NETWORKED PLATFORM GOVERNANCE: RECONCILING HORIZONTALS AND HIERARCHIES IN THE PLATFORM ERA

By

ROBYN CAPLAN

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Media Studies

Written under the direction of

Philip M. Napoli, Ph.D.

And approved by

New Brunswick, New Jersey

October, 2021

ABSTRACT OF THE DISSERTATION

Networked Platform Governance: Reconciling Horizontals and Hierarchies in the Platform Era

By ROBYN CAPLAN

Dissertation Director:

Philip M. Napoli, Ph.D.

Over the last several years, concerns about the credibility or trustworthiness of information online have been mounting. At the center of these concerns were questions about the role platform companies — particularly search and social media — should be playing in controlling access to information online. This dissertation provides three perspectives on the development of content standards for false information online. Through case studies told from the perspectives of the platform industry, from a media association, and from YouTube content creators, this dissertation explores the challenges and consequences of multistakeholderism in content policymaking by platforms. Though what constitutes trustworthy or credible information has always been a concern in communication systems, this study proposes that recent concerns about the spread of false information over platforms reveal a renegotiation of the boundaries between amateurs and experts in knowledge production that has been unfolding over the last several decades with the rise of platforms and social media. This study places this process of renegotiation within the broader context of a new area for media reform referred to as *platform governance*.

Acknowledgements

I began this dissertation (or rather, defended my dissertation proposal) almost four years ago in July 2017. Since that summer, I have gotten married, had a child, lost a parent, and, along with everyone else in the United States, lived through Donald Trump as President, as well as a global pandemic. There were multiple points over the last several years when I, and probably many of the members of my committee (rightly) thought the odds of me completing this dissertation were fairly slim. Truthfully, it would not have gotten done without the support, guidance, and endless patience of many mentors, friends, and family members.

To that end, I have many people to thank in this acknowledgement section. My sincerest thank you to my dissertation chair Philip M. Napoli who has spent countless hours reviewing my work, talking me through myriad professional and personal challenges over Skype and Zoom, and thinking through various contours of this debate with me. I am forever grateful for your dedication, patience, and scholarship (which very much inspired this work). I would also like to thank danah boyd, the most "internal external committee member that has ever existed" (in her words). Words cannot describe how grateful I am for your endless guidance, your friendship, and for giving me my intellectual home. I would also like to thank Susan Keith for her mentorship throughout my very long PhD journey; thank you for sticking with me and for answering the phone during particularly hard times. Thank you as well to Jeff Lane for your time and feedback, and for so readily jumping onto my committee four years ago. I would also like to thank past committee members Aram Sinnreich and Ingrid Erickson, who have since moved on to other universities. Thank you as well to the administrative staff and faculty administration at Rutgers

University. In particular, I would like to thank Danielle Lopez, and Jennifer Theiss for their guidance and support as they helped me through the many bureaucratic elements of doing a PhD.

I would like to thank my colleagues at Data & Society Research Institute for supporting me in completing my dissertation work, and for pushing my thinking through many productive debates and conversations. In particular, I would like to thank Janet Haven and Sareeta Amrute for giving me the space and stability I needed to complete this work. I would also like to thank long-time friends and colleagues past and present who were particularly influential in guiding my thinking and providing emotional support when I needed it, including Will Partin, Audrey Evans, Andrew Selbst, Angie Waller, Mark Latonero, Alex Rosenblat, Madeleine Claire Elish, Seth Young, Alexandra Mateescu, Monica Bulger, Caroline Jack, Francesca Tripodi, Patrick Davison, Gabby Lim, Jake Metcalf, Becca Lewis, Elizabeth Watkins, Mary Madden, Aiha Nguyen, Irene Chung, Amanda Lenhart, Kinjal Dave, and countless others (including fellows and post-docs). A special thank you to my daily writing partner, Emanuel Moss. Thank you for your daily encouragement and for inviting me into your Google hangout. I am so glad we are finishing this odyssey together.

I would like to thank friends and colleagues in the platform governance community. In particular, I'd like to thank Tarleton Gillespie for his continuous guidance and mentorship. He is the secret fifth member of my committee (and the secret moderator for my panels). Thank you to many other mentors throughout the years that have guided me at various points of my journey: Siva Vaidhyanathan, Alice Marwick, Mary Gray, Nancy Baym, Monroe Price, Anne Washington, Fenwick McKelvey, C.W. Anderson, Dan Bouk, Nate Matias, David Craig, Caroline Jack, Mark Ackerman, Angèle Christin, Nate Matias, and Anne Washington. Thank you as well to members of the Platform Governance Research Network who have kept me sane these last few months through group chats and Slack, particularly Robert Gorwa, Sonja Solomun, and Pranav Bidare, and to other wonderful friends and sources of support this last few years, always ready to answer a Twitter DM or text: Elena Maris, Christopher Persaud, Andrew Young, Molly Sauter, MC Forelle, Sophie Bishop, Julia Powles, and Kate Klonick. Thank you to my grad school bestie Vanessa Kitzie for getting me through the first few years of graduate school, and to life-time besties Nay Laywine, Xavier Scott, SJ Thiessen, and Eden Sela for being there always.

I want to thank my family for their love and support these last few years as I tried to balance many professional and personal milestones at once. To my Uncle Hugh — academia was not your first choice for me, but thank you for convincing me to keep going and finish. To my sisters and brother, Kim, Mikki, and Ben (and their spouses Tanner, Trevor, and Ida) thank you for being there for me, and for being 'there' when I could not be. To my dad, Frederick Caplan, thank you so much for keeping me on track, and for convincing me to go to sleep when I needed it. And to Ellen and Marc, thank you for all of your love and support over these last few difficult years.

Lastly, I want to thank my husband and son. Adam, thank you for being my partner in love and parenting, and for being my best friend. Thank you so much for giving me the space, freedom, and stability to write over these last several years, and for the many ways you pushed me intellectually, exposing me to ideas and theorists within political science that helped guide my thinking on this work. This finally finished dissertation belongs to both of us (so maybe your parents will stop telling you to finish yours). To my son Elliot, thank you for giving me daily joy, and for making my dream (which was always to be your mom) finally come true. And finally, my mother, Marion Caplan, passed away on May 16th 2020 after an eight-year battle with cancer. Thank you to her for encouraging me to take big risks, to work hard, and to finish what I started. This project is dedicated to her memory. In summer of 2020, my father was also diagnosed with cancer. This project is likewise dedicated to his fight.

Acknowledgement of Previously Published Work Contained Within This Dissertation

As this dissertation research progressed, parts of it were published as I worked toward completing the project in its entirety. Select portions of the literature review (Chapter 1) were published in the report *Dead Reckoning: Content Moderation After Fake News*, published by Data & Society Research Institute, which I co-authored with Lauren Hanson and Joan Donovan (2018). The portions that have been incorporated into the dissertation came entirely from my contributions to the piece. Chapter 4 of this dissertation has been published in a journal article I co-authored with Tarleton Gillespie (2020) titled *Tiered Governance: The Shifting Terