilizing social media in their work (Hiltz, Kushma, and Plotnick 2014;). The rumor issue may indeed be directly tied to the affordances of social media—e.g. how they enable extremely rapid information-sharing and re-sharing (Huang et al. 2015) and how re-sharing and re-mixing can cause tweets to lose context (boyd, Golder, and Lotan 2010), making it hard to identify the provenance and assess the credibility of a specific piece of information. However, rumoring in the context of crisis events is not unique to social media or the Internet.

4.3.3 *Rumoring as a Collective Sensemaking Process*

Social psychologists have been investigating the dynamics of rumor spread since at least the 1940s (Allport and Postman 1946). This work has often connected rumors to the crisis context, where information uncertainty and ambiguity along with individuals' anxiety about impacts and potential responses, act as drivers for the development and spread of rumors (Allport and Postman 1946; Shibutani 1966). Shibutani (1966) framed the telling of rumors—i.e. rumoring—as a collective sensemaking process whereby people come together and attempt to make sense of imperfect and incomplete information. In this perspective, rumors are not necessarily false, but can be true, false, or somewhere in between. Additionally, rumoring is not an inherently bad activity, nor is it driven primarily by ill-intentions. Aligned with Quarantelli's (1991) assertions that the majority of people experiencing disasters are pro-social and active, some rumor participants in this context are motivated by the potential of helping others (Huang et al. 2015). Others participate as a cathartic activity—to reduce their own anxiety about the events (Bordia and DiFonzo 2004; Rosnow 1991). Though researchers have explained rumoring as a natural feature of crises, it can be viewed as bad behavior and consequently there may be a social cost to passing along rumors (Rosnow, Esposito, and Gibney 1988).

4.3.4 *Online Rumoring and the Self-Correcting Crowd*

Previous studies have explored the online rumoring phenomenon through its digital traces, both from quantitative (Ratkiewicz et al. 2011b; Rosnow 1991; Spiro et al. 2010) as well as mixed-method and qualitative approaches (Andrews et al. 2016; Oh, Agrawal, and Rao 2013; Starbird et al. 2014). Some of these studies look at rumors generally (e.g. Ratkiewicz et al. 2011b), while others focus specifically on the context of political discourse (e.g. Rosnow, 1991) or the spread of rumors during crisis events (Castillo, Mendoza, and Poblete 2011; Mendoza, Poblete, and Castillo 2010; Starbird et al. 2014). Though much of the research in this space (Castillo, Mendoza, and Poblete 2011; Ratkiewicz et al. 2011b; Starbird et al. 2014; Zhao, Resnick, and Mei 2015) includes a stated goal of developing methods to automatically detect rumors, another common focus is on the causes or motivations of rumor-sharing. In a study based on statistical analysis of manually-coded tweets, Oh et al. explored online rumoring as a process of collective sensemaking and attempted to identify the causes of rumor propagation, showing that unclear information source, personal involvement and anxiety were factors in rumor spread (Oh, Agrawal, and Rao 2013). Tanaka et al. used an experimental study to assess factors related to retransmission of rumors during the aftermath of the 2011 Japan Earthquake, finding that users' determination of the information's importance—but not factors such as perceived credibility—were predictive of intention to pass along a rumor tweet (Tanaka, Sakamoto, and Matsuka 2012). These findings align with other work (Allport and Postman 1946; Huang et al. 2015) suggesting that a motivation to be helpful to others is a factor in spreading online rumors.

Several researchers have also specifically examined online corrections (Castillo, Mendoza, and Poblete 2011; Mendoza, Poblete, and Castillo 2010; Starbird et al. 2014; Zhao, Resnick, and Mei 2015), almost all with the stated aim of rumor detection. Mendoza et al. explore the dynamics of rumors in the crisis context, finding that rumors tended to be questioned or challenged more by the crowd than factual reports. Their work concludes with a hypothesis that rumors could be automatically detected by algorithms using a content-based approach that identifies questioning or challenging tweets. Castillo et al. follow up that work by presenting a machine learning approach for assessing the credibility of tweets (Castillo, Mendoza, and Poblete 2011). Zhao et al. build upon this idea of detecting rumors early in their lifecycle by identifying what they call 'enquiry' tweets—tweets that they define loosely as seeking more information or expressing skepticism about a story (Zhao, Resnick, and Mei 2015). Andrews et al. (2016) examine the role of official accounts in propagating and correcting rumors, demonstrating that official corrections could help to slow or stop the spread of a rumor.

Underlying much of the research in this area is the notion of the self-correcting crowd—a commonly-held perception that the online crowd will identify, challenge, and eventually correct misinformation. Journalists have referred to Twitter as a "self-cleaning oven" (Frere-Jones 2012) and a "truth machine" capable of "savage" corrections (Herrman 2012). This idea builds upon theories of collective intelligence (Lévy 1997; Woolley et al. 2010) and the popularized notion of "wisdom of crowds" (Surowiecki 2005), which claim that collections of individuals can exhibit intelligent behavior in their aggregated activities that exceeds the abilities of any single individual. Mendoza et al. invoke this principle when they claim that the Twitter crowd can act like a "collaborative filter" for information (Mendoza, Poblete, and Castillo 2010). However, subsequent research demonstrates that many false rumors do not get corrected by the crowd—at least not at the rates that they spread (Starbird et al. 2014). Moreover, few studies explore how members of the crowd choose to (or not to) correct online rumors, or the different strategies they use to do this work.

4.4 METHODS

This research is primarily based on fifteen interviews that we conducted with Twitter users who participated in the propagation or correction (or both) of two specific rumors that spread during two distinct crisis events. We draw from two distinct rumors/events not to draw sharp comparisons, but to identify convergent themes across rumor participants—i.e. to identify and articulate strategies and motivations of rumor-correcting behaviors that may exist across rumor and event types. This study is primarily qualitative, utilizing a grounded, interpretivist approach to gather and analyze interview data. However, we employ a range of other methods in this research—e.g. log analysis, content analysis, visual analysis, descriptive quantitative analysis—to identify these behaviors and show how they fit into the broader collective activity of online rumoring.

Our analysis of correction behavior on social media during crisis events begins by capturing digital traces on Twitter. These traces are collected by the research team in real-time using the Twitter Streaming API. We then identify specific subsets of tweets that are related to particular rumors from within these larger event-level collections. Next, we manually classify each unique tweet in each subset as to whether it affirms or denies the rumor. Following this, we generate a behavior pattern or 'signature' for each user according to their rumor affirming/denying actions over time. This allows us to identify individuals who demonstrate different patterns of corrective behavior (for instance, a person who switched from passing along lots of messages that spread the rumor, to ones that correct the rumor). Finally, we interview Twitter users who exhibited different patterns to better understand their motivations and rationales for the rumoring and correcting actions that they took. We unpack each of these steps below.

Step 1: Event Collections and Rumor Scoping

We focus on two false rumors from two crisis events. For both, we captured data using the Twitter Streaming API, executing forward-in-time collections based on keyword search terms selected and curated by our research team.

Case 1: Rumored Hijacking of WestJet Flight 2514

The first rumor is the rumored hijacking of WestJet flight 2514 during the afternoon of January 10, 2015. As the flight was approaching its destination, a flight-tracking site reported that the plane had broadcast a code indicating a hijacking. This rumor soon began to spread on Twitter, as users speculated about the implications of this report—whether or not the plane had indeed been hijacked and, if so, who the culprits were. Eventually, several official accounts including WestJet's became involved in the conversations. In the end, the aircraft arrived in Puerto Vallarta as scheduled and without incident.

Data collection began approximately 20 minutes after the first tweet, at 4:33pm MST on January 10 and stopped at 2pm the next day. We tracked the following terms: westjet, #WS2154, hijack, hijacked, and hijacking. After ending the collection, to reduce noise from the "hijacking" terms, we scoped the rumor to only include tweets that contained at least one other term related to the WestJet event. The rumor-related dataset for the WestJet Hijacking contains 18,506 total tweets. It is limited by the 20-minute delay in initiating the collection, but we did not experience rate-limiting or other data loss during this event.

Case 2: A Shooting at Les Halles during the Paris Attacks

On November 13, 2015, a series of coordinated terrorist attacks took place in Paris and its nearby suburb, Saint-Denis. At 21:20 CET, three suicide bombers struck near the Stade de France in Saint-Denis, after which suicide bombings and mass shootings took place at cafés, restaurants and the Bataclan Theatre. As these events developed and people attempted to make sense of imperfect and often conflicting information, several rumors began to spread. One erroneously identified the Forum des Halles, a commercial center and an iconic Paris location, as an affected location, claiming that there was a shooting there.

We began collecting tweets at 22:37 CET on November 13, more than an hour after the first attacks, and collected more than 10 million tweets over the next five days. The search term list includes dozens of different terms, including Paris, ParisAttack, ParisAttacks, and several other terms related to specific locations that were affected or rumored to be affected—e.g. Bataclan, Stade de France, and Les Halles.

As the event unfolded, researchers identified Les Halles as (first) a potentially affected site and (later) a false rumor. Once the collection was complete, we scoped the rumor to include only tweets that contained the term "Halles." This resulted in 36,505 tweets. Importantly, the Les Halles rumor did not begin to spread until after our event collection was initiated. However, we did experience substantial rate-limiting during its propagation window and several short periods (~1 minute) of data loss. In this paper, which focuses on interview data, we attempt to report within the constraints of these limitations.

Step 2: Categorizing Tweets

We then manually classify every tweet in the rumor subsets as one of five mutually exclusive codes: Affirm, Deny, Neutral, Unrelated, and Uncodable. We code tweets that support or pass along the rumor as Affirm, and tweets that correct or refute a rumor as Deny. The Neutral category is assigned to tweets that relate to the rumor, but do not take a stance on it. Tweets are labeled Uncodable if they contain words that cannot be deciphered by the researchers, including any non-English words. Significant for the research here, we include only English-speaking tweets in our analysis. Prior work describes this coding scheme and process in greater detail (Andrews et al. 2016; Arif et al. 2016, Starbird et al. 2014).

Step 3: Identifying Deletions

After the manual coding process, we then identify tweets that have likely been deleted by their author. Using the Twitter Search API, we execute a "status lookup" for the Tweet ID of each tweet in the rumor subset. If that lookup does not return a tweet, then we label the tweet as likely deleted. (Other reasons for it to be missing include a change in a user's privacy settings or account suspension.) For deleted tweets that are retweets, we attempt to determine (by looking up the original) if the deletion was made by the retweeting user or by an upstream author. For the WestJet rumor, deletion identification occurred ten weeks after the event. For the Les Halles rumor, with the goal of moving quickly to the interview phase, deletion identification occurred four days after the event.

Table 4.1 Descriptions of User Groups and List of Participants WJ = WestJet Interviewee; LH = Les Hall Interviewee

User Group	Behavior	Interviewed
Affirm-only	Users post one or more tweets affirming the rumor.	LH9
Deny-only	Users post one or more tweets denying the rumor.	WJ3, LH11
Affirm-Deny	Users post one or more tweets affirming the rumor and one or more tweets denying the rumor.	WJ1, WJ5, LH1, LH2, LH5, LH6, LH7, LH10

Affirm-Delete	Users post one or more tweets affirming the rumor and then delete one of those tweets.	
Affirm-Delete-Deny	Users post one or more tweets affirming the rumor, one or more tweets denying the rumor, and deleted at least one tweet.	WJ2, LH3, LH4, LH8

Step 4: Generating User Behavior Signatures

Next, we construct a user behavior signature for every user who shared a rumor-related tweet in one of the rumor subsets. We log three kinds of user actions relevant to corrective behavior: affirms, denies, and deletions. We use this log to create a behavior signature for each user that sequentially summarizes their recorded involvement in the rumor. For instance, a user who posted two tweets affirming the rumor, deleted them, and then added a tweet denying the rumor would have the signature "A(Del) A(Del) D" and be assigned to the Affirm+ Delete+ Deny+ Group. Table 4.1 offers an overview of these different user groups with the number of interviewees from each group.

Step 5: Interviewing Diverse Rumor Participants

To better understand how Twitter users experience and reflect upon their rumoring and rumor-correcting behaviors, we conducted interviews with people who had participated in one of these rumors. To gain insight into different kinds of user behaviors, we attempted to interview individuals with different types of user behavior signatures. The signatures therefore served as a mechanism for enhancing the diversity of our interview sample.

Interview Recruitment

For recruiting, we selected users from each user behavior group and reached out to them through Twitter. The selection process was random, though we removed selected users from the pool if their account profiles or recent tweets contained abusive or profane content. The initial contact tweet (which was public) was vague—i.e. we did not specifically mention rumoring—and requested follow-up communication through a private channel (DM/email).

Not surprisingly, the overall response rate was low. Of 185 total recruitment attempts, we interviewed fifteen participants. For the WestJet rumor, interviews occurred between three and four months after the event. For the Les Halles rumor, interviews were conducted between five and eight weeks after the event. There was also a significant selection bias in the respondents: those from the Affirm-only and Affirm-Delete groups were far less likely to respond to our interview requests.

Interview Protocol

We completed 15 interviews total (4 from WestJet and 11 from Les Halles). Of these, 8 were men and 7 were women. Surprisingly, though perhaps related to the self-selection bias, five self-identified as journalists. For the Les Halles rumor, four participants lived in Paris—and two were actually living near Les Halles at the time of the attacks.

We conducted in-depth, one-hour interviews with each participant. All were remote, conducted via Skype or phone. All but one were in English. For the final interview, LH11, a native French-speaking member of our research team interviewed the participant in French. Except for LH11, each interview was carried out by two researchers (with a third usually acting as a silent note-taker). All interviews were recorded and transcribed.

We used a semi-structured protocol designed to elicit participants' perspectives on their own corrective behavior. Participants were first asked about how they learned about the event, their impressions of the information space at the time, their motivations for participating in information sharing on Twitter and their intended audience (if any). Following this, we asked questions designed to help them speculate and reflect upon what kinds of things they would do in future events (and why) if they realized they had posted misinformation. At the midpoint of the interview, as a cue for more specific questions, we provided participants with a log of the tweets we had collected and used to identify them for recruitment. We then asked them to talk us through these tweets, to explain their motivations and intentions. Several also chose to browse through their social media history as a memory aid during the interview.

Step 6: Interview Data Analysis

In analyzing the interview data, we took a grounded, inductive approach to help us organize and surface patterns from the data. As a first step, our research team (eight individuals) carried out an open-card sort on the interview transcripts. Each was atomized into individual statements and printed onto a card, then the research team clustered these cards based on similarities. This clustering process yielded multiple emergent categories that were iteratively merged, removed or split based on the research team's perceptions of their usefulness, descriptive power and scope. After discussing and refining these categories, we settled on a small set of salient themes that seemed most relevant to our initial questions about corrective behaviors.

In a subsequent phase of focused-coding, we returned to the original transcripts to identify content related to our refined list of themes. For each thematic category, two researchers went through each transcript looking to identify each instance of that theme. Researchers also generated memos, articulating their ideas for how interview content connected to specific themes and how themes connected to each other. Additional sub-themes emerged during this phase as well.

Ethical Considerations: Identifying Deleted Tweets and Interviewing Rumor Tweeters

We encountered significant ethical challenges around working with deleted tweets as well as recruiting, interviewing and reporting results from online users who participated in a socially stigmatized activity: passing along rumors. For the deleted tweets, we followed the protocol outlined in (Maddock, Starbird, and Mason 2015)—removing from content analysis any tweet that had been deleted once we identified it as a deletion. We made an exception for the deleted tweets from the interviewees, who consented to participate in this study. We also attempted to reduce stigma during the interviews themselves by telling participants that rumors are a natural part of disaster events and, for the Les Halles rumor, that one of our researchers had also shared a rumor-affirming tweet. To reduce the risk of damage to participants' reputations, we have anonymized all usernames, changed some demographic data, and have not included any actual tweets in our reporting. Where we refer to tweet content, we have significantly altered the syntax and structure of the original tweet along with other details like the time, number of retweets, etc. to prevent discovery of its original author.

4.5 Trace Analysis Findings

4.5.1 *Rumor 1: WestJet Hijacking*

The WestJet rumor began with a notification on a flight-tracking website that Flight 2514 was "squawking" code 7500, the code for hijacking. Shortly thereafter, a user took a screenshot of that report and posted it to Twitter. A rumor that the flight had indeed been hijacked quickly began to propagate, as aviation fans and breaking news accounts spread the information to an increasingly wide audience. Peak volume exceeded 400 tweets per minute about thirty minutes after the initial report.

Table 4.2 Deletions by Tweet Type for WestJet

Code Category	Total Tweets	# Deleted	% Deleted	# Actively Deleted
Total	21,057	3662	17.4%	2651
Affirms	8438	2165	25.6%	1394
Denies	8064	852	10.6%	722
Neutral	1013	148	14.6%	140
Uncodable	2551	387	15.2%	299
Unrelated	991	110	11.1%	96

Unlike the majority of Twitter rumors where affirming tweets dominate (Starbird et al. 2014), in the WestJet rumor there are almost as many Denies as Affirms (46% of related tweets vs 48%). Most striking is a dramatic shift from mostly affirming tweets to mostly denying tweets. This shift occurred about 45 minutes into the rumor's lifecycle and immediately after WestJet's official Twitter account began to post rumor-denying tweets—official corrections that were widely retweeted. Previous research suggests that the both the shift and the relatively high volume of denials in this case were related to efforts by the official WestJet account to refute the rumor (Andrews et al. 2016).

Tweet deletions are another interesting feature here. Recent research suggests ~11% of tweets are deleted (Bhattacharya and Ganguly 2016). Aligning closely with that number, when we captured deletion information (two months after the event), 11.1% of Unrelated tweets and 10.6% of Deny tweets were missing. However, 25.6% of Affirm tweets (nearly twice the baseline rate) were missing, and 64% of those appeared to be "active" deletions—i.e. not the result of deletion cascades. This lends evidence to previous claims (Maddock, Starbird, and Mason 2015) that for false rumors, affirming tweets are more likely to be deleted than denying tweets.

In our dataset, there are 8963 users who shared a tweet related to the WestJet rumor (Table 4.3). The two largest groups of users exhibited the Affirm-only pattern (39%) and the Deny-only pattern (28%).

About one-third of rumor participants posted tweets demonstrating a shift from one rumor stance to another (e.g. from affirming to denying). Of these, 1728 users (19% of the total) sent at least one Affirm and one Deny with no deletions, 806 (9%) users sent one or more Affirms and actively deleted at least one of them, and 406 (4.5%) sent at least one Affirm, deleted at least one tweet, and also posted at least one Deny. Deletion was a prominent behavior in this rumor—more than 10% of users who participated in this Twitter rumor deleted at least one of their tweets.

Table 4.3 Accounts by User Behavior Signature

User Behavior Signature	Total Accounts for WestJet	Total accounts for Les Halles
Total	8963	4589
Affirm-only	3476	4097
Affirm-Deny	1728	13
Affirm-Del	806	283
Affirm-Del-Deny	406	3

Deny-only	2547	193

4.5.2 *Rumor 2: Les Halles Shooting during the Paris Attacks*

On November 13, 2015, a series of terrorist attacks took place in and around Paris. Though the attacks were coordinated, they took place at different times and some, like the siege at the Bataclan Theater, lasted for extended periods of time. As events developed, Twitter users responded with first-person accounts, attempts at providing material and social support, and information about the location of the attacks. The individuals we interviewed, who include both locals and remote onlookers, described this time period as one of high uncertainty and anxiety:

> LH7: "It's really hard to convey how little everyone knew. There were so many rumors flying around. Where there were attacks going on. How many attacks there were. Whether or not they were coordinated. Whether or not it was all a hoax or prank. No one knew anything for sure. The only thing people knew anything of at first was the thing that happened at the football stadium. But all the other little things happening in the different parts were kind of hearsay at first… things going around on Twitter about Les Halles, about the Louvre, about so many places in Paris that weren't at all targets as it turned out."

The rumor about Les Halles began at approximately 11 p.m. CET with a French language tweet stating that there was a shooting there. That tweet was highly retweeted, and was followed by a wave of similar messages claiming an attack at that site. Several mainstream media sources helped to spread the rumor through their broadcasts, websites and social media accounts. Europe 1 radio was an early source. France 24, BBC, SkyNews, Reuters, FoxNews and others also helped to spread the rumor in some capacity. The rumor peaked on Twitter about one hour after it began and then experienced a period of exponential decay, functionally disappearing about twelve hours later.

Overall, the denial signature for this rumor is weak—only 4.4% of related tweets were denials. Though small in relative volume (against affirms), the rate of Denies surged slightly just after 12am CET, due to tweets from (and retweets of) a few individuals on the ground near Les Halles. But despite these first-hand denial tweets and the activity around them, the denial rate never surpassed the Affirm rate (as we saw in WestJet). This denying activity does seem to correspond with a drop in the rate of affirming tweets—the tweet rate lost more than 50% of its volume during a 20-minute window when the denial surged. Research (Starbird et al. 2014) suggests that such patterns in volume—where affirms show a dramatic spike and then slowly fade away, and where denials never reach the same peak tweet rate as affirms—are common for rumors during crisis events. In this sense, the Les Halles rumor can be considered more 'typical' than the WestJet rumor.

Table 4.4 Deletions by Tweet Type for the Les Halles

Code Category	Total Tweets	# Deleted	% Deleted	# Actively Deleted
Total	36,505	4131	11.3%	2792
Affirms	4790	621	12.9%	308
Denies	224	6	2.7%	6
Neutral	37	3	8.1%	3
Uncodable	22,177	2296	10.4%	1302
Unrelated	9277	1205	13.0%	1173

We captured deletion information for this rumor four days after the event. At that time, 13% of affirms were missing and 50% of those appeared to be 'active' deletions. Conversely, only 2.6% of denies were missing and all six of those were considered to be active deletions. Though these data are not directly comparable with the WestJet data due to a shorter gap between the event date and the deletion identification, they reinforce the claim that Deny tweets are less likely to be deleted than Affirm tweets.

Table 4.3 provides a breakdown of 4589 users we identified having shared a tweet related to the Les Halles rumor. The vast majority only sent Affirm tweets, with 89% of users in the Affirm-only group and 6.2% in the Affirm-Delete group. Less than 5% of accounts sent a Deny tweet, and most of those were in the Deny-only group. We only identified thirteen users in the Affirm-Deny and three users in the Affirm-Delete-Deny groups. Interestingly, users who affirmed the rumor were much more likely to take the correcting action of deleting a tweet than sending a denial.

4.6 Interview Findings

4.6.1 *Corrective Objectives*

Our analysis of interviews with Twitter users who participated in rumoring and correcting rumors uncovered three different objectives for taking correcting actions: *correcting oneself*, *correcting the information space*, and *correcting another person* (or organization). For each objective, there are different types of actions that can be taken—for example, in correcting oneself, a user can choose to delete a rumor-affirming tweet or to post a correction. Our interviews revealed that, even within a single rumor, some Twitter users employed multiple types of corrective actions and often considered others that they elected not to use. Using the user behavior patterns as an initial guide, and the interviews to unpack those patterns, in this section we describe users' rationale for their corrective actions in relation to the three correction objectives.

4.6.1.1 Correcting Oneself

The first, and perhaps most obvious objective for taking corrective action related to Twitter rumoring is to correct oneself. In this case, a user has posted a rumor-affirming tweet and later becomes aware that the information they shared is either untrue or unconfirmed.

Affirm-Deny: One action a user can take to correct herself is to post a Deny tweet. One way to do this is an *explicit self-correction*, where the user specifically addresses the fact that she shared a rumor-affirming tweet. When asked about their general strategies for correcting themselves after passing along a rumor, four participants stated they would post a follow-up tweet to explicitly acknowledge their error and apologize. However, these explicit corrections are rare in the rumoring data we collected. Far more common among the users we interviewed were *implicit self-corrections*, where after sharing one or more rumor-affirming tweets, a user sends a subsequent tweet that contains information either 1) directly questioning or noting uncertainty regarding information in the original tweet; or 2) implicitly contradicting information in the original tweet.

Affirm-Delete: Another action a user can take after she realizes she posted a rumor-affirming tweet is to delete. A deletion removes a tweet from that user's history and the public timeline. Her followers are no longer able to see it and it will no longer appear in searches. The deletion action therefore can function as a correction of the information space (discussed below) or a self-correcting action.
As a self-correcting action, a deletion without a follow-up correction was considered by some of our interviewees as a form of hiding one's error. Seven shared negative opinions about this strategy, and though we attempted to recruit 52 users with this behavior pattern, only one responded to an interview request, and his recollection of his tweeting patterns suggest he may have had a different pattern.

Affirm-Delete-Deny: Some users choose a two-part correction strategy that involves both deleting the rumor-affirming tweet, and posting a denial tweet. Among our interviewees, four demonstrated this corrective action sequence. Two, including LH4, were self-identified journalists who tweeted during the Paris Attacks.

LH4 was a high-volume tweeter during the event. The account was actually operated by multiple people constituting a "new media" organization. This account posted two tweets related to the Les Halles rumor, early in its lifecycle. The first was an Affirm, the second a Deny. Both expressed uncertainty, calling attention to the ambiguity around this rumor. Then, about 45 minutes later, the account posted this tweet, clearly affirming the rumor and providing (false) evidence:

> "Photo from the shooting at Les Halles in Paris. <URL>"

According to the account operator we interviewed, within ten minutes, she deleted that tweet and posted a correction:

> "That image was not from Les Halles. Our previous tweet has been deleted. Sorry."

This corrective action sequence functions to 1) remove the false information from the broader information space; and 2) to draw attention to the fact that the information has been challenged or corrected. Interestingly, this denial tweet is the only explicit self-correction shared by any of the interviewees in this study. Though there are occasions where it might not be ideal, this action sequence can be considered both altruistic (in terms of removing false information) and honest/transparent (as the user openly admits to her mistake). We return to and build upon those distinctions in a subsequent section of this paper.

Affirm (only): Finally, users can choose to take no action. By far, the most common 'correcting action' across our data set (for those who sent a rumor-affirming tweet) was no action. 54% of users who shared the WestJet rumor and 93% of users who shared the Les Halles rumor sent only affirming tweets and did not take any direct action to correct them. We cannot make assumptions about how many came to know that the information they shared was false, but our interviews suggest that the absence of corrective action does not mean that someone did not become aware a rumor had been challenged or corrected.

One reason that participants gave for not correcting, especially in the case of the Les Halles rumor, was continued uncertainty about the rumor. Unlike the WestJet rumor, for which there was an official correction within an hour of its origin, the Les Halles rumor did not see such a quick or firm resolution. Six interviewees explained that even after many users, including some 'on the ground' in Les Halles, began to tweet denials, they still were not sure about the rumor's veracity. This ambiguity may have discouraged people from correcting. LH1 explained why someone might hesitate or choose not to correct in this situation, "[I would have to be] certain it was a definite false alarm before I would go back and say 'yes it's a false alarm', I think. I would have to be pretty definite."

Other users expressed they did not feel the need to self-correct. LH9 suggested that the burden of false information lies with the consumer. He sent hundreds of tweets (almost all retweets) related to the Paris Attacks. Four of those were affirmations of the Les Halles rumor. He did not delete or correct any of those tweets. He explained that since he was not tweeting this information to anyone in particular, he did not need to correct it. Acknowledging the role of imagined audience, he went on to say that if he had misinformed someone he knew personally, then he would have let them know. In other words, in his view, other people seeing his tweets are responsible for verifying this information for themselves, except for those to whom he has close ties. Implied in this rationale is an argument that downstream users should verify their sources, and that being a close tie is a form of verification.

4.6.1.2 Correcting the Information Space

A second objective for corrective behavior is to correct the information space. Almost all of the interviewees who took corrective action noted that, on some level, their motivations were not necessarily to correct a previous error they had made, but to make sure the information spreading through Twitter was as accurate as it could be. LH2 summed up this orientation, "I was concerned with trying to not allow rumors to spread. I wanted to make a modest contribution in which to clarify what was happening and not allow rumors and misinformation." Four correcting action sequences were associated with this objective.

Affirm-Deny: Several interviewees who exhibited the Affirm-Deny pattern explained their objective as correcting the information space, not themselves. LH7 is an interesting case. She was living near Les Halles during the Paris Attacks. Initially, she was gathering information through Twitter. At around 22:40 UTC, she saw tweets about the Les Halles rumor. She retweeted one of those tweets and then shared her own original tweet stating that there was an attack at Les Halles. Then she left her apartment and went out to verify for herself if that rumor was true. About ten minutes later, she shared two tweets similar to:

> "I'm in Les Halles. People don't understand the reports of a shooting in the area. Police on the scene have left. #ParisAttacks"

Her rationale for the rumor-denying tweets was not related to her previous affirming tweets. Instead, it "...was to correct the information that was out there. In the perspective of someone who is there, rather than the hearsay that was going around Twitter." Due to her location on the scene, she recognized that she had 'ground truth' information to share, and she wanted to use that position to get the best information out.

Affirm-Delete & Affirm-Delete-Deny: Some users also explained deletions of rumor-affirming tweets as a way of improving the information space—i.e. by deleting the tweet, the user takes it out of the public stream. People will no longer be able to see or retweet that post, and previous retweets of the original will be automatically deleted as well. The act of deleting can therefore be viewed as one of trying to stop a rumor from spreading.

LH4, the "new media" account whose rumor behavior is featured in the self-correcting section above, described a nuanced rationale for deletions, and related those directly to the potential impact the rumor-affirming tweet would have on the information space:

> LH4: “We very, very, very rarely delete tweets. It’s pretty much never will we do that, unless we are so worried about incorrect information getting out that we have to delete it. So this was one of those case where we were like, ‘I think we need to delete this because it’s gone really completely blatantly wrong and it’s going to be retweeted a lot because everything we [are] doing [is] getting retweeted a lot.’ So we decided to delete it.”

Similarly, WJ2 described the rationale for deleting in this way, “we found out what actually happened and I was like, alright there is no point in putting false information out there, no point in having everyone see it and come to conclusions. So I got the truth, and I am going to delete the false information that is not real because no need for other people to see it.”

Deny (only): Another user pattern in our data is one of only denial tweets. 28% of users in the WestJet rumor and 4% of users in the Les Halles rumor only posted Deny tweets (according to our data collection and coding). In these cases, a user is clearly not correcting themselves. Instead, they are often trying to contribute to a better information space. Previous research has identified information verification (and rumor challenging) as a core task taken on by digital volunteers during crisis events (Starbird and Palen 2011).

When asked about this behavior, WJ3, who sent four denial tweets of the WestJet rumor, positioned his actions as targeting the information space, not a specific individual.

> WJ3: “No I was not correcting anyone, I was just providing information…. I was not taking aim at the people who were speculating. I was trying to find the best, correct information that I could.”

WJ3 was not specifically correcting another person, but simply trying to share the best available information at the time. Other interviewees (with other user behavior patterns) shared similar sentiments about not directly correcting or challenging other users about their rumor-affirming tweets.

4.6.1.3 Correcting Another User

Direct, explicit corrections of other users were rare in our data set. One rough measure of this behavior is to identify denial tweets that begin with a “@” or a “.@”. This results in only 34 tweets within the WestJet rumor and 10 tweets in the Les Halles rumor. After our first round of interviews, noting an important gap in our participant pool related to this behavior, we purposefully recruited and interviewed a user who shared a Deny tweet in this form, LH11. LH11’s one denial tweet was addressed to a breaking news account:

> “.@<breaking new account> it isn’t at les halles, it’s at the bataclan, stade de France and the place de la République"

LH11 reported that he had been confident in posting this correction because he had firm evidence that this rumor was false—from calling a relative who was at Les Halles. LH11 held to the rationale that correcting others and oneself is the right thing to do if it means 'redirecting people' towards the right information.

4.6.2 *To Delete or Not to Delete*

An important consideration for users who realize that they shared a tweet related to a false rumor is whether to delete that tweet or not. Interviewees described extremely nuanced heuristics for this decision, noting a number of different conditions and factors. One theme that emerged repeatedly was the tension between maintaining an accurate personal tweet history (a reputation concern) and maintaining an accurate information space (an altruistic concern).

Multiple interviewees expressed a reluctance to alter the historical record by deleting their tweets. Their explanations suggested two related concerns: 1) deleting a tweet is a potential method for protecting reputation by hiding an error; and 2) this kind of correcting action is perceived as deceptive and dishonest. LH4 described how her colleagues and she tried to be "100% transparent" and that they "very rarely delete tweets." WJ1 explained that "...you don't want to make it seem like you are deleting your record of what you tweeted previously." And LH6 stated, "As a general policy I am not revising my history."

However, interviewees were also aware that there were cases, especially around rumoring tweets in this context of crisis events, where deleting a tweet might be a better—or more altruistic—strategy.

> WJ3: "There are some people who say you should never delete. I am not that type of person. For some, deleting might be the best practice. The worst thing is to have bad information out there, particularly on Twitter, because any individual tweet not seen in a stream is 'out of context.' If that tweet is seen and you do not see the correction tweet, then that is true."

In the above quote, WJ3 explains that there is danger in not deleting. Because of the way information propagates on Twitter (and other online spaces), simply sending a follow-up denial tweet may not stop a rumor from spreading. If the original affirming tweet is not deleted, it can continue to propagate, e.g. as retweets, and those downstream tweets may not retain a connection to the denial.

Interviewees described a detailed set of criteria upon which they based the deletion decision. One issue was timing. If the author thought the tweet had been out there for a while, they would not delete it. But if they felt it had just recently been sent, and that few people had seen it, they would delete it. A second concern was interactions—i.e. how many retweets or mentions their rumor-affirming tweet received. At first glance, interviewees offered seemingly contradictory heuristics here. Several rationalized that if a rumor-affirming tweet had received a lot of retweets or if it might receive a lot of retweets, then it should be deleted. These "active" tweets, as one participant termed them, are beneficial to delete because retweets will be deleted as well. LH6 expounded, "If I tweeted something that turned out to be erroneous, and it had 40-50 retweets or there is a lot of action happening on it, it would make more sense probably to delete the tweet." LH4 explained that the potential of being retweeted at a high volume, even if it was not yet happening, was a good rationale for deleting a rumor-affirming tweet. On the other hand, LH7 said that if a rumor-affirming tweet had several retweets, it would be a reason *not* to delete it. LH8 took a more relative approach, noting the deletion cascade effect and drawing a distinction between retweets and other kinds of interactions:

> LH8: "If it's retweeted and I delete it, I think it's deleted from all the other feeds. So I would definitely delete it. And if there is a conversation that's a bit tricky [...] If there is a conversation I don't know if I would delete it because sometimes you come to a conclusion that the information was partially wrong or partially right."

A final consideration that was shared by multiple interviewees involved the weight of the error—i.e. how much damage it might do. Three interviewees told us that minor errors, like typos, were okay to delete. At the other end of that spectrum, tweets that could cause major harm, for example to another person's reputation, were also considered okay to delete.

Considered together, these diverse rationales for whether or not to delete a tweet demonstrate how the imagined audience (boyd 2008; Litt 2012), including not just who that audience is and how it will perceive the user, but also how that audience will act upon the information, contributes to a rumor-tweeter's decision on whether and how to correct.

4.6.3 *Locus of Responsibility*

When asked to review and explain their tweeting actions, almost all participants provided rationalizations for why they were not fully responsible for having shared the rumor. Though likely a natural response to the question format, these rationalizations shed light on users' perceptions for how online rumoring takes place and what roles they play in this rumoring. During the initial card sort of our interview data, we noted the salience and diversity of these rationalizations and how they connected to different kinds of behaviors. We identified six distinct—though in places overlapping—perspectives on the "locus of responsibility" for sharing and correcting rumors.

4.6.3.1 Self: Taking Responsibility for Sharing/Correcting Rumors

Some participants expressed a sense of personal responsibility for posting the rumors—blaming themselves for not having verified adequately and noting concern about how their posts may have affected others.

> LH4: "This one was me and I was wrong. I had neglected to notice [how this information was not related to Les Halles]. My [colleague] came back online and said 'Hey, we already tweeted that earlier today. It's [not Les Halles]. I've verified that.' I'm like 'Oh my gosh, I'm so sorry.'"

As LH4's quote shows, when a person draws the locus of responsibility inward, it can be uncomfortable. They may feel ashamed for what they view as an error—and a public one. They may also feel responsible for causing others to see and share the rumor. At least two other interviewees expressed significant distress about their rumoring activity.

> WJ1: "I was left with a sense of anxiety after the whole thing was over, and I feared I had needlessly alarmed or frightened people, and I worry about that event. I feel uneasy about it. I am not sure it is possible to avoid this kind of feeling, but it left me with a feeling of unease."

WJ1 describes his anxiety as being rooted in a sense of responsibility to those who may have been misinformed by his tweet. Here he positions the relevant "downstream" individuals as people who may have had a loved one on the flight, but this sense of responsibility was also seen to apply to other Twitter users who simply read and passed along his tweet—those had been drawn into the rumor by his tweets. This demonstrates the role of imagined audience—in this case a concern for how members of ones' audiences will perceive ones' actions, as well as how those audiences will be affected by those actions.

4.6.3.2 Upstream User: "But I Heard it From <username>"

In many cases, participants were seen to place the locus of responsibility on the source of the information, whether a trusted news source or a friend. For example, LH3 had lived in Paris and had worked with one of the mainstream news outlets there. In explaining why she sent rumoring-affirming tweets about Les Halles, she stated that both were retweets of major news sources whom she trusted. LH5 also noted the role of mainstream news outlets in his rationale:

> LH3: "[I tweeted] because of this person citing this. … I probably should have been vague, but the fact of the matter is that when you see <news outlet> reporting it, I'm like 'OK'. It seemed much more real."

Like LH3 and LH5, interviewees who invoked this perspective deflecting responsibility to the source typically pointed the finger at trusted media outlets. This sentiment, which was shared to some extent by the journalists we interviewed, suggests that online rumor participants have different expectations for journalists compared to other members of the crowd. However, users assigning blame to upstream sources also directed that blame at trusted friends and other accounts they were following.

4.6.3.3 Downstream Users: "They Should Have Verified My Tweet"

A small set of interviewees placed the locus of responsibility on downstream users—their followers and others who might be reading and re-posting their tweets. In their interview responses, they rationalized their rumor-sharing behavior by suggesting that their audience should not be accepting their tweets as fact, but should be verifying this information themselves.

4.6.3.4 Crisis Events: "That's Just the Nature of These Events"

Other interviewees emphasized that rumoring is just a natural part of crisis events. They described their motivations as trying to help other people by getting information out quickly. They noted the uncertainty and ambiguity in the information space and how difficult it is to discern good information from bad, truth from rumor. Two participants highlighted a difficult trade-off: is it worse to pass along this rumor (in the case that it turns out to be false), or to not pass along this rumor (in the case that it turns out to be true)? Often, it seems, the default answer is tweet now and worry later.

Table 4.5 Locus of Responsibility Categories

Locus of Responsibility	**Description**
Self	Takes responsibility. Likely to correct.
Downstream Users	Says people should verify info themselves.
Upstream Users	Says they trusted source who got it wrong.
Crisis Events	Just the nature of crisis events.
Twitter	Affordances of the platform lead to rumors.
Crowd	Says the "crowd" should/will correct.

4.6.3.5 Twitter: "That's Just the Nature of Social Media"

Several interviews pushed some responsibility onto the platform mediating the rumor. There was a common perception across almost all of the interviewees that the real-time nature of Twitter was both a huge advantage for it as a place to seek information during disaster and a major contributor to the spread of false rumors. For some, this awareness was a cautionary point, something that one should take into account as they participate. But for others, this perspective could lead to more of an acceptance that these are just the limitations of Twitter, that you cannot expect to it be what it is not. LH8 touches on this:

> LH8: "I think most people, like me, they trust TV more than Twitter because when you're on Twitter you know that people post things that they have not checked before. That's why being a journalist is a job, because you checked your sources first, and this is not the case for Twitter. But when you know this, it is fine."

For some interviewees, this attitude included a hint of resignation and an abdication of responsibility. However, for the journalists in our set, this came with a new set of competing responsibilities. All were quite reflective about the challenges of balancing journalistic expectations with the pressures of keeping up with and staying relevant within real-time news.

4.6.3.6 The Crowd: "The Crowd Will Fix It"

Finally, several participants provided explanations placing the locus of responsibility for rumoring and rumor-correction on "the crowd." Some expressed an implicit trust in the crowd, using it to help verify information, for example through triangulation or by waiting to see if a story "has legs." Considering rumor corrections specifically, several interviewees expressed a more explicit trust in the crowd when it came to identifying and correcting rumors:

> LH7: "I think [rumoring is] part of Twitter and something we have to understand … that's going to happen. It's like a free information sharing tool. Everyone has freedom of speech (hopefully) and hopefully if someone is spreading false information, that information is quickly debunked through other people responding and giving their own insight to something."

Comments such as this can be linked back to the notion of the self-correcting crowd—i.e. that the online crowd will naturally identify, challenge and ultimately correct misinformation propagating among its members. This idea, which has been popularized in the press (Frere-Jones 2012, Herrman 2012), can be viewed as a kind of "folk theory" (Eslami et al. 2016; Kempton 1986) of how social media systems function.

4.7 DISCUSSION

Through the analysis of interviews and Twitter data related to two rumors in two significantly different crisis situations, this paper illustrates how online users engage in rumor-correcting behavior. In this section, we first synthesize the three components of the findings into a preliminary model of rumor correcting. Finally, we explore how imagined audiences (boyd 2008; Litt 2012; Marwick and boyd 2011) and the broader concept of folk theories (Eslami et al. 2016; Kempton 1986) guide the actions users take to correct online rumors.

4.7.1 *A Model of Decision Making for Crisis Rumor Correction*

After encountering conflicting information about the veracity of an online rumor, the decision of whether and how to correct has multiple inter-related factors. This paper identifies and explores three components of this decision-making process (see Figure 3.1).

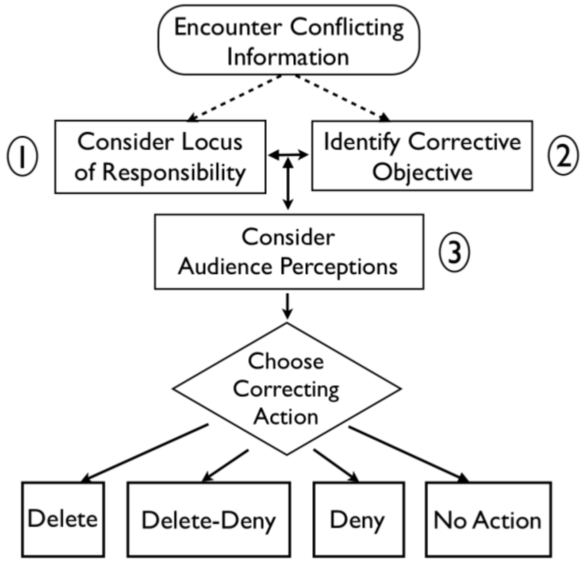

Encounter Conflicting Information
1
Consider Locus of Responsibility
Identify Corrective Objective
2
Consider Audience Perceptions
3
Choose Correcting Action
Delete
Delete-Deny
Deny
No Action

Figure 4.1 Decision Making for Twitter Rumor Correction

One salient component is the *locus of responsibility*. This includes consideration of who is to blame for the spread of false rumors as well as who has the power to correct them. It also includes how a user conceives of her role within that group. For example, if a user places responsibility in the crowd (espousing a belief in the "self-correcting crowd"), does she see herself as part of that crowd and as capable of playing a role in the correction? If she considers journalists to have a different set of obligations regarding rumor correction, does she see herself as a journalist? If so, then she may take different actions than someone who does not.

A second component is to identify the *corrective objective*—i.e. oneself, another user, or the information space. This consideration is shaped by where the user places the locus of responsibility as well as whether or not that user shared a rumor-affirming tweet. For example, a user who positions himself as the locus of responsibility and has shared a false rumor (e.g. WJ1, LH4) will often choose to correct himself. The corrective objective can also act to shape considerations of the locus of responsibility—for instance, LH7 provided a rationale suggesting her first order concern was to correct the information space and that she later came to realize the significance of her position within that space—as a local authority (Starbird and Palen 2010)—and herself at the locus of responsibility.

A third component is to consider whom one's audiences are—i.e. the imagined audiences (boyd 2008; Litt 2012; Marwick and boyd 2011). This is not just limited to conceptualizing the people with whom we are communicating and their immediate reactions, but also how they will act upon the information we share with them. For instance, one concern that arose repeatedly in the interviews was a perceived trade-off between accuracy and transparency: Deleting can be perceived as a sign of inauthenticity, but is it better to leave it out there where it may mislead others? Another concern was the social impact of explicitly correcting another user—something many users tried to avoid. Though we noted that this was primarily a downstream concern for our participants—likely due to the way that interviewees (and perhaps all users) rationalize their behavior after the fact—it is likely that reputational concerns have a shaping effect on corrective objectives and the locus of responsibility. For example, the perceived reputational impact of an explicit self-correction might cause an individual to revise his corrective objective (e.g. from himself to another) and subsequently reassess his position on locus of responsibility (e.g. from himself to an upstream source).

Finally, users take corrective action. For example, a user who positions herself as at least partially responsible for the spread of a false rumor and chooses to correct herself might post a denial in the form of an explicit self-correction; a user who thinks the crowd will correct and accepts some agency for himself as a crowd-member might choose to correct the information space by deleting his tweet (if he believes accuracy trumps transparency) or by posting an implicit denial (if transparency is more important). Some users choose to take no action. The decision-making process is shaped at each level by whether or not you affirmed the rumor—there are different options (at each level) for those who affirmed and those who did not.

This preliminary model, which emerged from our grounded analysis of interview data, aligns closely with Litt's (2012) model demonstrating the relationship between imagined audiences and online action. This sets the stage for a discussion about how users' conceptions of their audiences—as well as their understandings of the broader dynamics of social media—play a role in shaping their corrective behavior.

4.7.2 *How Imagined Audiences Shape Corrective Behavior*

Research suggests the importance of imagined or perceived audiences in shaping a social media user's actions (boyd 2008; Litt 2012; Marwick and boyd 2011). When asked directly, most participants in our study did not acknowledge reputational concerns or attending to their audiences. However, in explaining their rationale for taking certain actions (and not taking others) during the crisis event, interviewees revealed underlying awareness and concerns about their various audiences. A few (LH4 and LH7) expressed sentiments suggesting they were acutely aware of a growing audience, due to their position of influence within the information stream around the event, and several interviewees talked about the perceived trade-off between maintaining an accurate information space (an altruistic concern) and being perceived as trying to hide something (a reputational concern). This evidence suggests that many online rumor-participants are indeed aware of their audiences, and adjusting their behavior according to what they see as the expectations of that audience.

Building upon Giddens' (1984) theory of structuration, Litt (2012) presents a model demonstrating how imagined audiences—which emerge from users' interactions with the system—shape online behavior. Indeed, we see evidence of the shaping role of imagined audiences in our data. Superficially, we can see in our model how considerations of who one's audience is and how they will perceive one's actions play a role in guiding the choice of if and how to correct a Twitter rumor (Figure 4.1, #3). However, we hypothesize that conceptions of imagined audiences is more complex than that.

Discussions of imagined audience have often focused on a user's conceptions of who an audience is and how that audience might perceive them through their online actions (boyd 2008; Litt 2012; Marwick and boyd 2011). This kind of dynamic shows up in our data—for example around the interviewees' rationale for not wanting to be perceived as trying to hide an error through a deletion. However, rationale presented by interview participants demonstrates that people are not only trying to understand the size and make-up of their audience, but are also trying to piece together how their audience is acting, both individually and collectively, upon what they share.

For example, WJ1 worried about "potentially causing [others] anxiety or making them pay attention to something that they shouldn't have to think about." Another participant noted that he acted the way he did because he was worried that his tweets would create a panic: "…so if you looked at the tweets that I retweeted... I picked and chose as carefully as I could, because while [the situation] was concerning…I want to not suddenly cause a panic. I have enough of a reach that I probably could have caused one, so I was just being very picky about it." Embedded in this concern about panic was an assumption that, under certain conditions, the audience might propagate their messages to a large volume of people. Similarly, considerations about whether and how to correct a rumor-related tweet included theories about how different members of ones' audience might encounter and choose to propagate (or not) the original or the correcting tweet later.

These conceptualizations of not just who an audience is, but how that audience works (in conjunction with system features), are possibly more akin to the "folk theories" that people have about how social media systems function (Eslami et al. 2016; Kempton 1986) . We can see evidence for these folk theories at work within the locus of responsibility categories that emerged in this study (Figure 4.1, #1). For example, positioning responsibility on the "crowd" reflects the use of the popularized notion of the self-correcting crowd—which takes into account how the "audience" acts upon information and how it reacts to others' actions within the system. Similarly, assigning Twitter (or crisis events) as a locus of responsibility, which many interviewees did at least to some extent, also reflects the impact of folk theories—e.g. about the intersection of technical affordances and human behavior—on the structure that guides decision-making in this context.

This research demonstrates that folk theories guide rumor-correcting actions, and that these folk theories consist of reasoning related not just to how the algorithms work (Eslami et al. 2016), but to how the broader system—including the technological platform or platforms with their affordances, interfaces and algorithms, as well as the other human (and non-human) actors in the system—functions.

4.7.3 *Limitations and Future Work*

This study has several limitations. The Twitter data we used is incomplete (due to rate limits) and biased (due to the terms we tracked). Though we attempted to utilize those digital traces to identify people who exhibited different kinds of rumor-correcting behavior, self-selection bias shaped the participant sample towards individuals who were more invested in actively correcting the rumor. Additionally, though we asked interviewees to discuss their larger patterns of use across other sites and platforms, the digital trace data, recruiting strategy, and interview protocol render this study highly-focused upon online rumoring within one platform—Twitter. And finally, as with any retrospective study, our interview responses were likely affected by misremembering and post-hoc rationalizations.

However, despite these inherent biases, by connecting actual digital traces to recruitment strategies and interviews, we were able to 1) recruit interviewees who displayed several different correcting behaviors; and 2) provide them with assistance in remembering their actual behaviors (their tweets and deletions). Though some rationalizations (e.g. around why to delete or not) are closely tied to the specific affordances of Twitter and the context of online rumoring, the broader findings about the role of imagined audiences and folk theories of how those audiences interact with data are likely to apply to other platforms and contexts.

The model of rumor correcting presented here is illustrative and functional, but likely incomplete. We introduce it here to synthesize findings, to show how the different constructs fit together, and to provide a foundation for the major theoretical contribution of this paper—demonstrating how the shaping role of imagined audiences in online behavior includes not just who those audiences are but also how they react to and interact with the information we share and the actions we take. Future work may reveal additional considerations and help to further unpack and link together the three constructs presented here.

4.8 ACKNOWLEDGEMENTS

This research is a collaboration between the emCOMP lab and DataLab at the University of Washington and was supported by NSF Grant 1420255. We also wish to thank the UW SoMe Lab for providing infrastructure support as well as the many students who provided significant assistance to this project.

(NOTE: This marks the end of the original publication)

4.9 Reflections: Lessons from taking a close look at the self-correcting crowd

By analyzing 57,562 tweets and 15 interviews across two crisis events, this project has provided insights along a number of dimensions. Its findings shed light on the relationship between misinformation and corrections on Twitter during mass-disruption events. More broadly, it increases our understanding of how the technical affordances of social media and people's practices around these mediums can interact in complex ways to shape the spread of online misinformation.

Among the findings of this research is the evidence of crowd-correction for each rumor, but with generally smaller proportions of correction. From a normative or ethical perspective, this evidence points to a simple truth: good habits or actions are always rarer and take more time to emerge and stabilize in social practices (Vallor 2016, 186). Yet when confronted with the problem of fostering better behaviors in online settings, we need not see the practical obstacles as insuperable; nor allow them to cloud the need for action or to make the perfect the enemy of the good. Instead, we can turn to the concrete question of how to strengthen corrective behaviors in online settings.

But why should we focus on strengthening these behaviors? On one level, the answer is simple: because it's good for us and the information environments we inhabit. On a more pragmatic note, one thing that the concept of sensemaking highlights, and this study clearly illustrates, is that 'misinformation' is often exposed retrospectively. What we come to label as misinformation actually unfolds as a series of approximations and attempts to discover an appropriate response in difficult conditions. And because it unfolds this way, out of an error-ridden activity, it contributes to a fog of contingencies that makes it difficult to reliably identify misleading information at scale. These contingencies and their obscuring effects, which one might call *information opacity*, might simply be far too numerous to resolve. What self-correcting crowds can offer is not a solution to information opacity, but a human-centered strategy of developing certain skills that can aid us in coping, and even thriving, in the midst of the uncertainties posed by this opacity. In that sense, this project can help us gain a more accurate and deeper understanding of how to pursue this strategy by examining how some social media users viewed their own activities around misinformation. I will draw forward three implications of this research in that regard.

4.9.1 *Supporting more constructive engagements with uncertainty*

One reason that several participants gave for not correcting rumors had to do with uncertainty and risk. In the words of one participant (LH1):

> "To be honest with you, if I thought it was an emergency I would still be bashing tweets out…If I thought that it was false then I would stop. But until I knew for sure, I would probably keep doing it until I realized it was not the thing to be doing… you don't know how things are working in terms of how terrorists or people who are saying are terrorists are dealing with this. They might be involved in it [the information space]. You never know."

This participant (and others) viewed the risk of letting the rumor go on and being wrong as less than the risk of negating the rumor and having it turn out to be true[24]. Researchers have argued that uncertainty in the information environment (Cassa et al. 2013) contributes to misinformation, but here, we can also appreciate how it can act as a barrier to correcting it.

This is interesting because even though social media platforms can place us in situations where we have to work with uncertainty, they don't seem to support engaging with that uncertainty in ways that help us step outside of it and be intentional with it. For example, the character limits imposed by Twitter can function to disincentivize users from expressing uncertainty during a breaking-news situation. This is compounded by how the platform does not presently allow its users to edit or publicly retract their tweets. It only allows for deletions, which can appear inauthentic to audiences (Marwick and boyd 2011). In these ways, platforms might be co-shaping environments where people are encouraged to handle uncertainty by normalizing it out of existence. To change this, and help reduce the spread of 'organic' misinformation, designers might wish to consider exploring ways of helping social media users express uncertainty (a banal example might be an additional checkbox or button that allows users to rate how confident they are in the information they are tweeting when they employ a disaster related hashtag).

More broadly, if we want to support people in self-correcting misinformation they pass along, then we have to weaken the tendency to normalize uncertainty away. If people fixate on their first plausible story and stop there (as arguably LH1 did), then they do have a sense of sorts, but only one that holds together if newer cues and consequences are ignored. If instead, people perceive themselves as making sense of an unfolding situation and trying to craft a fuller story, they might be more open to revision, self-correcting and responsive, and with more of their rationale being transparent. The creative use of technology can aid in this effort, but cultivating such views also speaks to the intersections of new media practices and education, which is a theme I will discuss further below.

[24] The perspective of LH1 is that of an individual working at the edge of codified knowledge, using social media, in an effort to make sense and protect lives. This illustrates how 'misinformation' or 'corrections' are often exposed retrospectively.

4.9.2 *Refining folk theories*

One of the significant findings of this project was that people's abstract ideas about how social media works guides rumoring behaviors[25]. Notably, these ideas consisted of reasoning related not only to their 'technological imaginaries' (Suchman 2014), but also their imagined audiences. This empirical result, built on the concept of folk theories, suggests an important area for potential growth in interface design: shifting designs to give users a greater sense of agency over, and responsibility towards, their information environment and promoting learning about that environment using seamful design.

Influenced by Mark Weiser's (1994, 1) notions of seamlessness and 'seamfulness', and in contrast to his proposition that "A good tool is an invisible tool", seamful designs emphasize mechanism[26] (Chalmers 2003). They do this by having "seams," visible interface hints disclosing aspects of automation operations, that help users develop and refine their theories of how systems work (Eslami et al. 2016). In this way, seamful designs acknowledge how technological systems exist in tension with folk theories and try to shape their evolution. Eslami writes:

> "A seamful design makes system infrastructure elements visible when the user actively chooses to understand or modify that system. Such design emphasizes experience and reflection, inviting the user to explore and discover connections in the system through manipulation, comparison, and feedback." (Eslami et al. 2016, 2373)

In the context of rumoring on social media, seams could be incorporated, for example, to help users explore and discover how information diffusion takes place on platforms. To brainstorm further on this example, we can imagine users being provided a view that walks them through the provenance of a tweet, tracing its diffusion across the network to help them appreciate how

[25] These ideas can also reveal something about our ethical comportment towards social media. One participant who had shared misinformation extensively in our data collection captured this in a remark about how he did not bother to verify or correct information: "No, I left that up to the journalists to fact check. You figure, as time goes on, before the sun sets, fact and fiction, rhyme and reason will have sorted itself out." Variants of this heedless viewpoint of 'I don't have to verify this, because others will do that; the crowd should fix it', were expressed by three participants.

[26] This differs from the popular approach of "seamless design" in which the "technology is hidden," and where the goal is to reduce or even eliminate folk theories (Wenneling 2007). A risk with seamless design in relation to the challenges of misinformation on social media is that it can make it more difficult for people to predict the consequences of their actions. This in turn makes it harder for people to identify and seek behaviors that are more conducive to human flourishing. Of course, other factors (e.g. rapid technological change) contribute to this as well. I would direct readers to the work of Shannon Vallor (2016) for further reading on this topic.

individual actions can shape the trajectory of a rumor. An alternative view might show them the *potential* effects of their own actions by visualizing the reach of a message before they share it based on algorithm outputs. The uncertainty stemming from such seams could support corrective behaviors. lead users towards deeper thinking and increase human agency in complex systems.

We can also translate this implication to other sites of professional practice. For instance, the insight that folk theories mediate corrective behaviors offers a conceptual starting point for generating new programming ideas for media literacy efforts. Specifically, it suggests we might design learning experiences to help learners acquire the skills to not just distinguish reliable from misleading information (a moving target), but also to retract information if they make a mistake (a strategy for coping with moving targets). One way to design learning experiences this way could be to focus on helping learners refine their understandings of not just what misleading information *is*, but also of how it flows through our sociotechnical systems. We could even imagine deploying seamful technologies to aid in this endeavor to help people construct insights for themselves.

4.9.3 *Mindfulness and Reflection*

Before concluding this chapter, I want to enlarge its analysis and bring in the theme of mindfulness. Although every interviewee in this study expressed an inclination towards wanting to do the right thing, e.g. to share useful information and not spread false rumors, some of them explained that this commitment was complicated by their desire to keep sharing information at a fast pace. LH3 succinctly captured this kind of behavior when she explained that “sometimes, I do retweet things without reading them.”

These participants associated their activity with having fragmented attention and reduced intentionality due to the speed of information on Twitter. For instance, LH7 described a kind of tunnel vision effect as a contributing factor to not being able to stop retweeting: “...it’s just the adrenaline keeps you going and then the tweets. It’s all this information. There’s just so much going around your head with all this information that I just didn’t stop”. Another participant who was a professional journalist talked about how she tweeted out misinformation because she was getting “really really tired” while covering the Paris attacks:

> LH7: “This is what happens when you’re working at a very fast pace sometimes… I didn’t take enough time to look back and [I] had sent hundreds of tweets at this juncture, and I just had forgotten…I did not do well with those [verification] checks that are normally on the list”.

These views suggest that social media can direct our attention in ways that contribute to the spread of misleading information. In the case of Twitter, which positions itself as a source of real-time news, the ability to amplify information rapidly via retweets can serve to facilitate sensemaking during time-sensitive crises. However, this same speed and convenience can also encourage sharing information on ‘autopilot’, or ways of thinking that are more prone to phenomena such as stereotyping, normalizing, confirmation bias etc (Cook and Woods 1994; Weick, Sutcliffe, and Obstfeld 2005). When this speed is coupled with the ephemeral quality of tweets, users can lose their connection to what they (and others) had posted earlier. Several participants (like LH8 for instance) remarked how “it’s hard to go back and find your old tweets”. This lack of temporal durability to tweets could contribute to a more fast-food like approach to information, where producing and consuming it take precedence over refinement and corrections.

Helping people be more attentive might aid them in making better use of social media—to be more healthily and effectively engaged with and through them, and to reduce the spread of misleading information. Finding design opportunities to support more mindful behaviors *within* such social mediums— without compromising the very qualities that make them so effective for crisis communication in the first place—is not a straightforward task. One basic idea could involve turning to existing algorithms that moderate spam and usage limits, adapting them to help nudge or remind users to be more mindful around certain topics or information channels during crisis events. For example, the concept and algorithms behind “Twitter Jail”—a commonly used term to describe how Twitter will disable a user’s ability to tweet for a set time period if their tweet rate exceeds a certain threshold (Twitter n.d.)—could possibly be retuned and repurposed to incentivize users to engage in less repetitive behaviors around certain hashtags or keyword terms that have been detected as being relevant to ongoing events.

Another insight that emerged from this work suggests an alternative way forward. Often, at the conclusion of my interview sessions, participants would thank me and my fellow researchers for the opportunity to simply review and reflect on what they had tweeted during these events. Some of them also hinted at the idea of learning from their experiences. LH3, for instance, shared: “I think this event did teach me about retweeting…I think next time I would be even slower in retweeting things”. Comments like this one are clearly visible in the data, suggesting that reflection can help individuals refine their approach to sensemaking and corrections over time.

This direction suggests future opportunities for HCI researchers and designers to explore the diverse ways that technologies might encourage or support user practices for reflection and learning within this context. For example, one can imagine providing users with their tweeting histories in certain time- windows to help them reflect back on their behaviors with the goal of informing future actions (or even just feeling more accountable to what they posted). This could be in a similar vein to earlier work like Jones et al. (2012) who showed users their mouse accuracy and found it reduced exaggerated movements; or Odom et al.'s (2014) Photobox which was a slow technology that provoked users into thinking about their engagements with digital photographs. Further, new systems could help members of the crowd conduct collective or individual after-action analyses of sorts on how they grappled with events containing uncertainty. Research on computer-supported reflection (e.g. Cheng et al. 2011; Li, Forlizzi, and Dey 2010; Mathur and Karahalios 2009; Zhao, Ng, and Cosley 2012) and revisiting one's digital content (Malacria et al. 2013; Odom et al. 2014) could be leveraged in support of this direction.

Taking this direction in education could involve using contemplative pedagogies in support of current media literacy efforts. By contemplative pedagogy, I mean approaches to teaching and learning that encourage deep learning through focused attention, reflection, and heightened awareness (Hart 2004; Barbezat and Bush 2013; Zajonc 2013). Designing learning experiences that promote a contemplative approach to misinformation[27] could involve, for example, exercises that ask learners to investigate the quality of the attention or assumptions they bring to social media (Levy 2016). The premise here being that once learners notice what they are doing and the effect that is having upon their life and information environment, the possibility arises of choosing to behave differently.

To sum up, I have suggested that supporting self-correcting behaviors might be a useful strategy for coping with the challenges posed by information opacity. I have argued that supporting this strategy requires cultivating spaces for a certain kind of sociotechnical education and practice. I invoked seamful design to highlight how social media has the potential to be shaped and used in ways that reinforce, rather than impede our efforts to become better technological citizens. I brought in media literacy to call attention to how supporting corrective behaviors can be a recursive procedure, in which educational techniques are deployed to generate the motivation to design and adopt corrective practices that shape our information environments in more constructive ways. Used as mutually reinforcing handholds, this interweaving of educational and technological design work can serve as a practical and powerful strategy for shaping behaviors that might reduce the spread of misleading information.

[27] I owe a great deal to the work of David Levy here for his work on *Mindful Tech* (2016) and for our conversations that have shaped my thinking around how one might design exercises that help learners directly observe their own behaviors and motivations in online settings.

Chapter 5. STUDY 2: ACTING THE PART

This chapter provides a study[28] of misleading information that is created and spread intentionally. It casts light on the work of *disinforming,* examining how the Internet Research Agency in St. Petersburg opportunistically exploited sensemaking efforts to further its objectives (Research Question 2). I encountered the workings of the agency's work by accident, as I was helping my co-author, Leo G. Stewart, answer questions he had about other online phenomena — concerning the different narratives that were being mobilized to make sense of the #BlackLivesMatter movement (Stewart et al. 2017). This trajectory is further described in the study itself, so I will focus on supplying three other pieces of context to help readers engage with this study.

First, this investigation was conducted when social media companies had released very limited information about the Internet Research Agency's activities. In fact, they had removed all content that could be linked to the agency, making its activities opaque to researchers, journalists and other members of the public. I initiated this research partially in response to this silencing of history. Fortunately, this state of affairs did not persist and social media companies eventually released more complete datasets than the one I used to publish this study (e.g. Gadde and Roth 2018). I will briefly discuss what these new data help reveal towards the end of the chapter.

Second, it is worth noting that this research was written in 2016, prior to certain developments. Since the killings of Ahmaud Arbery, Breonna Taylor and George Floyd, public perceptions of law-enforcement in the United States, the #BlackLivesMatter movement, and the reactionary counter-stances of #BlueLivesMatter and #AllLivesMatter have all undergone significant shifts. Hashtags like #ACAB (short for the slogan 'all cops are bastards') that seemed to be closer to the fringe of public discourse in 2016, and were being promoted by the Internet Research Agency, have since been normalized — not because of the agency, but because of people's efforts to shift the cultural ground. These developments do not negate this research, but they are worth keeping in mind as we proceed.

[28] This study is previously published work. To cite material from this Chapter, please cite this original work as well as the dissertation:

Arif, Ahmer, Leo G. Stewart, and Kate Starbird. 2018. "Acting the Part: Examining Information Operations within #BlackLivesMatter Discourse." *Proceedings of the ACM on Human-Computer Interaction* 2, No. CSCW: 1-27. https://doi.org/10.1145/3274289.

Finally, this study introduces and uses the term *information operations* in place of *disinformation campaigns*. The term has been adopted by social media companies to describe the same phenomena (coordinated communicative activity intended to influence large numbers of people). I use it in this investigation to help frame the study of disinformation for CSCW audiences, and to further nuance our understanding of disinformation.

Acting the Part: Examining Information Operations Within #BlackLivesMatter Discourse

Published in Proceedings of the 2018 ACM Conference on Computer Supported Cooperative Work

Authors: Ahmer Arif, Leo G. Stewart & Kate Starbird

5.1 ABSTRACT

Information campaigns that seek to tap into and manipulate online discussions are becoming an issue of increasing public concern. Social media companies are now problematizing some campaigns, specifically those that intentionally obscure their origins, as 'information operations'. This research examines how social media accounts linked to one such operation—allegedly conducted by Russia's Internet Research Agency—participated in an online discourse about the #BlackLivesMatter movement and police-related shootings in the U.S. during 2016. We study the interactions of these accounts within the online crowd using interpretative analysis of a network graph based on retweet flows in combination with a qualitative content analysis. Our empirical findings show how these accounts imitated ordinary users to systematically micro-target different audiences, foster antagonism and undermine trust in information intermediaries. Conceptually, this research enhances our understanding of how information operations can leverage the interactive social media environment to both reflect and shape existing social divisions.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing • Human-centered computing → Social media**

Keywords

Social media; Twitter; Information Operations; Disinformation; Media Manipulation; Black Lives Matter

ACM Reference format: Ahmer Arif, Leo G. Stewart, and Kate Starbird. 2018. Acting the Part: Examining Information Operations Within #BlackLivesMatter Discourse. In *Proceedings of the ACM on Human-Computer Interaction,* Vol. 2, CSCW, Article 20 (November 2018). ACM, New York, NY. 27 pages. https://doi.org/10.1145/3274289

5.2 Introduction

Although the advent of social media was initially met with enthusiasm for more democratic information systems, our evolving information practices are now forcing us to think about how these new points of access can be manipulated. This has become a more urgent consideration in recent years as social media platforms have allowed misinformation—as well as disinformation, and political propaganda—to spread and engage audiences in new ways. Recently, social media companies have acknowledged that their platforms have become sites for *information operations*, i.e. actions taken by governments or organized non-state actors to manipulate public opinion (Weedon, Nuland, and Stamos 2017; Twitter 2018; Tumblr Help Center 2018). Though information operations are not new, their intersection with social media is not well understood.

This study focuses on inauthentic social media accounts as a component of information operations to consider how they harness the sociotechnical infrastructure of social media platforms for their benefit. The accounts that we analyze were publicly suspended by Twitter for being affiliated with the Internet Research Agency (RU-IRA), a Russian organization based in St. Petersburg that has been formally indicted by the U.S. government for engaging in professional propaganda, including hiring 80 full-time employees to use social media accounts while pretending to be U.S. citizens (U.S. Justice Department 2018). Despite mounting allegations, the tactics used by the social media accounts linked to these efforts have not yet been systematically examined.

We investigate how these RU-IRA affiliated accounts participated in an online discourse about the #BlackLivesMatter movement and shootings in the U.S. during 2016. We did not select this discourse or collect our initial data with the intent to study information operations. Instead, we had previously scoped and analyzed this data in work examining "framing contests" within politically charged discourse on Twitter (Stewart et al. 2017). Later, when Twitter released a list of RU-IRA affiliated accounts during formal hearings with the U.S. House of Representatives Select Committee on Intelligence (2017), we recognized several accounts from our earlier work. This led us to ask ourselves: were more of these accounts present in the data we collected, and if so, with whom did they interact, and what were they doing?

We approach these questions from a CSCW perspective, adapting methods from the field of crisis informatics (Maddock et al. 2015; Palen and Anderson 2016; Arif et al. 2017) to analyze both the large-scale interactions between these accounts and other members of these online communities, and the specific online actions that the operators of these accounts took as they worked to infiltrate and influence these communities. To answer the first of our questions—if Russian information operations were active in the #BlackLivesMatter discourse—we used a network graph of retweets to learn that at least 29 of these accounts did have a meaningful presence within the information flows of this discourse. The graph also revealed that different RU-IRA accounts were participating on both 'sides' of the conversation—within two structurally distinct communities. Then, to understand what these RU-IRA accounts were doing, we launched a multi-sited qualitative investigation into the messages, personas, and interactions of these accounts. As we immersed ourselves in their content, our questions about what these accounts were doing evolved. We asked: Who did these accounts attempt to mimic? What did these accounts do to produce and maintain their personas? What were these personas used to model and project in the discourse that we studied? To what extent did these 'performances' seem to adhere to a common script or set of constraints and where did they deviate from each other?

Addressing these questions contributes to a fuller account of the dynamics that emerge between information operations and those who use social media platforms for cooperative work such as grassroots political organizing (e.g. Savage, Monroy-Hernandez, and Höllerer 2016), disaster response (e.g. Dailey and Starbird 2017; Kogan, Palen, and Anderson 2015; Wong-Villacres, Velasquez, and Kumar 2017), and more broadly the collective activity to consume and elevate breaking news (Wang and Mark 2017). Our findings suggest that information operations were occurring in this context and that while social media platforms may intend to bring us together, at least some of these platforms are being targeted, deliberately, to pull us apart. On another level, this research helps us see that the 'work' these accounts were doing to facilitate information operations goes beyond publishing biased information. The work can also be seen as an improvised performance being carried out by an account operator (or, perhaps, a small team of operators) to try and 'inspire' the online communities they target. These performances can involve connecting to cultural narratives that people know, enacting stereotypes, and modeling how to react to information. This has implications for platform designers as they consider the strategies they will use—or more specifically, the policies they will create to guide the strategies they will use—to address information operations.

5.3 Literature Review

In this literature review, we first provide background on information operations generally and on their emerging use in the online sphere. Within that accounting, we highlight a specific (theorized) goal of information operations related to the concept of *disinformation* that is relevant to the study presented here, and explain how our research contributes to better understanding that goal and the tactics used to achieve it. Finally, we explain how approaching this topic from a CSCW lens helps to conceptualize the activities of these accounts as a type of online 'work' conducted by an information operator (or agent) in interaction with an online crowd.

5.3.1 *Information Operations*

Information operations is a term employed by the U.S. intelligence community to describe actions taken to disrupt the information streams and information systems of a geopolitical adversary (U.S.

Joint Chiefs of Staff 2014). These actions focus on degrading the decision-making capabilities of others through non rational means (e.g. deception and psychological warfare) (Armistead 2004; Jowett and O'Donnell 1999). Unlike 'information warfare' which is generally conducted during actual combat, information operations can be carried out in peacetime environments to influence civil affairs (Armistead 2004). Consequently, these operations are increasingly considered a 'soft' yet formidable alternative to 'hard power' or 'hard warfare', targeting perception and cognition rather than launching physical attacks on infrastructure (Burns and Eltham 2009; Pomerantsev and Weiss 2014; Lin and Kerr 2017).

Some academics (e.g. Faris et al. 2017; Lin and Kerr 2017; Ong and Cabañes 2018) and journalists (e.g. Pomerantsev and Weiss 2014) have theorized that a primary or secondary goal of many information operations is not necessarily to convince someone of something, but to strategically direct discourse in ways that "kill the possibility of debate and a reality-based politics" (Pomerantsev and Weiss 2014, 16). By eliciting confusion, division, disenchantment, and paranoia, information operations can potentially serve to silence political dissent, enable historical revisionism, and hinder collaboration (Faris et al. 2017; Lin and Kerr 2017; Woolley and Howard 2017). Both journalists and former intelligence professionals have suggested that such efforts can be tied to historical strategies of *dezinformatsiya* (Bittman [1983] 1985; Pomerantsev and Weiss 2014; Snyder 1995), a Russian term that translates to disinformation and describes the intentional spread of false or inaccurate information meant to mislead others about the state of the world.

Disinformation can therefore be viewed as a specific form of information operation that has its historical roots in tactics initially developed and deployed by the Soviet Union (Pomerantsev and Weiss 2014; Snyder 1995). These tactics have been characterized as having an 'ideological fluidity' allowing them to overlap with a range of oppositional political groups—with the goal of fostering social division (Paul and Matthews 2016). The core of these tactics involves harnessing existing public discontent by amplifying reductive social interpretations that confirm existing beliefs, support desired conclusions, or prompt certain strong emotions regarding groups of people and events (Faris et al. 2017; Lin and Kerr 2017). By strategically and opportunistically tapping into latent social fractures—as in cases surrounding the Ku Klux Klan as well as the AIDS and Ebola epidemics—trust in civil institutions and information intermediaries can be undermined (Bittman [1983] 1985; Lin and Kerr 2017; Pomerantsev and Weiss 2014).

The clandestine nature of information operations means that our current understanding of the relationship between existing social rifts and disinformation tactics remains speculative. Our work empirically examines this relationship by systematically exploring what RU-IRA affiliated accounts were doing in a discourse that is already deeply segregated in terms of politics and race.

5.3.2 *Information Operations on Social Media*

The announcements by Facebook, Twitter and Tumblr reveal that social networking sites have become a front for information operations—a front that can be accessed from nearly anywhere in the world, by nearly anyone, and where users may be particularly vulnerable (Weedon, Nuland, and Stamos 2017; Twitter 2018; Tumblr Help Center 2018). Researchers have noted that the interactivity afforded by these social computing systems can allow information operations to produce emergent and self-reinforcing effects (Burns and Eltham 2009; Prier 2017). Moreover, this new media ecosystem is dominated by increasingly partisan news sources (Gentzkow and Shapiro 2011), political homophily (Grevet, Terveen, and Gilbert 2014; Lazer et al. 2010), and algorithmically derived newsfeeds being skimmed by audiences that are trying to cope with the cascades of information before them. These structural issues can contribute to the effectiveness of information operations, including disinformation. At the same time, increasing protection against information manipulation on these platforms risks undermining the free speech and open discourse foundational to democracies (Lin and Kerr 2017; Woolley and Howard 2017).

5.3.3 *Information Operations as Collaborative Work*

Researchers have noted that the ’work’ of information operations on social media is, in principle, collaborative in the sense that high-level digital marketing strategists and political clients work together to design campaign objectives which are then implemented and shaped by a multitude of different actors (Ong and Cabañes 2018). Tucker et al. (2018) partially capture the complexity of this assemblage by noting how bots, fake-news websites, conspiracy theorists, trolls, highly partisan media outlets, the mainstream media, influential bloggers, and ordinary citizens are now all playing overlapping—and even competing—roles in producing and amplifying propaganda in the social media ecosystem. Relevant here, these authors (Tucker et al. 2018) note that hired trolls or anonymous influencers that use fake online profiles to support disinformation campaigns are a relatively understudied set of actors partially due to the difficulties involved in identifying them. Our research helps to address this gap.

Although impersonating others to spread harmful narratives is an old practice (e.g. the forged 1903 pamphlet, Protocols of the Learned Elders of Zion that was used to justify anti-Semitic agendas), its intersection with the networked media environment is not well understood (Jowett and O’Donnell 1999). What we do know is that impersonation is now being used to amplify racist narratives (Farkas, Schou, and Neumayer 2018a; ibid 2018b) and mobilize digital workers being paid to act like grassroot activists in a variety of work arrangements. For instance, Rongbin Han’s (2015) research on the digital political operations of China’s “fifty-cent army” surfaces efforts to incentivize state-sponsored workers to act like “spontaneous grassroots supporters” in online discussion boards. In contrast to Han’s (2015) study—which found rigid work arrangements producing unnatural bot-like activity—Ong and Cabañes’s (2018) research in the Philippines context revealed how a hierarchized group of professional political operators used fake online personas in ways that emphasized individualization and flexibility to conduct an information operation.

In our research, we analyze this phenomenon of coordinated impersonation within an online discourse or activist community from a CSCW perspective—considering this activity as a type of online 'work' conducted by an information operator (or agent) in interaction with an online crowd. This lens allows us to conceptualize how this collective activity includes other collaborating agents as well as more sincere activists who may not recognize that they are interacting with political agents. It also allows us to reveal this work as an improvised performance that both reflects and shapes the discourse within which it is embedded.

5.4 Background

Our initial data for this study was not collected with the advance intent of studying information operations in relation to the #BlackLivesMatter movement. Rather, the seed data for this research was collected to facilitate prior related work that studied this discourse to learn about how digital activists frame events and competing social movements (Stewart et al. 2017). Just weeks after publication of that work, we realized that the communities we had studied had been targeted for online information operations. This motivated us to return to this dataset to better understand how the work of those information operators intersected with the activities of online activists within that conversation.

5.4.1 *Black Lives Matter and Blue Lives Matter Discourse in 2016*

As boyd (2017b), and Wardle and Derakhshan (2017) have argued, the production of online propaganda cannot be understood in isolation from its social, political, technological, and cultural context. This research examines the production of online propaganda on Twitter in a context that intersects with issues of race, partisanship, gun violence, digital activism, and the failures of public institutions. Specifically, we investigate the activities of one set of actors in an online discourse about the #BlackLivesMatter movement and shootings in the U.S. during 2016.

The hashtag #BlackLivesMatter was first coined in a Facebook post by Patrice Cullors and Alicia Garza in 2013 in response to the acquittal of George Zimmerman in the shooting death of Trayvon Martin (Guynn 2018). The post and correspondingly the hashtag spread virally across social media platforms and crystallized in an on- and offline social movement that brought conversations on race into mainstream discourse, particularly shootings of African-American men by police officers. On their webpage, the BLM organizers describe BLM as "an ideological and political intervention in a world where Black lives are systematically and intentionally targeted for demise" (Black Lives Matter n.d.). Over time, a counter-movement took shape on social media, specifically critiquing the BLM movement for deprioritizing other lives (#AllLivesMatter) and being founded in a "false narrative" that vilifies police officers (#BlueLivesMatter) (Blue Lives Matter 2017). This counter-movement gained momentum in 2016, after shootings of police officers in Baton Rouge, Louisiana and Dallas, Texas prompted a spike in the volume of tweets related to counter-frames, for example about #BlackLivesMatter activists allegedly advocating for violence towards police (Anderson and Hitlin 2016; Stewart et al. 2017).

5.4.2 *Public Announcements Regarding Information Operations in 2017*

This discourse was also taking place during a time (2016) when Russian information operations in the US were particularly active, prior to the congressional investigations to highlight the problem (U.S. House of Representatives Permanent Select Committee on Intelligence 2017; U.S. Senate Select Committee on Intelligence 2017) and the actions taken by the social media companies to address it (Stamos 2018; Twitter 2018). In an April 2017 report, Facebook acknowledged that their platform had been used for "information operations" by both state (i.e. Russia) and non-state (i.e. Wikileaks-affiliated) actors to influence the 2016 U.S. Presidential election (Weedon, Nuland, and Stamos 2017). After Facebook's announcement, representatives from other social media companies including Twitter, Tumblr, and Reddit also came forward to acknowledge that their platforms had been utilized for information operations by the previously mentioned Internet Research Agency (RU-IRA), an entity known to be a Russian 'troll farm'.

In response to speculation surrounding the role of the RU-IRA in the 2016 presidential election, Twitter released a list of 2,752 RU-IRA affiliated troll accounts in November 2017 (U.S. House of Representatives Permanent Select Committee on Intelligence 2017; U.S. Senate Select Committee on Intelligence 2017). After identifying these accounts and presumably to protect other users from further deception, Twitter suspended the RU-IRA accounts, removing their account profile and tweet history from public view. This illustrates how social media content associated with clandestine activities can be challenging to gather and study due to its ephemerality. Our research team was able to overcome the ephemerality issue in this case because we had already curated, visualized, and intensely analyzed the relevant data described here.

Since the release of the initial list, Twitter has announced the suspension of more RU-IRA accounts (although the details of these accounts have not been released) and investigative reporting has provided a clearer image of how RU-IRA troll accounts operated (Silverman 2018; Troianovski 2018). These reports indicate that the RU-IRA employed carefully-vetted individuals with strong knowledge of American pop culture and fluency in English to pose as Americans on social media and engage in conversations surrounding American social issues. Journalists have specifically noted that the online conversation around BlackLivesMatter and BlueLivesMatter was a significant point of access for these information operations (e.g. Silverman 2018). Though these industry reports and journalistic accounts provided rapid and needed insight, there is still a need to more systematically understand what these strategies are and how they interact with online discourse communities.

5.5 METHODS

Our interpretivist mixed-methods research iteratively analyzes our data by drawing on the guidelines and perspective of Charmaz's ([2006] 2014) constructivist grounded theory to render a nuanced and flexible explanation of the activities enacted by RU-IRA affiliated Twitter accounts. Acknowledging the scale and multi-sited nature of the networked discourse in which we study these accounts, we extend methods for conducting research on large-scale, online social interactions (Geiger and Ribes 2011; Howard 2002; Palen and Anderson 2016; Rottman et al. 2012) and analyzing the spread of online misinformation (Arif et al. 2017; Maddock et al. 2015) during crisis events. We start by generating a network graph of retweets that reveals structurally distinct communities in the politicized discourse we are studying. This guides our inquiry by allowing us to harness structural data (behavioral network ties) to narrow down our case-selection for in-depth qualitative research. We do this by cross-referencing a list of 2,752 suspended RU-IRA affiliated accounts and systematically selecting the 29 accounts that were well integrated into the information network (the 'who'). We then conduct a qualitative analysis through bottom-up open coding on the digital traces left by these accounts (i.e. tweets, profiles, linked content and websites), writing analytical memoes, and reflecting on the research process to consolidate observations of how they were participating in this discourse (the 'what'). Juxtaposing these fragmented micro-level observations with the network graph—which illuminates the sub-networks these accounts were integrated with (the 'where')—helps us build up into a more macro-understanding of how these accounts worked to support an information operation.

5.5.1 *Data Collection and Filtering*

Our initial dataset consisted of 58.8M tweets that were posted and collected between December 31st 2015 and October 5th 2016. We collected these tweets by tracking shooting-related keywords like "gun shot", "gunman", "shooter" and "shooting" using the Twitter Streaming API.

We further filtered this set to tweets containing the terms "BlackLivesMatter", "BlueLivesMatter", or "AllLivesMatter" ("*LM") in the text. The resulting dataset of 248,719 tweets was used in prior work which established divergent and competing frames tied to the #BlackLivesMatter and #BlueLivesMatter hashtags (Stewart et al. 2017). This curated dataset—i.e. limited to *LM tweets with shooting terms—enabled us to explore the role played by RU-IRA affiliated accounts in a politically-charged online discussion related to activist movements and counter-movements in the U.S. in 2016. Importantly, this dataset is not representative of the broader BlackLivesMatter discourse but is focused on discourse related to violent offline events that included shootings of African Americans by police officers and shootings of police officers by an African American.

To focus our investigation on accounts that demonstrated some level of sustained engagement and influence in the conversation, our final filtering step involved limiting our analysis to accounts with a retweet degree (sum of how many times an account was retweeted and how many times an account retweeted other accounts) greater than one. This final step produced 22,020 accounts, who were responsible for 89,437 of the tweets in our "*LM" dataset.

5.5.2 *Network Analysis*

We iteratively visualized retweet flows between the 22,020 accounts by constructing a network graph (see Figure 5.1 and Figure 5.2) in which we defined nodes to be Twitter accounts and directed edges to be retweets between accounts. We used the Force Atlas 2 layout in Gephi (Bastian, Heymann, Jacomy 2009) to determine the visual layout of this graph. The retweet flows between these accounts consisted of 58,698 retweets. To formalize structural observations of the network, we used the Infomap optimization of the map equation to systematically detect communities in the graph, ultimately producing two main communities ("clusters") (Edler and Rosvall n.d.; Rosvall, Axelsson, and Bergstrom 2009). We examined the effect of tuning Infomap parameters such as the inclusion of nested subclusters and overlapping modules; however, these did not significantly alter the extreme separation of the two main communities of the graph, and we thus ran the Infomap analysis specifying a directed graph with all other parameters at the default setting. To categorize and contextualize these clusters, we applied methods used in our prior work (Stewart et al. 2017), examining the most frequently appearing hashtags in the account descriptions and supplementing this with the most-followed accounts in each cluster. This established that the two clusters could be categorized as roughly divided across American political lines (Right-leaning and Left-leaning). Finally, we located the RU-IRA accounts in the graph. More details on this process and its results are included in the Findings section.

5.5.3 *Identifying RU-IRA Accounts*

Having established the broader context of the retweet graph, we next looked for the RU-IRA accounts. To identify RU-IRA-affiliated accounts in this dataset, we relied on a list of 2,752 suspended RU-IRA accounts released by Twitter in November 2017 as part of their testimony before the U.S. House of Representatives Permanent Select Committee on Intelligence (U.S. House of Representatives Permanent Select Committee on Intelligence 2017; U.S. Senate Select Committee on Intelligence 2017).

In the initial keyword-filtered dataset, cross-referencing with Twitter's list revealed that 96 RU-IRA accounts from Twitter's list were present in the data—the subset of RU-IRA troll accounts who tweeted at least once with #BlackLivesMatter, #BlueLivesMatter, or #AllLivesMatter. After filtering by retweet degree and limiting to the two large communities as described above, the number of RU-IRA accounts in our dataset was reduced to 29. We can summarize this subset as the RU-IRA accounts who participated via retweeting or being retweeted at least twice in the network. As described above, the purpose of this filtering was to find those accounts that were relatively well integrated into the information network, meaning that this subset of RU-IRA accounts generally interacted more with the network surrounding them. Though this limited the number of RU-IRA accounts we examined, it allowed us to focus our subsequent qualitative analysis on those accounts that likely had greater visibility and perhaps greater potential for influence within the network.

5.5.4 *Qualitative Analysis*

After examining the position of known RU-IRA accounts in relation to other accounts in the network, we began an analytic accounting of how these 29 accounts participated in *LM discourse. These accounts produced 109 tweets (retweeted 1,934 times) in our *LM collection, which we used as an initial sample in our qualitative inquiry. This data helped us develop some initial interpretations, but our constructivist grounded approach required further data collection via theoretical sampling to check, fill out and extend our theoretical categories.

We therefore supplemented our analyses using data from the Internet Archive's Wayback Machine, a free and open-source internet archive that save webpages (The Internet Archive n.d.) through a variety of web crawls being run by different programs. Searching this archive, we were able to manually retrieve 234 timeline snapshots—including profile content as well as 4,682 tweets and retweets—for these accounts. While timelines for these accounts are not systematically preserved, this content provides a window into the RU-IRA trolls' digital presence in ways that mitigate the limitations of keyword sampling and thus complement our other data. The snapshots also allow us to see how each account presented itself, including elements like profile images that were otherwise unavailable since Twitter had suspended the account.

We considered three main units of analysis (in addition to the network graph). First, we examined profile data—i.e. the display pictures, background images and profile descriptions of the RU-IRA accounts. Second, we considered tweets with a focus on the original content produced by these accounts, including embedded images such as memes. We also paid close attention to cases in which these accounts retweeted each other. Third, we considered the external websites, social platforms and news articles these accounts linked to in an effort to "follow the person" (Marcus 1995) to attain a more holistic understanding of the disinformation campaign we were studying.

Each of these types of data was examined, segmented and summarized through an initial round of open coding. Our codes focused on actions visible in the data and leveraged our prior contextual knowledge from having studied this particular #BlackLivesMatter-related discourse. These initial codes which fragmented the data were then drawn together through analytical memoing and clustering to form themes and categories.

5.5.5 *Methodological Challenges*

This study confronted three main methodological challenges that must be understood to interpret our findings correctly. First, the seed Twitter data we used to generate our network graph is both incomplete (due to rate limits) and biased (because of the shooting related terms we tracked). As a result, our findings are not intended to be representative of the overall #BlackLivesMatter conversation. Rather, we have a portion of a particular online discourse that invokes the movement in conjunction with incidents of violence during 2016. Similarly, due to the incomplete nature of our data, we cannot and do not seek to quantitatively assess the impact RU-IRA activities and contributions had on even this one discourse. Our goal is to understand how RU-IRA content was designed to interact with this discourse—which we already understand to be polarized and made up of a heterogenous web of actors who are speaking to different interests and values.

Second, it is important to note that the identification and suspension of RU-IRA affiliated accounts is likely part of an evolving and ongoing effort at social media companies. We do not have access to Twitter's methodology for identifying these accounts, but we do know that at least one of the 2,752 accounts was revealed to be a false positive (i.e. unaffiliated with the Internet Research Agency) (Matsakis 2017). Moreover, Twitter has identified additional RU-IRA accounts since the release of this initial list but has not made information on these accounts publicly available to our knowledge (Twitter 2018). Independently, we have tracked more accounts being suspended in both clusters—but particularly on the right—since we conducted this analysis (although we cannot infer that these accounts were RU-IRA affiliated). Consequently, we wish to caution readers from drawing any false equivalencies from the fact that we located and subsequently examined 22 RU-IRA accounts in the left-leaning cluster and 7 in the right-leaning cluster.

Third, despite the generally presumed persistence of social media content, the content associated with clandestine activities is prone to ephemerality, creating challenges for research (Farkas, Schou, Neumayer 2018a; Shein 2013). Our multi-sited research approach—using of Internet Archive data, examining linked websites and considering the activities of these accounts on other social platforms—attempts to address these challenges by acknowledging that information operations on these platforms are interconnected and interrelated activities.

5.6 Findings

5.6.1 *Structural Analysis: Positioning Across Political Lines*

We now return to the accounts in the dataset identified in section 4.1 which both tweeted with an *LM keyword and were well-integrated into the retweet network. Figure 5.1 illustrates each step of our analysis of the information flow graph, where the 22,020 Twitter accounts are nodes and the 58,698 retweets between these accounts are directed edges. In our first step, we visualized the structure of the graph, noting that the majority of nodes are concentrated in two relatively distinct clusters. This observation suggests homophily in the accounts retweeting each other. To solidify this, our next step was to use a community detection algorithm to systematically identify clusters. Specifically, we used the Infomap algorithm, an optimization of the Map Equation that assigns nodes to a community using a greedy algorithm that optimizes flow (in this case retweets) between nodes. The results of this step supported our earlier observation of structural homophily: 91.7% (20,192) of the nodes are grouped in two large clusters in the center of graph containing 48.5% and 43.2% of the nodes. We focus our remaining investigation on these two clusters (colored pink and green in Figure 5.1).

Our final step was to understand who was in the clusters. To do this, we used salient account characteristics—the top 10 hashtags in the accounts' profile descriptions as well as the most-retweeted accounts by cluster—to classify and contrast the two clusters (shown in Table 5.1). In both clusters, the number of accounts with a hashtag in the user description ranged from 31.6% to 34.2%. This analysis revealed that our graph was roughly divided along political lines. The most frequently occurring hashtags in the pink community bios were #BlackLivesMatter, #ImWithHer (expressing support for Democratic presidential candidate Hillary Clinton), and #BLM (a shortening of #BlackLivesMatter). #BlackLivesMatter is the top hashtag by a significant amount. We also see that left-leaning journalist and activist @ShaunKing and pro-BLM news account @trueblacknews are in the top ten most-retweeted accounts of this community. Therefore, we categorize this cluster as broadly Left-leaning on the U.S. political spectrum. In contrast, the most frequent hashtags in the green community were #Trump2016, #MAGA, and #2A, where #Trump2016 and #MAGA indicate support for Republican presidential candidate Donald Trump and #2A indicates support for the right of private citizens to own guns. Nearly 7% of the accounts in this cluster had #Trump2016 in their user descriptions. We categorize this cluster as broadly Right-leaning on the U.S. political spectrum. Building upon previous work (Stewart et al. 2017), we infer that these two communities held divergent and competing frames surrounding officer-involved shootings and the Black Lives Matter and Blue Lives Matter movements.

Next, we identify accounts from within our data that were associated with the RU-IRA and examine their location within the retweet network graph. In total, there were 96 RU-IRA accounts within our dataset but only 29 of these appeared in our retweet network graph (limited to accounts with a retweet degree of at least two and within the two clusters). 22 of these accounts were in the left (pro-BLM) cluster and 7 of these accounts were in the right (anti-BLM) cluster.

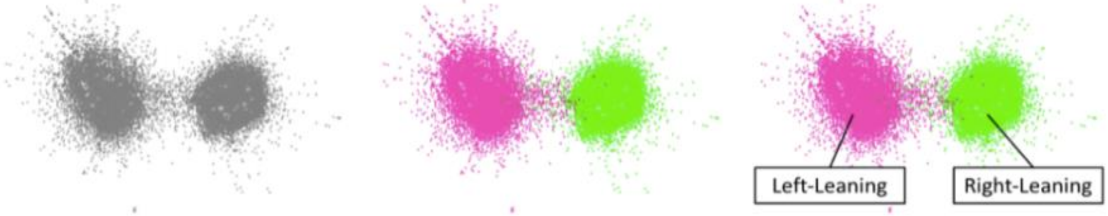


Figure 5.1 From left to right: using Force Atlas 2 to visualize retweet flows, identifying clusters with Infomap, and using cluster characteristics to label communities

Table 5.1 Overview of Accounts in the Two Clusters

Color	Top 10 hashtags in account descriptions	Number of accounts	Top 10 accounts by retweet count
Pink	blacklivesmatter (8.529%), imwithher (1.442%), blm (1.105%), uniteblue (1.039%), feelthebern (1.021%), allblacklivesmatter (0.721%), bernieorbust (0.599%), neverhillary (0.571%), nevertrump (0.571%), freepalestine (0.524%)	10681	trueblacknews (3773), YaraShahidi (2108), ShaunKing (1553), ShaunPJohn (1214), BleepThePolice (692), Crystal1Johnson (573), DrJillStein (524), meakoopa (409), kharyp (387), tattedpoc (307)
Green	trump2016 (6.615%), maga (6.099%), 2a (5.237%), tcot (2.787%), trump (2.776%), neverhillary (2.524%), makeamericagreatagain (2.461%), nra (2.229%), trumptrain (1.998%), bluelivesmatter (1.872%)	9509	PrisonPlanet (4945), Cernovich (1704), LindaSuhler (1034), MarkDice (789), DrMartyFox (758), _Makada_- (591), andieiamwhoiam (510), LodiSilverado (500), BlkMan4Trump (458), JaredWyand (447)

These 29 accounts also demonstrated a wide range of engagement: @BleepThePolice was retweeted 692 times by 614 distinct accounts on our graph while six RU-IRA accounts were not retweeted at all. The top-ten most prominent RU-IRA accounts by retweet count—such as @BleepThePolice, @Crystal1Johnson, and @BlackNewsOutlet on the left and @SouthLoneStar, @TEN_GOP, and @Pamela_Moore13 on the right—are highlighted in Table 5.2. Cross-referencing Table 5.1 and Table 5.2, we note that in the left cluster, two RU-IRA accounts (@BleepThePolice and @Crystal1Johnson) are among the left cluster's most-retweeted accounts.

Figure 5.2 highlights the trajectories of retweets of RU-IRA accounts (orange) in the rest of the graph (blue). Of the 58,698 total retweet edges on the graph, 1,960 (3.33%) were retweets of RU-IRA accounts. We do not attempt to tackle the question of the influence of RU-IRA accounts with this graph, but rather to illustrate their position in the ecosystem. While we cannot speak to their impact, we can use this graph to examine where their content circulated and, in tandem with qualitative analysis, identify their tactics and apparent coordination practices and situate these within our current knowledge of information operations.

An initial—and striking—observation is that there were clearly RU-IRA accounts embedded in both clusters, meaning that RU-IRA content was retweeted on both 'sides' of the conversation. Furthermore, we can see that while RU-IRA content spread throughout each community—and in some cases was relatively highly retweeted—it very rarely moved between them. Informed by prior work examining divergent framing (Stewart et al. 2017), this suggests an effort by the RU-IRA to purposefully embed themselves in two distinct communities on either side of a highly charged framing conflict.

Table 5.2 Prominent RU-IRA Accounts Ordered by Cluster and Number of Retweets

Handle	Cluster (Left or Right)	Number of Tweets in Dataset	Number of Retweets in Cluster	Follower Count
@BleepThePolice	L	18	692	11,926
@Crystal1Johnson	L	14	573	16,510
@BlackNewsOutlet	L	2	60	4,723
@gloed_up	L	15	53	17,876
@BlackToLive	L	2	47	7,072
@nj_blacknews	L	2	35	1,992
@blackmattersus	L	2	34	5,841
@SouthLoneStar	R	2	225	15,612
@TEN_GOP	R	1	45	18,451
@Pamela_Moore13	R	1	23	9,289

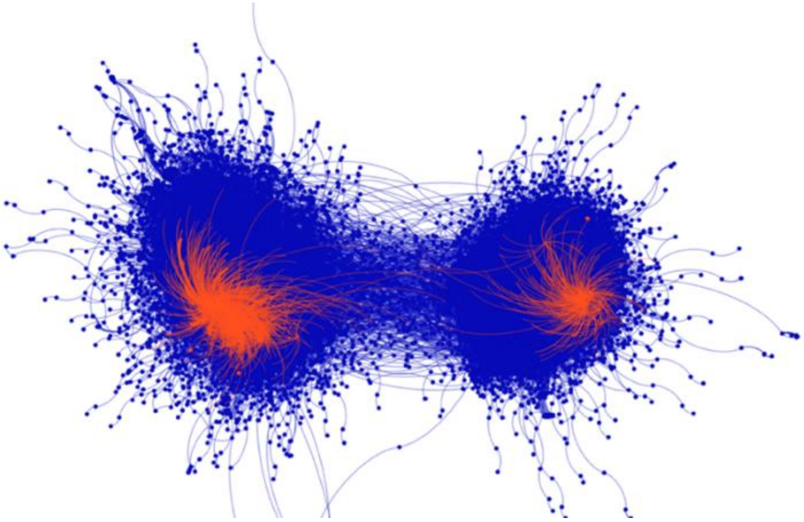


Figure 5.2 Highlighting retweets of known RU-IRA accounts (orange) compared to retweets of the rest of the graph (blue).

We can summarize these findings by stating that while RU-IRA content was clearly broadcast to both clusters, the RU-IRA content that circulated in each cluster originated from two distinct groups of RU-IRA accounts. With the inference that these communities hold oppositional and incompatible beliefs surrounding officer-involved shootings and race, this suggests that the RU-IRA accounts tailored content to each community. This aligns with previous literature claiming that current disinformation tactics are ideologically fluid and seek to exploit social divides (Paul and Matthews 2016; Pomerantsev and Weiss 2014).

We also note that while the presence of orange nodes and edges appears larger in the left-leaning cluster, the limitation of our original dataset and the curated list of RU-IRA accounts provided by Twitter prevent any quantitative comparisons between the two sides. In other words, this graph provides a window into RU-IRA activity and patterns but does not determine relative impact.

5.6.2 *Production of Inauthentic Identities*

Our network analysis reveals that RU-IRA affiliated accounts interacted with two different networked audiences in this large-scale discourse (politically left leaning and right leaning). For the remainder of our analysis we will focus on the orange nodes in Figure 5.2 to understand the nature of these interactions and how these accounts adapted to fit within the two structurally distinct communities. We begin by considering how these accounts presented themselves. This helps us understand how processes of feigning authenticity have evolved and adapted to social media environments, which contain less static and more user-driven content production and a networked architecture that blurs the lines between contexts like entertainment and news consumption. This also helps us triangulate the extent to which the RU-IRA accounts in Figure 5.2 intentionally targeted different audiences, since how the operators of these accounts attempted to portray themselves reflects their *imagined audience*

(Litt 2012; Marwick and boyd 2011) —i.e. the mental pictures people construct about others to guide self-presentation. Just as writers imagine media audiences appropriate to their topic and form and use textual cues to invoke those audiences into being (Ong 1975), the differences and similarities across RU-IRA profiles reveals who these accounts were attempting to write to and deceive.

5.6.2.1 Profiles: Like many other social media participants, RU-IRA affiliated Twitter accounts constructed user profiles to portray both an interesting and authentic self. These profiles were reproduced on other platforms like Facebook and Tumblr, suggesting an effort to build and maintain consistent online personas.

We observed four systematic patterns of forged profiles. The first two were the establishment of 'the proud African American' as a political identity, on the one hand, and the articulation of 'the proud White Conservative', on the other. These two patterns consisted of accounts that presented themselves as the personal Twitter accounts of real and ordinary citizens within their communities. These accounts used cultural, linguistic, and identity markers in their Twitter profiles to align themselves with the shared values and norms of either the left- or right-leaning clusters. For instance, accounts in the left-leaning cluster that fell in this category consistently used display pictures to present themselves as African Americans coming from locations such as Chicago, New Jersey, and Richmond, Virginia with profile descriptions such as:

> @TrayneshaCole: "Love for all my people of Melanin. Your BLACK is BEAUTIFUL! #MyPussyMyChoice #BlackGirlsMagic #BlackLivesMatter"
>
> @Crystal1Johnson: "It is our responsibility to promote the positive things that happen in our communities."
>
> @4MySquad: "no black person is ugly #BlackLivesMatter #StayWoke"

Accounts in the right-leaning cluster tended to use photographs to present themselves as white men and women living in Texas or other southern states who were interested in firearms and the right to bear them, using profile descriptions like:

> @TheFoundingSon: "Business Owner, Proud Father, Conservative, Christian, Patriot, Gun rights, Politically Incorrect. Love my country and my family #2A #GOP #tcot #WakeUpAmerica"
>
> @Pamela_Moore13: "Southern. Conservative. Pro God. Anti Racism"
>
> @USA_Gunslinger: "They won't deny us our defense! Whether you're agree with me or not, you're welcome here! If you don't want to be welcomed, go f*ck yourself."

These profiles can appear to be the online personas of real African and White Americans because they appeal to creative self-expression and caring for others. Another part of what can make these personas intuitively 'fit' comes from how they invoke stereotypical thinking by articulating African and White Americans as binary groups that are internally homogenous with respect to politics. In the past, such dichotomizations have been directly and indirectly constructed by media portrayals elsewhere (Dijk 2015; Downing and Husband 2005). But by exploiting the participatory and interactive nature of social media, imaginary *others* can be brought to life in new ways by information operations in order to sustain and amplify these dichotomizations (Farkas, Schou, and Neumayer 2018b).

The third and fourth patterns mirrored the first two, but enacted organizational accounts for grassroots political and media groups from these respective 'sides.' For instance, accounts in the right-leaning cluster adopted names like @tpartynews, using a "Tea Party" teapot logo in the colors of the American flag and acting as a conservative news source. Similarly. @TEN_GOP, a well-known RU-IRA affiliated account (Griffin and O'Sullivan 2017) that appeared in our dataset, described itself as the "Unofficial Twitter of Tennessee Republicans. Covering breaking news, national politics, foreign policy and more. #MAGA #2A." In the left-leaning cluster, these accounts presented themselves as alternative media sources for racial justice. These accounts emphasized localness, frustration with mainstream media, and crowd participation, respectively, with profile descriptions like:

> @nj_blacknews: "Latest and most important news about New Jersey black community"
>
> @Blackmattersus: "I didn't believe the media so I became one."
>
> @BlackToLive: "We want equality and justice! And we need you to help us. Join our team and write your own articles! DM us or send an email: BlackToLive@gmail.com"

These accounts often linked back to their own websites, which suggests an attempt to undermine traditional media in favor of alternative media websites that might have been setup to support the information operation. For instance, the account @dontshootcom links to the domain donotshoot.us, which describes itself as a tool for empowering grassroots activists:

> "Don't Shoot is a community site where you can find recent videos of outrageous police misconducts, really valuable ones but underrepresented by mass media. We provide you with first-hand stories and diverse videos. Our mission is to improve the situation in the US and the lives of its citizens, to do our best to help end inhumane and biased acts. We are here to empower you, give you a voice and help you get justice with all our might."

Figure 5.3 summarizes how RU-IRA accounts used profile display pictures to foster identities that could attract and command attention from audiences with different political alignments and news consumption habits. Viewing these images collectively in this manner reveals both convergence and divergence in the production dynamics governing how these identities were crafted. The consistent and similar nature of these fake identities (within any one of the single 'quadrants' below) suggests convergence: that perhaps a common script, manual or 'brand bible' (Ong and Cabañes 2018) may have been used to delineate the political stances, social background and personality traits of these accounts. Ensuring this kind of brand or identity consistency aligns with professional practices of micro-targeting in marketing and American political campaigning that have evolved to take advantage of the capabilities of social media platforms (Murray and Scime 2010).

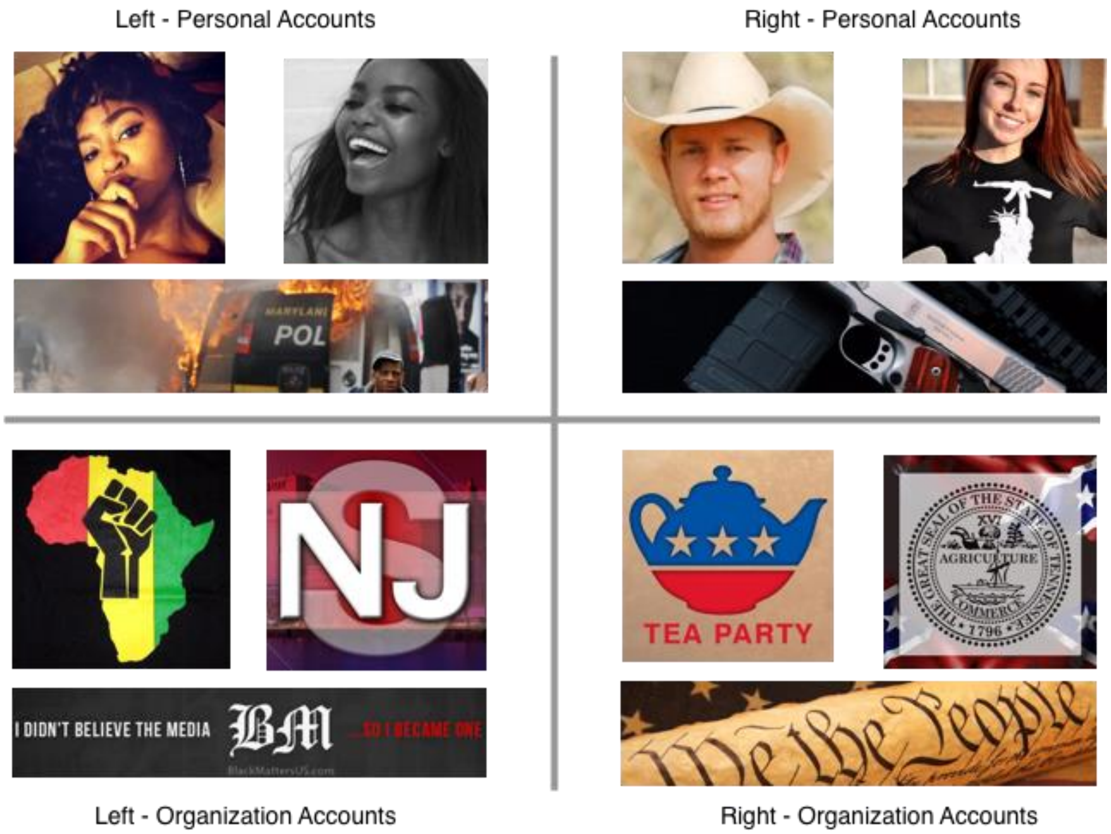

Left - Personal Accounts
Right - Personal Accounts
POL
NJ
I DIDN'T BELIEVE THE MEDIA
BM
...SO I BECAME ONE
BlackMattersUS.com
TEA PARTY
We the People
Left - Organization Accounts
Right - Organization Accounts

Figure 5.3 Display pictures of RU-IRA accounts arranged by categories.

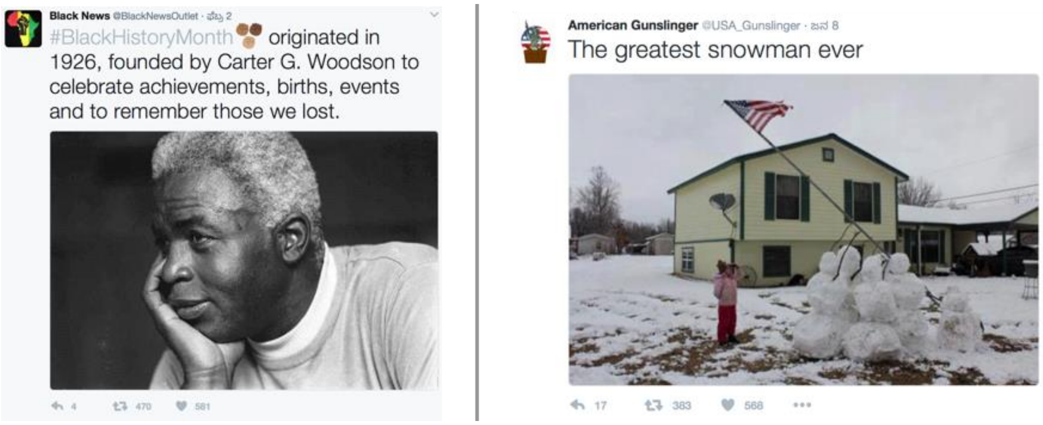
Simultaneously, the differences in these identities (between the left/right or upper/lower sides of Figure 5.3) suggests efforts to engage in audience segmentation and having multiple audience touchpoints. For instance, by delivering either a personal identity or a more organizational one, RU-IRA accounts collectively took advantage of how social contexts 'collapse' together on sites like Twitter to promote messages to audiences through different points of access. Researchers have noted that trying to balance these contexts through a single account opens the possibility of appearing inauthentic to one's followers (Marwick and boyd 2011)—a risk the RU-IRA mitigated by having accounts specialize in different roles.

5.6.2.2 Tweets: Beyond creating a fake profile, the RU-IRA accounts produced tweets containing commentary, images, news and videos that helped shape, reproduce and solidify the political identities they enacted. RU-IRA accounts with both 'personal' and 'organizational' profiles in the left-leaning cluster frequently tweeted to uphold the accomplishments and culture of African Americans and share positive feelings around the Black Lives Matter movement. For instance, @Crystal1Johnson maintained a pinned tweet about how Muhammad Ali's Hollywood Walk of Fame Star is unique for 'hanging on a wall, not for anyone to step on' and actively celebrated Black History Month by tweeting regularly about topics like African American women's hairstyles and accomplishments in education. Similarly, accounts like @TrayneshaCole, @gloed_up, @BlackToLive, @RobertEbonyKing and @BlackNewsOutlet tweeted in support of entrepreneurship projects by African Americans and locating missing Black persons. The expression of personal opinions on events, and the use of humor and entertainment also featured prominently as these accounts also tweeted about music by African American artists and joked around movies like Black Panther and Hidden Figures in which African Americans played prominent roles.

Similarly, accounts in the right leaning cluster tweeted to celebrate traditional American holidays, the American flag, and military service. For instance, @TheFoundingSon maintained a pinned tweet for #PearlHarborRemembranceDay as "a reminder to the rest of the world that American people cannot be easily broken." Similarly, @SouthLoneStar also pinned a tweet that told the personal story of "Nick [who] was paralyzed by an IED in Afghanistan. Wendy met him in VA hospital and became his caregiver full-time. Now these 2 heroes are married." Moreover, just as left-leaning RU-IRA accounts tweeted about certain movies and occasions like Black History Month, these accounts made it a point to celebrate traditional American holidays like Thanksgiving and Easter while commenting on television shows with hashtags like #TheWalkingDead. Another example from @SouthLoneStar is illustrative here:

> "Today is National Peace Officer Memorial Day. We honor those that paid the ultimate sacrifice #BlueLivesMatter"

Other accounts like @USA_Gunslinger and @KarenParker93 followed similar patterns and used hashtags like #WednesdayWisdom to tweet pictures of snowmen holding up an American flag (see Figure 5.4) and children pretending to be police officers.



Black News @BlackNewsOutlet · ಫೆಬ್ರ 2

#BlackHistoryMonth originated in 1926, founded by Carter G. Woodson to celebrate achievements, births, events and to remember those we lost.

American Gunslinger @USA_Gunslinger · ಜನ 8

The greatest snowman ever

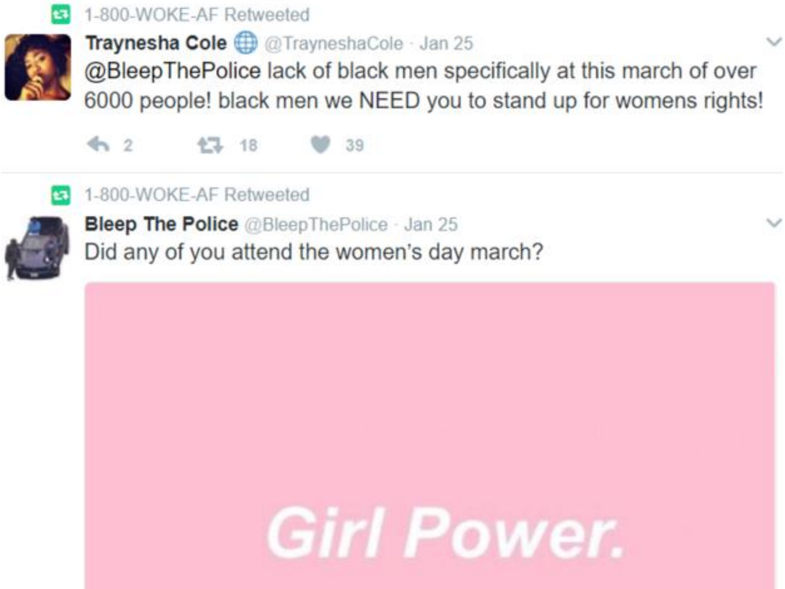
Figure 5.4 Sample tweets circulated by RU-IRA accounts in separate clusters to cultivate trust.

Other accounts like @USA_Gunslinger and @KarenParker93 followed similar patterns and used hashtags like #WednesdayWisdom to tweet pictures of snowmen holding up an American flag (see Figure 5.4) and children pretending to be police officers.

These examples highlight how information operations can invoke content that is not always amenable to fact-checking nor straightforward to problematize. The activities of these accounts included not only acts of 'rational' political persuasion like presenting arguments and true or false claims. They also involved representing and affirming the personal experiences, shared beliefs and cultural narratives of their audiences. This could help these accounts blend into the communities they targeted, and it could also help them tap into the social and emotional literacies that often guide people's engagement with the public sphere.

Although the consistency of this content speaks to a certain level of rigid arrangements (e.g. accounts on the left ought to celebrate Black History Month), the content also serves to illustrate a level of spontaneity. For instance, multiple accounts demonstrated the ability to understand the nuances of American pop-culture and creatively adapt to trending topics to 'build their brand' (e.g. opining about movies, music and television shows). Aligning with investigative interviews with former RU-IRA employees (Troianovski 2018), we would suggest that these dynamic behaviors are a signal that these accounts were not fully automated bots—and that the workers operating these accounts had at least some agency to 'improvise' as part of their work.

5.6.2.3 Coordination to Build Trust: On social media, interacting with streams of user-generated content produced by one's personal network is central to exhibiting 'evolving connectivity' (Papacharissi 2009) and cultivating trust (Farkas, Schou, and Neumayer 2018a). We did not observe explicit interaction between RU-IRA accounts when they were in different clusters, but we did observe accounts from within the same cluster mentioning and retweeting each other over a variety of topics. For instance, for a researcher reading their content, the users @gloed_up, @BleepThePolice and @TrayneshaCole gave the impression that they were part of a social clique. Their occasional, casual interactions projected authenticity while also enabling them to better manage their audience's attention by generating 'buzz' around certain topics such as protests or other news items. Figure 5.5 below furnishes an example that succinctly captures the flavor of interactions between these accounts.



1-800-WOKE-AF Retweeted

Traynesha Cole @TrayneshaCole · Jan 25

@BleepThePolice lack of black men specifically at this march of over 6000 people! black men we NEED you to stand up for womens rights!

2 18 39

1-800-WOKE-AF Retweeted

Bleep The Police @BleepThePolice · Jan 25

Did any of you attend the women's day march?

Figure 5.5 Three RU-IRA accounts retweeting each other.

In this example, @BleepThePolice tweeted out a graphical meme touting "Girl Power," celebrating the march and asking if anyone is attending, perhaps with the goal of getting responses—and therefore engagement—from that account's audiences. @TrayneshaCole answers that call with a tweeted reply message pleading for black men to get more involved in women's rights. Later, @gloed_up—whose screen name is 1-800-WOKE-AF—retweets both tweets. This example shows the three RU-IRA accounts interacting with each other to create the illusion of organic engagement.

Retweet flows provide an incomplete picture of how RU-IRA accounts supported each other's activities. A richer window into understanding how the RU-IRA coordinated and provided mutual support to each other to appear as authentic activists and influencers comes from @BlackMattersUS. A website associated with this account was promoted on Twitter by @Crystal1Johnson, and the site in turn credits Crystal Johnson as a writer who interned at NBC:

> "Crystal Johnson has been with Black Matters since October 2014. Her passion is giving voice to the community. During her undergrad, Crystal took an internship with the local NBC affiliate WEYI. In 2014 she moved to Atlanta to help start a new project called BlackMatters. She is among the most active members of BlackMatters."

Aligning with journalistic investigations by Craig Silverman, we also observed that @BlackMattersUS took the step of creating and promoting multiple meetups, possibly to create links—or project the illusion of having links—with real, local organizing groups. These meetup related efforts were also supported by accounts like @Crystal1Johnson who recruited volunteers and @Blacktivists who set up a ‘Black Unity March’.

> @BlackMattersUs: “If you are against #policebrutality #racism #incarceration #oppression take part in #BlackLivesMatterMarch <link to meetup>”
>
> @BlackMattersUs: “Support Black Owned Small business at this one stop shop expo event!!! #BLM #BlackLivesMatter <link to meetup>”
>
> @Crystal1Johnson: “We’re looking for good people who are ready to help us in organizing events around the country. DM for more info”

The BlackMattersUS website also put together a podcast on SoundCloud called ‘SKWAD 55’ to ‘gather strong Black voices’ [29] (Black Matters US n.d.b; Sound Cloud n.d.), which was promoted by accounts like @4MySquad which positioned themselves as interested in rap music. These examples illustrate how RU-IRA accounts collaborated to feign legitimacy via multiple channels and platforms.

5.6.3 *RU-IRA Participation in #*LM Discourse*

We have described how RU-IRA accounts carefully constructed fictitious identities as people and organizations with ethno-cultural backgrounds that systematically shifted depending on whether the account was embedded within the politically left- or right-leaning cluster. In this section we will summarize RU-IRA content related to #BlackLivesMatter, #BlueLivesMatter and #AllLivesMatter. We organize this content into three different patterns to show how a seemingly diffuse set of individual actors on social media worked together to amplify certain messages.

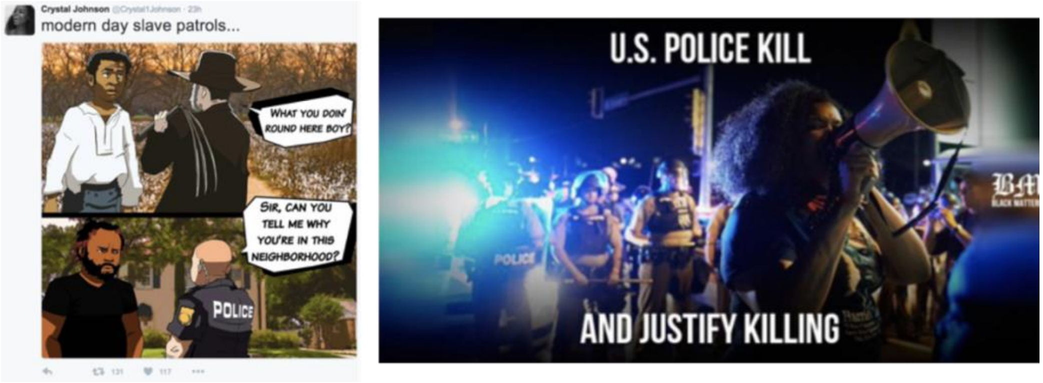
5.6.3.1 Modeling the ‘anti-Police’ #BlackLivesMatter protestor: Each RU-IRA account that we examined in the left-leaning cluster connected their African-American identity to being a #BlackLivesMatter activist by tweeting extensively about police officers shooting unarmed African American men and women, including disabled persons and minors. These tweets frequently linked to stories from established media sources such as Fox News (2016) and the New York Times (Moynihan 2016) but also alternative media sources including conspiracy theory and RU-IRA affiliated sites such as TheFreeThoughtProject (Agorist 2016) and BlackMattersUS. The process of mixing ‘traditional’ and alternative media sources into a single content stream is notable because it can elevate the image and

[29] For the purposes of this dissertation, I have chosen to separate academic references from references to content that I have problematized as misleading or potentially triggering. I have placed references to the latter at the end of the chapter (and to the former in the Works Cited section towards the end of the dissertation). This has been done to avoid driving traffic to it and blending it with more credible information. The references to problematic information are formatted as superscripts to distinguish them.

content of the more alternative sites, particularly for audiences that skim headlines to cope with high volumes of information.

These accounts also used their political identities of African-American #BlackLivesMatter activists to model an exuberant anti-police stance via tweets, profile background images, and occasionally account names. Accounts like @Bleepthepolice, @gloed_up, and @4mysquad combined hashtags like #BLM and #BlackLivesMatter, with #ACAB (short for all cops are bastards), #Amerikkka, #BadCop, #BleepThePolice, #CowardCops, #HateIt, #KillerCops and #riot:

> @4MySquad: "they don't hire anyone with an iq of over 100' #StayWoke #Police #dumb #AllCopsAreBad #ACAB"
>
> @GloedUp: "French #police are too corrupt, incompetent to fight terrorism #BlackTwitter #BlackToLive #BlackLivesMatter #acab"
>
> @Crystal1Johnson: "Blue's a job, that shit don't matter! #BlackLivesMatter!"



Crystal Johnson @Crystal1Johnson · 23h

modern day slave patrols...

131 117

WHAT YOU DOIN' ROUND HERE BOY?
SIR, CAN YOU TELL ME WHY YOU'RE IN THIS NEIGHBORHOOD?
POLICE
U.S. POLICE KILL
POLICE
BM
BLACK MATTERS
AND JUSTIFY KILLING

Figure 5.6 Example memes circulated by RU-IRA accounts in the left cluster.

Figure 5.6 above also illustrates how the memes these accounts presented favored an uncompromising and adversarial stance towards law enforcement. The use of these charged messages and vocabulary of hashtags in conjunction with the central political tag of #BlackLivesMatter suggests an attempt by RU-IRA accounts to connect with both existing discontent and amplify it by proliferating certain meanings around the #BlackLivesMatter tag—similar to the phenomenon of hashtag drift (Booten 2016).

This activity feeds directly into attempts to frame #BlackLivesMatter as an anti-police hate-group. From prior research (Black Live Matter Vermont 2017) we know that such framings were actively resisted and addressed by #BlackLivesMatter activists while being proliferated within anti-BlackLivesMatter discourse. By tapping into this larger reservoir of antagonistic discourses proliferating in American politics, these accounts amplified toxicity in public discussions. This is further supported by how these accounts invoked the competing hashtags #BlueLivesMatter and

#AllLivesMatter to attack them. 'Calling out' these hashtags illustrates how these accounts did not just speak to the communities that they were pretending to be a part of, but also aimed to communicate an antagonistic representation of those communities to others.

> @BleepThePolice: "#BlueLivesMatter is BS"
>
> @TrayneshaCole: "And y'all not saying #AllLivesMatter when y'all are shooting up schools now are you?"

Finally, it is significant that not all of the stories about police misconduct that were circulated by these accounts were verified or grounded in fact. One notable example in our data that highlights the creativity of these accounts, and which has been decisively debunked elsewhere (Silverman 2018), relates to @4mysquad circulating gifs with the description "Shocking video shows Black teenage girl being sexually assaulted by NYPD officer." These gifs were framed as surveillance video footage showing a black teenager being assaulted by a white police officer, and they were also presented on @4mysquad's Tumblr account. Following these gifs going viral, members of the online crowd began to refute and debunk this story. At this point BlackMattersUS tweeted and published a website article that linked to the gifs and attempted to refute the corrections (Black Matters US n.d.a) (Silverman 2018). @4mysquad ultimately went on to issue an apology, stating:

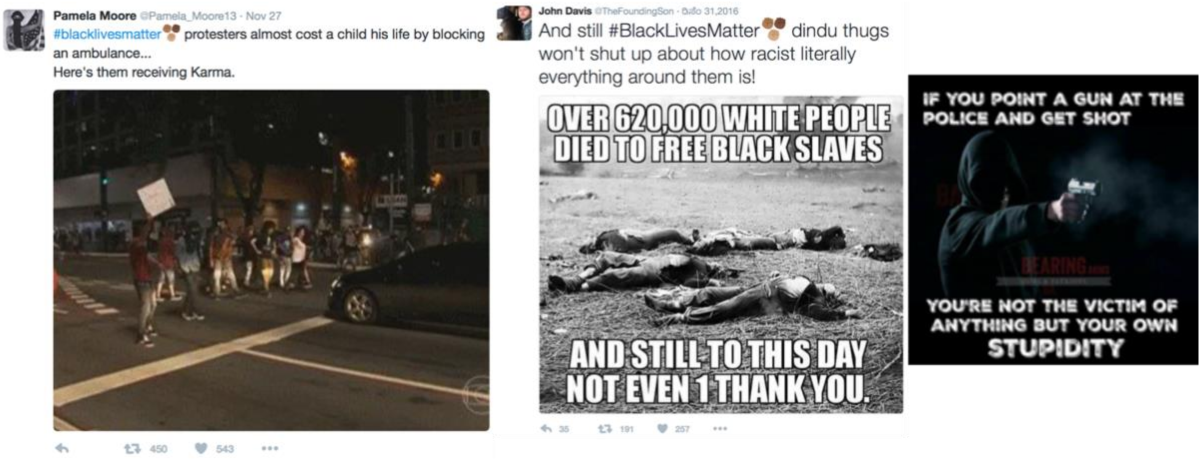
> "it was absolutely insensitive of me to make those gifs. I was furious and stoned...originally I've got dis anonymous message asking me to make a post…" (quoted in Silverman 2018)

This example represents a creative and intentional attempt to inject false information into the #BlackLivesMatter discourse. The apology suggests again that these accounts were not fully automated 'throw-away' bots since they were managing their 'brand', disguise, and audience by monitoring and responding to feedback. The involvement of the BlackMattersUS website illustrates how RU-IRA accounts worked to sow anger and confusion over multiple channels and platforms. Examined as a two-part act, the video incident functioned both to further stoke anti-police sentiments on the left and, once it was debunked, increase anti-BlackLivesMatter sentiments on the right.

5.6.3.2 Promoting anti-BlackLivesMatter discourse: Diverging from their counterparts, RU-IRA accounts in the right leaning cluster tweeted to both support #BlueLivesMatter and #AllLivesMatter and denigrate #BlackLivesMatter. These tweets delegitimized the #BlackLivesMatter movement by equating the meaning of the movement with propaganda and anti-police activities. @tpartynews and @TEN_GOP, for instance, engaged in this type of framing by tweeting out stories around the 2016 Baton Rouge and Dallas shootings of police officers with titles like "Mother of police shooting suspect blames #BlackLivesMatter," and "WATCH: #BlackLivesMatter supporters interrupt a moment of silence for fallen police officers!" The personal category of RU-IRA accounts in this cluster also attacked #BlackLivesMatter more directly.

> @Pamela_Moore13: "Black Lives Matter is a political construct, a hateful destructive ideology. It's never been about black life."
>
> @KarenParker93: "RT: If U Point A Gun At A Cop & Get Shot, Who's Stupid #BlueLivesMatter"
>
> @TheFoundingSon: "Black man intentionally drives through 3 cops. That is hate that #BLM and Obama created #BlueLivesMatter"

The additional examples provided in Figure 5.7 also highlight how these accounts made heavy use of aggressive memes and images. Overall, these tweets play a complementary role with the content RU-IRA accounts were propagating in the left leaning cluster. Supporters and followers of the #BlackLivesMatter hashtag could potentially see this charged content and use it in forming their perceptions of others and the possibility of civil dialogue. Simultaneously, critics of the #BlackLivesMatter movement could see RU-IRA content that focused more on attacking police and less on the movement's core messages. Both groups of users were also being selectively presented with news and information from these accounts that possibly played to pre-existing beliefs and biases (e.g. #BlackLivesMatter affiliated protesters behaving as looters and executing police officers / police officers sexually assaulting black citizens). In summary, RU-IRA accounts were acting as both information distributors and antagonistic stereotypes of ethno-cultural others.



Pamela Moore @Pamela_Moore13 · Nov 27
#blacklivesmatter protesters almost cost a child his life by blocking an ambulance...
Here's them receiving Karma.

450 543

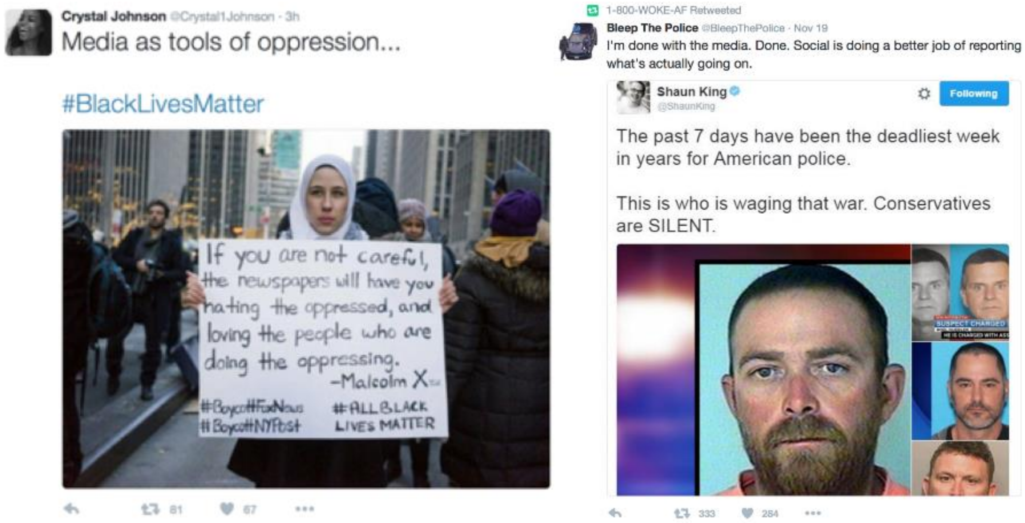
John Davis @TheFoundingSon · ธ.ค. 31,2016
And still #BlackLivesMatter dindu thugs won't shut up about how racist literally everything around them is!

35 191 257

OVER 620,000 WHITE PEOPLE DIED TO FREE BLACK SLAVES
AND STILL TO THIS DAY NOT EVEN 1 THANK YOU.
IF YOU POINT A GUN AT THE POLICE AND GET SHOT
YOU'RE NOT THE VICTIM OF ANYTHING BUT YOUR OWN STUPIDITY

Figure 5.7 RU-IRA content about #BlackLivesMatter in right-leaning cluster.

5.6.3.3 Converging to attack the 'mainstream' media: RU-IRA accounts in both clusters converged by using #BlackLivesMatter discourse and their constructed political identities to criticize the 'mainstream media'. The @BlackmattersUS profile description and website slogan of "I didn't believe the media so I became one" effectively summarizes this message, which was also carried forward by personal style RU-IRA accounts on the left. These accounts mixed content that A) expressed frustration with how older traditional media institutions cover issues like officer related shootings and the #BlackLivesMatter movement itself; and B) equated these long-standing institutions with tools of oppression. Figure 5.8 illustrates more and less direct versions of this message. The second tweet in this example shows @BleepThePolice (boosted by another RU-IRA account) repurposing a message by @ShaunKing to hold up social media as a viable alternative to "the media."



Crystal Johnson @Crystal1Johnson · 3h
Media as tools of oppression...

#BlackLivesMatter

1-800-WOKE-AF Retweeted
Bleep The Police @BleepThePolice · Nov 19
I'm done with the media. Done. Social is doing a better job of reporting what's actually going on.

Shaun King @ShaunKing
The past 7 days have been the deadliest week in years for American police.

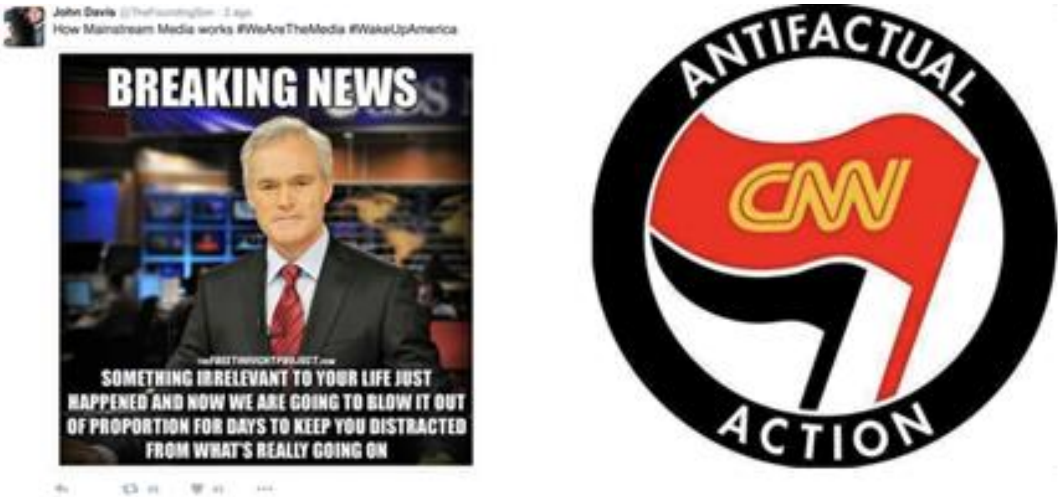
This is who is waging that war. Conservatives are SILENT.

If you are not careful, the newspapers will have you hating the oppressed, and loving the people who are doing the oppressing.
-Malcolm X
#BoycottFoxNews
#BoycottNYPost
#ALL BLACK LIVES MATTER
81
67
Following
SUSPECT CHARGED
333
284

Figure 5.8 Examples of 'left' RU-IRA tweets criticizing traditional media.

RU-IRA accounts in the right-leaning cluster echoed their counterparts in the left cluster using hashtags like #FakeNews, #WeAreTheMedia, #WakeUpAmerica and #CNNisISIS. "Propaganda is everywhere," warned one account, after sending out a series of tweets criticizing mainstream media outlets for being the partisan mouthpieces of a corrupt global elite. The examples in Figure 5.9 illustrate how the RU-IRA accounts took advantage of the fragmented media landscape in the U.S. by framing traditional outlets for being irrelevant distractions. Accounts in this cluster further appropriated #BlackLivesMatter as a vector for such messages by linking the movement to globalist conspiracies.

> @Pamela_Moore13: "If we don't stop George Soros now, he will continue to drive divisive race baiting MSM narratives & riots to undermine Trump! #LockHimUp"
>
> @TheFoundingSon: "While the NYT tells you how Soros fights hate crimes his agenda incites hate towards police officers which results in tragedies #KeithScott"



John Davis

How Mainstream Media works #WeAreTheMedia #WakeUpAmerica

BREAKING NEWS
SOMETHING IRRELEVANT TO YOUR LIFE JUST HAPPENED AND NOW WE ARE GOING TO BLOW IT OUT OF PROPORTION FOR DAYS TO KEEP YOU DISTRACTED FROM WHAT'S REALLY GOING ON
ANTIFACTUAL
CNN
ACTION

Figure 5.9 Examples of 'right' RU-IRA tweets criticizing traditional media.

In summary, RU-IRA accounts among both the left and right leaning clusters converged to position traditional media outlets as institutions which manufacture a false reality for masses of people. This aligns with previous speculations (Pomerantsev and Weiss 2014) suggesting that undermining trust in established media sources can be a characteristic of disinformation, with the end goal of further destabilizing democratic discourse.

5.7 Discussion

5.7.1 *Information Operations as Collaborative Improvisations*

Information operations—including political propaganda, disinformation, and other forms of manipulation—on online platforms are a growing concern for political officials, platform designers, and the public at large. Journalists, intelligence professionals, and researchers from diverse fields are converging to examine this phenomenon. In this paper, we analyze an extended campaign of information operations from a CSCW perspective, applying a methodological approach that emerged from research on online interactions and collaborations in crisis events (Palen and Anderson 2016; Starbird and Palen 2012; Maddock et al. 2015) to examine these operations not simply as messages broadcast to audiences, but as interactions between an account operator and their audience—or, more fittingly, as a performance by one or more actors, on and through multiple social media accounts, from within and in interaction with an online community. Our research suggests that these performances are not simply automated or even scripted, but are instead like an improvisation in the sense that an actor is given a set of constraints, but then dynamically adapts their performance in interaction with the crowd.

Considering the limits of our data, we cannot see how this work is explicitly coordinated within the Internet Research Agency itself, but from our perspective we can see how the accounts enact particular kinds of online personas, how they interact with each other in the online sphere, and, to some extent, how they interact with the online communities that they infiltrated. This view allows us, both as researchers and as people who participate in these online conversations, to better understand these tactics, revealing some of the mechanisms they use to manipulate people and what some of their larger goals are, in terms of shaping online political discourse (specifically in the United States). It also illuminates some of the challenges that social media platforms face in attempting to defend against these operations.

5.7.2 *Nurturing Division: Enacting Caricatures of Political Partisan Accounts*

Our findings show RU-IRA agents utilizing Twitter and other online platforms to infiltrate politically active online communities. Rather than transgressing community norms, these accounts undertook efforts to connect to the cultural narratives, stereotypes, and political positions of their imagined audiences. Understanding this performative aspect of RU-IRA accounts is critical for understanding how the work of information operations not only includes activities of disseminating true or false information on social media, but also activities to reflect and shape the performances of other (not RU-affiliated) actors in these communities. Taking a perspective based on the theory of structuration (Giddens 1984), the impact of these accounts cannot be considered in a simple cause and effect type model, but instead should be examined as a relationship of mutual shaping or resonance between the affordances of the online environment, the social structures and behaviors of the online crowd, and the improvised performances of agents that seek to leverage that crowd for political gain.

Importantly, this activity did not limit itself to a single 'side' of the online conversation. Instead, it opportunistically infiltrated both the politically left-leaning pro-#BlackLivesMatter community and the right-leaning anti-#BlackLivesMatter community. Though the tone of content shared varied across different accounts, in general these accounts took part in creating and/or amplifying divisive messages from their respective political camps. In some cases (e.g. @BleepThePolice), the account names and content shared reflected some of the most highly charged and morally questionable content. Together with the high-level dynamics revealed in the network graph (Figure 5.2), this observation suggests that RU-IRA operated-accounts were enacting harsh caricatures of political partisans that may have functioned both to pull like-minded accounts closer and to push accounts from the other 'side' even further away. Though we cannot quantify the impact of these strategies, our findings do support theories developed in the intelligence field that suggest one goal of specifically Russian (dis)information operations is to "sow division" within a target society (Lin and Kerr 2017; Pomerantsev and Weiss 2014). This study also offers some insight into how such an effort works, by leveraging the affordances and social dynamics of online social media.

5.7.3 *The Challenge of Regulating through Authenticity*

As social media platforms (e.g. Twitter, Facebook) begin to acknowledge the problem of information operations and to devote resources and attention towards addressing it (Stamos 2018), one repeated refrain has been that these companies do not want to be "arbiters of truth" or seen as censoring political content. This is likely because they are wary of removing posts by ideological believers of that content. This is important here, because the vast majority of accounts in the conversations described in this research—the nearly 22,000 other accounts in our Twitter collection—would likely fall into the category of ideological believers (not RU-IRA agents).

Reluctant to take on the role of deciding what kinds of ideologies are valid and/or appropriate, the platforms are therefore faced with a challenge of developing other criteria for determining what kinds of activities to promote, allow, dampen, or prevent on their platforms. One recent focus has been on "authenticity" (Stamos 2018) —which could be defined as whether an account is who it pretends to be and whether the account believes the content it is sharing and/or amplifying. The RU-IRA invested considerable time in developing online personas for their operations, yet these accounts do not qualify as authentic by these criteria. So, this developing strategy demonstrates a potential way forward that allows the platforms to walk the fine line between criticisms of rampant manipulation and concerns about censorship.

Still, our research suggests that those wishing to deceive are working hard to establish the appearance of "authenticity." To underscore that point, personas featured in this research were "authentic" enough for @jack (Twitter's CEO) and at least one of our researchers to retweet, and we assume it will be challenging for platforms to determine authenticity for the vast number of active accounts. We do not know how difficult or easy it was for Twitter to identify the RU-IRA accounts featured here, but we can assume that developing mechanisms for determining authenticity—and even refining the criteria for what authenticity means—represents an important and challenging direction for future work.

5.7.4 *Information Operations and the Challenges Ahead*

Through interactions with and reactions from other users and the connections displayed by linking to their own network of websites, the RU-IRA accounts developed unique and individual profiles. Discerning between a legitimate social media profile and one constructed by the RU-IRA is a complicated—and emotionally fraught—task. Our own experiences of conducting this research have taught us that calling out and problematizing accounts as impersonators or information operators can be challenging, especially when those accounts align closely with one's own values and worldviews. Despite having a certain level of critical awareness, an understanding of the context, knowledge of populist rhetoric, and an 'official' list of suspended accounts, we found ourselves experiencing doubt when linking some of these accounts with pejorative terms like ‘trolling’ and ‘propaganda’. This was especially true when we immersed ourselves with RU-IRA data in the ways that most closely resemble how an ordinary social media user would encounter their content.

Crucially, we observed that our own biases made it difficult to problematize certain RU-IRA accounts in the left-leaning cluster when we were analyzing their tweets. This highlights how the ways in which we make sense of information is significantly impacted by our self-identity and the ‘tribes’ (Haidt 2012) we associate with. Since these accounts tried to present themselves as members of our ‘tribe’ and speak to our truths (i.e. using information laden with progressive values shared by members of our research team), we were sometimes left in a state of doubt and confusion as to whether these left-leaning accounts were bad actors at all. We would express doubts concerning Twitter’s methodology for identifying these accounts, requesting each other to rerun certain analyses, and generally searching for anchors to ground us and give us certainty. At one level, this provides another small piece of evidence to suggest that these tactics are effective at what many have argued they intend to do—sowing doubt, creating confusion.

It also raises important questions for researchers and educators: What kinds of emotional and critical literacies do we need to cultivate to accurately evaluate credible profiles on social networks and effectively challenge information operations? How can we help users look past their individual interactions with inauthentic accounts to see the larger patterns of activity behind information operations? How can users become more critical of information produced through aggressive and reductive messages? While we support efforts by social media companies to take responsibility to curb propaganda on their platforms, we also feel that it is important for researchers to ‘intervene’ in the sense of helping to call attention to these forms of manipulation and to help the public (and social media companies) understand these phenomena, including how and where users are being targeted. CSCW researchers, specifically, can help by furnishing conceptual frameworks for better understanding the activities of information operations as interactive, and in some ways collaborative efforts that enlist the online crowd (often without their knowledge) in their campaigns.

5.8 CONCLUSION

This study examined the online activities of social media accounts affiliated with an organization that has been accused of functioning as part of the Russian government’s intelligence and media apparatus U.S. Justice Department 2018; U.S. House of Representatives Permanent Select Committee on Intelligence 2017). We focus on the activities of these accounts—i.e. their *information operations*—within #BlackLivesMatter discourse during 2016, during the lead-up to the U.S. presidential election. Our research demonstrates how these accounts presented themselves as ‘authentic’ voices on both sides of a polarized online discourse, modeling pro- and anti-BlackLivesMatter agendas respectively. We also show how these accounts converged to undermine trust in information intermediaries like ‘the mainstream media'. This work conceptually sheds light on how information operations use fictitious identities to reflect and shape social divisions. We conclude by highlighting both the need and the challenges of evaluating authenticity within social computing environments.

5.9 ACKNOWLEDGEMENTS

This research is a collaboration between the emCOMP lab and DataLab at the University of Washington and was supported by Office of Naval Research Grants N000141712980 and N000141812012 as well as National Science Foundation Grant 1749815. We also wish to thank the UW SoMe Lab for providing infrastructure support.

(NOTE: This marks the end of the original publication)

5.10 REFERENCES TO PROBLEMATIC OR MISLEADING CONTENT IN THIS CHAPTER

For the purposes of this dissertation, I have chosen to separate academic references from references to content that I have problematized as misleading or potentially triggering. I have placed references to the latter here and to the former in the Works Cited section towards the end of the dissertation. This has been done to avoid driving traffic to it and blending it with more credible information.

Agorist, Matt. 2016. "Disturbing Video Shows Cops Shoot Suspect, Then Walk Up to His Hostage and Execute Her." *The Free Thought Project*, April 6, 2016. https://thefreethoughtproject.com/disturbing-video-shows-cops-shoot-suspect-walk-hostage-put-4-rounds/.

Black Matters US. n.d.a. "Major Mismatches in the Story of the White Cop Raping 15-year-old Black Girl." Accessed September 6, 2018. https://blackmattersus.com/17023-major-mismatches-in-the-story-of-white-cop-raping-15-yo-black-girl/.

Black Matters US. n.d.b. "Meet the First Skwad 55 Podcast." Accessed September 6, 2018. http://blackmattersus.com/15026-meet-the-first-skwad-55-podcast/.

Sound Cloud. n.d. "Skwad55." Accessed September 6, 2018. https://soundcloud.com/skwad55.

5.11 Summary and Takeaways

Concentrating on the discussions on Twitter around the #BlackLivesMatter movement, this research shows how the Internet Research Agency fostered antagonism and undermined trust in authorities. It describes tactics which entangle orchestrated action with organic activity, highlighting complex technological, social, and cognitive vulnerabilities. And it examines how these tactics were employed for different tasks like generating anger and engagement for those most likely to support then-candidate Donald Trump and creating disillusionment and disengagement among Left-leaning and Black communities.

Additional data released by social media companies has aligned with and added context to the findings of this research (Gadde and Roth 2018). This new evidence has helped confirm that the Internet Research Agency was conducting a protracted campaign on social media that was leveraged, in part, to influence political views in the United States leading up to the 2016 election (Jamieson 2020; National Intelligence Council 2017). It also supports the finding that not all of the agency's efforts targeted people on the political 'right'; a concerted effort was also made to target the 'left', and the #BlackLivesMatter hashtag was a significant focal point of these efforts (DiResta et al. 2019). The new data also reveals that the social media accounts examined in this research were some of the most prolific ones employed by the agency in terms of audience engagement, and that this engagement rose significantly from past years after the accounts shifted from more bot-like behaviors to the ones this research describes (Starbird 2018).

One finding from this research I will build on and stay with over the next few chapters is that the information the agency spread wasn't always objectively false. Depending on one's worldview and understanding of power, much of it might not even seem particularly problematic. And yet it was clearly intended to reinforce tribalism and to normalize points of view strategically advantageous to the Russian government on a range of social issues. It was designed to exploit societal fractures, and sow doubt and suspicion of media entities and the information environment, of government, and of each other.

We often equate disinforming someone with giving them inaccurate information, but here we can appreciate it more broadly as *sensegiving*—as a series of actions taken to influence the sensemaking and meaning construction efforts of others (Gioia and Chittipeddi 1991). In other words, disinformation is contextual, not just a matter of facts but a matter of intentions. This complicates the widespread sentiment that we can fact-check our way out of this conundrum, but it does suggest another design direction.

This other direction involves judging not only content but also its source. If we can identify and weed out ‘bad actors’ — those websites and social media accounts operating with dubious intentions — then we can build a healthier information environment while avoiding some of the complications that occur when these actors layer their messages with accurate content. I believe this direction is a valuable one. After all, this research would not have been possible without Twitter’s efforts to identify and weed out the Internet Research Agency’s social media accounts. I also think that in the context of new media literacies, there is some merit in designers focusing on interventions that help learners look past the accuracy of content to focus on that content’s provenance and location in larger patterns of communicative activity. But I also think this direction has some limitations. Intentions, as I noted in Chapter 2 for instance, can shift and be difficult to identify, while people with good intentions can be manipulated. The next chapter will show how these limitations play out in the context of a more sophisticated disinformation campaign.

Returning to the above finding, we can also catch a glimpse of how the act of critiquing powerful actors can be strategically perverted. Allies and members of social movements like #BlackLivesMatter often critique institutions (e.g. law enforcement, media outlets, academia) as a way of inviting people to look for new explanations (Benford 1993). But the hole that opens up, that invites people to look for new explanations, can be opportunistically exploited by disinformation campaigns in deeply problematic ways. For example, the ‘left’ RU-IRA tweets attacking traditional media (Figure 5.8) co-opt critiques meant to combat injustice and oppression for plausible deniability and to promote dubious, reactionary, or even repressive aims. This example strongly suggests that we need to carefully assess the role of tools that help people engage in fault finding and negative judgment of information and information sources. The next two chapters will help develop these ideas further.

Chapter 6. STUDY 3: ECOSYSTEM OR ECHO-SYSTEM?

In this study[30], which offers additional insight towards Research Question 2, we continue to explore the structures and activities of online disinformation—in this case shifting our attention to a campaign targeting the White Helmets, a volunteer humanitarian response group that works in rebel held areas of Syria. Looking beyond the activity of Twitter accounts being operated by one organization, this research[31] investigates how disinformation campaigns commingle with a broader range of actors to shape the information ecosystem surrounding a particular discourse.

[30] This study is previously published work. To cite material from this Chapter, please cite this original work as well as the dissertation:

Starbird, Kate, Ahmer Arif, Tom Wilson, Katherine Van Koevering, Katya Yefimova, and Daniel Scarnecchia. 2018. "Ecosystem or Echo-System? Exploring Content Sharing across Alternative Media Domains." In *Proceedings of International AAAI Conference on Web and Social Media*, 365-374. https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17836.

[31] This study is part of a larger research effort examining disinformation campaigns / information operations in this context. So my observations in this chapter are informed by some related investigations led by my colleagues, including Tom Wilson and Kate Starbird. Following conventions, I will cite the relevant work where appropriate. But I wish to highlight this connection upfront so that readers have a clearer view of my standpoint and can quickly look up the related works if they are inclined to (Wilson, Zhou, and Starbird 2018; Starbird 2017).

Ecosystem or Echo-System? Exploring Content Sharing across Alternative Media Domains

Published in Proceedings of the 2018 AAAI Conference on Web and Social Media (ICWSM)

Authors: Kate Starbird, Ahmer Arif, Tom Wilson, Katherine Van Koevering, Katya Yefimova & Daniel Scarnecchia

6.1 ABSTRACT

This research examines the competing narratives about the role and function of Syria Civil Defence, a volunteer humanitarian organization popularly known as the White Helmets, working in war-torn Syria. Using a mixed-method approach based on seed data collected from Twitter, and then extending out to the websites cited in that data, we examine content sharing practices across distinct media domains that functioned to construct, shape, and propagate these narratives. We articulate a predominantly alternative media 'echo-system' of websites that repeatedly share content about the White Helmets. Among other findings, our work reveals a small set of websites and authors generating content that is spread across diverse sites, drawing audiences from distinct communities into a shared narrative. This analysis also reveals the integration of government-funded media and geopolitical think tanks as source content for anti-White Helmets narratives. More broadly, the analysis demonstrates the role of alternative newswire-like services in providing content for alternative media websites. Though additional work is needed to understand these patterns over time and across topics, this paper provides insight into the dynamics of this multi-layered media ecosystem.

6.2 INTRODUCTION

In September 2016, a documentary featuring the Syrian Civil Defense, a volunteer response group in Syria who are also known as the White Helmets (WH), was released to wide acclaim. Later, the film would win an Oscar for Best Documentary. It was set within the context of the then five-year-old civil war in Syria, troubling claims about the brutality of Syrian President Bashar al-Assad who had been accused of bombing civilians and medical workers (Fouad et al. 2017), the rise of the Islamic State (IS) and other extremist groups in areas outside the government's control, and a massive exodus of refugees escaping the violence (BBC News 2016). The documentary and subsequent sympathetic articles by mainstream media outlets worldwide brought global awareness to the plight of Syrian people, especially those resisting the Assad government.

This narrative, which promoted the role of the WH as a humanitarian response organization in Syria, functioned to grow solidarity between many Western audiences and the WH—as well as Syrian people from rebel-held areas who were seen as victims of the Assad regime. However, this narrative was not aligned with other views of the complex geopolitical landscape of the conflict. In particular, representatives and supporters of the Syrian government and its allies in that conflict (including Russia and Iran) resisted this sympathetic perspective. In response, a counter-narrative took shape and eventually spread to other online communities. Critics argued that the group, which is funded by Western governments, was a propaganda construct, supported by mainstream media, and employed as a tool of NATO interests in Syria[32] (Russia Today 2017). Some claimed that the WH aided and in some cases were themselves active in terrorist organizations (Beeley 2016). Supporters of the WH, in turn, accused their critics of orchestrating a propaganda campaign to spread a "conspiracy theory" about the group (e.g. Ellis 2017). These contested narratives, which are still active, are spread through and shaped by various media—including mainstream news articles, alternative media articles, blog posts, social media posts and interactions, etc. The resulting information space is an evolving and multi-layered one.

This paper explores the dynamics of a subsection of the media ecosystem that was active around Twitter conversations about the WH during a three-month period in the summer of 2017. Using a mixed-method approach based on seed data collected from Twitter, and then extending out to the websites cited in that data, we examine content sharing practices across distinct media domains—articulating an alternative media, and, to a lesser extent, a mainstream media 'echo-system' of websites that repeatedly share content about this topic. Among other findings, our analysis reveals a small set of source domains (and authors) generating content that is spread across diverse domains, drawing audiences from distinct communities into a common narrative. This analysis also reveals the integration of government-funded media (RT, SputnikNews) and geopolitical think tanks (GlobalResearch) as source content for anti-WH narratives. More broadly, the analysis demonstrates the role of alternative 'newswire' services in providing content for alternative media websites. Though more study is needed to understand these patterns over time and across topics, this paper provides insight into the dynamics of this multi-layered media ecosystem.

6.3 BACKGROUND

6.3.1 *The White Helmets and the Syrian Civil War*

The ongoing conflict in Syria has taken more than 400,000 lives and displaced millions more. Armed conflict began during the 2011 Arab Spring, when anti-government protests calling for President

[32] For the purposes of this dissertation, I have chosen to separate academic references from references to content that I have problematized as misleading or potentially triggering. I have placed references to the latter at the end of the chapter (and to the former in the Works Cited section towards the end of the dissertation). This has been done to avoid driving traffic to it and blending it with more credible information. The references to problematic information are formatted as superscripts to distinguish them.

Bashar al-Assad to step down escalated into full scale civil war between the Syrian government and those opposing Assad's rule (International Committee of the Red Cross 2012). It has since evolved into an internationalized, multi-sided conflict. Militants from the IS took advantage of the conflict to capture much of Eastern and Northern Syria (Whewell 2013). Russia and Iran initially supported the Assad government with materiel and financial assistance, and later committed troops, targeting Syrian opposition forces and IS militants (Associated Press 2013; Kramer and Barnard 2015). In 2014, the United States and Gulf League states began bombing IS in Syria and providing assistance to Kurdish forces fighting IS and Syrian opposition groups (Gordon 2014). The UK joined this coalition in 2015.

The WH are a group of trained volunteer rescuers that operate throughout Syria's opposition-controlled areas to assist civilians affected by the violence. According to the organization's website (http://syriacivildefense.org), they abide by the fundamental principles of the Red Cross and work in accordance to Article 61 of the Additional Protocol I, which defines the activities that constitute civil defense: protecting civilians from hostilities or disasters, aiding recovery in the aftermaths of such events, and providing conditions for the survival of the civilian population. WH activities include search and rescue, evacuating buildings, firefighting, medical aid, providing emergency shelter and supplies (Pictet 1979).

6.3.2 *Online Propaganda & Disinformation in 2017*

This research took place during a period (June 2017-January 2018) of global attention to the threat of misinformation, disinformation, and political propaganda, and the role of technology in facilitating their spread. In prior years, researchers optimistically noted the rise of citizen journalism (Gillmor 2004), and also acknowledged a related "crisis in journalism" (Fuller 2010) as economic, production and distribution models for "news" became disrupted. As traditional journalists and media outlets attempted to adapt to these new conditions, citizen journalists and media outlets outside the mainstream worked to establish their legitimacy—resulting in an information space where new voices were heard in new ways. Simultaneously, it became increasingly difficult for information consumers to assess the validity of the information they saw. The rise of partisan "fake news" sites and the subsequent appropriation of that term to challenge "mainstream" outlets (Qiu 2017) corresponded with record-low levels of trust in media and information (Swift 2016; Barthel and Mitchell 2017).

Additionally, for decades critics have explicated systemic biases in "mainstream" media[33] —e.g. towards neoliberal, pro-Western, and colonialist/imperialist ideologies (e.g. Herman and Chomsky 1988). As a prominent example, relevant here, were claims within the New York Times (Ravi 2005)

[33] The term "mainstream media" has historically been used as a pejorative, especially by those who identify with "alternative" perspectives or media, to criticize the agenda-setting power of mass media. However, the "mainstream" and "alternative" terms have also become a common way to distinguish between media types. We employ these terms here for efficiency, but also with acknowledgment of their political roots and the tension between them.

that Saddam Hussein had weapons of mass destruction, a premise used to garner public support for the 2003 U.S. invasion of Iraq. These examples and the arguments constructed around them likely contribute to diminished trust in mainstream media.

6.3.3 *Information Warfare and Online Disinformation*

While these claims suggest strategies tied to traditional means of news production, recent evidence suggests others—including state and non-state actors—are working to leverage online technologies to forward their geopolitical goals (e.g. Weedon, Nuland, and Stamos 2017). These actors are using a mix of automation and human curation to intentionally spread misleading information using online technologies such as social media (Woolley and Howard 2017).

In particular, Russia has been accused of conducting an "information war" that extends long-standing tactics of disinformation to new Internet-enabled channels (Pomerantsev and Weiss 2014). Though researchers and intelligence communities are still working to understand these strategies, evidence suggests that Russia and others are utilizing social media in conjunction with other channels to spread their messages (Weedon, Nuland, and Stamos 2017; Paul and Matthews 2016). The Russian government also utilizes its own media apparatus, including RT and SputnikNews, to forward their geopolitical aims. RT receives the vast majority of its funding from its government (Tetrault-Farber 2014), and its editor-in-chief has argued that RT uses that funding to support "information warfare … against the Western world" (DFRLab 2018).

Pomerantsev and Weiss (2014) define Russian disinformation as intended not to simply convince, but to confuse—to sew doubt and distrust across a society. The idea is that doubt can act to reduce agency. In other words, if we are not sure about what the truth is, we cannot choose the best action to take, and therefore will take no action. For this reason, disinformation campaigns do not need to rely on a single narrative or counter-narrative, but can work by presenting diverse and even contradictory narratives.

6.3.4 *Information Operations and Non-Governmental Organizations (NGOs)*

The Russian government has previously taken issue with—and actively worked to undermine the mission of — NGOs which they see as a threat to its geopolitical interests (Ambrosio 2007). In the early 2000s, Western NGOs supported "pro-democracy" civil society groups in Russia and its neighboring states and played a role in facilitating a shift away from Russia-aligned governments and policies. Following the color revolutions that took place in the former Soviet states, the Russian government argued that these activities represented unfair interference. Vladimir Putin (2012) specifically called out and criticized "pseudo-NGOs" funded by foreign governments and corporations for their role in destabilizing other countries. In this study, we can see an extension of that criticism to a humanitarian response organization working—both through its efforts to assist affected people and to garner attention for their cause—against the geopolitical interests of Russia and its ally, the Syrian government.

6.3.5 *Conducting Research on Information Operations*

This paper is a small component of a larger research effort examining contested narratives involving the WH. Over the course of several months, our team spent hundreds of hours analyzing this data at multiple levels. This information space can be intensely *disorienting*. Our researchers repeatedly use this word to describe how the qualitative analyses affect us. The arguments and evidence presented in support of narratives on both sides are often compelling. Despite, or because of, deep engagement with this content, our researchers are often left in a state of confusion about what and whom to believe. In this study, we do not speak directly to this question. Instead, we focus on describing the media ecosystem surrounding these conversations—especially the dynamics of content-sharing practices—with the goal of gaining insight into how these narratives and counter-narratives are produced and disseminated.

6.4 METHODS

6.4.1 *Data Collection and Processing*

Our White Helmets dataset (WH dataset) consists of tweets posted between May 27 and September 9, 2017. We created this collection using the Twitter Streaming API, initially tracked various keyword terms related to the Syrian conflict including geographic terms of affected areas. Later, we scoped this data to tweets that contained “white helmet” or “whitehelmet”, resulting in 135,827 tweets.

To understand the role played by external websites in Twitter conversations about the WH, we examined the links embedded within these tweets. 52,903 tweets contained a URL link. To process this data, we expanded shortened links, removed HTML parameters, and filtered out duplicates. We also removed links (approximately 35% of the total) that resolved to social media domains (i.e. Twitter, Facebook, Youtube) and newsreaders (feedproxy.google.com).

The resulting set of 3410 distinct URLs was used to extract articles in a structured and automated fashion via a tool built using Newspaper, a python library designed for full-text and article metadata retrieval (Ou-Yang n.d.). Due to how some web servers and content delivery networks were configured, we were unable to automatically scrape content from 111 domains. Subsequent analysis suggested that some of these domains play a prominent role in this information space (e.g. GlobalResearch), so we manually captured content from the top-10 most tweeted of these missing domains by traversing the URL to the article and copy-pasting its content. However, 322 URLs from the other 101 domains were omitted from the analysis.

Next, we passed the articles through an algorithm for detecting article similarity. This was done by computing the term frequency–inverse document frequency (tf-idf) statistic (Salton, Fox, and Wu 1983) for each article and obtaining the cosine distance between the tf-idf vector for each pair of articles. The resulting matrix of similarity scores was used to identify duplicate articles across domains. We selected a threshold of >=85% to identify two articles as containing shared content. This level of similarity generally captured identical articles without being overly sensitive to small changes in image captions and article bylines.

After collapsing links to similar articles within the same domain, there were 1680 distinct news articles. From there, we identified 558 articles that had significant (>=85%) overlap with another article within another domain in our set. Interestingly, nearly two-thirds (63%) of tweets with URLs cited one of these 558 articles that appeared on more than one domain.

Next, we constructed 'paths' for each article—tracing all URLs in our dataset where that article appeared. We identified 135 paths, which provide insight into how content was shared across domains in the media ecosystem. We used these 135 paths to construct a network graph (Figure 6.1, Figure 6.2, Figure 6.3) where two nodes (domains) are connected if they appear in the same path—e.g. if one domain hosted an article that had >=85% similarity with an article in the other domain. The edge weight represents the number of similar articles shared by the two domains. These edges do not encode directionality, but merely reflect similar content. Using manual analysis to determine the original source of each article, we labeled each node as primarily a *source* (publishes original content), an *amplifier* (republished content from others), or a *hub* (published original and borrowed content). Nodes are sized by the number of tweets in the WH data that link to that domain, therefore representing the salience of this domain—and its articles—in the Twitter conversation. We use the ForceAtlas2 algorithm to determine the visual layout of the graphs and the Louvain method to detect communities for Figure 6.1.

6.4.2 *Interpretative, Mixed-Method Analysis*

We conducted interpretive, mixed method analysis of this data, expanding upon methods developed for the investigation of online rumors in the context of crisis events (Maddock et al. 2015). This approach iteratively blends quantitative and qualitative analyses—in this case generating a network graph to see larger patterns of content sharing across domains, and then using that representation as an entry point for a closer examination of both the practices of content sharing and the influential domains within this ecosystem. For the qualitative analysis, we focused primarily on the content within each domain, including its home page, about page, and the content-sharing practices visible within the specific articles cited in the WH dataset.

6.4.3 *Note on Data and Privacy*

In this paper, we identify a small number of prominent authors within the alternative media ecosystem. We considered anonymizing these names, but chose to publish real names because these authors are self-identified journalists and their patterns of activity—both within our data set and before/after—are important for understanding the nature of content sharing in this ecosystem. Several articles from both 'sides' of this conversation are cited in the references.

6.5 FINDINGS

Figure 6.1 shows the complete content-sharing domain network graph for the entire WH dataset (nodes sized by tweet volume and colored by Louvain-detected community). From a high level, this graph has several key features: two large clusters—Cluster A in pink on the left and Cluster B in blue on the right, with some connective tissue between them; and a small distinct community (Cluster C, in yellow) loosely connected with Cluster B. There are also a large number of small, distinct clusters that are unconnected to the other clusters (in grey). One of these (Cluster D, in red) is interesting because it contains a highly tweeted, but disconnected, domain (see Table 6.1).

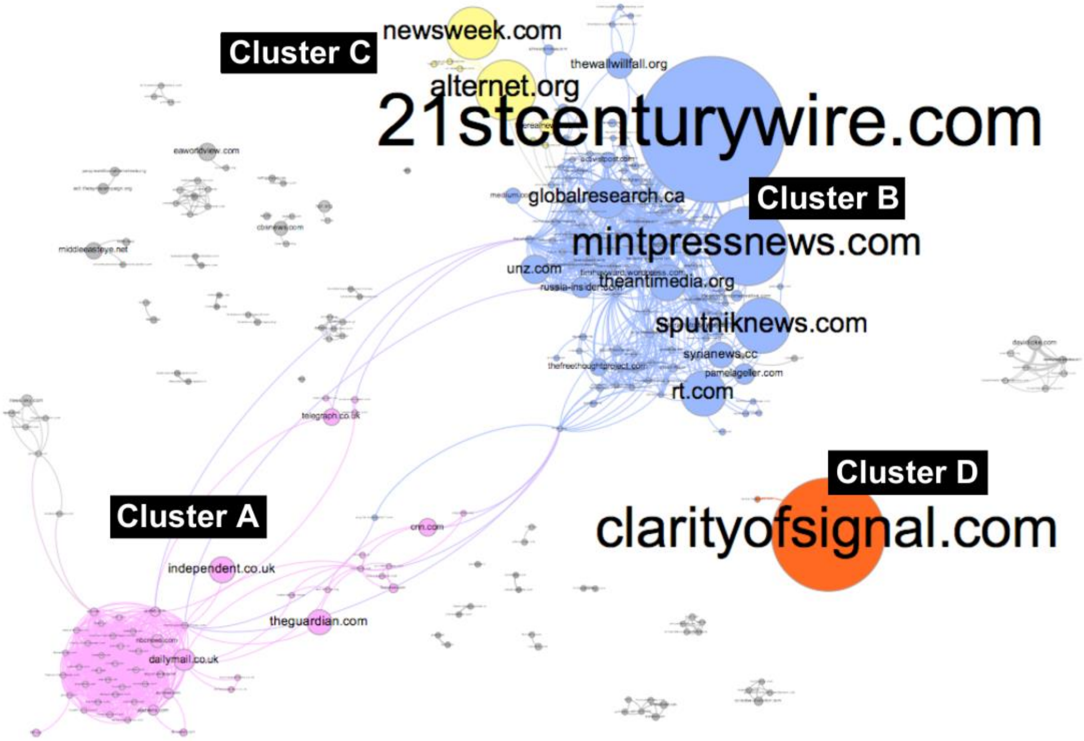


Figure 6.1 WH Content Sharing Domain Network Graph

6.5.1 *Cluster A: An Associated Press News Cluster*

Cluster A (Figure 6.2) is a relatively large component of 40 nodes, consisting of several western "mainstream" media sites (i.e. wsj.com, dailymail.co.uk, apnews.com), news outlets from the Arab world (aljazeera.com, arabnews.com, english.alarabiya.com), and other local and alternative media outlets from around the world. There are also a few news aggregators in this cluster, including castwb.com. Most of the edges in this cluster have a weight of one, representing >=85% overlap of a single WH-related article within both web domains.

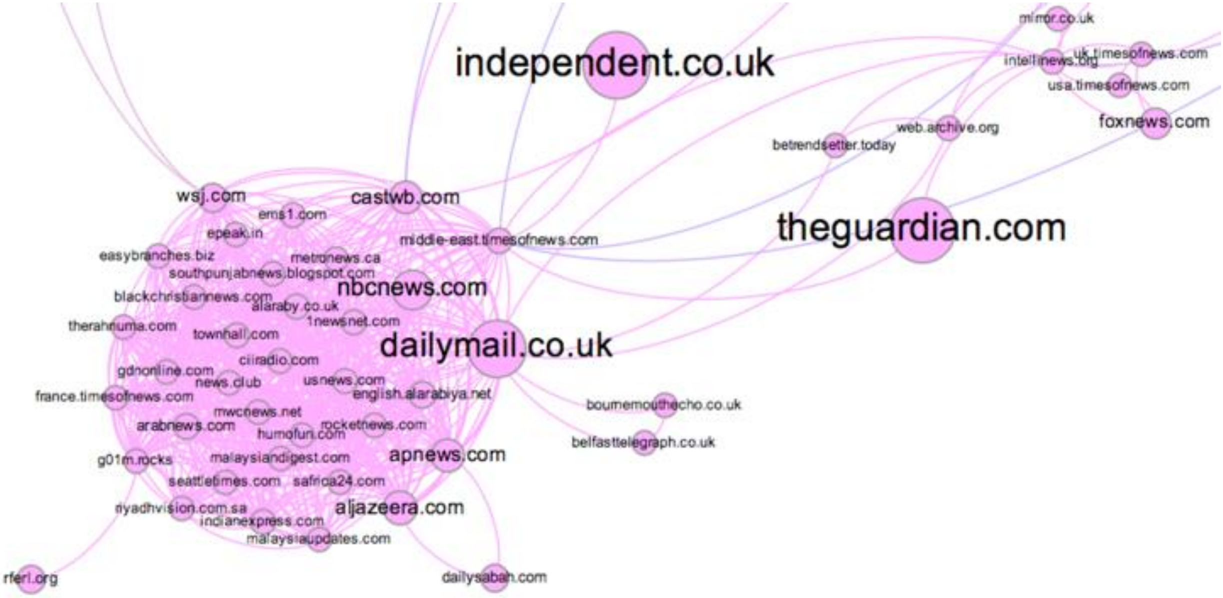


Figure 6.2 Close-up of Cluster A

Almost all of the connections in this cluster were related to a single article published in August 2017 (Mroue 2017), describing the murder of seven WH volunteers at their office in Idlib, Syria. This article was sympathetic to the WH, presenting them as "first-responders who have been known to risk their lives to save people from the civil war."

This article constituted the largest path in our data—its content appeared in 44 different domains. Its original source, almost always cited in the downstream articles, was the Associated Press (AP). The AP operates as an international, non-profit news cooperative and allows its partner news outlets to reuse its content. These partners pay to use the AP's content in their own newspaper or website. This content-sharing model allows media outlets to provide coverage of diverse topics across the globe. Another similar agency, Agence France-Presse (AFP), appears within a small isolated cluster elsewhere in the graph. This model, which is a long-standing one that pre-exists the Internet (Fenby 1986), results in content sharing across news outlets in our data, including many that are considered "mainstream."

However, the relative scarcity of paths other than this AP path suggests that intensive content sharing about the WH was not observed in this set of websites.

A small number of nodes serve to connect—over one to three degrees—Cluster A (the AP cluster) to Cluster B (on the right of the graph). Within this connective tissue are a few "mainstream" media domains including the Telegraph, Independent, CNN, and the Guardian. These domains' content was re-published by news aggregators (i.e. intellinews.org, f3nws.com) that connected those websites to other mainstream and alternative media domains. Most edges in this section have an edge weight of one—a single article, of which the mainstream media domain was the original source. The articles featured in this 'connective tissue' area were generally supportive of the WH—promoting narratives that featured the WH as courageous volunteers who were risking their lives to rescue and provide medical assistance to Syrians who were injured by Syrian government and Russian military operations.

6.5.2 *Cluster B: The Alternative Media Ecosystem*

Most of the volume represented in the complete content-sharing network graph (Figure 6.1)—in terms of tweets, articles, and distinct domains—resides in Cluster B (Figure 6.3). The articles cited within these domains were highly critical of the WH. This cluster contains 110 nodes or web domains. 10,821 tweets in the WH data included a URL that linked to Cluster B, compared to only 2526 in Cluster A. Unlike Cluster A, which exhibits a consistent, nearly symmetrical structure, Cluster B is more heterogeneous in terms of both node size and edge weight. Edges vary in strength from one to seven articles. Thicker edges represent more consistent content sharing patterns over time.

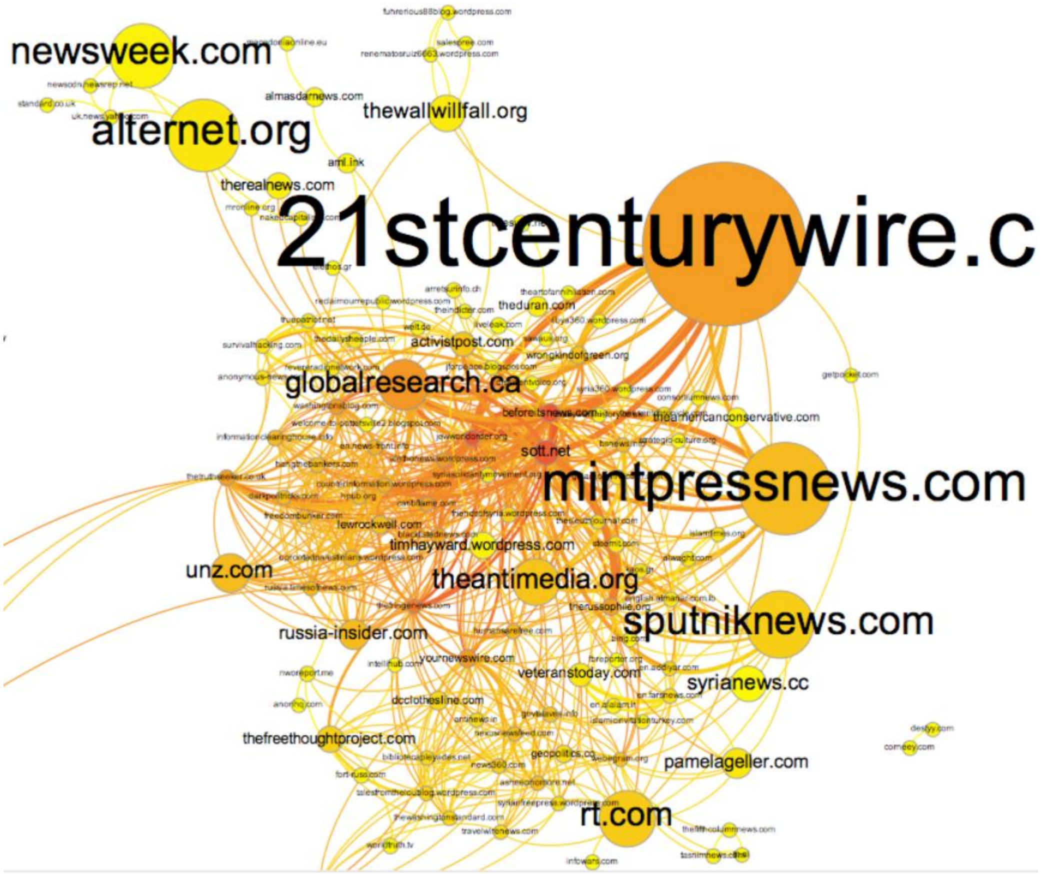


Figure 6.3 Close-up, expanded view of Clusters B and C;

Structural analysis on the domain graph in combination with content analysis of the articles within the content-sharing paths and the domains that hosted them reveal a few salient categories of domains: a small set of prominent alternative media 'hub' domains that produce source content for the rest of the graph and occasionally re-publish each other's articles (21stCenturyWire, MintPressNews, GlobalResearch); two Russian government-funded outlets (RT, SputnikNews) that provide source content and occasionally amplify articles from the prominent hub domains; and a diverse set of alternative news aggregators that consistently amplify content from the peripheral sources.

As evidenced by their size in the graph, Cluster B includes many of the most highly tweeted domains in the data. Table 6.1 lists Top 10 domains within the WH collection in terms of tweet volume. Seven are located in Cluster B, and each of these was cited for multiple articles that were critical of the WH. Table 6.1 also provides the number of WH tweets that link to each domain and the 'degree' of each domain in the graph—e.g. the number of other domains in the graph that are cited in the WH tweets for an article that has high similarity to one of the articles cited from this domain. We also note whether the domain is a source, a hub, or primarily an amplifier of content in this ecosystem.

Table 6.1 Top 10 Most Tweeted Domains in WH Dataset

Domain	Tweets	Degree	Network Role
21stcenturywire.com	3119	30	Central Hub
clarityofsignal.com	2391	1	Isolated
mintpressnews.com	1630	22	Central Hub
alternet.org	1219	6	Peripheral Hub
sputniknews.com	1110	16	Central Source
newsweek.com	1046	2	Peripheral Source
rt.com	879	17	Central Source
globalresearch.ca	707	33	Central Hub
theantimedia.org	682	19	Amplifier
unz.com	512	22	Central Source

6.5.2.1 Central Hubs in the Content Re-sharing Ecosystem

One key finding is that three of the most-tweeted domains (21stCenturyWire, MintPressNews, and GlobalResearch) generated the majority of source content for the re-sharing practices reflected in Cluster B. Interestingly, these domains were not exclusively source domains, but also borrowed content from each other, and published original content from some of the same authors. One author, Vanessa Beeley—a British journalist and leading critic of the WH—had original articles on each of the three domains, content which later appeared on one or more of the others. These three domains are central hubs of content re-sharing in the anti-WH conversation.

21stCenturyWire is by far the most cited domain in the WH collection—3119 tweets link to 26 distinct articles within this web domain. 13 of these articles appear elsewhere in the ecosystem, republished in all or large part on other domains. Most of these articles were written by Beeley. 21stCenturyWire was founded by Patrick Henningsen, whose Guardian byline includes affiliations to RT and alternative news site Infowars. The website positions itself as grassroots, independent media that provides "news for the waking generation." In the WH data, 21stCenturyWire is often the source of content that spreads across other prominent domains in the cluster (i.e. MintPressNews and theAntiMedia) and across several less trafficked domains (i.e. BeforeItsNews, YourNewsWire, and

JewWorldOrder). 21stCenturyWire also re-publishes content that originally appears elsewhere in the graph, including articles from GlobalResearch, MintPressNews, and RT.

MintPressNews (MPN) is the third most-tweeted domain in the WH data—1630 tweets, 12 distinct articles. This domain is connected to 22 different domains in the graph, including many of the same domains as 21stCenturyWire. MPN describes itself as an "independent watchdog journalism organization" that features original reporting "through the lens of social justice and human rights." Its home office is located in Minnesota (USA), but the website covers both national topics and foreign affairs. In our data, their content is strongly pro-Syrian government and critical of the WH. Apart from their original content, MPN has multiple news and syndication partners whose content they frequently re-publish. Their original articles appear on other prominent domains (21stCenturyWire and GlobalResearch) as well as common amplifiers (i.e. theAntiMedia.org, YourNewsWire, BeforeItsNews, Sott.net, and JewWorldOrder). They also re-share content from other domains, including 21stCenturyWire, ActivistPost, and TheAmericanConservative.

GlobalResearch is an influential hub within the content sharing network that appears 8th on our most tweeted list, being tweeted 707 times for 17 different articles. Seven of these had high similarity with other articles in the WH data. GlobalResearch is operated by the Centre for Research on Globalization, "a non-profit independent research and media organization" that describes itself as a think tank on economic and geopolitical issues. The center is operated by Michel Chossudovsky, Professor Emeritus at the University of Ottawa. Chossudovsky [(2015)] has previously published claims of conspiracies related to world events—including that the September 11, 2001 attacks were not perpetrated by Islamic terrorists. In our data, GlobalResearch is both a source and an amplifier. Underscoring its role in supporting this information ecosystem, of the top-10 most-tweeted domains, GlobalResearch has the highest degree, sharing articles whose content overlaps with 33 different domains. Though it has a strong, multi-article connection to 21stCenturyWire and shares several overlapping amplifiers with 21stCenturyWire and MintPressNews, its content also reaches a subset of domains outside of that subnetwork (i.e. LewRockwell.com, WashingtonsBlog, and FreedomBunker).

6.5.2.2 Government-Funded News Outlets

Two government-funded news outlets (SputnikNews and RT) are also within the Top 10 most-tweeted domains and Cluster B. **RT** is a Russian government-funded media outlet that provides content to international audiences. Founded in 2005 with the stated purpose of improving Russia's image abroad, it has been accused of spreading disinformation and its U.S.-based affiliate has been forced to register as a foreign agent (Stubbs and Gibson 2017). Its current tagline is "Question More," and its content often encourages readers to question western and mainstream narratives of world events. RT.com was tweeted 879 times for 16 different articles which all take a critical perspective of the WH. Within our content-sharing paths, RT is primarily a source domain. Its content is re-shared entirely or in large excerpts across 17 other domains. Interestingly though, all of its edges have the weight of one (article). The graph shows a large number of domains borrowing a single RT WH article (not always the same one), rather than consistently re-publishing their content. In addition to content-sharing that we can see through the similarity graph, several articles within other domains embed videos from RT in their content.

SputnikNews, founded in 2013 as the replacement for the "Voice of Russia," is another Russian government-funded media outlet that features radio, television, and online content. Like RT, they have also been accused, primarily by western governments and media, of spreading disinformation and political propaganda that is favorable to the current Russian government (Dearden 2017). They are slightly more highly tweeted than RT in our WH data—1110 tweets for 15 articles, all critical of the WH. They are also primarily a source domain in this set. Their content is re-shared across 16 domains. Their most common amplifiers are Sott.net, theRussophile.org, and en.Addiyar.com (a Lebanese news outlet with a pro-Syrian government leaning). Each of those websites re-shared multiple articles from SputnikNews in our data.

6.5.2.3 Central Source Domains

In addition to RT and SputnikNews, there are two other central source domains—UNZ.com and ActivistPost.com. **UNZ** is an alternative media outlet founded by Ron Unz, a former (conservative) political candidate in California. The outlet's national security editor, Philip Giraldi, was the author of an article arguing that the WH are "a fraud." This article, hosted on the UNZ website, was tweeted 512 times within our data. It was also re-published on 21 other domains that appear in the WH data. UNZ therefore performed as a central source domain, though solely through sharing of this single article.

Another central source domain is **ActivistPost,** an alternative independent media outlet whose tagline is "propaganda for peace, love, and liberty." Their WH-related content is consistently critical, echoing many of the common narratives, claiming that they are a propaganda construct of mainstream media and western government interests. In terms of tweet count and compared to the more visible hubs, ActivistPost is relatively small—the domain was only tweeted 213 times. However, their role in the content sharing is significant. They are the source domain for eight different "paths" in the graph—e.g. eight of their original articles were re-published in all or part by other domains in the graph. In total, their content appeared in 18 domains, including central hub domains GlobalResearch and MintPressNews. All of their articles were authored by Brandon Turbeville, and include a Creative Commons license that enables the free distribution of the work.

6.5.2.4 Alternative News Aggregators

Another core component of Cluster B is a large number of Alternative News Aggregators that repeatedly share content that originally appears elsewhere in the graph. The most prominent of these aggregator domains, in terms of tweet volume, is theAntiMedia.org. This website positions itself as the "homepage for the independent media movement," claiming to be a "non-partisan, anti-establishment news publisher and crowd-curated media aggregator." **TheAntiMedia** functions in part as a news aggregator, pulling in articles from other alternative and independent media outlets and mixing those with its own original articles. In the WH data, theAntiMedia was tweeted 682 times for one original and three borrowed articles (from MintPressNews and 21stCenturyWire). Most of these tweets link to a single article, re-shared from MintPressNews.

A large number of other domains in the graph function exclusively as amplifiers. In Figure 6.3, domains are colored by their degree (number of edges), from yellow (few edges) to red (many edges). Many of the most connected web domains (in red) are primarily content borrowers that repeatedly republish content from other websites in the graph. **Sott.net**, **theRussophile.org**, **JewWorldOrder.org**, and **BeforeItsNews.com,** are the domains with the highest degree in the graph, which also have thick, multi-article edges with the three central hub sites. All are exclusively amplifiers in this conversation—serially reposting content that first appeared elsewhere. Two other domains, **YourNewsWire** and **FringeNews**, are slightly less connected, but serve similar roles in a subnetwork in the lower-left-center of Cluster B. Many of these exclusively amplifier domains receive far fewer tweets for this content than other domains in our graph.

Another core component of the graph are the small (in terms of tweet volume) domains that are connected via thin edges to a relatively small number of domains. These domains typically appear in the graph for re-publishing one or two WH-related articles. Their relative positioning, near certain hubs and not others, may reflect a particular type of ideological targeting. For example, in the lower-left of the graph, near RT, TheFreeThoughtProject (a small source domain) and YourNewsWire, are domains like **ASheepNoMore**, **GovtSlaves**, and **HumansAreFree** which promote content questioning many mainstream narratives and suggesting large-scale geopolitical conspiracies. And in the upper left, near GlobalResearch and UNZ are a collection of libertarian-leaning domains (**LewRockwell**, **FreedomBunker**, **HangTheBankers**). Most of these political and/or ideology-centered domains appear in the graph for a single article re-shared from one of the source or hub domains. The domains do not necessarily amplify everything in the ecosystem, but may pick and choose content to reshare, as their focus is not necessarily on the WH, but on a specific worldview that these anti-WH narratives reflect.

6.5.3 *Cluster C: A Peripheral Hub: Reframing Mainstream Content for the Alternative Ecosystem*

Cluster C is a small (in terms of number of domains), distinct community that is loosely connected to Cluster B. Two of the Top 10 domains are in this peripheral cluster: Newsweek and Alternet. Newsweek, a "mainstream" media outlet, was cited for six articles in the WH data. However, we only found evidence of one of these articles being re-shared on other domains. Rather, Newsweek appears in Cluster C, and is peripherally connected to Cluster B, through the Alternet domain—due to one highly-tweeted article that described how a WH volunteer was caught on video (and subsequently fired for) disposing of the mutilated bodies of Syrian soldiers. Alternet, an alternative news site that both aggregates content and posts its own articles, re-published this article with attribution. Alternet's version, however, uses a different title, re-framing the original content to suggest that this incident was part of an ongoing pattern of misbehavior by WH volunteers, which aligns with other Alternet content critical of the WH. In this case, Alternet functioned as a peripheral hub, borrowing source content from "mainstream" media and re-framing it to fit the predominant narrative of Cluster B.

6.5.4 *Content Remixing Practices and Echo Effects*

Though there are hundreds of distinct URLs in our tweet data, a significant percentage of the linked-to content is authored by a small number of prolific authors whose content is often re-shared and re-mixed elsewhere in the ecosystem. Beeley is the author of record for at least a dozen articles that appear in 'paths' within the WH data—shared across several domains in Cluster B. In our tweet data, she was cited for original content in at three source domains: 21stCenturyWire, MintPressNews, and TheWallWillFall. She also repeatedly appears as a secondary source in articles by other authors through quotes, excerpts, and embedded videos of interviews. Similarly, Turbeville, who primarily publishes in ActivistPost, authored eight articles that were source articles for multiple content-sharing paths across the ecosystem.

The following example traces a single, short path that includes both Beeley and Turbeville and illustrates several of the diverse content remixing practices and echo effects that manifest in this ecosystem. On May 2 2017, ActivistPost published an article by Turbeville titled "Photos from Syria Show White Helmets and Nusra/Qaeda Are The Same Organization" (2017). This article used photos and videos that Beeley captured while in Syria and posted on her Facebook account. Later in the article there is a textual excerpt, citing Beeley, that describes the content in one of the videos. At the foot of the article there are seven links to other articles about the WH: four authored by Turbeville and published on ActivistPost, and three authored by Beeley published on 21stCenturyWire.

This same article (including Beeley's photos, videos, and excerpt) is published on MintPressNews on the same day, citing Turbeville and ActivistPost, but removing the links to related content on 21stCenturyWire and ActivistPost. Thirteen days later, on May 15 2017, Beeley publishes the same ActivistPost article on TheWallWillFall, her personal blog. In this version the article is titled "WHITE HELMETS: Living next door to Al Qaeda in Aleppo" and Beeley is listed as the author. However, below an additional image that did not appear in the original version, Turbeville at ActivistPost is cited as the author, followed by the original article in its entirety—including the photos, videos and the quote from Beeley. There are now 13 links to other related White Helmet-related articles—ten of them on 21stCenturyWire, plus TruthDig, WrongKindofGreen, and Wikipedia. These circular citations and remix practices create another kind of echo effect within this system.

6.6 Discussion: Alternative Media Echo-System

In this research, we explored content sharing practices across media domains, using URL links in tweets to capture domains that were active in an online conversation, and an article similarity metric to determine domains that shared articles with high similarity. The conversation we focused on—views of the WH in relation to the ongoing civil war in Syria—is a highly contested one with geopolitical significance. Using content similarity, we generated a network graph of shared content, and utilized that network graph to conduct a mixed-method, interpretative analysis of the structure and dynamics of content sharing across active domains.

Our analysis uncovered sharing practices among both mainstream and alternative media domains, as well as a few aggregator domains that bridged the two. Articles that originated (or echoed) within mainstream media (Cluster A) were largely supportive of the WH, reflecting some of the critique (from those in Cluster B) that the WH are favored by Western, mainstream media. Our graph shows a couple of clear examples of content sharing of and by mainstream media. In particular, Cluster A represents (primarily) a single 'path' of an article originally posted by the AP and re-published by dozens of domains, including global and local mainstream media outlets and news aggregators. This activity is not insignificant, and clearly demonstrates that 1) mainstream media were participating in the WH conversation, primarily through the production and diffusion of pro-WH narratives; and 2) content-sharing is a component of mainstream news distribution.

However, for the WH conversation happening on Twitter during the summer of 2017, the vast majority of content production and amplification occurred on and through the alternative media domains represented in Cluster B. The content shared within these domains was strongly critical of the WH, promoting several related narratives that framed them as a propaganda construct and accused them of aiding, working with, or being terrorists. Using tweets as seed data, we were able to unwind trajectories of content-sharing across domains, articulating an alternative media ecosystem—or "echo-system"—of 130 distinct domains that provided source content for, or re-published existing content from, another domain in Cluster B. Analysis of this content-similarity structure, the domains that played significant roles within it, and the practices of content-sharing and remixing across domains, provides several interesting insights and leads to additional questions about how and why this echo-system has these properties.

6.6.1 *Explicit Critique of Mainstream Media*

One widespread theme within the domains that constituted the alternative media ecosystem in Cluster B is criticism of mainstream media and skepticism or outright rejection of its narratives. Messaging across many domains suggests that the mainstream media is lying to members of the public who should come to this website to get the truth. While some of these domains are extremely conspiratorial in nature (a synergistic worldview to the anti-media arguments), others are more focused on questioning the motives of western governments, in some cases specifically around conflicts in the Middle East, and positioning mainstream media as tools of those governments in those conflicts.

6.6.2 *Support of Russian Government*

Perhaps not surprising, considering the position of Russia as an ally of the Syrian government (which views the WH as assisting rebel forces), many of the domains in this ecosystem are explicitly supportive of the Russian government. Beyond RT and Sputnik, there are a few other sites that focus on Russia-related topics from a point of view favorable to the current regime: Russia-Insider.com, Russophile.org (Russia News Now), and Fort-Russ.com. Many of the other domains in Cluster B feature content supportive of Russian geopolitical positions (abroad) and specifically resistant to accusations that Russia had an impact on recent elections in the U.S. and elsewhere.

6.6.3 *Shared Content across Ideologically Diverse Sites*

But perhaps more interesting—or more impactful—than the commonalities across domains are the differences between them. Superficially, many of the domains in the alternative media echo-system articulated here appear to promote different ideologies. Consider a selection of domains that appear in this echo-system—i.e. MintPressNews, JewWorldOrder, LewRockwell, FreedomBunker, UprootedPalistinians, TruePatriot, TheDailySheeple, TheFringeNews, Anonymous-News, MakeWarHistory, ActivistPost, and TheRussophile. This list includes websites with strong political themes reflecting distinct (and in some cases, seemingly conflicting) ideologies—including anti-imperialist left, libertarian, conservative and alt-right; as well as other more niche ideological leanings, including explicit anti-Semitism. These websites are publishing the same content, but inside very different wrappers. The content itself is not necessarily tailored for each community (though each domain may select the articles most likely to resonate with its audience), but it is packaged up for them, appearing within a domain that features other material that may appeal to a reader's existing ideology. The effects of this kind of sharing may be to draw people from diverse, niche, political and ideological communities into a set of common narratives. We may think of these niche communities as being isolated and distinct, but here they are connected (in terms of common content) with other quite different communities.

6.6.4 *Future Work to Understand the Drivers and Impacts of the Information Echo-System*

Although in our Twitter seed data we found hundreds of domains and dozens of articles, we discovered that a small number of authors are responsible for a large proportion of highly-cited articles. We uncovered instances of circular attributions when authors cite themselves from other sources. We also demonstrated how similar content is repeated across in some cases vastly different domains. While such practices could reflect a more coordinated strategy, our evidence suggests that this complex ecosystem both has organic properties and is strongly influenced by a small set of politically, ideologically, and financially motivated actors and organizations.

Prior work has found that exposure to repetition of misinformation (which is not necessarily intentional) leads to a fluency effect—as people become familiar with claims they are more likely to judge them as true (Nyhan and Reifler 2012). By disseminating unverified or falsified stories to audiences through various channels, and from multiple sources, people may begin to assume they are true, regardless of the credibility of the individual sources (Paul and Matthews 2016). Although this research does not provide evidence of a coordinated strategy, the distribution of content across these seemingly distinct domains resembles a kind of intentional "astroturfing" campaign (Ratkiewicz 2011a) meant to exploit these cognitive biases. Future work is needed both to better understand the mechanisms underlying these patterns of content sharing and their effects on online audiences.

6.6.5 *Limitations*

One limitation of this work is the use of tweet data to 'seed' the investigation of the surrounding ecosystem. This method had the advantage of allowing us to measure the impact of these domains on the online conversation, but the disadvantage of having our view of the content-sharing shaped by the contours of a single social media platform. The vast majority of tweets with URLs in our dataset link to domains that were critical of the WH and appear in Cluster B in the graph—the area that we have termed the alternative media echo-system. Additionally, and perhaps consequently, the majority of content-sharing 'paths' across domains are in this area of the graph as well. It is likely that this tweet-seeded method resulted in a better view of content-sharing practices across alternative media than across mainstream media. Other less significant limitations include the loss of some URLs (and the articles/domains they pointed to) that we were unable to resolve, and the exclusion of article content that was not publicly available when we completed our automatic and manual scraping (in November 2017).

6.7 ACKNOWLEDGEMENTS

This research was supported by U.S. NSF #1715078 and ONR N00014-17-1-2980 and N00014-18-1-2012. We wish to thank students who provided significant assistance to this project, including Ostin, Vera, and Gordon.

(NOTE: This marks the end of the original publication)

6.8 REFERENCES TO PROBLEMATIC OR MISLEADING CONTENT IN THIS CHAPTER

For the purposes of this dissertation, I have chosen to separate academic references from references to content that I have problematized as misleading or potentially triggering. I have placed references to the latter here and to the former in the Works Cited section towards the end of the dissertation. This has been done to avoid driving traffic to it and blending it with more credible information.

Beeley, Vanessa. 2016. "Exclusive: The REAL Syria Civil Defence Exposes Fake 'White Helmets' as Terrorist-Linked Imposters." *21st Century Wire*, September 23, 2016. Accessed December 18, 2020. https://21stcenturywire.com/2016/09/23/exclusive-the-real-syria-civil-defence-expose-natos-white-helmets-as-terrorist-linked-imposters/.

Chossudovsky, Michel. 2015. "Saudi Arabia's Alleged Involvement in the 9/11 Attacks and the 28 Pages: 'Red Herring', Propaganda Ploy." *Global Research*, April 14, 2015. Accessed December 18, 2020. https://www.globalresearch.ca/saudi-arabias-alleged-involvement-in-the-911-attacks-red-herring/5442545.

Russia Today. 2017. "UK's Shadowy £1bn Conflict Fund Being Kept Secret from MPs." March 7, 2017. Accessed December 18, 2020. http://www.rt.com/uk/379702-conflict-fund-secretive-spending/.

Turbeville, Brandon. 2017. "Photos from Syria Show White Helmets And Nusra/Qaeda Are The Same Organization." *Activist Post*, May 2, 2017. Accessed December 18, 2020. http://www.activistpost.com/2017/05/photos-from-syria-show-white-helmets-and-nusraqaeda-are-the-same-organization.html.

6.9 Summary and Takeaways

This research examines the content and structure of discourse about the White Helmets, showing how a disinformation campaign leveraged 'independent' media outlets and think tanks to delegitimize the group. These outlets work alongside others explicitly funded and/or heavily influenced by the Russian (RT, Sputnik), and Syrian (syrianews.cc) governments to produce anti-White Helmets content which is then distributed across a diverse array of "activist sites that work tirelessly to bring attention to social justice issues at home and abroad"[34]. By citing one another and creating a dense network of interlocutors, this ecosystem is able to evoke a constant narrative aimed at discrediting the humanitarian group.

This ecosystem obstructs and reduces the practical value of trying to determine who might be a sincere but 'unwitting' actor (Bittman [1972] 1981) or a paid agent, complicating the strategy I mentioned in the previous chapter of addressing disinformation campaigns by de-platforming 'bad' actors (Stamos 2018). Disinformation campaigns do not just spread their messages by setting up 'fake news' websites or bot and 'troll' accounts, they can also try to engage existing online communities invested in different causes to participate in spreading their narratives. In the setting I studied here, these participatory dynamics contributed to anti-White Helmets content dominating other perspectives on Twitter, being cited by upwards of three times as much and by forty percent more accounts, many of which appear to have been drawn in from other causes— including anti-war activism and support for an independent Palestinian state (Starbird, Arif, and Wilson 2019).

This study also serves to extend the analysis from the previous chapter by illuminating how a disinformation campaign mobilized a wider assemblage of actors (e.g. 'independent-investigative journalists') to bend critical perspectives on powerful actors towards their own ends. For example, anti-imperialist and postcolonialist frames were mobilized to promote suspicion against the White Helmets and their coverage in Western media. The next chapter will explore this dimension of disinformation campaigns in further detail.

[34] This statement was used by MintPressNews, an 'independent watchdog journalism organization' to describes its partnerships. See:

MintPress News. n.d. "About MintPress News." Accessed December 20, 2020. https://www.mintpressnews.com/about-mint-press-news/.

Chapter 7. NETWORKS OF SUSPICION

I will now turn to consider how disinformation campaigns invite us to engage with information and to what ends (Research Question 3). Let me specify at the start that this chapter is conceived as a humanistic essay. It relies on the previous two chapters which presented scientific reports, but compared to those its structure is more fluid and idiosyncratic, while its tone is more conversational, inclusive, and provisional[35]. I have chosen this rhetorical form because my intent is to shift how we think about something, which does not require proof so much as provocation — a way to entice you into my thinking space. I am also not after agreement in the way that scientific reports often demand from readers through their stringent integration of methods, findings, and implications. To chase after such agreement is to invite the skeptic's question of whether I have arrived at the truth. Given the terrain we'll be navigating, I believe there is some value in setting aside that question, and instead sitting alongside to ponder the truth with me. I find that this collaborative author-reader relationship works better for questions of how best to think about what we know and for questions that reflexively challenge our habits of mind.

My position in this essay is that we mistake our object if we think of disinformation as consisting simply of a series of propositions or intellectual arguments. Disinformation also links to practices of suspicious interpretation. Technology designers and scholars in the information sciences often bound these practices using terms like 'conspiracy theorizing', but after studying how disinformation campaigns perform their arguments, I believe we are sorely in need of a more comprehensive and compelling vocabulary. So, what I am trying to do here is clarify an important aspect of disinformation campaigns and to de-center how we think about them—thereby freeing up designers to embrace a wider range of interventions. In particular, I want to put you in touch with a possible strategy for addressing disinformation that has to do with noticing the internal experience of suspicion and going beyond that. To begin, let's turn to a story with some humanistic depth to introduce a few ideas.

7.1 INFECTED BY DOUBT

On 31st August 2020, the Washington Post published an article titled *'Infected by Doubt'* to tell the personal story of how a 26-year-old aspiring filmmaker named Micah Conrad was slowly influenced by anti-vaccine ideology (Jamison 2020). While trying to sort out how he should respond to COVID-19 and make ends meet, Micah got a job editing videos for the Health Freedom Summit, a project organized by a network of influential anti-vaccine activists and coronavirus skeptics. Micah recognized that these skeptics' arguments are rejected by an overwhelming

[35] As the Bardzells' (2015, 68) have noted, these are some of the key characteristics of the humanistic essay.

majority of scientists and doctors, but how they *performed* these arguments infected him with doubt. The skeptics encouraged people to challenge received truths and consider ideas for themselves. This resonated with Micah, partly because he was raised to value individualism and to question authority. Unsure of what to make of these mixed feelings, he turned to his online social circle, where he was met with polarizing views. Ultimately, his trust in established information intermediaries eroded, and he made the following post on Facebook:

> I've had countless moments where I completely stopped work, and just listened to what these people had to say about topics like COVID-19, the government setting America up for mandatory Vaccinations, and how we are so brainwashed by what the mass media is telling us day in and day out, that we don't have enough moments where we stop and think for ourselves.
>
> My main take away from working on this project is that WE NEED TO GET OUR INFORMATION ELSEWHERE.
>
> Check WHERE you are getting your facts when it comes to COVID-19.
>
> Before you make a potential life changing decision about your health and personal freedoms, ask yourself WHO and WHAT ultimately convinced you. Challenge your own facts and reasoning. (quoted in Jamison 2020)

External judgements by fact-checkers on the information he was engaging with only served to deepen rather than dispel his doubt. For example, the article notes his response when Facebook labeled the Health Freedom Summit's posts as misinformation: "It didn't make me concerned about what I was doing, it made me motivated to get it out" (Jamison 2020). This highlights how combating misleading information is not a simple matter of tallying up false propositions. It is also about querying the entrenchment of a certain mindset of skepticism.

Even if you disagree with Micah's denial of science, you can no doubt recognize something of his shape of thought. After all, part of what Micah is inviting us to do here is to be wary of where we get our information from and to keep an open and critical mind. It is commonplace to hear that refrain in conversations and classrooms. In that sense, Micah's post articulates something of the struggle we[36] are all now engaged in, as individuals and as a society. That struggle is one of

[36] Lest we view Micah as intrinsically different from ourselves, I should highlight that researchers studying these information spaces can be infected by doubt as well. My two studies on disinformation, for instance, explicitly noted

navigating an information landscape where practices of reading against the grain and between the lines — practices that have served as an important tool for exposing, interrogating, and undoing structures of power, exclusion, and abuse — have been strategically perverted by other people. And as noted by boyd (2017a), frequently these 'other people' are those who believe themselves to be resisting the same powerful actors that we normally seek to critique!

In highlighting these connections, I join a growing groundswell of voices who are suggesting that disinformation campaigns can manipulate critical perspectives promoted within and outside of academia to pull apart the common ground or 'consensus reality' needed for civil society to flourish (boyd 2017a; Caulfield 2018; Pomerantsev and Weiss 2014). An earlier salvo was fired in this direction by Bruno Latour (2004) in his essay, '*Why Has Critique Run out of Steam?*,' in which he contends that tactics forged by progressives— skepticism about the objectivity of facts, surfacing the motives of scientists— now serve the agendas of more reactionary forces, evident in matters like climate change denial. Latour's general call to the social sciences and humanities is to develop methods of critique devoted not only to debunking but also to "repair, take care, assemble, reassemble, stitch together" (2010, 475).

Latour's argument is rich in potential but in an as yet unclear or muddled way in the context of modern disinformation campaigns, especially in the field of HCI. If we conceive of practices like fact-checking, questioning sources and debunking as exercises of critique, then clearly many of the interventions currently being designed to support these practices have the potential to become self-sabotaging. When an overdose of cynicism is part of the problem, it's not so clear that we can rely on critique as a solution.

That many researchers and practitioners want to address mis- and disinformation by developing more tools to critique information (i.e. to label it after finding fault with it) is understandable. Critique is pervasive and would not be such a charismatic mode of engagement, after all, if it didn't gratify and reward its practitioners. But if the only response we can muster to the perspective Micah absorbed and shared is to try and refute it, then I would contend something is limiting our collective imagination—because that did not work as far as the story goes.

that disorientation was part of the research process. There are several layers to this disorientation, but part of it has to do, I think, with how disinformation campaigns soak us in an atmosphere of suspicion. For example, while studying the disinformation campaign targeting the White Helmets, I had my research team write reflections and discuss them, which helped us discover that each of us, in our own ways, was stuck with feeling noticeably more wary and skeptical of mainstream media outlets, western governments and nonprofits. I should also note that having open and candid conversations about these uneasy feelings would become more challenging if we approached them with the frame of conspiracy theorizing.

If our creativity in this clash of interpretive forces seems constrained, then I suspect part of it has to do with how we often treat Micah's mode of thought as something special and different. Consider how the terminology of paranoia has established itself as a ready-to-hand label to orient us around misleading information. We use terms like *conspiracism*, *conspiracy theories* and *the paranoid style* (Hofstadter 1952) to discuss stories like Micah's which involve being suspicious of established knowledge authorities. For example, the full title of the story covering him is '*Infected by doubt: A 26-year-old film editor's descent into coronavirus vaccine conspiracy theories*'. This terminology is a way to both describe perspectives and to disqualify them by yoking them to a diagnostic category associated with irrationalism, obsession, and monomania. Such language can help us productively engage in 'boundary work' (Gieryn 1983) when we are faced with interpretations that seem to consistently exclude contingencies, embrace tautologies, and evade counterarguments.

Past a certain point though, this language doesn't take us further because it also does the work of estrangement, rupturing continuities and severing attachments. Too often, the above terms are essentially "more sophisticated ways of calling someone a crackpot" (Bratich 2008, 5), which strikes me as a rather bleak and limiting prognosis for those of us committed to doing human-centered design. The prospect of addressing disinformation using design principles built on empathy, trust, and meeting people where they are starts rapidly receding when we position people's mode of thought as maladjusted. No matter how precise or metaphorical, the effect of speaking of paranoia is to cast a pathological shadow over people's sensemaking.

So, what if instead of diagnosing or refuting, we try reframing? If we frame Micah's mode of interpretation as critique, we downgrade its specialness by linking it to a larger history of suspicious interpretation. I want us to entertain the idea that disinformation campaigns ask audiences to be critical in much the same ways that we encourage it in other settings (e.g. higher education). My conviction is that grasping some of these continuities could help us productively recalibrate the prospects for a wide range of interventions against misleading information today.

7.2 Critique and Postcritique

By now we are overdue for some qualifications and definitions. Before delving further, we need to understand 'critique' in a way that could be fruitful for scholars interested in disinformation. On a basic level, we can see critique as a style or genre[37] of interpretation (Anker and Felski 2017, 3).

[37] Anker and Felski (2017, 4) use genre in the Wittgensteinian ([1953] 2009) sense of family resemblances. It is meant to acknowledge that the word critique points to a wide and fluid constellation of practices that, although

To go further and understand some of the characteristic modalities of critique, I want to draw on certain influential perspectives from the humanities that have advocated for a 'postcritical' sensibility. After unpacking this perspective, I will use it to interpret some findings from the previous two studies to illustrate what it can bring into view. Afterwards, we will discuss some additional implications and consequences of this position for researchers and designers.

To understand critique, I will use the work of Rita Felski, especially her book, *The Limits of Critique* (2015). Felski, whom I will quote at length and with pleasure, has deftly sketched the core convictions and character of critique by building on the ideas of Bruno Latour (2004; 2005; 2010). She does not seek to critique critique, so much as position it as a dominant mode of thought to help us expand our horizons beyond it. Rather than dissecting and dismissing critique (as critical readings often do), she tries to deeply explore its limits. While her writing is focused on literary and cultural studies many of her arguments[38] have found broader purchase, being adopted by scholars working in fields like sociology (Jensen 2014), social justice (Anker 2017), international relations (Austin, Bellanova, and Kaufmann 2018) and critical race studies (Nishikawa n.d.).

Felski (2015) describes critique as a version of what philosopher Paul Ricoeur (1965) called the *hermeneutics of suspicion*. To read[39] something suspiciously is to insist its real meaning lies hidden — shrouded by the 'intentions' of the broader social contexts which produced it. It is, "to grapple with the oversights, omissions, insufficiencies, or evasions in the object one is analyzing. It is to tabulate a limit, to discern a lack, to heave a sigh of disapproval or disappointment" (Felski 2015, 127). By foregrounding this negativity and suspicion, Felski makes it clear that critique is not an exclusively intellectual[40] activity. The skepticism and nonchalant detachment which often grant

impossible to account for in a comprehensive manner, are overlapping and related in some ways. Influential objections to critique like Felski's have pointed out that the distinctions and discontinuities between different practices of critique can matter, but that "an exclusive fixation on these differences prevents us from asking equally important questions: What do forms of critique have in common?" (Felski 2015, 20).

[38] It should go without saying that my precis cannot do justice to Felski's subtle arguments. What I'm trying to offer here is a 'just-enough' explanation to kick-start a conversation (while trying to avoid piecemealism). I also want to acknowledge that this 'just-enough' explanation owes something to the writings of Elizabeth S. Anker (2017) and Matthew Mullins (2015), both of whom have implied that Felski's ideas can be adopted by intellectuals in other disciplines and have summarized them in several venues with wonderful insight and clarity.

[39] I have chosen to use the words 'read' and 'interpret' fairly interchangeably in this section. The same is true for 'texts' and 'information'. My intention is not to efface any distinctions between these terms but to facilitate dialog between different discourses.

[40] Felski's position challenges a certain academic self-image that tries to place critique firmly in the hands of academics and out of the hands of broader publics.

critique its unique and exceptional status come to “mark it as a resolutely emotional endeavor” (Mullins 2015). To practice reading and thinking in this manner, to become the critic, brings its own affective and cognitive payoffs. The critic gets to be the knowing reader who stands above the attachments holding back others from deducing what a text or piece of information really means – whether it be a novel, a news story, or a statistic. Casual observers are like the hapless Watson: seeing but not observing, they are easily bamboozled by the appearances of things. The critic on the other hand gets to be a member of an exclusive club of sleuths, like Sherlock, who can press below distracting surfaces to detect concealed truths.

What Felski is doing here helps us carefully unlock the *mood* of critique. Moods, as described by Heidegger, are less our inner, subjective states of Being and more “like an atmosphere, in which we are steeped” (1996, 67; quoted in Felski 2015, 20; cf. Blattner 2006). Moods “set the tone” for our interactions with the world by suffusing our thoughts and modulating them, shaping how things appear to us (Felski 2015, 20). They are also something ambient and lingering, pervasive and slow to change; they are hard to bring into focus and to examine. Part of Felski's argument is that styles of interpretation call up an ambient mood, orienting their practitioners in ways that inflect or accent their activities (e.g. whether they approach a piece of information with a stance of trust or mistrust). In this sense, thinking or reading critically is not an absence of mood, but one manifestation of it. When we educate people (say, students or social media users) to be more critical of information (e.g. by encouraging them to look for a fault to find or gap to fill), we are not just teaching them a method, but initiating them into a certain sensibility. And like any other repeated activity, the more we engage in critique, the more we become acclimated[41] to a certain mood, which starts seeming less alien and more like a “reassuring rhythm of thought. Critique inhabits us, and we become habituated to critique” (Felski 2015, 21).

Felski (2015) offers two spatial metaphors to describe critique’s modes of operation. First, critics can *dig down* to excavate a repressed or hidden aspect of reality. To dig down is to assume that a text/piece of information possesses qualities of “interiority, concealment, penetrability, and depth; it is an object to be plundered, a puzzle to be solved, a hieroglyph to be deciphered” (Felski 2015, 53). Critiquing this way involves stooping to dissect and decode texts, mistrusting their surfaces. The second metaphor is to *stand back* which swaps discovery for defamiliarization. Rather than burrowing for hidden layers, this mode of practice involves drawing away from texts to position them within larger structures of power. Felski writes: “Insight, we might say, is achieved by distancing rather than by digging, by the corrosive force of ironic detachment rather than intensive

[41] This is to recall the adage that to a fish, water just is.

interpretation. The goal is now to 'denaturalize' the text, to expose its social construction by expounding on the conditions in which it is embedded" (Felski 2015, 54). That which is taken for granted must be scrutinized.

Felski (2015) associates digging down with traditions of Freudian and Marxist[42] thought and stepping back with poststructuralism. Her point is that even though the methodologies of critique might encompass a wide range of theories, frameworks, as well as disciplinary and political commitments, they are tied together at the level of mood. Both digging down and standing back try to pinpoint and evade misperceptions by assuming that the text or its source is guilty of some crime requiring suspicious interpretation; both prize a stoic detachment and resistance to a text's address. Felski (2020) notes that these stances feed an *epistemological asymmetry* that is problematic: "the promptness to explain what other people want and do (but never one's own desires and actions!) in terms of hidden structures they fail to understand".

To understand critique's confidence and charisma, its reassurance that its mistrust is warranted, it becomes important to recognize that when we do critique, we are not alone. Like any practice, there is a collective 'we' stretching across time and space that we draw strength from. Critique, Felski (2015, 49) writes, "creates imagined or real communities around a sensibility, ethos, and practice of reading". To illustrate this, she surveys four different historical strands of suspicion to consider how they mediate critique. For example, she considers *vernacular suspicion* (derived from the ingrained wariness and cynicism of downtrodden and disenfranchised peoples) and *philosophical suspicion* (as espoused by thinkers like Descartes and Kant who upheld critique as a tool to overcome modes of unenlightened captivity). She argues that these strands of suspicion imbue critique with an ethical promise of opposing injustice[43]. This can help critics feel reassured that they are engaged in a marginal, oppositional, or radical practice — whether it's justified or not. Any questioning of critique risks being automatically equated with being uncritical and a meek acquiescence towards the status quo.

[42] Felski (2015, 61) is coming at this with the perspective that "psychic repression and political oppression can be seen as two sides of the same coin—in both cases, something is being forced down, restrained, and muffled by a controlling force". Understandably. some interpreters of Felski's work also connect Marxist thought to the stance of standing back (e.g. London 2016).

[43] This view of critique as a tool to undermine authority (rather than enforcing it) helps explain why the term 'critique' might feel awkward when we use it in reference to, say, a fact-checking intervention designed by a social media company. I confess this is something I chafe against and must continue thinking about. It might be more prudent to think of such fact-checking interventions as a narrow operationalization of critique (since the posture of suspicion and negation does carry over somewhat).

Recalling Latour, Felski (2015, 45) however insists that there is nothing “automatically progressive about a stance of suspicion”. Although critique has served as an important tool for battling injustice, it can also deride and write off the very lines of human experience and expertise that are crucial to human flourishing. Moreover, the oppositional and negative attitude of critique often feed into a vicious and inexhaustible cycle. Suspicion has a way of intensifying our drive to dissect things. It primes us to be alert and vigilant, to second-guess possible motives and watch out for lurking dangers that have yet to emerge (Felski 2015, 37; see also Shand 1922). Thus, suspicion drives us to engage more fervently and furiously in sensemaking, digging down, standing back and so forth. One limitation of this overriding concern with putting motives to question and exposing wrongdoing is that it inhibits generosity and our ability to understand many of the reasons that draw people to some text or information in the first place (e.g. feeling of belonging, increased self-understanding, emotional consolation etc.).

A significant challenge of responding to critique’s self-propagation is that it can be very difficult, almost disorienting, to question: “To object to or disagree with critique is to be caught in the jaws of a performative contradiction; in the act of disagreeing with certain ways of thinking, we cannot help being drawn into the negative or oppositional attitude we are trying to avoid” (Felski 2015, 192). To critique critique is to subject a certain style of thought to its own methods, redoubling it. This meta-suspicion informs Felski’s sound choice to not position her work as a polemic against critique. Instead, she invites us to try and be ‘postcritical’—i.e. to explore how we might expand our critical moods and enlarge our intellectual repertoire beyond the deflationary work of digging down or standing back. Gaining traction, her work has contributed to the emergence of the related term, *postcritique*, as scholars in various fields have started exploring new possibilities and intellectual alternatives to a suspicious hermeneutics (Anker and Felski 2017). Felski acknowledges that the “post-” is as cumbersome here as it is in many other words that have proliferated in the last few decades. But it is also appropriate because it signals an inevitable dependency on the practices of critique. It views critique as being valuable in certain situations, but inadequate in others. To be postcritical is thus not to be un- or anti- critical but about supplementing critique with new interpretive practices when the situation calls for it.

What might these supplemental interpretive practices look like? Felski is careful not to overprescribe the forms that interpretation should take. But let us listen to one direction she suggests that is relevant to our purpose here: “Rather than looking behind the text—for its hidden causes, determining conditions, and noxious motives—we might place ourselves in front of the text, reflecting on what it unfurls, calls forth, makes possible” (2015, 12). That is, we could be asking how we might help people not just cultivate detachment but think carefully about their own attachments when they engage critically with information/texts. We will return to these ideas later.

But the key point here is that we could ask how we might support ways of reading that still bring out the very real pleasures and rewards of critical thinking but are not blinkered by suspicion.

Before we proceed, a final point to note is that Felski considers the vagueness of postcritique to be "its singular strength, allowing it to serve as a placeholder for emerging ideas and barely glimpsed possibilities" (2015, 173). This allows it to be an experimental and open space — postcritique denotes a conversation that is present, vital and ongoing. Felski is offering a provocation that I believe the HCI community can take inspiration from and build upon constructively in its own ways, particularly in the context of misleading information.

Let's explore this direction further by bringing some of the continuities between critique and disinformation into focus. One way to do this is by taking some of the concepts that have been introduced to describe the properties of critique and seeing how they might also be used to describe the properties of disinformation campaigns[44]. Some of the specific concepts and ideas we've been introduced to include *mood*, *suspicion*, *digging down*, *standing back*, *the thrill of detection*, *the claim to being radical*, *drawing strength from communities*, and *epistemological asymmetry*. Placing these concepts, which we inherit from Felski's deeply humanistic perspective, into conversation with the understanding that we've been developing over the last few chapters, can help us arrive at a more holistic understanding of disinformation campaigns.

7.3 CRITIQUE AND DISINFORMATION

In this section, I will pull forward a few examples from the discourses I examined in the previous two chapters[45]. I will use these examples to illustrate that disinformation campaigns can invite us to be critical in the ways that Felski (2015) describes. Surfacing these connections can help us see that these campaigns are as much a matter of affect and rhetoric as of information, and suggest some fresh directions for design interventions. To start, we can consider the disinformation campaign that targeted the White Helmets. The following is an illustrative quote from the 'About Us' page of the most cited domain in the data, 21stCenturyWire [46] (2012):

[44] And again, more specifically ones that might use the tactics associated with *dezinformatsiya*.

[45] Note: some of these examples were analyzed but not described in the original publications due to considerations of space and focus.

[46] For the purposes of this dissertation, I have chosen to separate academic references from references to content that I have problematized as misleading or potentially triggering. I have placed references to the latter at the end of the chapter (and to the former in the Works Cited section towards the end of the dissertation). This has been done to avoid driving traffic to it and blending it with more credible information. The references to problematic information are formatted as superscripts to distinguish them.

> The 21st Century is the beginning of a new information epoch and you, the reader, are the freshman class of free and critical thinkers in a new and dynamic information age...
>
> In spite of the Internet Revolution, the corporate Mainstream Media (MSM) is still clinging to a predictable and monolithic format which it regularly uses to disseminate government, military and corporate disinformation, as well as outright lies and propaganda. This fact is the primary reason why the alternative media is increasingly its share of audiences attention every day [sic]. We believe the essence of a free Internet is really about choice, interaction, discourse, dialogue and debate – and not about being force-fed a narrow band of information and 'consensus reality' view points, constantly packaged and re-packaged by a few major media moguls and legacy news outlets. To realize the benefit of humankind's opportunity in the 21st century, then you, the surfer of the information waves, cannot take information for granted. You must go out and dig to find the forum, the blog or the video that you are looking for. That's why you are here right now – you are here because you are not getting your primary news and analysis from the likes of the New York Times, Washington Post, The Huffington Post, CNN, FOX NEWS or TIME Magazine. If you try explaining that one to an Ivory Tower full of Big Media execs, they simply won't get it, as too few of them have really been able to grasp what has happened in the this space over the last 20 years, much less where the next decade is heading… (21st Century Wire 2012)

The parallels between this stance and Felski's account of critique are strikingly obvious. Here we can see the claim to hold knowledge, truth and evidence in high esteem, an invitation to engage in intellectual exploration, as well as the argument that it is established knowledge authorities who are 'post-truth'. The call to go out and dig, which echoes both Felski's (2015) metaphor and Russia Today's slogan of 'questioning more', is particularly significant: we often assume that disinformation targets the unsavvy or the uninformed, but here we perceive it speaking to a spirit of blistering disenchantment—a desire to puncture illusions, topple idols, and demystify hidden truths.

By reframing these messages as an invitation to practice critique, we can start focusing on what makes them enticing and tricky to repudiate. For example, we can glimpse the thrill of being a detective and the affective rewards that come from standing above more susceptible others (that same thrill that sometimes drives us in higher education). We can also see a stance of being oppositional and radical. Opaque social forces must be skewered and any other move, by contrast, must fall into the camp of compliance. Felski's perspective helps us recognize that this conviction that those at odds with the status quo see better and farther than others is not merely a hallmark of conspiracy theory, but also part of the iconoclastic mood of critique. Bringing these connections into focus can provoke us to think about new questions. Questions like: What do we make of the idea that those accused of spreading disinformation and conspiracies may be seeking the same thrill often pursued in academia? What do we make of the power and energy that can come from knowing one's efforts to be marginal, oppositional and radical?

Before we can turn to such questions, we should note that, as a solitary object, this excerpt from 21stCenturyWire has little strength. Its real impact comes from how messages like it are situated and repeated throughout the larger discourse. Drawing on the structural analysis from the previous chapter, it is clear that part of the website's popularity can be traced to a far wider network of content creators. We can appreciate this network in light of Felski's remark that critique draws strength by forging connections — from gathering people into 'interpretive communities' (Fish 1980; Felski 2020) of sorts where certain concerns, beliefs and ways of reading become naturalized. Here, some of the more influential content creators (or members of this community) include 'independent' journalists and academics who actively call upon their audiences to dig down and stand back.

One of the ways they do this is by providing a certain kind of education to help their audiences make sense of the media landscape surrounding the Syrian Civil War (see Figure 7.1 below). This education involves, for example, introducing audiences to Herman and Chomsky's (1988) propaganda model (21st Century Wire 2020; Pacheco 2017); discussing ideas like Gaslighting (Beeley 2016) and Neocolonialism (21st Century Wire 2017); framing academia as Pavlovian conditioning (Dyer 2016); positioning humanitarian NGOs as a tool of oppression (21st Century Wire 2016); and running a #FakeNewsWeek campaign (21st Century Wire n.d.) (to help readers learn about the deliberate lies told by mainstream media)[47]. Audiences are schooled to view established knowledge authorities as the grinding machinery of regulation—with resistance being the only conceivable escape hatch. And over and over, audiences are reminded that resistance requires being critical— they must be the

[47] It is not necessarily straightforward or easy to problematize actions like promoting awareness about U.S. foreign policy fiascos or the corporatization of higher education. What I've found helpful is to consider how these actions are fitting into larger patterns of *coordinated* activity rather than thinking of them in isolation. In the White Helmets case study for example, appreciating these larger patterns of coordinated activity helped ground me and my colleagues by helping us understand how these actions were being taken towards destructive ends (i.e. tearing down the White Helmets and anyone who might support them).

Of course, this is not entirely satisfying. Patterns of coordination can be very difficult to establish and mentally hold on to. And on a deeper level, my life in Pakistan has taught me that problematizing activities by claiming coordination can risk harming progressive causes. This is part of my reasons for exploring the limits of critique and negation in this domain.

ones who dismantle, disassemble, and unravel the threads of explanation and judgement woven by existing knowledge authorities. Moreover, by offering a panoramic view of systems of discourse and grids of power, these content creators are able to argue for the validity and need for the alternative points of view that they provide. To resist, one only has to "consult reliable sources and to employ critical thinking" (21st Century Wire 2017).

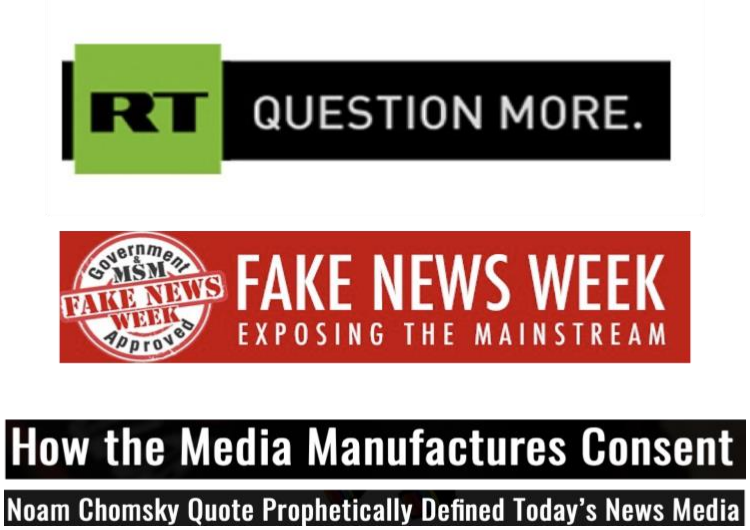

RT
QUESTION MORE.
Government & MSM
FAKE NEWS WEEK
Approved
FAKE NEWS WEEK
EXPOSING THE MAINSTREAM
How the Media Manufactures Consent
Noam Chomsky Quote Prophetically Defined Today's News Media

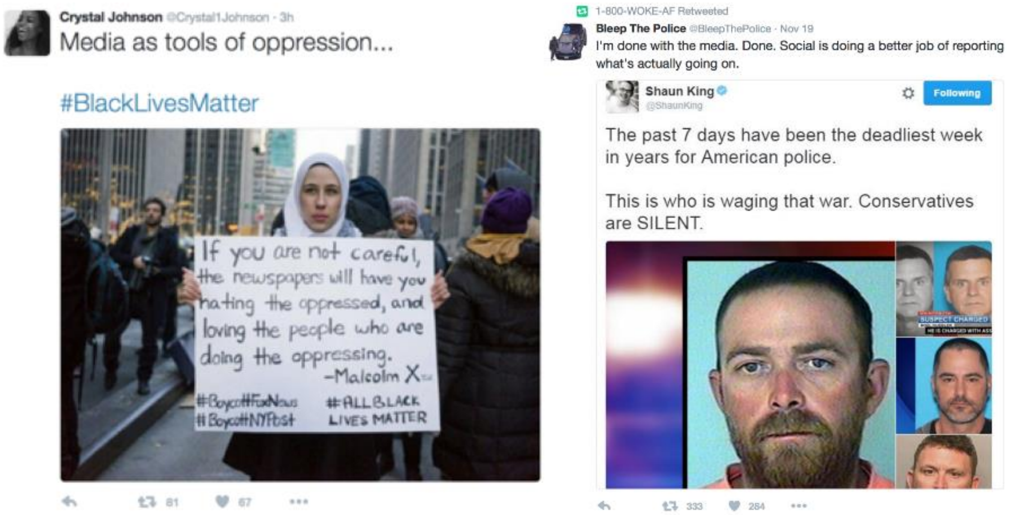
Figure 7.1 Sample graphics to illustrate how these websites promote 'critical thinking'. From top to bottom: RT.com's slogan; 21stCenturyWire's graphical header for 'Fake News'; and a headline from MintPressNews.com.

Self-styled Western investigative journalists like Vanessa Beeley and Eva Bartlett capitalize on this framing to aggressively challenge pro-White Helmets content and mobilize their followers (e.g. through tweets, articles, videos) to drown it out in the information space. Beeley and Bartlett, who themselves gained influence partly through amplification from state-sponsored media outlets (e.g. broadcast interviews on RT, sponsored speaking tours etc) (BBC News 2018; Starbird, Arif, and Wilson 2019) frequently perform the metaphorical act of *standing back* to critique the authenticity of views sympathetic towards the White Helmets. Rather than engaging with the substance of these views (matters of fact), they prefer instead to defamiliarize them (i.e. make them strange). For example, in articles like "*Channel 4, BBC, The Guardian – Architects of 'Humanitarian' War*" and "*What to Expect From BBC Panorama and Guardian's Whitewash of UK Gov't Funding Terrorists in Syria*," Beeley reduces the import of mainstream media sources to the Western corporate and government interests she claims they are embedded in. This is, as Latour (2004, 243) remarks, the critic's gesture of reducing something to dust by showing it is made up. It's also, as Felski remarks, the smug gesture of "scholars to impute hidden causes and unconscious motives to the arguments of others, while exempting themselves from the same charge: 'I speak truth to power, while you are a pawn of neoliberal interests!'" (Felski 2015, 186). We can see this as an example of critique's epistemological asymmetry.

These content creators also uphold *digging down*. Their videos and articles frequently call upon practicing 'independent research' by drilling down into pictures, newspaper articles, and scientific claims to counter the false consciousness promoted by Western media. For example, in articles like "*Propaganda Alert: Madaya Media Fabrications, Recycled Photos*", Eva Bartlett(2016) persistently scrutinizes photos, videos and maps to make the claim that White Helmets media workers use crisis actors to play the role of victims — a claim that has been debunked (Palma 2016).

Another prominent case is the *Working Group on Syria, Propaganda and Media* (n.d.) — comprising three professors, two lecturers and three postgraduate researchers at British universities who have been accused by public officials, intelligence analysts and journalists for spreading pro-Assad conspiracy theories (e.g. York 2020; Haynes 2019). These academics promote independent and intense acts of digging down with a veneer of scientism. For example, McKeigue (Hayward 2017a; Hayward 2017b) employs Bayes' theorem and Hempel's paradox to evaluate various hypotheses regarding the events in Ghouta and Khan Sheikhoun (during 2017), and to "show you, the reader, how to evaluate evidence for yourself using simple back-of-the-envelope calculations based on probability calculus" (Hayward 2017a). He shows readers how to construct a chain of reasoning that concludes there is "overwhelming" evidence that the chemical attacks by the Assad regime were actually "a managed massacre of captives [involving opposition media], with rockets and sarin used to create a trail of forensic evidence that would implicate the Syrian government in a chemical attack" (Hayward 2017b). This absurd example begins to illustrate how this network of content creators can frame politics through research and research through politics, while at the same time critiquing academic research that has political aims.

Critique is also salient in the context of the Internet Research Agency's activities within #BlackLivesMatter discourse (Study 2). Compared to the White Helmets study, we cannot see the same dense network of interlocutors, but we can still capture one or two important ideas. For example, Figure 7.2 below shows two tweets I highlighted in Chapter 5 that demonstrate how the Agency's social media accounts on the 'left' critiqued traditional and mainstream media.



Crystal Johnson @Crystal1Johnson · 3h

Media as tools of oppression...

#BlackLivesMatter

1-800-WOKE-AF Retweeted

Bleep The Police @BleepThePolice · Nov 19

I'm done with the media. Done. Social is doing a better job of reporting what's actually going on.

Shaun King @ShaunKing

The past 7 days have been the deadliest week in years for American police.

This is who is waging that war. Conservatives are SILENT.

Figure 7.2 Some examples from Chapter 5 that show how ‘left’ RU-IRA tweets criticizing traditional media (See Figure 5.8 and also Figure 5.9 for the original context).

These examples show, perhaps more clearly in an American context, that disinformation campaigns can invoke critique in a way that is linked to progressive causes and is authorized by being rooted in the experiences of those who have been historically deprived of authority (the traditions of *vernacular suspicion* noted by Felski [2015]). Perceiving this, we might start to *feel* the limitations of addressing disinformation by critiquing it. Weeding out ‘false’ information or ‘bad’ arguments doesn’t seem quite right here. Critiquing these critiques (i.e. problematizing narratives rooted in suspicion) can give rise to a discomfort — a jarring dissonance — the moment one recognizes that people who have been disenfranchised are entitled to their suspicion.

In short, the invitation to critique is prominent in the messages promoted by disinformation campaigns. Invoking critique grants these campaigns an immediately recognizable, widely applicable claim to serious thought. It also casts a protective shield over their endeavors, making them resistant to the application of critique as a counter. Let us unpack this idea further in the next section.

7.4 Postcritique and Disinformation

7.4.1 *Some limitations of critique as a response to disinformation*

So far we have used Rita Felski’s (2015) work to understand critique as a wide constellation of practices that are nevertheless tied together through their shared stance of suspicion. Having brought this stance into view, we have also seen that it is mirrored in the messages and arguments promoted by disinformation campaigns. We can now turn to consider what this humanistic perspective affords us.

Part of the value of this perspective, I think, lies in how it alerts us to some of the deeper liabilities of critique as a response to disinformation (i.e. unmasking bad actors, interrogating false information, and deconstructing poor arguments). Here, I wish to underscore that in talking of such liabilities I am not trying to construct a screed against critique, disagreement or negative judgment (I have engaged in all these activities in the preceding pages of this dissertation). I am not advocating for us to do away with critical responses to disinformation. I am trying to help us recognize the limitations of such responses when we face sophisticated efforts to disinform. When these efforts usurp the tools we consider reliable — that we have traditionally counted on to make sense of the information landscape around us — then the overwhelming focus on negation that lies at the heart of our critical responses can start to become counterproductive.

The crux of the matter here is that critiquing acts of critique — a move Felski would humorously call ‘critique squared’ (2015, 148) — draws us further into a suspicious mood, and we find ourselves caught in an endless regress of questioning and condemnation. To respond to the critiques promulgated by disinformation campaigns with ‘critique squared’ is to catch ourselves mimicking the very move that we are trying to subdue (e.g. questioning messages that ask us to question more, claiming insight into the motives of those who claim to have insight into ours) (Felski 2015, *35*; see also Sedgwick 1997). One problem with this is that the underlying logics of *dezinformatsiya* can remain intact. Let me illustrate this point in two ways so that you can follow my chain of reasoning.

A picture of society

In the previous section, we observed that disinformation campaigns critique the positions or information sources that they oppose by invoking some larger frame (e.g. western imperialism). In doing so, they call up a certain picture of society. In this picture, we live within a prison house of opaque social forces that we can only break out of by engaging in radical acts of deconstruction, defamiliarization and demystification. Drawing on Felski (2015), we can note that this picture is hardly original; it’s the one we routinely paint as part of doing critique — when we invoke power as a determinant of social meaning. We can also acknowledge that there are important and complex reasons for this picture to exist — as one example, consider how feminist or postcolonial perspectives have served as a tool to critique structures of oppression. By taking some intellectual shortcuts, disinformation campaigns can tacitly link to this larger history to claim a moral and political high ground.

Keeping this picture in mind is important because it is part of the ethics of the situation. If we focus narrowly on labeling false information or bad actors while failing to engage with this larger frame, we risk lending it strength with our actions. For example, if a social media company like Facebook deploys a fact-checking intervention, it can be used to confirm the power of Western corporate interests to silence dissent (a particularly powerful argument in non-Western contexts!)[48]. Given the participatory dynamics of disinformation campaigns, it is important to avoid playing Goliath to their David. If we do, disinformation campaigns might be better able to enlist and maintain audiences who know themselves as being embattled, oppositional and part of a radical resistance.

At the same time, it can be difficult to engage with this picture because the tools critique brings to hand are limiting here. Recall that even though this view of society can be used by inauthentic actors towards malicious ends, it has layers of truth that connect to larger histories of suspicion. Negating this view by critiquing it makes for a weirdly self-canceling and ethically fraught argument. When we try to interrogate and deconstruct this view, we are in danger of placing ourselves in the untenable position of clashing with real and legitimate histories of suspicion (e.g. histories of colonial exploitation and racial inequality). Again, interventions that attempt to do ‘critique squared’ risk propagating problematic status quos and contributing (or being perceived as contributing to) the stifling forces of oppression.

Severing attachments

While disinformation campaigns claim to speak on behalf of ordinary people, we’ve also seen them assume the guise of an alternative intelligentsia doing serious intellectual work. This intelligentsia assign themselves and their followers the vantage point of an unswerving and sharp-eyed thinker while refusing to extend that same capacity to others. We can glimpse this move, for example, in 21stCenturyWire’s invitation to join “the freshman class of free and critical thinkers” who discern matters more clearly than the “Ivory Tower”[A]. As Felski notes, this move is powerful and seductive. It relies on a juxtaposed binary of, ‘those of us who are not critical are doomed to be uncritical’ — a form of *othering* (Said 1978). Who doesn’t want to feel like they are critical or intellectual? And if one rejects such an invitation, who wants to be associated with the stench of the uncritical?

These are part of the affective rewards and punishments of critique that complicate our efforts to address disinformation merely by labeling it. When we set out to reduce different lines of reasoning to whether they are ‘accurate’ or ‘inaccurate’, we overlook what makes them attractive or

[48] Similarly, on a more individual level, we can imagine how a person’s efforts to debunk a false claim about vaccines can feed the smug satisfaction that they are naively oblivious to the machinations of big pharmaceutical companies.

unattractive, which is significant given the participatory dynamics of modern disinformation. This relates to critique's affective inhibition which both Latour (2004) and Felski (2015) have commented on. Critique's overriding concern with questioning motives and exposing wrongdoing results in a mindset that makes it hard to account for the emotional investiture people bring to their beliefs. Latour likens such critical responses to a sledge hammer, with which, "you can do a lot of things: break down walls, destroy idols, ridicule prejudices, but you cannot repair, take care, assemble, reassemble, stitch together" (Latour 2010, 475).

Another limitation of the 'critique squared' approach is the risk of exerting authority over epistemology, defining what qualifies as 'critical thinking'. Critique has a contagious way of inviting us to mark the boundaries of what counts as serious thought. The temptation arises to restrict critical thinking to one side of the engagement, where we can get to "behave as if [we] were 'critical,' 'reflexive,' and 'distanced' enquirers meeting a 'naïve,' 'uncritical,' and 'unreflexive' actor" (Latour 2005, 57). If part of the logics of *dezinformatsiya* involve dehumanizing others, fostering antagonism, enabling authoritarianism, and promoting divisions in society, then this kind of *othering* ought to give us pause.

7.4.2 *Turning to Postcritique*

So what exactly am I proposing? The question can no longer be put off. Having laid out, to the best of my ability, some of the limitations of addressing disinformation by developing more tools to simply interrogate and indict information, I want to move on and expand our solution space. I believe that disrupting the critiquiness — the suspicious hermeneutics and negativity — promoted by disinformation campaigns requires us to attempt, along with Felski and others, to broaden our repertoire of critical moods. This involves designing to help us exercise our critique-muscle in ways that make wise use of not just our conscious reasoning but our emotional and social resources as well.

Obviously, I am taking my cues and inspiration here from postcritique while attempting to extend it to a new domain of concern. Again, the "post-" here suggests that interventions which try to address disinformation by critiquing information and those that try to make a postcritical turn can and should co-exist. But here are some things that a postcritical design intervention might forgo: subjecting information or its source to interrogation; diagnosing its hidden motivations; scoring points by showing that it is socially constructed; categorizing it as accurate or inaccurate. Here are some tasks it might help us with instead: reflecting on what makes a certain line of reasoning personally attractive or repulsive to us; contemplating our negative reactions when a friend belittles an article we shared; making sense of our reactions when an article or tweet we trusted is accused of being misleading; noticing the effects our information environment is having upon our mood.

Conceptually, the idea behind ushering in these alternative ambitions is to become more attuned to the complex ties that shape how we evaluate information. In contrast to the detached view that critique prioritizes, the postcritical position brings forward what Felski (2015; 2020) calls our attachments, which can be physical—my personal connection to the place where a mass-disruption event is unfolding; cognitive—the website that gave me a new way of looking at things; ethical — the moral commitments that inform my response to a tweet. The vocabulary of critique tends to be quite thin when it comes to such attachments, inviting us to either dismiss them as subjective sentimentality or view them suspiciously as chains of coercion, control and discipline. Felski (2020), however, draws on actor-network theory to present them as “a non-negotiable aspect of our being in the world”. In this view, the attachments we bring to our interactions with information become as decisive as the information itself. It is these attachments — our affective engagements — that are the very means by which information (faulty or otherwise) is able to reach and reshape our perceptions. What this conveys is that in addition to designing tools that help us critique information, we can try designing for our entanglements with that information.

An illustration of these ideas

Let me furnish an example of how some of these ideas[49] might be brought into conversation with education and design. In 2018 and 2019, I led two research groups to give students the opportunity to learn about the issues raised by disinformation campaigns and to design both technological and education interventions to address these issues. The groups consisted of eight to ten undergraduate and graduate design students, and I ran them like fairly typical humanities seminars (the group met

[49] Given that I owe an important debt to Felski’s work, let me briefly explain the stages in which her ideas met with my own. In 2017, it started becoming very clear to me that addressing online disinformation would require a cultural change in how we understand our own role in grappling with information. This was due to my empirical research and life experiences in Pakistan, where I have encountered the fallibility of two different activities: a hermeneutics of differentiation (Bratich 2008) that tries to feverishly sort out rationality into the authentic and the copy; and a hermeneutics of suspicion (Ricoeur [1965] 2008) that tries to target and excise those styles of thought that it deems unredeemable.

I started to search for ways to develop these ideas and write about them which led me to organize the directed research groups I am discussing here. While I remained unaware of Felski’s work at this time, I was drawn to Latour’s (2004) essay, “*Why Has Critique Run out of Steam?*”, which came to guide my thinking. It took a recommendation from my friend and colleague, Os Keyes, to change that in 2020. I am indebted to Felski (2015) for the theoretical vocabulary she has provided and the actual language that she has used to construct her arguments, which has helped me frame the work I’d been already doing. It was a further surprise and joy to learn Felski herself — unbeknownst to me — was also using classrooms as a place to work out how to put some of these ideas into practice.

once every week for ten weeks, readings were assigned and discussed, I used seminar-style teaching practices etc). However, I also ran them as “living laboratories” for design-based research (Evans and Karvonen 2011; see also Kafai 2005). Half the sessions focused on topics that helped students recognize the fault lines of the information landscape around them (e.g. the role of digital advertising, the asymmetrical polarization supported by right-wing media, the disruptions in the journalism industry and how that is exploited by media manipulators). These topics were a pleasure to teach and gave the students more fluency in examining the information landscape critically (yes, inviting them to be suspicious).

To bring the groups more in line with my evolving concerns about disinformation campaigns however, the other half of the sessions focused on topics that we usually do not associate with networked propaganda: the positions that attract or repulse us; our personal habits of reading and sharing the news; the things that absorb our attention in these settings; the moments we shared something that embarrassed us; empathy and sympathy. Such topics were chosen to help the student-researchers recognize how these aspects (and the attachments they represent) can mediate our interactions with information in complicated ways. The student-researchers began to recognize their own fault-lines[50] and appreciate that disinformation is not simply a matter of facts or poor judgement.

For instance, one activity the student-researchers tried out involved exploring how their own emotional makeup and values triggered certain reading habits in them when they encountered information from sources they had moral differences with. With careful scaffolding, each student-researcher repeatedly engaged with a news item of their choice that was repulsive and/or troubling to them (e.g. some learners self-selected the State of the Union address). They carefully reflected on their flow of emotions over time and some tried to empathize with those who they imagined to be the media’s target audience. Each student-researcher described that this activity was challenging but also immensely rewarding because it helped them learn something about themselves —e.g. certain habits, thoughts or emotions that had previously gone unnoticed. It also helped us as a group recognize that a human-centered design approach to address disinformation is not necessarily about some distant user who is uncritical and unreasonably paranoid but about ordinary people who are drawn to narratives because of their attachments. These attachments are not beyond reproach but none of us are immune to them.

[50] It was often hard for students not to be self-critical and judgmental about the discoveries that they made. But inviting them to halt this critical machinery for a moment brought out a certain energy. I heard a collective sigh of relief from the students at encountering a space and analytical frame that valued simply noticing and appreciating their experiences of engagement — as something more than user errors or signs of bias that get in the way of hard-nosed critical thinking.

Devoting time to this postcritical stance led the student-researchers to grapple with tough questions. How are our responses to a news item shaped by our individual histories and the expectations we place on ourselves? What can we do in education to deter learners from focusing on proving others wrong to prove themselves right? How do we design interventions that address people's psychological needs after we correct them about something they shared? What are the cognitive payoffs of reading something we disagree with? (One of the student-researchers noticed that reading something he disagreed with produced more active engagement while agreement drew him into a more passive mood). How can technology be an ally in promoting inclusive dialogue, and how can it be incentivized to do so?

Such questions motivated the students to prototype a variety of interventions. One first-year undergraduate student prototyped a series of tactile experiences using Lego blocks to help users create physical representations of their information diet and social network to facilitate further reflection on these attachments. Another student prototyped a tool that used the tweets promoted by the Internet Research Agency to help users notice what encounters with inauthentic content might set alight in them. As a group, the student-researchers chose to design an online tool to help users explore what it can be like to read and critique news in an alternative way—one that afforded more peer-support and opportunities to practice 'slow' thinking. They prototyped the tool in Google Docs and tested it in 16 semi-structured pairwise chats about the news. In these chats, student-researchers either played the role of readers or 'critical friends.' Readers brought in some information or text about current events that they wanted to think about more deliberately. Critical friends, conversely, played supportive listeners, asking provocative questions and offering alternative perspectives. The researchers improvised these questions during their chats, with the aid of a handbook that they had co-constructed using tenets from Non-Violent Communication (Rosenberg and Chopra 2015). In this way, critical friends could help readers[51] reassess not only how they read and critique but also why (part of the agenda of postcritique).

Disrupting Disinformation

The above examples help illustrate a design direction that we can associate with postcritique. This direction involves addressing disinformation by testing out alternate ways of reading and thinking. It seeks to refresh our perceptions about how we *relate* to information, taking the act of critiquing information (something we can never fully do away with) and trying to infuse it with new gestures, attitudes and habits of thought. To reprise Latour, it is to indicate the direction of critique away

[51] Part of the value of peer-support lies in how the individual providing the support or teaching also has learning gains. So it might be unsurprising to hear that one of the project's findings was that the critical friends were also getting to explore and learn about how to do 'slow critique'.

from lifting “the rugs from under the feet of the naïve believers” and towards offering “the participants arenas in which to gather” (2004, 246).

This direction might undermine the logic of disinformation in some important ways. First, as Kathryn Fleishman (2019) has remarked, postcritique “reclaims affect as an integral part of intellect.” In doing so, we reveal the seemingly objective stance of ‘questioning more’ as an emotionally charged one — one limited and fallible position among others. Disinformation campaigns often shroud their rejection of existing knowledge authorities by positioning their own interpretations as purely detached and non-complicit with any power structures. If we recognize that interpretation is never purely detached in the first place, we weaken this presumption of epistemological and political privilege. As Felski remarks, critique thus becomes “a less muscular and macho affair than it is often made out to be” (2015, 10).

A willingness to acknowledge and more fully engage our attachments can open up new directions for design. In addition to efforts that help their users refute information, we can, as I’ve already suggested, also promote interventions that help them approach their information environment with an embodied mode of attentiveness (i.e. supporting acts of noticing, feeling, reflecting). We can think of such design interventions as cognitive strengthening exercises that can help their users be more intentional as they engage in acts of sensemaking. To put it in the language of Harvard psychologist Robert Kegan (1994), we can help people *have* their attachments rather than be *had* by them. Instead of being fused with our emotions, values and epistemology, we can start to step outside of these things to reflect on and be responsible for them. Part of what makes this direction attractive is that it honors the *emergent* nature of the problems we now face with disinformation. As media manipulation efforts continue to grow more sophisticated (consider, for example, deep fakes), it might be prudent to focus on the skills that will help us be *adaptive* going forward rather than playing catch up with only short-term solutions.

There is another way that postcritique might strike a blow against *dezinformatsiya*. When we reign in the critical impulse to destroy and deconstruct, we constrain the tribal logic of defense that disinformation campaigns promote and rely upon, making more room to rebuild trust in institutions and that frayed social fabric we sometimes call ‘consensus reality’. In the course of my work, I have noticed that focusing on our own attachments to information has a funny way of pushing us to see the humanity in ourselves and in others. We can find ourselves rejecting the premise of a radical asymmetry between the mode of interpretation that Micah (from the beginning of the chapter) was using and our own, perhaps seeing him less as hopelessly paranoid and more as someone trying to do sensemaking in an unstable world in ways that are personally fulfilling. This does not mean that his views are correct—we need not admit any such thing; but when his views are wrong, we're less likely to declare them as being wrong in ways that are utterly different to our own vulnerabilities. When we take on a more capacious and democratic vision of what counts as valid sensemaking, we find ourselves in a better position to honor the legitimate feelings of alienation and disenfranchisement that we've seen disinformation campaigns try to exploit. This is how postcritique orients us: towards "a politics of relation rather than negation, of mediation rather than co-option, of alliance and assembly rather than alienated critique" (Felski 2015, 10). It is this orientation, I think, that will put us on the road to addressing disinformation in more human-centered ways.

7.5 LOOKING FORWARD

In this essay, I have taken a position on how disinformation campaigns invite us to engage with information (Research Question 3). Let me outline the schema of this position in five points. My position is that (1) sophisticated efforts to spread disinformation can encourage a stance of suspicion towards established information intermediaries. To support more accessible, sustainable and democractic responses to this suspicion, we need (2) to try and understand it in ways that are not only diagnostic but also relational. In other words, we need to make room for an understanding of this suspicion that sidesteps the image of a clinician scrutinizing his patient for signs of pathology. For this, I decenter the terminology of conspiracy theories by directing our attention towards (3) forms of reading and interpretation that can be associated with the methods of critique. This is where some of Bruno Latour's (2004; 2005; 2010) contributions, enlarged and developed recently by Felksi (2015), have been so crucial. I draw on the findings of the studies presented in chapters five and six to illustrate how critique is salient in the messages put forward by disinformation campaigns. And by delving into the lines of reasoning they promote, I highlight (4) some of the limitations of using critique as a response to counter disinformation. To help address some of these limitations, I offer (5) a design direction that is aligned with the sensibilities of postcritique, and briefly explain some of the ways it could provide a positive orientation in this domain for designers that are motivated by humanistic thought.

Naturally, there are places where my argument gets shaky. For instance, I did not address issues of scale while sketching out this postcritical design direction. This is not only because I don't have all the answers (nobody could, actually), but also because I think addressing such issues requires this direction to be fleshed out into a more substantial design space. Mapping out a space could help us see and compose connections to existing lines of work in research and practice. There are certain affinities, for example, between the direction I have presented and research on supporting reflection in CHI and CSCW (e.g. Cheng et al. 2011; Odom et al. 2014; Zhao, Ng, and Cosley 2012), as well ongoing efforts related to 'dialog journalism' (Morell 2018). Creating spaces where these lines of work could be gathered and placed into fruitful dialog with one another might raise new issues while helping us approach existing ones (like those related to scale) quite differently.

On a conceptual level, I've taken some liberties with Felski's (2015) ideas by grafting some of the properties of critique she has described onto the discourses I have been studying. Her work, after all, is concerned with ways of thinking while the data that I call upon in this essay includes media content that lacks humanistic depth. The wager is that any awkwardness inflicted by this move on my readers would be outweighed by the insights it helped me to bring to them. Caught up in the spirit of critique, I have struggled with the urge to deconstruct my own position for its reliance on digital traces. That is certainly a limitation shared with approaches trying to come to grips with disinformation by plumbing the depths of whatever data might be gathered via computational tools. But in the spirit of postcritique, I would like to recast this fault in my argument as a direction for future work: the insights I have constructed could be made richer and more satisfying by triangulating them through other data sources (e.g. interviews) and methods (e.g. participant observation). If we can improve the fit of humanistic perspectives with HCI's growing interest in taking on the problems posed by misleading information, then we can develop fresh design agendas and help keep the field more honest and reflective; and help all of us imagine better futures and forms of life worth living.

Such, anyway, is my hope.

7.6 References to Problematic or Misleading Content in this Chapter

For the purposes of this dissertation, I have chosen to separate academic references from references to content that I have problematized as misleading or potentially triggering. I have placed references to the latter here and to the former in the Works Cited section towards the end of the dissertation. This has been done to avoid driving traffic to it and blending it with more credible information.

21st Century Wire. 2012. “About 21st Century Wire.” August 12, 2012. Accessed December 18, 2020. https://21stcenturywire.com/about/.

———. 2016. “An Introduction: Smart Power & The Human Rights Industrial Complex.” *21st Century Wire*, April 19, 2016. Accessed December 18, 2020. https://21stcenturywire.com/2016/04/19/an-introduction-smart-power-the-human-rights-industrial-complex/.

———. 2017. “Victorious Syria: The New Dawn of Resistance Against Imperialism.” June 23, 2017. Accessed December 18, 2020. https://21stcenturywire.com/2017/06/23/victorious-syria-the-new-dawn-of-resistance-against-imperialism.

———. 2020. “Sunday Screening: ‘Manufacturing Consent' (1992)." January 26, 2020. Accessed December 18, 2020. https://21stcenturywire.com/2020/01/26/sunday-screening-manufacturing-consent-1992/.

———. n.d. “Fake News Week.” Accessed December 18, 2020. https://21stcenturywire.com/tag/fake-news/.

Bartlett, Eva. 2016. “Propaganda Alert: Madaya Media Fabrications, Recycled Photos.” *Dissident Voice* (blog), January 15, 2016. Accessed December 18, 2020. https://dissidentvoice.org/2016/01/propaganda-alert-madaya-media-fabrications-recycled-photos/.

Beeley, Vanessa. 2016. “Gaslighting: State Mind Control and Abusive Narcissism.” *21st Century Wire*, May 26, 2016. Accessed December 18, 2020. https://21stcenturywire.com/2016/05/26/gaslighting-state-mind-control-and-abusive-narcissism/.

Dyer, Jay. 2016. “Modern Education is Pavlovian Conditioning.” *21st Century Wire*, September 19, 2016. Accessed December 18, 2020. https://21stcenturywire.com/2016/09/19/modern-education-is-pavlovian-conditioning/.

Hayward, Tim. 2017a. “How to Weigh a Mountain of Evidence: Guest Blog by Professor Paul McKeigue (Part 1).” *Tim Hayward* (personal blog). August 11, 2017. Accessed December 18, 2020. https://timhayward.wordpress.com/2017/08/11/how-to-weigh-a-mountain-of-evidence-guest-blog-by-professor-paul-mckeigue-part-1/.

Hayward, Tim. 2017b. “Who is Responsible for Chemical Attacks in Syria? Guest Blog by Professor Paul McKeigue (Part 2).” *Tim Hayward* (personal blog). August 31, 2017. Accessed December 18, 2020. https://timhayward.wordpress.com/2017/08/31/who-is-responsible-for-chemical-attacks-in-syria-guest-blog-by-professor-paul-mckeigue-part-2/.

Pacheco, Marta. 2017. “Aleppo and the War on the Right to Know.” *Katoikos*, January 9, 2017. Accessed December 18, 2020. https://katoikos.world/analysis/aleppo-and-the-war-on-the-right-to-know.html.

Working Group on Syria, Propaganda and Media. n.d. “About.” Accessed December 18, 2020. https://syriapropagandamedia.org/about.

Chapter 8. CONCLUSION

Let me now pull together the various contributions of this dissertation and offer some conclusions. I will do this by summarizing my answers to the research questions that I set out to address and by explaining some selected implications of those answers.

To recap: in this dissertation, I examined some of the information activities that occur on social media in the wake of mass disruption events like disasters and large-scale protests. I studied these activities by analyzing social media data, alternative news websites and interviews. And I conceptualized them using the literature on sensemaking, extending it to consider how sensemaking can be disrupted and exploited in online settings. I specifically focused on the activities of: a) social media users who spread and corrected rumors as they engaged in sensemaking during two different crisis situations; and b) two online disinformation campaigns that opportunistically exploited such sensemaking efforts.

8.1 SUMMARY OF RESEARCH CONTRIBUTIONS

I have developed several knowledge-contributions by analyzing these activities. These contributions are intended to help researchers and designers who are invested in constructivist and human-centered approaches to technology (e.g. the HCI and CSCW communities), and who wish to better understand and address the spread of misleading information in online settings. These contributions help by: 1) making the structure and dynamics of misleading information more legible as it manifests and spreads across social media and online domains during mass-disruption events; and 2) potentially broadening our perspectives on misleading information and how we might address it.

I address the first goal by describing some: i) emergent practices for correcting misleading information; ii) entanglements between disinformation campaigns and other actors; and iii) rhetoric that is used by disinformation campaigns. I address the second goal by placing the grounded understanding I have developed around my empirical data into conversation with concepts drawn from the literature on sensemaking and postcritique. By bringing certain diffuse properties of misleading information into clearer focus (e.g. how disinformation campaigns promote suspicion), I enlarge our theoretical repertoire for thinking about the relationships between misleading information, people and technology, and furnish some potentially fresh directions for design and research in this domain. I will now turn to the research questions that this dissertation addresses in order to explain these contributions.

8.1.1 *Research Question 1: How do well-intentioned members of the online crowd understand their own actions when they unwittingly circulate misleading information on social media while using it for sensemaking?*

To respond to this question, I presented a study in Chapter 4 called *A Closer Look at the Self-Correcting Crowd*. In the study, I examined the actions taken by journalists and ordinary citizens in the online crowd to correct rumors they had helped spread during the 2015 Paris Attacks and a rumored plane hijacking. I analyzed social media data to identify five patterns of tweeting activity ("behavioral signatures") and interviewed individuals who exhibited these patterns to understand how their reasoning aligned with their activity. This helped me characterize six folk-theories that individuals held about how misinformation spreads over social media, and which colored their reasoning about their own rumoring behaviors.

More broadly, I highlighted that there are a multitude of design opportunities for building tools to support, leverage, and learn from this activity. I argued that promoting some of these behaviors could offer a human-centered strategy for coping with the fog of contingencies that make it difficult to proactively identify misleading information at scale. I discussed how we could support these behaviors by helping social media users express uncertainty, refine their folk-theories, and be more mindful and reflective. Foreshadowing the design direction I offer in chapter 7, these are ways of addressing the widespread dissemination of misleading information that focus on promoting positive human skills over protecting people from ‘bad’ information.

These contributions advance our understanding of how human agency and technical affordances mutually shape the spread of misleading information during periods of mass-disruption. They help illuminate a relatively understudied information activity (corrections), which can inform our perspective on how we might shape our sociotechnical systems to be more conducive to civic flourishing. They also help us pay more attention to the complex situational factors that influence people’s behaviors with regards to misleading information in online settings, shifting emphasis away from ‘bad’ people making ‘poor’ decisions to people struggling to make meaning in difficult circumstances. Collectively, these contributions can help researchers and designers be more constructive, self-aware and generous when they intervene in this area of inquiry.

8.1.2 *Research Question 2: How do state-affiliated actors opportunistically exploit these sensemaking efforts to spread disinformation?*

I presented two investigations to answer this question. In *Acting the Part* (chapter 5), I examined how a known Russian troll farm used Twitter and other platforms to influence a highly charged conversation in 2016 about police violence in America and the #BlackLivesMatter movement. This work showed that these actors did not limit themselves to a single 'side' of the online conversation, but rather cultivated personas (all but indistinguishable from other participants in these spaces) to appeal to the values of both right-leaning anti-BLM voices and left-leaning pro-BLM voices. In the second investigation, *Ecosystem or Echo-System* (chapter 6), I showed how the White Helmets, a volunteer humanitarian group working in the Syrian conflict zone, became the subject of a disinformation campaign that undermined the group's image by leveraging a diverse network of alternative media websites. This work illuminated how these websites — which are integrated with Russian government-funded media — draw together different audiences (with perhaps different values) into a shared story that the White Helmets are criminals and a propaganda construct supported by Western imperialists.

These investigations trace some of the ways in which disinformation campaigns manipulate our efforts to organize online on social media during mass-disruption events. They show how sophisticated campaigns can insert themselves into online activism and sensemaking efforts in ideologically fluid ways, crafting information channels and messages to micro-target different audiences. Actors affiliated with these campaigns are then effectively positioned to engage in related information activities like: opportunistically amplifying certain messages (that may or may not have been devised by them); encouraging audiences to do the same (making them unwitting collaborators); sharing stories that layer accurate information with inaccurate information; promoting interpretations that are not amenable to fact-checking but function to foster doubt[52] and division (and driving audiences to do further sensemaking). In short, disinformation campaigns exploit organizing efforts on social media by eagerly participating in them and these studies have supplied evidence that makes these dynamics more legible.

[52] This observation complicates an idea that I put forward in chapter 4 that was about helping people express uncertainty. If disinformation campaigns promote uncertainty, then there can be something self-sabotaging about giving users more tools to express it. We need to think carefully about the differences between helping people notice and express the existing uncertainty that they are feeling about some information during an ambiguous situation (the potentially positive activity that I highlighted in chapter 4) and actually adding more uncertainty to the information environment (which is a problematic activity). It is precisely the complicated work of putting feelings into words and *constructive* actions that we need to understand how to support (and that includes figuring out what those constructive actions might look like).

Towards the end of each of these studies, I also pulled forward certain insights to help undo some misconceptions about the spread of disinformation. For example, we often think disinformation is a matter of inaccurate or false information that can be addressed through fact-checking. But in *Acting the Part*, I showed how disinformation is often contextual and that campaigns can promote messages that are not necessarily problematic on the surface. Similarly, there is a widespread sentiment that disinformation is mainly propagated by bad actors like bots and trolls. But in *Ecosystem or Echo-System*, I showed how sophisticated disinformation campaigns can integrate with online activism in ways that are difficult to disentangle. One implication of these insights that I chose to focus on involved how these tactics work to obstruct and reduce the value of interventions that prioritize sorting and labeling information and information sources.

In the larger picture, this research has contributed to our understanding of the complicated relationships that emerge between disinformation campaigns and other members of the online crowd during mass-disruptions events. In some ways, it is an act of translation: it took up insights about disinformation that arose in offline contexts — and which primarily relied on the testimony of former intelligence professionals like Bittman — and contributed new evidence to support and help refresh these insights for our current information landscape. As an act of translation, this research has also fed other work by my colleagues and myself that invites the CSCW community to view the production and dissemination of disinformation as a collaborative activity. In these ways, the contributions of this research can support researchers and designers appreciate issues pertaining to the spread of disinformation with greater nuance.

8.1.3 *Research Question 3: How do disinformation campaigns invite their audiences to make sense of the information landscape through a lens of suspicion?*

I addressed this question by writing an essay (chapter 7). In this essay, I proposed that disinformation campaigns call upon the rhetoric and tools of critique to invite audiences to join them in questioning established evidence and knowledge authorities. To support my position, I provided an analysis of the rhetoric promoted by disinformation campaigns that combines some of the findings from my studies in chapters 5 and 6. This analysis drew on postcritical theory, which offered a foundational point for my efforts to better understand certain analytical gestures and sentiments of suspicion that I had observed in my data during my investigations. By synthesizing my empirical findings and placing them in conversation with extant theory, I helped to clarify how disinformation campaigns entice audiences into their sphere of influence, disrupt established understandings, and advocate for new ones.

I also drew forward some implications for how we might respond. For example, I highlighted some of the limitations of addressing these campaigns by designing more tools that help us with the work of critiquing information. After tracing some of the destructive impacts of the hermeneutics of suspicion (critique as described by Rita Felski [2015]), I turned to the reconstructive task of exploring responses to disinformation not rooted in suspicion. I sketched out a design direction aligned with the values of postcritique and outlined how it might undermine the logics of disinformation. The ideas I offered, while nascent and tender, help establish the theoretical underpinnings of a potentially more human-centered approach to address disinformation.

By addressing this question, this research has contributed new resources to help us think about how disinformation campaigns foster doubt and division, and how we might structure our responses. It has also provided further support and clarification on how certain styles of thinking that are promoted within academia can be manipulated, a concern that has been raised by several intellectuals (boyd 2017a; Rid 2020; Pomerantsev and Weiss 2014). By highlighting some of these tensions and portraying how the act of critiquing powerful actors can be strategically perverted, this research can potentially help others conduct more socially responsible research and design work in this emerging area. As an exercise in interdisciplinary research, it has connected a theory from the humanities with HCI research to raise new questions and provide a more holistic view on the dynamics of disinformation. In some ways, it has also tried to give back to these ideas by providing further evidence of the need to design alternatives to the standard operations of critique, and by calling upon the HCI community to contribute to this endeavor. In that sense, it has taken a small part in the work of building the collaborations that we will need to build a more just and informed society.

8.2 FINAL THOUGHTS

Let me end with a personal statement to expand on some of the potential directions this inquiry has brought forward. In developing the analysis I have presented here, I have been opening up questions not only about what we need to be doing to bring about healthier information environments, but also about the ‘we’ that can be involved in this doing. I’ve gone about this by highlighting different activities that could be promoted to address the toxicities of our information environment and by trying to make room for more diverse theoretical orientations (e.g. mindfulness and postcritical thinking).

I have done this partly out of a desire to be generative and inclusive. This is important because I believe that the issues I have studied will require concurrently maneuvering around technology, policy, and education in ways that are not just about reaffirming existing arrangements, but also expanding and reimagining them. Fortunately, the area of mis- and disinformation studies is emergent, rapidly evolving and holds promise to welcome such interdisciplinary dialogues. That is the spirit in which I have dispensed my implications, and it is in that sense that I think a certain type of collective work is unfolding.

At the same time, it is important not to overlook the politics of this collective work. As scholars like Ferguson (1990) remind us, ‘we’ do not all share the same interests, and to pretend otherwise risks bringing out exactly the sort of utopian thinking that can disguise highly partial and interested interventions as disinterested, universal or inherently benevolent. For instance, whatever the accomplishments of large technology corporations or government agencies, little about their general mode of conduct would justify envisioning a collective ‘we’ that’s working towards human growth and flourishing.

So, in a more accurate sense, this dissertation has been about what should *we* scholars and intellectuals invested in human-centered design do about misleading information? To the extent that there are shared values and commitments, this becomes an intelligible question and a “we” becomes real. My experiences suggest that many design-oriented students, practitioners and researchers who are concerned about misleading information today do broadly share certain values. They value being participatory and egalitarian. They tend to have democratic commitments and seek to advance those by taking up projects of empowerment and social transformation. For those intellectuals who share in these commitments and who wish to help shape a healthier information environment, the question of ‘what should we do’ is indeed a real one. But answers to this question must entail, even if only implicitly, a theory of the nature of this reparative work and which actors are empowered or disempowered in the course of doing it.

For anyone who shares in the values I have been discussing, making critique and fact-checking the primary form of their endeavors would seem to signal a view that healthier environments come about by detecting and bracketing off information that is ‘bad’. No doubt, there are ample situations where such approaches are effective. But, as this dissertation has shown, there are also limitations to this way of theorizing the solution space and making it the only way we design for misleading information.

Such approaches are arguably more information-centered than human-centered — they risk overlooking the messy, yet very real attachments humans bring to their engagements with information. They also seem to imply that democracy and empowerment are to be worked for and brought about by the benevolent intervention of those actors who can harness and control large volumes of information (e.g. social media companies). Again, there may be specific contexts where that happens. But experience suggests that identifying the interventions of such actors with progress and reform can be circular (consider how social media was once positioned as a technology to aid in the righteous pursuit of liberty). It can also facilitate the dismissal or even suppression of oppositional views (consider how social media companies are coming under fire for repressing dissent at the behest of governments [BBC News 2020]). Hoping to prevent others from being misled, it can become rather easy to enter into complicity with organizations that, in all but the most extraordinary of situations, serve the interests of local or global hegemony.

If, as I have suggested, 'critique' based interventions are not the only way for human-centered designers to address misleading information, then what are some additional directions we might pursue? An important direction involves researching and designing more human-centered information infrastructures to address misleading information. By this, I mean, designing for practices and ways of thinking that can help us reflect on our attachments and engage with them across moral and epistemic divides. This direction does not occupy the same space as 'critique' based interventions; it is not meant to challenge or replace them. What it does perhaps offer is a form of channeling one's intellectual energy towards the work of addressing misleading information in a manner that is consistent with the commitments I've described above.

Open questions remain. We might, for example, consider how specific social media practices hinder or promote the development of ethical traits (such as honesty, patience, generosity etc.) that people will need to use and shape social media to reliably and effectively address mis- and disinformation. Focusing on these practices could serve as a ground for developing better empirical methods to study the actual impact they have on the spread of misleading information. Further, there are a multitude of design opportunities to support these practices (e.g. through seamful design, designing to promote reflective self-examination). At first, exploring such directions might not be straightforward. But it is possible to imagine a network of human-centered designers and researchers forging the links necessary to take up this work.

REFERENCES

Acar, Adam, and Yuya Muraki. 2011. "Twitter for crisis communication: lessons learned from Japan's tsunami disaster." *International Journal of Web Based Communities* 7, no. 3: 392-402.

Ackerman, Mark S. 2000. "The Intellectual Challenge of CSCW: The Gap between Social Requirements and Technical Feasibility." *Human–Computer Interaction* 15, no. 2-3: 179-203. https://doi.org/10.1207/S15327051HCI1523_5.

Allport, Gordon W. 1985. "The Historical Background of Social Psychology." In *The Handbook of Social Psychology*, edited by Gardner Lindzey and Elliot Aronson, 1-46. New York: Random House.

Allport, Gordon W., and Leo Postman. 1946. "An Analysis of Rumor." *Public Opinion Quarterly* 10, no. 4: 501-17.

———. 1947. *The Psychology of Rumor*. New York: Henry Holt.

Ambrosio, Thomas. 2007. "Insulating Russia from a Colour Revolution: How the Kremlin Resists Regional Democratic Trends." *Democratization* 14, no. 2: 232-52.

Anderson, Monica, and Paul Hitlin. 2016. *Social Media Conversations About Race: How Social Media Users See, Share and Discuss Race and the Rise of Hashtags Like #BlackLivesMatter*. Pew Research Center. https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2016/08/PI_2016.08.15_Race-and-Social-Media_FINAL.pdf.

Andrews, Cynthia, Elodie Fichet, Yuwei Ding, Emma S. Spiro, and Kate Starbird. 2016. "Keeping Up with the Tweet-Dashians: The Impact of 'Official' Accounts on Online Rumoring". In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 452–65. https://doi.org/10.1145/2818048.2819986.

Anker, Elizabeth S. 2017. "Postcritique and Social Justice." *American Book Review* 38, no. 5: 9-10.

Anker, Elizabeth S., and Rita Felski, eds. 2017. *Critique and Postcritique.* Durham: Duke University Press.

Arif, Ahmer, John J. Robinson, Stephanie A. Stanek, Elodie S Fichet, Paul Townsend, Zena Worku, and Kate Starbird. 2017. "A Closer Look at the Self-Correcting Crowd: Examining Corrections in Online Rumors". In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 155–68. https://doi.org/10.1145/2998181.2998294.

Arif, Ahmer, Kelley Shanahan, Fang-Ju Chou, Yoanna Dosouto, Kate Starbird, and Emma S. Spiro. 2016. "How Information Snowballs: Exploring the Role of Exposure in Online Rumor Propagation". In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 466-477. https://doi.org/10.1145/2818048.2819964.

Arif, Ahmer, Leo Graiden Stewart, and Kate Starbird. 2018. "Acting the Part: Examining Information Operations within #BlackLivesMatter Discourse." *Proceedings of the ACM on Human-Computer Interaction* 2, No. CSCW: 1-27. https://doi.org/10.1145/3274289.

Armistead, Leigh. 2004. *Information Operations: Warfare and the Hard Reality of Soft Power*. Lincoln, NE: Potomac Books, Inc.

Asmolov, Gregory. 2018. “The Disconnective Power of Disinformation Campaigns.” *Journal of International Affairs* 71, no. 1.5: 69–76. http://www.jstor.org/stable/26508120.

Associated Press. 2013. “Russia to Keep Helping Syria If It’s Attacked.” September 6, 2013. https://apnews.com/article/bf91ce0b8b3f4302af884d459654bf9c.

Austin, Jonathan Luke, Rocco Bellanova, and Mareile Kaufmann. 2018. *Doing and Mediating Critique: An Invitation to Practice Companionship*. New York: SAGE Publications.

Barbezat, Daniel P., and Mirabai Bush. 2013. *Contemplative Practices in Higher Education: Powerful Methods to Transform Teaching and Learning*. Hoboken, NJ: John Wiley & Sons.

Bardzell, Jeffrey, and Shaowen Bardzell. 2015. *Humanistic HCI.* San Rafael, CA: Morgan & Claypool Publishers.

Bardzell, Shaowen, and Jeffrey Bardzell. 2011. "Towards a feminist HCI methodology: social science, feminism, and HCI." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 675-684. https://doi.org/10.1145/1978942.1979041.

Barnes, Luke. 2018. "Parkland Shooting Survivors are being Smeared as 'Crisis Actors' and it's Going Viral." *Think Progress,* February 20, 2018. https://www.thinkprogress.org/parkland-conspiracies-going-viral-092288a904b6/.

Barthel, Michael, and Amy Mitchell. 2017. *Americans' Attitudes about the News Media Deeply Divided along Partisan Lines*. Pew Research Center. http://www.journalism.org/2017/05/10/americans-attitudes-about-the-news-media-deeply-divided-along-partisan-lines/.

Barthel, Michael, Amy Mitchell, and Jesse Holcomb. 2016. *Many Americans Believe Fake News Is Sowing Confusion*. Pew Research Center. http://www.journalism.org/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion/.

Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. 2009. "Gephi: an open source software for exploring and manipulating networks." In *Proceedings of the Third AAAI International Conference on Web and Social Media* 8: 361-362. http://aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

BBC News. 2016. "Syria: The Story of the Conflict." March 11, 2016. https://www.bbc.com/news/world-middle-east-26116868.

———. 2018. "Syria War: The Online Activists Pushing Conspiracy Theories." April 18, 2018. https://www.bbc.com/news/blogs-trending-43745629.

———. 2020. "Vietnam: Facebook and Google 'complicit' in censorship." December 1, 2020. https://www.bbc.com/news/world-asia-55140857

Benford, Robert D. 1993. "Frame Disputes within the Nuclear Disarmament Movement." *Social Forces* 71, no. 3: 677–701.

Bergstrom, Carl T., and Jevin D. West. 2020. *Calling Bullshit: The Art of Skepticism in a Data-Driven World*. New York: Random House.

Bernstein, Michael S., Eytan Bakshy, Moira Burke, and Brian Karrer. 2013. "Quantifying the Invisible Audience in Social Networks". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 21–30. https://doi.org/10.1145/2470654.2470658.

Bhattacharya, Parantapa, and Niloy Ganguly. 2016. "Characterizing Deleted Tweets and Their Authors." In *Proceedings of the Tenth International AAAI Conference on Web and Social Media*, 547–50. http://aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/viewPaper/13133.

Bittman, Ladislav. (1972) 1981. *The Deception Game.* Syracuse, NY: Syracuse University Research Corporation. Reprint, New York: Ballantine Books.

———. (1983) 1985. *The KGB and Soviet Disinformation: An Insider's View*. Reprint of the first edition with a foreword by Roy Godson. Oxford, UK: Pergamon-Brassey's. Citations refer to the second ed.

Black Lives Matter. n.d. "Herstory." Accessed April 4, 2018. https://blacklivesmatter.com/about/herstory/.

Black Lives Matter Vermont. 2017. "FAQ." Accessed September 15, 2018. http://blacklivesmattervermont.com/wp-content/uploads/2017/01/FAQ.pdf.

Blattner, William. 2006. *Heidegger's 'Being and Time': A Reader's Guide*. London: A & C Black.

Blomberg, Jeanette, Mark Burrell, and Greg Guest. 2009. "An Ethnographic Approach to Design." In *Human-Computer Interaction*, edited by Andrew Sears and Julie A. Jacko, 71–94. Boca Raton, FL: CRC Press.

Blomberg, Jeanette, and Helena Karasti. 2013. "Reflections on 25 Years of Ethnography in CSCW." *Computer Supported Cooperative Work: CSCW: An International Journal* 22, no. 4–6. https://doi.org/10.1007/s10606-012-9183-1.

Blue Lives Matter. 2017. "About Us - Blue Lives Matter." Accessed April 4, 2018. http://bluelivesmatter.blue/organization/.

Bodhi, Bhikkhu. 2000. *A Comprehensive Manual of Abhidhamma: The Philosophical Psychology of Buddhism*. Kandy, Sri Lanka: Buddhist Publication Society.

Boje, David M. 1991. "The Storytelling Organization: A Study of Story Performance in an Office-Supply Firm." *Administrative Science Quarterly*: 106-26.

Booten, Kyle. 2016. “Hashtag Drift: Tracing the Evolving Uses of Political Hashtags over Time.” In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2401–05. https://doi.org/10.1145/2858036.2858398.

Bordia, Prashant, and Nicholas DiFonzo. 2004. “Problem Solving in Social Interactions on the Internet: Rumor as Social Cognition.” *Social Psychology Quarterly* 67, no. 1: 33–49.

boyd, danah. 2008. “Why Youth ♥ Social Network Sites: The Role of Networked Publics in Teenage Social Life." In *Youth, Identity, and Digital Media*, edited by David Buckingham, 119–142. The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning. Cambridge, MA: MIT Press. http://ssrn.com/abstract=1518924.

———. 2017a. “Did Media Literacy Backfire?” *Journal of Applied Youth Studies* 1, no. 4: 83–89.

———. 2017b. “Google and Facebook can’t just make fake news disappear.” *Points* (blog), *Data & Society*. March 27, 2017. https://points.datasociety.net/google-and-facebook-cant-just-make-fake-news-disappear-48f4b4e5fbe8.

———. 2018. "You Think You Want Media Literacy… Do You?" *Points* (blog), *Data & Society*. March 9, 2018. https://points.datasociety.net/you-think-you-want-medialiteracy-do-you-7cad6af18ec2.

———. n.d. “What’s in a name?” http://www.danah.org/name.html. Accessed December 15, 2020.

boyd, danah, and Kate Crawford. 2012. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication & Society* 15, no. 5: 662-79.

boyd, danah, Scott Golder, and Gilad Lotan. 2010. "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter". In *Proceedings of the 43rd Hawaii*

International Conference on System Sciences, 1–10. http://doi.org/10.1109/HICSS.2010.412.

Bratich, Jack Z. 2008. *Conspiracy Panics: Political Rationality and Popular Culture*. Albany, NY: SUNY Press.

Bruns, Axel, Jean Burgess, Kate Crawford, and Frances Shaw. 2012. *#qldfloods and @QPSMedia: Crisis Communication on Twitter in the 2011 South East Queensland Floods.* ARC Centre of Excellence for Creative Industries & Innovation. http://eprints.qut.edu.au/48241/1/floodsreport.pdf.

Burns, Alex, and Ben Eltham. 2009. "Twitter Free Iran: An Evaluation of Twitter's Role in Public Diplomacy and Information Operations in Iran's 2009 Election Crisis." In *Communications Policy & Research Forum,* 298-310. http://networkinsight.org/verve/_resources/Burns_Eltham_file.pdf

Burns, Ryan. 2015. "Rethinking Big Data in Digital Humanitarianism: Practices, Epistemologies, and Social Relations." *GeoJournal* 80, no. 4: 477–90.

Bytwerk, Randall L. 2010. "Grassroots Propaganda in the Third Reich: The Reich Ring for National Socialist Propaganda and Public Enlightenment." *German Studies Review* 33, no. 1: 93-118. www.jstor.org/stable/40574929.

Cassa, Christopher A., Rumi Chunara, Kenneth Mandl, and John S Brownstein. 2013. "Twitter as a Sentinel in Emergency Situations: Lessons from the Boston Marathon Explosions." *PLoS Currents* 5. http://doi.org/10.1371/currents.dis.ad70cd1c8bc585e9470046cde334ee4b.

Castillo, Carlos, Marcelo Mendoza, and Barbara Poblete. 2011. "Information Credibility on Twitter." In *Proceedings of the 20th International Conference on World Wide Web,* 675–84. https://doi.org/10.1145/1963405.1963500.

Caulfield, Michael. 2018. "Media Literacy Is About Where to Spend Your Trust. But You Have to Spend It Somewhere." *Hapgood* (blog). February 23, 2018. http://hapgood.us/2018/02/23/media-literacy-is-about-where-to-spend-your-trust-but-you-have-to-spend-it-somewhere/. Accessed October 19 2020.

Chalmers, Matthew. 2003. "Seamful Design and Ubicomp Infrastructure". In *Proceedings of Ubicomp 2003 Workshop at the Crossroads: The interaction of HCI and systems issues in Ubicomp*, 577-84.

Charmaz, Kathy. (2006) 2014. *Constructing Grounded Theory*. New York: SAGE Publications. Revised 2nd ed, New York: SAGE Publications. Citations refer to the second ed.

Chen, Adrian. 2015. "The Agency." *The New York Times Magazine*, June 2, 2015. https://www.nytimes.com/2015/06/07/magazine/the-agency.html.

Cheng, Justin, Akshay Bapat, Gregory Thomas, Kevin Tse, Nikhil Nawathe, Jeremy Crockett, and Gilly Leshed. 2011. "GoSlow: Designing for Slowness, Reflection and Solitude." Abstract. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, 2011, 429-438. *https://*doi.org/10.1145/1979742.1979622.

Choo, Chun Wei. 1996. "The Knowing Organization: How Organizations Use Information to Construct Meaning, Create Knowledge and Make Decisions." *International Journal of Information Management* 16, no. 5: 329–40.

Cook, Richard I., and David D. Woods. 1994. "Operating at the Sharp End: The Complexity of Human Error." *Human Error in Medicine* 13: 225–310.

Corley, Kevin G., and Dennis A. Gioia. 2004. "Identity Ambiguity and Change in the Wake of a Corporate Spin-Off." *Administrative Science Quarterly* 49, no. 2: 173-208.

Corvey, William J., Sarah Vieweg, Travis Rood, and Martha Palmer. 2010. "Twitter in Mass Emergency: What NLP Can Contribute." In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, 23-24. https://www.aclweb.org/anthology/W10-0512.pdf

Couldry, Nick. 2013. "Why Media Ethics Still Matters." In *Global Media Ethics: Problems and Perspectives,* edited by Stephen J. A. Ward, 13–28. Hoboken, NJ: Wiley-Blackwell.

Couldry, Nick, and Ulises A. Mejias. 2019. "Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject." *Television & New Media* 20, no. 4: 336-49.

Crawford, Kate, and Megan Finn. 2015. "The Limits of Crisis Data: Analytical and Ethical Challenges of using Social and Mobile Data to Understand Disasters." *GeoJournal* 80, no. 4: 491-502.

CSCW 2021. n.d. "ACM SIGCHI CSCW, 2021". Accessed October 19, 2020. https://cscw.acm.org/2021.

Currie, Graeme, and Andrew D. Brown. 2003. “A Narratological Approach to Understanding Processes of Organizing in a UK Hospital.” *Human Relations* 56, no. 5: 563–86.

Dailey, Dharma. 2020. “*Social Media as Local Crisis Infrastructure: The Interconnected Work of Citizens, Responders, and Journalists in the Social Media Crowd*.” PhD diss., University of Washington.

Dailey, Dharma, and Kate Starbird. 2017. “Social Media Seamsters: Stitching Platforms & Audiences into Local Crisis Infrastructure.” In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1277-1289. http://doi.org/10.1145/2998181.2998290.

Dearden, Lizzie. 2017. “NATO Accuses Sputnik News of Distributing Misinformation as Part of 'Kremlin Propaganda Machine'.” *The Independent,* February 11 2017. https://www.independent.co.uk/news/world/europe/sputnik-news-russian-government-owned-controlled-nato-accuses-kremlin-propaganda-machine-disinformation-syria-brexit-refugee-crisis-a7574721.html.

Dervin, Brenda. 1983. “An Overview of Sense-Making Research: Concepts, Methods and Results to Date.” Paper presented at the *International Communications Association Annual Meeting*, May 1983. http://faculty.washington.edu/wpratt/MEBI598/Methods/An%20Overview%20of%20Sense-Making%20Research%201983a.htm

Dewey, John. (1922) 2012. *Human Nature and Conduct*. New York: Henry Holt and Company. eBook edition, Project Gutenberg. Citations refer to the eBook ed. http://www.gutenberg.org/files/41386/41386-h/41386-h.htm.

DFRLab. 2018. “Question That: RT’s Military Mission. Assessing Russia Today’s Role as an ‘Information Weapon’.” *Digital Forensic Research Lab (DFRLab),* January 7, 2018. https://medium.com/dfrlab/question-that-rts-military-mission-4c4bd9f72c88.

Diamond, Larry. 2010. “Liberation Technology.” *Journal of Democracy* 21, no. 3: 69–83. https://doi.org/10.1353/jod.0.0190.

DiFonzo, Nicholas, Prashant Bordia, and Ralph L. Rosnow. 1994. "Reining in rumors." *Organizational Dynamics* 23, no. 1: 47-62. https://doi.org/10.1016/0090-2616(94)90087-6.

Dijk, Teun A. Van. 2015. *Racism and the Press*. Abingdon, UK: Routledge.

DiResta, Renee, Kris Shaffer, Becky Ruppel, David Sullivan, Robert Matney, Ryan Fox, Jonathan Albright, and Ben Johnson. 2019. "The Tactics & Tropes of the Internet Research Agency." *New Knowledge*. https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1003&context=senatedocs.

Donovan, Joan, and Brian Friedberg. 2019. “Source Hacking: Media Manipulation in Practice.” *Data & Society*, September 4, 2019. https://datasociety.net/library/source-hacking-media-manipulation-in-practice/.

Dourish, Paul. 2006. “Implications for Design.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 541–550. https://doi.org/10.1145/1124772.1124855.

Downing, John D. H., and Charles Husband. 2005. *Representing Race: Racisms, Ethnicity and the Media*. New York: SAGE Publications.

Drabek, Thomas E. 1970. “Methodology of Studying Disasters: Past Patterns and Future Possibilities.” American Behavioral Scientist 13, no. 3: 331–43.

Drazin, Robert, Mary Ann Glynn, and Robert K. Kazanjian. 1999. “Multilevel Theorizing about Creativity in Organizations: A Sensemaking Perspective.” *Academy of Management Review* 24, no. 2: 286–307.

Dretske, Fred. 2008. “Epistemology and Information.” *Philosophy of Information* 8, https://resources.illc.uva.nl/HPI/Draft_Epistemology_and_Information.pdf

Dynes, Russell Rowe. 1970. *Organized Behavior in Disaster*. Lexington, MA: Heath Lexington Books.

Edler, Daniel, and Martin Rosvall. n.d. "The MapEquation Software Package." Accessed December 14, 2020. https://www.mapequation.org.

Ellis, Emma Grey. 2017. “Inside the Conspiracy which Turned Syria’s First Responders into Terrorists.” *Wired,* April 30, 2017. https://www.wired.com/2017/04/white-helmets-conspiracy-theory/.

Eslami, Motahhare, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. "First I ‘like’ it, then I hide it:

Folk Theories of Social Feeds." In *Proceedings of the 2016 CHI conference on Human Factors in Computing Systems*, 2371-2382. https://doi.org/10.1145/2858036.2858494.

Evans, James, and Andrew Karvonen. 2011. “Living Laboratories for Sustainability: Exploring the Politics and Epistemology of Urban Transition.” In *Cities and Low Carbon Transitions,* edited by Harriet Bulkeley, Simon Marvin, Vanesa Castan Broto and Mike Hodson, 126–41. Abingdon, UK: Routledge.

Fallis, Don. 2015. “What Is Disinformation?” *Library Trends* 63, no. 3: 401–26.

Faris, Robert, Hal Roberts, Bruce Etling, Nikki Bourassa, Ethan Zuckerman, and Yochai Benkler. 2017. "Partisanship, propaganda, and disinformation: Online media and the 2016 US presidential election." *Berkman Klein Center for Internet & Society Research Paper.* http://nrs.harvard.edu/urn-3:HUL.InstRepos:33759251.

Farkas, Johan, Jannick Schou, and Christina Neumayer. 2018a. “Cloaked Facebook Pages: Exploring Fake Islamist Propaganda in Social Media.” *New Media & Society* 20, no. 5: 1850–67.

———. 2018b. “Platformed Antagonism: Racist Discourses on Fake Muslim Facebook Pages.” *Critical Discourse Studies* 15, no. 5: 463–80.

Felski, Rita. 2015. *The Limits of Critique*. Chicago: University of Chicago Press.

———. 2020. “Postcritical.” In *Further Reading,* edited by Matthew Rubery and Leah Price. eBook version. Oxford: Oxford University Press. http://doi.org/10.1093/oxfordhb/9780198809791.013.11.

Fenby, Jonathan. 1986. *The International News Services*. New York: Schocken Books.

Ferguson, James. (1990) 1994. *The Anti-Politics Machine: Development, Depoliticization, and Bureaucratic Power in Lesotho*. Cambridge, UK: Cambridge University Press. New edition, Minneapolis, MN: University Of Minnesota Press.

Festinger, Leon. 1957. *A Theory of Cognitive Dissonance*. Palo Alto: Stanford University Press.

Fichet, Elodie S., John Robinson, Dharma Dailey, and Kate Starbird. 2016. “Eyes on the Ground: Emerging Practices in Periscope Use during Crisis Events.” In

Proceedings of the International Conference on Information Systems for Crisis Response and Management, 1–10.

Finn, Megan. 2018. *Documenting Aftermath: Information Infrastructures in the Wake of Disasters*. Cambridge, MA: MIT Press.

Fish, Stanley Eugene. 1980. *Is There a Text in This Class?: The Authority of Interpretive Communities*. Cambridge, MA: Harvard University Press.

Fiske, John. 1986. "Television: Polysemy and Popularity." *Critical Studies in Media Communication* 3, no. 4: 391–408.

Fleishman, Kathryn. 2019. "The Statue and the Veil: Postcritique in the Age of Trump." *Post45,* January 16 2019. https://post45.org/2019/01/the-statue-and-the-veil-postcritique-in-the-age-of-trump/. Accessed December 16 2020.

Fouad, Fouad M., Annie Sparrow, Ahmad Tarakji, Mohamad Alameddine, Fadi El-Jardali, Adam P. Coutts, Nour El Arnaout, Lama Bou Karroum, Mohammed Jawad, Sophie Roborgh, Aula Abbara, Fadi Alhalabi, Ibrahim AlMasri, and Samer Jabbour. 2017. "Health workers and the weaponisation of health care in Syria: a preliminary inquiry for The Lancet–American University of Beirut Commission on Syria." *The Lancet* 390, no. 1011: 2516-2526. https://doi.org/10.1016/S0140-6736(17)30741-9.

Fox, Christopher. 1983. *Information and Misinformation. An Investigation of the Notions of Information, Misinformation, Informing, and Misinforming.* Westport, CT: Greenwood Publishing Group.

Fox News. 2016. "Chicago Police to Take Second Look at Deadly Shooting of Teen with Antique Gun.", June 28, 2016. http://www.foxnews.com/us/chicago-police-to-take-second-look-at-deadly-shooting-of-teen-with-antique-gun.

Frere-Jones, Sasha. 2012. "Good Things About Twitter." *The New Yorker,* March 21, 2012. https://www.newyorker.com/culture/sasha-frere-jones/good-things-about-twitter.

Friggeri, Adrien, Lada A. Adamic, Dean Eckles, and Justin Cheng. 2014. "Rumor Cascades." In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 101–110, http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8122.

Fritz, Charles E. (1961) 1965. "Disaster." In *Contemporary Social Problems: An Introduction to the Sociology of Deviant Behavior and Social Disorganization*, edited by Robert K, Merton and Robert A. Nisbet, 651-94. Reprint, London: Rupert Hart-Davis Ltd.

Fritz, Charles E., and John H. Mathewson. 1957. “Convergence Behavior in Disasters: A Problem in Social Control.” Special Report. National Academy of Sciences-National Research Council 476.

Fuller, Jack. 2010. *What is Happening to News: The Information Explosion and the Crisis in Journalism*. Chicago: University of Chicago Press.

Gadde, Vijaya, and Yoel Roth. 2018. “Enabling Further Research of Information Operations on Twitter.” *Twitter Blog*, October 17, 2018. https://blog.twitter.com/official/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter.html. Accessed December 15, 2020.

Gao, Huiji, Geoffrey Barbier, and Rebecca Goolsby. 2011. “Harnessing the Crowdsourcing Power of Social Media for Disaster Relief.” *IEEE Intelligent Systems* 26, no. 3: 10–14. http://doi.org/10.1109/mis.2011.52.

Geiger, Stuart R., and David Ribes. 2011. “Trace Ethnography: Following Coordination through Documentary Practices.” In *Proceedings of the 2011 44th Hawaii International Conference on System Sciences*, IEEE, 1–10.

Gellately, Robert. 2002. *Backing Hitler: Consent and Coercion in Nazi Germany.* Oxford: Oxford University Press.

Gentzkow, Matthew, and Jesse M. Shapiro. 2011. “Ideological Segregation Online and Offline.” *The Quarterly Journal of Economics* 126, no. 4: 1799–839.

Giddens, Anthony. 1984. *The Constitution of Society: Outline of the Theory of Structuration*. Berkeley: University of California Press.

Gieryn, Thomas F. 1983. "Boundary-Work and the Demarcation of Science from Non-Science: Strains and Interests in Professional Ideologies of Scientists." *American Sociological Review*: 781-95.

Gillmor, Dan. (2004) 2006. *We the Media: Grassroots Journalism by the People, for the People*. Sebastopol, CA: O’Reilly Media.

Gioia, Dennis A., and Kumar Chittipeddi. 1991. "Sensemaking and Sensegiving in Strategic Change Initiation." *Strategic Management Journal* 12, no. 6: 433-48.

Glaser, Barney G. 1998. *Doing Grounded Theory: Issues and Discussions*. Mill Valley, CA: Sociology Press.

Glaser, Barney G., and Anselm L. Strauss. (1967) 2017. *Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine Atherton. New edition, Abingdon, UK: Routledge. Citations refer to the Routledge edition.

Goebbels, Joseph. 1933. “The Radio as the Eight Great Power.” Translated by Randall Bytwerk. In *German Propaganda Archive*. http://durenberger.com/wp-content/uploads/2018/08/GOEBBELS.pdf. Accessed December 15, 2020.

Gordon, Michael R. 2014. “US Strikes Follow Plea by Syrian Opposition Leader on Behalf of Kurds.” *The New York Times*, September 22, 2014. https://www.nytimes.com/2014/09/23/world/middleeast/syrian-opposition-leader-calls-on-us-to-strike-militants-from-air.html.

Grevet, Catherine, Loren G. Terveen, and Eric Gilbert. 2014. “Managing Political Differences in Social Media.” In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 1400–08. https://doi.org/10.1145/2531602.2531676.

Griffin, Drew, and Donie O'Sullivan. 2017. “The Fake Tea Party Twitter Account Linked to Russia and Followed by Sebastian Gorka.” *CNN*, September 22, 2017. https://www.cnn.com/2017/09/21/politics/tpartynews-twitter-russia-link/index.html. Accessed 19 Oct. 2020.

Guynn, Jessica. 2018. "Meet the Woman Who Coined #BlackLivesMatter." *USA TODAY,* April 4, 2018. https://www.usatoday.com/story/tech/2015/03/04/alicia-garza-black-lives-matter/24341593/.

Hagar, Christine, and Caroline Haythornthwaite. 2005. "Crisis, Farming and Community." *Journal of Community Informatics* 3: 41.

Haidt, Jonathan. 2012. *The Righteous Mind: Why Good People are Divided by Politics and Religion*. New York: Vintage Books.

Han, Rongbin. 2015. "Manufacturing Consent in Cyberspace: China's ‘Fifty-Cent Army’." *Journal of Current Chinese Affairs* 44, no. 2: 105-34.

Haraway, Donna. 1988. “Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective.” *Feminist Studies* 14, no. 3: 575–599. http://www.jstor.org/stable/3178066.

———. (1997) 2018. *Modest_Witness@Second_Millennium.FemaleMan_Meets_OncoMouse: Feminism and Technoscience*. Baltimore, MD: John Hopkins University Press. New edition, with an introduction by Thyrza Nicholas Goodeve, Abingdon, UK: Routledge. Citations refer to the Routledge edition.

———. 2016. *Staying With the Trouble: Making Kin in the Chthulucene.* Durham, NC: Duke University Press.

Harding, Sandra G. 1987. *Feminism and Methodology: Social Science Issues*. Bloomington, IN: Indiana University Press.

Hart, Tobin. 2004. "Opening the Contemplative Mind in the Classroom." *Journal of Transformative Education* 2, no. 1: 28-46.

Haynes, Deborah. 2019. “'Highly likely' GRU hacked UK institute countering Russian fake news.” *Sky News*, March 6, 2019. https://news.sky.com/story/highly-likely-moscow-hacked-uk-agency-countering-russian-disinformation-11656539.

Heidegger, Martin. 1996. *Being and Time: A Translation of Sein Und Zeit*. Translated by Joan Stambaugh. Albany, NY: SUNY Press.

Henwood, Karen, and Nick Pidgeon. 2003. “Grounded Theory in Psychological Research.” In *Qualitative Research in Psychology: Expanding Perspectives in Methodology and Design,* edited by P. M. Camic, J. E. Rhodes, & L. Yardley, 131-155. Washington, DC: American Psychological Association. https://doi.org/10.1037/10595-008.

Herman, Edward S, and Noam Chomsky. 1988. *Manufacturing Consent: The Political Economy of the Mass Media*. New York: Pantheon Books.

Herrman, John. 2012. “Twitter Is A Truth Machine.” *BuzzFeed News*, October 30, 2012. https://www.buzzfeednews.com/article/jwherrman/twitter-is-a-truth-machine.

Heverin, Thomas, and Lisl Zach. 2012. “Use of Microblogging for Collective Sense-making during Violent Crises: A Study of Three Campus Shootings.” *Journal of the American Society for Information Science and Technology* 63, no. 1: 34–47.

Hiltz, Starr Roxanne, Jane A. Kushma, and Linda Plotnick. 2014. “Use of Social Media by US Public Sector Emergency Managers: Barriers and Wish Lists.” In *Proceedings of the 11th International Conference on Information Systems for Crisis Response and Management*. https://10.13140/2.1.3122.4005.

Hofstadter, Richard. (1952) 2012. *The Paranoid Style in American Politics*. Reprint, New York: Vintage Books.

Howard, Philip N. 2002. “Network Ethnography and the Hypermedia Organization: New Media, New Organizations, New Methods.” *New Media & Society* 4, no. 4: 550–74.

Huang, Y. Linlin, Kate Starbird, Mania Orand, Stephanie A Stanek, and Heather T Pedersen. 2015. "Connected through Crisis: Emotional Proximity and the Spread of Misinformation Online". In *Proceedings of the 18th ACM conference on Computer Supported Cooperative Work & Social Computing,* 969–80. https://doi.org/10.1145/2675133.2675202.

Hughes, Amanda L., and Leysia Palen. 2012. “The Evolving Role of the Public Information Officer: An Examination of Social Media in Emergency Management.” *Journal of Homeland Security and Emergency Management* 9, no. 1.

Im, Jane, Eshwar Chandrasekharan, Jackson Sargent, Paige Lighthammer, Taylor Denby, Ankit Bhargava, Libby Hemphill, David Jurgens, and Eric Gilbert. 2020. “Still out There: Modeling and Identifying Russian Troll Accounts on Twitter.” In *Proceedings of the 12th ACM Conference on Web Science,* 1–10. https://doi.org/10.1145/3394231.3397889.

International Committee of the Red Cross. 2012. “Syria: ICRC and Syrian Arab Red Crescent Maintain Aid Effort amid Increased Fighting.” July 17, 2012. https://www.icrc.org/en/doc/resources/documents/update/2012/syria-update-2012-07-17.htm.

Jack, Caroline. 2017. “Lexicon of Lies: Terms for Problematic Information.” *Data & Society,* August 9, 2017. https://datasociety.net/library/lexicon-of-lies/.

James, William. 1890. *The Consciousness of Self, Principles of Psychology*. New York: Dover. http://library.manipaldubai.com/DL/the_principles_of_psychology_vol_I.pdf.

Jamieson, Kathleen Hall. 2020. *Cyberwar: How Russian Hackers and Trolls Helped Elect a President: What We Don’t, Can’t, and Do Know*. Oxford: Oxford University Press.

Jamison, Peter. 2020. “Infected by Doubt: A 26-year-old Film Editor’s Descent into Coronavirus Vaccine Conspiracy Theories.” *The Washington Post,* August 31, 2020. https://www.washingtonpost.com/dc-md-va/2020/08/31/covid-19-vaccine-conspiracy-theories-public-support.

Jensen, Casper Bruun. 2014. "Experiments in Good Faith and Hopefulness: Toward a Postcritical Social Science." *Common Knowledge* 20, no. 2: 337-62.

Jones, Jasmine, Steven Hall, Mieke Gentis, Carrie Reynolds, Chitra Gadwal, Amy Hurst, Judah Ronch, and Callie Neylan. 2012. "Visualizations for Self-Reflection on Mouse Pointer Performance for Older Adults". In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility*, 287–88. https://doi.org/10.1145/2384916.2384996.

Jowett, Garth S., and Victoria O’Donnell. 1999. *Propaganda and Persuasion.* New York: SAGE Publications.

Kabat-Zinn, Jon. 2003. "Mindfulness-based Interventions in Context: Past, Present, and Future." *Clinical Psychology: Science and Practice* 10, no. 2: 144-56.

Kafai, Yasmin B. 2005. “The Classroom as ‘Living Laboratory’: Design-Based Research for Understanding, Comparing, and Evaluating Learning Science through Design.” *Educational Technology* 45, no. 1: 28–34. https://www.jstor.org/stable/44429186.

Kaplan, Andreas M., and Michael Haenlein. 2010. “Users of the World, Unite! The Challenges and Opportunities of Social Media.” *Business Horizons* 53, no. 1: 59–68.

Karlova, Natascha A., and Jin Ha Lee. 2011. "Notes from the Underground City of Disinformation: A Conceptual Investigation." *Proceedings of the American Society for Information Science and Technology* 48, no. 1: 1-9.

Kegan, Robert. 1994. *In Over our Heads: The Mental Demands of Modern Life*. Cambridge, MA: Harvard University Press.

Kempton, Willett. 1986. "Two Theories of Home Heat Control." *Cognitive Science* 10, no. 1: 75-90.

Kendra, James, and Tricia Wachtendorf. 2003. “Creativity in Emergency Response to the World Trade Center Disaster.” Preliminary Paper. Presented at the 9th Annual

Conference of The International Emergency Management Society. http://udspace.udel.edu/handle/19716/733.

Kirgan, Harlan. 2014. “Parish Official: Text Alert of Toxic Fume Warning Believed to Be Hoax.” *StMaryNow.com - Banner-Tribune Daily Review*, September 11, 2014. http://web.archive.org/web/20140912004309/www.banner-tribune.com/local/parish-official-text-alert-toxic-fume-warning-believed-be-hoax.

Kitchin, Rob. 2014. "Big Data, New Epistemologies and Paradigm Shifts." *Big Data & Society* 1, no.1. https://doi.org/10.1177/2053951714528481.

Klein, Gary, Brian Moon, and Robert R. Hoffman. 2006. “Making Sense of Sensemaking 1: Alternative Perspectives.” *IEEE Intelligent Systems* 21, no. 4: 70–73.

Kogan, Marina, Leysia Palen, and Kenneth M. Anderson. 2015. “Think Local, Retweet Global: Retweeting by the Geographically-Vulnerable during Hurricane Sandy.” In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 981–93.

Kramer, Andrew E., and Anne Barnard. 2015. "Russian Soldiers Join Syria Fight." *The New York Times*, October 5, 2015. https://www.nytimes.com/2015/10/06/world/middleeast/russian-soldiers-join-syria-fight.html.

Langer, Ellen J. 2000. “Mindful Learning.” *Current Directions in Psychological Science* 9, no. 6: 220–23.

Latour, Bruno. 2004. “Why Has Critique Run out of Steam? From Matters of Fact to Matters of Concern.” *Critical Inquiry* 30, no. 2: 225–48.

———. 2005. *Reassembling The Social: An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.

———. 2010. “An Attempt at a ‘Compositionist Manifesto’.” New Literary History 41, no. 3: 471–90.

Lazer, David, Brian Rubineau, Carol Chetkovich, Nancy Katz, and Michael Neblo. 2010. "The Coevolution of Networks and Political Attitudes." *Political Communication* 27, no. 3: 248-74. https://doi.org/10.1080/10584609.2010.500187.

Levy, David. 2016. “Mindful Tech: Developing a More Contemplative and Reflective Relationship with Our Digital Devices and Apps.” *The Journal of Contemplative Inquiry* 3, no. 1: 35-50. https://journal.contemplativeinquiry.org/index.php/joci/article/view/111.

Lévy, Pierre. 1997. *Collective Intelligence: Mankind's Emerging World in Cyberspace*, translated by Robert Bononno. New York: Perseus Books.

Lewis, Rebecca. 2018. “Alternative Influence: Broadcasting the Reactionary Right on YouTube.” *Data & Society*, September 18, 2018. https://datasociety.net/library/alternative-influence/.

Li, Ian, Jodi Forlizzi, and Anind Dey. 2010. “Know Thyself: Monitoring and Reflecting on Facets of One’s Life.” In *Proceedings of CHI’10 Extended Abstracts on Human Factors in Computing Systems*, 4489–92.

Lin, Herbert S., and Jaclyn Kerr. 2017. “On Cyber-Enabled Information/Influence Warfare and Manipulation.” *SSRN*. https://ssrn.com/abstract=3015680.

Litt, Eden. 2012. “Knock, Knock. Who’s There? The Imagined Audience.” *Journal of Broadcasting & Electronic Media* 56, no. 3: 330–45.

Locke, Karen, Karen Golden-Biddle, and Martha S. Feldman. 2008. “Perspective—Making Doubt Generative: Rethinking the Role of Doubt in the Research Process.” *Organization Science* 19, no. 6: 907–18.

London, Daniel. 2016. “Ideas of Attachment: What the “Postcritical Turn” Means for the History of Ideas.” *Journal of the History of Ideas* (blog), November 28, 2016. https://jhiblog.org/2016/11/28/ideas-of-attachment-what-the-postcritical-turn-means-for-the-history-of-ideas/.

Lopate, Phillip. 1998. *Totally, Tenderly, Tragically: Essays and Criticism from a Lifelong Love Affair with the Movies*. New York: Anchor Books.

Lotan, Gilad, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce, and danah boyd. 2011. “The Revolutions Were Tweeted: Information Flows during the 2011 Tunisian and Egyptian Revolutions.” *International Journal of Communication* 5: 31, https://ijoc.org/index.php/ijoc/article/view/1246.

Love, Heather. 2010. “Close but Not Deep: Literary Ethics and the Descriptive Turn.” *New Literary History* 41, no. 2: 371–91.

———. 2016. “The Temptations: Donna Haraway, Feminist Objectivity, and the Problem of Critique.” In *Critique and Postcritique.* Edited by Elizabeth S. Anker and Rita Felski. Durham: Duke University Press.

Lukito, Josephine, Jiyoun Suk, Yini Zhang, Larissa Doroshenko, Sang Jung Kim, Min-Hsin Su, Yiping Xia, Deen Freelon, and Chris Wells. 2020. "The Wolves in Sheep’s Clothing: How Russia’s Internet Research Agency Tweets Appeared in US News as Vox Populi." *The International Journal of Press/Politics* 25, no. 2: 196-216. https://journals.sagepub.com/doi/full/10.1177/1940161219895215.

Lyons, Tessa. 2019. “Replacing Disputed Flags With Related Articles.” *Facebook Newsroom*, November, 7 2019. https://newsroom.fb.com/news/2017/12/news-feed-fyi-updates-in-our-fight-against-misinformation/.

Maddock, Jim, Kate Starbird, Haneen Al-Hassani, Daniel E Sandoval, Mania Orand, and Robert M Mason. 2015. "Characterizing Online Rumoring Behavior using Multi-Dimensional Signatures". In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 228–41. https://doi.org/10.1145/2675133.2675280.

Maddock, Jim, Kate Starbird, and Robert M. Mason. 2015. “Using Historical Twitter Data for Research: Ethical Challenges of Tweet Deletions.” Workshop paper. *CSCW 2015 Workshop on Ethics for Studying Sociotechnical Systems in a Big Data World.* http://faculty.washington.edu/kstarbi/maddock_starbird_tweet_deletions.pdf.

Madrigal, Alexis. 2013. “#BostonBombing: The Anatomy of a Misinformation Disaster.” *The Atlantic*, April 19, 2013. https://www.theatlantic.com/technology/archive/2013/04/-bostonbombing-the-anatomy-of-a-misinformation-disaster/275155/.

Maitlis, Sally, and Marlys Christianson. 2014. “Sensemaking in Organizations: Taking Stock and Moving Forward.” *Academy of Management Annals* 8, no. 1: 57–125.

Maitlis, Sally, Timothy J Vogus, and Thomas B Lawrence. 2013. “Sensemaking and Emotion in Organizations.” *Organizational Psychology Review* 3, no. 3: 222–47.

Malacria, Sylvain, Joey Scarr, Andy Cockburn, Carl Gutwin, and Tovi Grossman. 2013. "Skillometers: reflective widgets that motivate and help users to improve performance." In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*, 321-330. https://doi.org/10.1145/2501988.2501996.

Marcus, George E. 1995. "Ethnography in/of the World System: The Emergence of Multi-Sited Ethnography." *Annual Review of Anthropology* 24, no. 1: 95–117.

Marwick, Alice E. 2018. "Why Do People Share Fake News? A Sociotechnical Model of Media Effects." *Georgetown Law Technology Review* 2, no. 2: 474–512. http://georgetownlawtechreview.org/why-do-people-share-fake-news-a-sociotechnical-model-of-media-effects/GLTR-07-2018/.

Marwick, Alice E., and danah boyd. 2011. "I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience." *New Media & Society* 13, no. 1: 114–33.

Marwick, Alice, and Rebecca Lewis. 2017. "Media Manipulation and Disinformation Online." *Data & Society*, May 15, 2017. https://datasociety.net/library/media-manipulation-and-disinfo-online/.

Mathur, Pooja, and Karrie Karahalios. 2009. "Using Bookmark Visualizations for Self-Reflection and Navigation." In *Proceedings of CHI'09 Extended Abstracts on Human Factors in Computing Systems*, 4657–62.

Matsakis, Louise. 2017. "Twitter Told Congress This Random American Is a Russian Propaganda Troll." *Motherboard Vice,* November 3, 2017. https://motherboard.vice.com/en_us/article/8x5mma/twitter-told-congress-this-random-american-is-a-russian-propaganda-troll.

Mendoza, Marcelo, Barbara Poblete, and Carlos Castillo. 2010. "Twitter under Crisis: Can We Trust What We RT?" In *Proceedings of the First Workshop on Social Media Analytics*, 71–79.

Morell, Ricki. 2018. "Can Dialogue Journalism Engage Audiences, Foster Civil Discourse, and Increase Trust in the Media?" *Nieman Reports*, October 23, 2018. https://niemanreports.org/articles/can-dialogue-journalism-engage-audiences-foster-civil-discourse-and-increase-trust-in-the-media/.

Moynihan, Colin. 2016. “10 Black Employees at New York Fire Department Cite Bias.” *The New York Times,* October 12, 2016. https://www.nytimes.com/2016/10/13/nyregion/10-black-employees-at-new-york-fire-dept-cite-bias.html.

Mroue, Bassem. 2017. “7 White Helmets Medics Killed in Syria’s Idlib.” *Associated Press,* August 12, 2017. https://apnews.com/article/b86ec7c7584a4e2a8ca1e113aa29a098.

Muller, Michael. 2014. "Curiosity, Creativity, and Surprise as Analytic Tools: Grounded Theory Method." In *Ways of Knowing in HCI,* 25-48. New York: Springer.

Mullins, Matthew. 2015. “Are We Postcritical?” Review of *Limits of Critique,* by Rita Felski. *Los Angeles Review of Books,* December 27, 2015. https://lareviewofbooks.org/article/are-we-postcritical/.

Murray, Gregg R., and Anthony Scime. 2010. “Microtargeting and Electorate Segmentation: Data Mining the American National Election Studies.” *Journal of Political Marketing* 9, no. 3: 143–66.

Nagar, Yiftach. 2012. “What Do You Think? The Structuring of an Online Community as a Collective-Sensemaking Process.” In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work,* 393–402.

National Intelligence Council. 2017. “Assessing Russian Activities and Intentions in Recent US Elections.” Office of the Director of National Intelligence, National Intelligence Council. https://www.dni.gov/files/documents/ICA_2017_01.pdf.

Nigam, Amit, and William Ocasio. 2010. “Event Attention, Environmental Sensemaking, and Change in Institutional Logics: An Inductive Analysis of the Effects of Public Attention to Clinton’s Health Care Reform Initiative.” *Organization Science* 21, no. 4: 823–41.

Nishikawa, Kinohi. n.d. “Race, Thick and Thin.” *Arcade* (blog). Accessed December 15, 2020. https://arcade.stanford.edu/content/race-thick-and-thin.

Ntuen, Celestine A., Eui H. Park, and Kim Gwang-Myung. 2010. "Designing an information visualization tool for sensemaking." *International Journal of Human–Computer Interaction* 26, no. 2-3: 189-205.

Nyhan, Brendan, and Jason Reifler. 2012. “Misinformation and Fact-Checking: Research findings from Social Science.” *New America Foundation*. https://www.dartmouth.edu/~nyhan/Misinformation_and_Fact-checking.pdf.

Odom, William T., Abigail J Sellen, Richard Banks, David S Kirk, Tim Regan, Mark Selby, Jodi L Forlizzi, and John Zimmerman. 2014. "Designing for Slowness, Anticipation and Re-Visitation: A Long Term Field Study of the Photobox". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1961–70. https://doi.org/10.1145/2556288.2557178.

Oh, Onook, Manish Agrawal, and H. Raghav Rao. 2013. “Community Intelligence and Social Media Services: A Rumor Theoretic Analysis of Tweets During Social Crises.” *MIS Quarterly* 37, no. 2: 407–426. https://doi.org/10.25300/misq/2013/37.2.05.

Ong, Jonathan Corpus, and Jason Vincent A. Cabañes. 2018. “Architects of Networked Disinformation: Behind the Scenes of Troll Accounts and Fake News Production in the Philippines.” Public Report. *The Newton Tech4Dev Network.* https://newtontechfordev.com/wp-content/uploads/2018/02/ARCHITECTS-OF-NETWORKED-DISINFORMATION-FULL-REPORT.pdf.

———. 2019. “When Disinformation Studies Meets Production Studies : Social Identities and Moral Justifications in the Political Trolling Industry.” *International Journal of Communication* 13: 5771–90. https://ijoc.org/index.php/ijoc/article/view/11417/2879.

Ong, Walter J. 1975. “The Writer’s Audience Is Always a Fiction.” *Publications of the Modern Language Association of America* 90, no. 1: 9–21. https://doi.org/10.2307/461344.

Orlikowski, Wanda J., and Daniel Robey. 1991. “Information Technology and the Structuring of Organizations.” *Information Systems Research* 2, no. 2: 143–69.

Ou-Yang, Lucas. n.d. “Newspaper3k: Article Scraping & Curation.” *Github*. Accessed on December 15 2020. https://github.com/codelucas/newspaper.

Oxford English Dictionary, s.v. "crowd, n.3." Accessed October 21, 2020. https://www.oed.com/view/Entry/45034.

———, s.v. "disinformation, n." Accessed October 21, 2020. https://www.oed.com/view/Entry/54579.

Paarlberg, Michael. 2017. “Why Verrit, a pro-Clinton Media Platform, Is Doomed to Fail.” *The Guardian*, September 8, 2017. http://www.theguardian.com/commentisfree/2017/sep/08/verrit-pro-clinton-media-platform-doomed-failure.

Palen, Leysia, and Kenneth M. Anderson. 2016. “Crisis Informatics—New Data for Extraordinary Times.” *Science* 353, no. 6296: 224–25.

Palen, Leysia, Kenneth M. Anderson, Gloria Mark, James Martin, Douglas Sicker, Martha Palmer, and Dirk Grunwald. 2010. “A Vision for Technology-Mediated Support for Public Participation & Assistance in Mass Emergencies & Disasters.” In *Proceedings of the 2010 ACM-BCS Visions of Computer Science Conference*, 1–12. https://dl.acm.org/doi/10.5555/1811182.1811194.

Palen, Leysia, and Amanda L. Hughes. 2018. "Social Media in Disaster Communication." In *Handbook of Disaster Research,* 497-518. New York: Springer. https://doi.org/10.1007/978-3-319-63254-4_24.

Palma, Bethania. 2016. “Syrian War Victims Are Being ‘Recycled’ and Al Quds Hospital Was Never Bombed?. *Snopes,* December 14, 2016. https://www.snopes.com/fact-check/syrian-war-victims-are-being-recycled-and-al-quds-hospital-was-never-bombed/.

Papacharissi, Zizi. 2009. “The Virtual Geographies of Social Networks: A Comparative Analysis of Facebook, LinkedIn and ASmallWorld.” *New Media & Society* 11, no. 1–2: 199–220.

Paul, Christopher, and Miriam Matthews. 2016. “The Russian ‘Firehose of Falsehood’ Propaganda Model.” *Rand Corporation*. https://doi.org/10.7249/PE198.

Peirce, Charles Sanders. 1960. *Collected Papers of Charles Sanders Peirce*. Cambridge, MA: Harvard University Press.

Penney, Joel. 2017. *The Citizen Marketer: Promoting Political Opinion in the Social Media Age*. Cambridge, MA: Oxford University Press.

Persily, Nathaniel. 2017. “Can Democracy Survive the Internet?” *Journal of Democracy* 28, no. 2: 63–76. https://doi.org/10.1353/jod.2017.0025.

Phillips, Whitney, and Ryan M. Milner. 2018. *The Ambivalent Internet: Mischief, Oddity, and Antagonism Online*. Hoboken, NJ: John Wiley & Sons.

Pictet, Jean. 1979. “The Fundamental Principles of the Red Cross.” *International Review of the Red Cross Archive* 19, no. 210: 130–49.

Pohjonen, Matti, and Sahana Udupa. 2017. "Extreme Speech Online: An Anthropological Critique of Hate Speech Debates." *International Journal of Communication* 11. https://ijoc.org/index.php/ijoc/article/view/5843.

Polkinghorne, Donald E. 1988. *Narrative Knowing and the Human Sciences*. Albany, NY: SUNY Press.

Pomerantsev, Peter, and Michael Weiss. 2014. “The Menace of Unreality: How the Kremlin Weaponizes Information, Culture and Money.” Report. New York: Institute of Modern Russia. https://imrussia.org/media/pdf/Research/Michael_Weiss_and_Peter_Pomerantsev__The_Menace_of_Unreality.pdf.

Prasad, Jamuna. 1935. “The Psychology of Rumour: A Study Relating to the Great Indian Earthquake of 1934.” *British Journal of Psychology* 26, no. 1.

Pratt, Michael G. 2000. “The Good, the Bad, and the Ambivalent: Managing Identification among Amway Distributors.” *Administrative Science Quarterly* 45, no. 3: 456–93.

Prier, Jarred. 2017. “Commanding the Trend: Social Media as Information Warfare.” *Strategic Studies Quarterly* 11, no. 4: 50–85.

Putin, Vladimir. 2012. “Vladimir Putin on Foreign Policy: Russia and the Changing World.” *Valdai Discussion Club*, February 27, 2012. https://valdaiclub.com/a/highlights/vladimir_putin_on_foreign_policy_russia_and_the_changing_world/.

Qazvinian, Vahed, Emily Rosengren, Dragomir Radev, and Qiaozhu Mei. 2011. "Rumor has it: Identifying misinformation in microblogs." In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1589-1599. http://www.aclweb.org/anthology/D11-1147.pdf.

Qiu, Linda. 2017. "Fact Check: Trump Blasts ‘Fake News’ and Repeats Inaccurate Claims at CPAC." *The New York Times*, February 24, 2017. https://www.nytimes.com/2017/02/24/us/politics/fact-check-trump-blasts-fake-news-and-repeats-inaccurate-claims-at-cpac.html.

Qu, Yan, Philip Fei Wu, and Xiaoqing Wang. 2009. “Online Community Response to Major Disaster: A Study of Tianya Forum in the 2008 Sichuan Earthquake.” In *Proceedings of the 42nd Hawaii International Conference on System Sciences*, 1–11.

Qu, Yan, Chen Huang, Pengyi Zhang, and Jun Zhang. 2011. “Microblogging after a Major Disaster in China: A Case Study of the 2010 Yushu Earthquake.” In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, 25–34. https://doi.org/10.1145/1958824.1958830.

Quarantelli, Enrico Louis. 1991. “Radiation disasters: Similarities to and Differences from Other Disasters.” In *The Medical Basis for Radiation-Accident Preparedness III: The Psychological Perspective*, edited R. Ricks, M. Berger and F. O'hara Jr., 15-24. Amsterdam, Netherlands: Elsevier Science Publishers.

Ratkiewicz, Jacob, Michael D. Conover, Mark Meiss, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer Menczer. 2011a. "Detecting and Tracking Political Abuse in Social Media". In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 297-304. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.646.5073&rep=rep1&type=pdf.

Ratkiewicz, Jacob, Michael D. Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. 2011b. "Truthy: mapping the spread of astroturf in microblog streams." In *Proceedings of the 20th International Conference Companion on World Wide Web*, 249-252. https://doi.org/10.1145/1963192.1963301.

Ravi, Narasimhan. 2005. “Looking Beyond Flawed Journalism: How National Interests, Patriotism, and Cultural Values Shaped the Coverage of the Iraq War.” *Harvard International Journal of Press/Politics* 10, no. 1: 45–62. https://doi.org/10.1177/1081180X05275765.

Reuter, Christian, and Marc-André Kaufhold. 2018. “Fifteen Years of Social Media in Emergencies: A Retrospective Review and Future Directions for Crisis Informatics.” *Journal of Contingencies and Crisis Management* 26, no. 1: 41–57.

Ricoeur, Paul. (1965) 2008. *Freud and Philosophy: An Essay on Interpretation*. New Haven: Yale University Press. New edition, Delhi: Motilal Banarsidass Publishers. Citations refer to the Banarsidass edition.

Rid, Thomas. 2020. *Active Measures: The Secret History of Disinformation and Political Warfare*. New York: Farrar, Straus and Giroux.

Rosenberg, Marshall B., and Deepak Chopra. 2015. *Nonviolent Communication: A Language of Life: Life-Changing Tools for Healthy Relationships*. Encinitas, CA: PuddleDancer Press.

Rosnow, Ralph L. 1980. *Psychology of Rumor Reconsidered.* Washington, DC: American Psychological Association.

———. 1991. “Inside Rumor: A Personal Journey.” *American Psychologist* 46, no. 5: 484.

Rosnow, Ralph L., James L. Esposito, and Leo Gibney. 1988. “Factors Influencing Rumor Spreading: Replication and Extension.” *Language & Communication* 8, no. 1: 29-42. https://doi.org/10.1016/0271-5309(88)90004-3.

Rosnow, Ralph L., and Allan J. Kimmel. 2000. “Rumor.” *Encyclopedia of Psychology* 7: 122-123.

Rosvall, Martin, Daniel Axelsson, and Carl T. Bergstrom. 2009. “The Map Equation.” *The European Physical Journal Special Topics* 178, no. 1: 13–23.

Rotman, Dana, Jennifer Preece, Yurong He, and Allison Druin. 2012. “Extreme Ethnography: Challenges for Research in Large Scale Online Environments.” In *Proceedings of the 2012 IConference*, 207–14. https://doi.org/10.1145/2132176.2132203.

Said, Edward. 1978. *Orientalism*. New York: Pantheon Books.

Salancik, Gerald R. 1977. “Commitment and the Control of Organizational Behavior and Belief.” *New Directions in Organizational Behavior* 1.

Salton, Gerard, Edward A. Fox, and Harry Wu. 1983. “Extended Boolean Information Retrieval.” *Communications of the ACM* 26, no. 11: 1022–36.

Sandberg, Jörgen, and Haridimos Tsoukas. 2015. “Making Sense of the Sensemaking Perspective: Its Constituents, Limitations, and Opportunities for Further Development.” *Journal of Organizational Behavior* 36, no. S1: S6–S32. https://doi.org/10.1002/job.1937.

Savage, Saiph, Andres Monroy-Hernandez, and Tobias Höllerer. 2016. “Botivist: Calling Volunteers to Action Using Online Bots.” In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 813–22. https://doi.org/10.1145/2818048.2819985.

Sawyer, Jeffrey K. 1990. *Printed Poison: Pamphlet Propaganda, Faction Politics, and the Public Sphere in Early Seventeenth-Century France*. Berkeley, CA: University of California Press.

Schneider, Susan C. 1987. “Information Overload: Causes and Consequences.” *Human Systems Management* 7, no. 2: 143–53.

Sedgwick, Eve Kosofsky. 1997. *Paranoid Reading and Reparative Reading, or, You’re so Paranoid, You Probably Think This Introduction Is about You*. Durham, NC: Duke University Press.

Shand, Alexander F. 1922. “Suspicion.” *British Journal of Psychology. General Section* 13, no. 2: 195–214.

Sharma, Nikhil. 2008. “Sensemaking Handoff: When and How?” *Proceedings of the American Society for Information Science and Technology* 45, no. 1: 1–12.

Shearer, Elisa, and Jeffrey Gottfried. 2017. *News use Across Social Media Platforms*. Pew Research Center. https://www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017/.

Shein, Esther. 2013. “Ephemeral Data.” *Communications of the ACM* 56, no. 9: 20–22. https://doi.org/10.1145/2500468.2500474.

Shibutani, Tamotsu. 1966. *Improvised News: A Sociological Study of Rumor*. London: Ardent Media.

Silverman, Craig. 2018. “Russian Trolls Ran Wild On Tumblr And The Company Refuses To Say Anything About It.” *BuzzFeed News*, February 6, 2018. https://www.buzzfeed.com/craigsilverman/russian-trolls-ran-wild-on-tumblr-and-the-company-refuses?utm_term=.ad65gb5jz#.rdwOw8O6Z.

Snook, Scott A. (2000) 2002. *Friendly Fire: The Accidental Shootdown of US Black Hawks Over Northern Iraq*. Princeton, NJ: Princeton University Press.

Snyder, Alvin A. (1995) 1997. *Warriors of Disinformation: American Propaganda, Soviet Lies, and the Winning of the Cold War: An Insider's Account*. New York: Arcade Publishing.

Søe, Sille Obelitz. 2016. “The Urge to Detect, the Need to Clarify: Gricean Perspectives on Information, Misinformation and Disinformation.” PhD diss., Københavns Universitet, Det Humanistiske Fakultet.

Spiro, Emma S., Sean Fitzhugh, Jeannette Sutton, Nicole Pierski, Matt Greczek, and Carter T. Butts. 2010. "Rumoring during extreme events: A case study of Deepwater Horizon 2010." In *Proceedings of the 4th Annual ACM Web Science Conference*, 275-283. https://doi.org/10.1145/2380718.2380754.

Sprague, Joey. (2005) 2016. *Feminist Methodologies for Critical Researchers: Bridging Differences*. Walnut Creek, CA: Alta Mira Press. Second edition, Lanham, MD: Rowman & Littlefield. Citations refer to the second edition.

Stamos, Alex. 2018. “Authenticity Matters: The IRA Has No Place on Facebook.” *Facebook Newsroom,* April 3 *2018.* https://newsroom.fb.com/news/2018/04/authenticity-matters/. Accessed December 15, 2020.

Star, Susan Leigh, and Geoffrey C. Bowker. 2007. "Enacting Silence: Residual Categories as a Challenge for Ethics, Information Systems, and Communication." *Ethics and Information Technology* 9: 273-280. https://doi.org/10.1007/s10676-007-9141-7.

Starbird, Kate. 2012. “Crowdwork, Crisis and Convergence: How the Connected Crowd Organizes Information during Mass Disruption Events.” PhD diss., University of Colorado.

———. 2017. “Examining the Alternative Media Ecosystem through the Production of Alternative Narratives of Mass Shooting Events on Twitter.” In *Proceedings of the 2017 AAAI International Conference On Web and Social Media*, 230–39.

———. 2018. “A First Glimpse through the Data Window onto the Internet Research Agency’s Twitter Operations.” *Medium* (blog), October 17, 2018. https://medium.com/@katestarbird/a-first-glimpse-through-the-data-window-onto-the-internet-research-agencys-twitter-operations-d4f0eea3f566.

Starbird, Kate, Ahmer Arif, and Tom Wilson. 2019. “Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations.” *Proceedings of the ACM on Human-Computer Interaction* 3: 1–26. https://doi.org/10.1145/3359229.

Starbird, Kate, Ahmer Arif, Tom Wilson, Katherine Van Koevering, Katya Yefimova, and Daniel Scarnecchia. 2018. "Ecosystem or Echo-System? Exploring Content Sharing across Alternative Media Domains." In *Proceedings of the International AAAI Conference on Web and Social Media*, 365-374. http://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17836.

Starbird, Kate, Jim Maddock, Mania Orand, Peg Achterman, Robert M. Mason. 2014. “Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter after the 2013 Boston Marathon Bombing.” In *Proceedings of the 2014 iConference*, 654 - 662. http://hdl.handle.net/2142/47257.

Starbird, Kate, and Leysia Palen. 2010. “Pass it on?: Retweeting in mass emergency.” In *Proceedings of the 7th International Conference on Information Systems for Crisis Response and Management*, 1-10.

———. 2011. "’Voluntweeters’ self-organizing by digital volunteers in times of crisis." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1071-1080. https://doi.org/10.1145/1978942.1979102.

———. 2012. “(How) Will the Revolution Be Retweeted? Information Diffusion and the 2011 Egyptian Uprising.” In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, 7–16. https://doi.org/10.1145/2145204.2145212.

Starbird, Kate, Leysia Palen, Amanda L. Hughes, and Sarah Vieweg. 2010. “Chatter on the Red: What Hazards Threat Reveals about the Social Life of Microblogged

Information.” In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, 241–250. https://doi.org/10.1145/1718918.1718965.

Starbird, Kate, Emma S. Spiro, Isabelle Edwards, Kaitlyn Zhou, Jim Maddock, and Sindhuja Narasimhan. 2016. “Could This Be True? I Think so! Expressed Uncertainty in Online Rumoring.” In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 360–71.

Stewart, Leo Graiden, Ahmer Arif, A. Conrad Nied, Emma S. Spiro, and Kate Starbird. 2017. "Drawing the lines of contention: Networked frame contests within# BlackLivesMatter discourse." *Proceedings of the ACM on Human-Computer Interaction* 1: 1-23. https://doi.org/10.1145/3134920.

Stubbs, Jack, and Ginger Gibson. 2017. "Russia's RT America Registers as ‘Foreign Agent' in US." *Reuters*, Nov 13, 2017. https://www.reuters.com/article/us-russia-usa-media-restrictions-rt/russias-rt-america-registers-as-foreign-agent-in-u-s-idUSKBN1DD25B.

Suchman, Lucy. 2014. “Mediations and Their Others.” In *Media Technologies: Essays on Communication, Materiality, and Society*, 129–39. Cambridge, MA: MIT Press.

Surowiecki, James. 2005. *The Wisdom of Crowds:* New York: Anchor Books.

Swift, Art. 2016. “Americans' Trust in Mass Media Sinks to New Low.” *Gallup*, September 14, 2016. https://news.gallup.com/poll/195542/americans-trust-mass-media-sinks-new-low.aspx.

Tanaka, Yuko, Yasuaki Sakamoto, and Toshihiko Matsuka. 2012. “Transmission of Rumor and Criticism in Twitter after the Great Japan Earthquake.” In *Proceedsings of the Annual Meeting of the Cognitive Science Society*, 2387-2392. https://escholarship.org/uc/item/0h97h07x.

Tetrault-Farber, Gabrielle. 2014. “Looking West: Russia Beefs Up Spending on Global Media Giants.” *The Moscow Times*, September 23, 2014. https://www.themoscowtimes.com/2014/09/23/looking-west-russia-beefs-up-spending-on-global-media-giants-a39708.

Thatcher, Jim, David O’Sullivan, and Dillon Mahmoudi. 2016. “Data Colonialism through Accumulation by Dispossession: New Metaphors for Daily Data.” *Environment and Planning D: Society and Space* 34, no. 6: 990–1006.

The Internet Archive. n.d. “About the Internet Archive.” Accessed April 17, 2018. http://archive.org/about/.

Thera, Nyanaponika. 1996. *The Heart of Buddhist meditation*. Cape Neddick, ME: Samuel Weiser.

Thordsen, Marvin L., and Gary A. Klein. 1989. “Cognitive Processes of the Team Mind.” In *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, 46–49.

Tierney, Kathleen J. 2007. “From the Margins to the Mainstream? Disaster Research at the Crossroads.” *Annual Review of Sociology* 33: 503–25.

Tripodi, Francesca. 2018. “Searching for Alternative Facts.” *Data & Society,* May 16, 2018. https://datasociety.net/library/searching-for-alternative-facts/.

Troianovski, Anton. 2018. “A Former Russian Troll Speaks: ‘It Was Like Being in Orwell’s World’.” *Washington Post*, February 17, 2018. https://www.washingtonpost.com/news/worldviews/wp/2018/02/17/a-former-russian-troll-speaks-it-was-like-being-in-orwells-world/.

Tucker, Joshua, Andrew Guess, Pablo Barbera, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. “Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature.” *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3144139.

Tufekci, Zeynep. 2014. “Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls.” In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, 505-514. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.885.3471&rep=rep1&type=pdf.

Tumblr Help Center. 2018. “Public Record of Usernames Linked to State-Sponsored Disinformation Campaigns.” https://tumblr.zendesk.com/hc/en-us/articles/360002280214.

Twitter. 2018. “Update on Twitter's Review of the 2016 U.S. Election.” Last modified January 31, 2018. https://blog.twitter.com/official/en_us/topics/company/2018/2016-election-update.html.

———. n.d. "About Twitter Limits." Help Center. Accessed December 14, 2020. https://help.twitter.com/en/rules-and-policies/twitter-limits.

U.S. House of Representatives Permanent Select Committee on Intelligence. 2017. "Exhibit B." Accessed December 14, 2020. http://democrats-intelligence.house.gov/uploadedfiles/exhibit_b.pdf.

U.S. Joint Chiefs of Staff. 2014. "Information Operations - Joint Publication 3-13." Accessed December 15, 2020. https://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/jp3_13.pdf.

U.S. Justice Department. 2018. "USA v. IRA et al. Case 1:18-cr-00032-DLF." https://www.justice.gov/file/1035477/download. Accessed December 15, 2020.

U.S. Senate Select Committee on Intelligence. 2017. "Testimony of Sean J. Edgett." November 1, 2017. https://www.intelligence.senate.gov/sites/default/files/documents/os-sedgett-110117.pdf. Accessed December 15, 2020.

Vallor, Shannon. 2016. *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford: Oxford University Press.

Verbeek, Peter-Paul. 2005. *What Things Do: Philosophical Reflections on Technology, Agency, and Design.* University Park, PA: Penn State Press.

———. 2011. *Moralizing Technology: Understanding and Designing the Morality of Things*. Chicago: University of Chicago Press, 2011.

Walczyk, Jeffrey J., Mark A. Runco, Sunny M. Tripp, and Christian E. Smith. 2008. "The creativity of lying: Divergent thinking and ideational correlates of the resolution of social dilemmas." *Creativity Research Journal* 20, no. 3: 328-342.

Walker, Charles J., and Bruce Blaine. 1991. "The Virulence of Dread Rumors: A Field Experiment." *Language & Communication* 11, no. 4: 291-297. https://doi.org/10.1016/0271-5309(91)90033-R.

Wang, Yiran, and Gloria Mark. 2017. "Engaging with Political and Social Issues on Facebook in College Life." In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 433–45. https://doi.org/10.1145/2998181.2998295.

Wanless, Alicia, and Michael Berk. 2017. “Participatory Propaganda: The Engagement of Audiences in the Spread of Persuasive Communications.” In *Proceedings of the Social Media & Social Order, Culture Conflict 2.0 Conference*.

Wardle, Claire, and Hossein Derakhshan. 2017. *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making*. Council of Europe, September 27, 2017. https://tverezo.info/wp-content/uploads/2017/11/PREMS-162317-GBR-2018-Report-desinformation-A4-BAT.pdf.

Weedon, Jen, William Nuland and Alex Stamos. 2017. “Information Operations and Facebook.” *Facebook Newsroom*, April 27, 2017. https://about.fb.com/br/wp-content/uploads/sites/3/2017/09/facebook-and-information-operations-v1.pdf. Accessed December 15, 2020.

Weick, Karl E. 1988. "Enacted Sensemaking in Crisis Situations." *Journal of Management Studies* 25, no. 4: 305-17. https://doi.org/10.1111/j.1467-6486.1988.tb00039.x.

———. 1993. “The Collapse of Sensemaking in Organizations: The Mann Gulch Disaster.” *Administrative Science Quarterly* 38, no. 4: 628–52.

———. 1995. *Sensemaking in Organizations*. New York: SAGE Publications.

———. 1998. "Introductory essay—Improvisation as a Mindset for Organizational Analysis." *Organization Science* 9, no. 5: 543-55.

Weick, Karl E., and Ted Putnam. 2006. "Organizing for Mindfulness: Eastern Wisdom and Western Knowledge." *Journal of Management Inquiry* 15, no. 3: 275-87.

Weick, Karl E., and Kathleen M. Sutcliffe. 2006. "Mindfulness and the Quality of Organizational Attention." *Organization Science* 17, no. 4: 514-24.

———. (2001) 2011. *Managing the Unexpected: Resilient Performance in an Age of Uncertainty*. Hoboken, NJ: John Wiley & Sons. Reprint, San Francisco: Jossey-Bass. Citations refer to the Jossey-Bass edition.

Weick, Karl E., Kathleen M. Sutcliffe, and David Obstfeld. 2005. "Organizing and the Process of Sensemaking." *Organization Science* 16, no. 4: 409-21.

Weiser, Marc. 1994. “The World Is Not a Desktop.” *Interactions* 1, no. 1: 7–8.

Wenneling, Oskar. 2007. “Seamful Design–The Other Way Around.” In *Proceedings of the Scandinavian Student Interaction Design Research Conference*, 14-16.

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.95.7739&rep=rep1&type=pdf.

Westrum, Ron. 1993. "Cultures with Requisite Imagination." In *Verification and Validation of Complex Systems: Human Factors Issues*, 401-416. New York: Springer.

———. 1997. “Social Factors in Safety-Critical Systems.” In *Human Factors in Safety Critical Systems*, edited by Felix Redmill and Jane Rajan, 233–56. Oxford: Butterworth-Heinemann.

Whewell, T. 2013. “Syrian activists flee abuse in al-Qaeda-run Raqqa.” *BBC News*, November 15, 2013. https://www.bbc.com/news/av/world-24958179. Accessed December 15, 2020.

Wilson, Tom, Kaitlyn Zhou, and Kate Starbird. 2018. “Assembling Strategic Narratives: Information Operations as Collaborative Work within an Online Community.” *Proceedings of the ACM on Human-Computer Interaction* 2, no. CSCW: 1–26. https://doi.org/10.1145/3274452.

Wittgenstein, Ludwig. (1953) 2009. *Philosophical Investigations*. Translated by G. E. M. Anscombe, P. M. S. Hacker, and Joachim Schulte. Revised fourth edition, Hoboken, NJ: John Wiley & Sons.

Wong-Villacres, Marisol, Cristina M. Velasquez, and Neha Kumar. 2017. “Social Media for Earthquake Response: Unpacking Its Limitations with Care.” *Proceedings of the ACM on Human-Computer Interaction* 1, no. CSCW: 1–22. https://doi.org/10.1145/3134747.

Woolley, Anita Williams, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. 2010. "Evidence for a collective intelligence factor in the performance of human groups." *Science* 330, no. 6004: 686-688. https://doi.org/10.1126/science.1193147.

Woolley, Samuel C., and Philip Howard. 2017. *Computational Propaganda Worldwide: Executive Summary*. Oxford, UK: The Computational Propaganda Project. http://blogs.oii.ox.ac.uk/politicalbots/wp-content/uploads/sites/89/2017/06/Casestudies-ExecutiveSummary.pdf. Accessed December 15, 2020.

Wulf, Volker, Kaoru Misaki, Meryem Atam, David Randall, Markus Rohde. 2013. "'On the Ground' in Sidi Bouzid: Investigating Social Media Use during the Tunisian Revolution." In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 1409–18.

York, Chris. 2020. "The 'Useful Idiots': How These British Academics Helped Russia Deny War Crimes At The UN." *The Huffington Post*, January 29, 2020. https://www.huffingtonpost.co.uk/entry/the-useful-idiots_uk_5e2b107ac5b67d8874b0dd9d.

Zajonc, Arthur. 2013. "Contemplative Pedagogy: A Quiet Revolution in Higher Education." *New Directions for Teaching and Learning* 134: 83–94.

Zeier, Kristin, and Francisco Perez. 2016. "Social Media and Breaking News: Keep Calm and Don't Retweet Everything You See." *Deutsche Welle*, July 23, 2016. https://dw.com/p/1JUdL.

Zhao, Ou Jie, Tiffany Ng, and Dan Cosley. 2012. "No Forests without Trees: Particulars and Patterns in Visualizing Personal Communication." In *Proceedings of the 2012 iConference*, 25–32.

Zhao, Zhe, Paul Resnick, and Qiaozhu Mei. 2015. "Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts." In *Proceedings of the 24th International Conference on World Wide Web*, 1395–405.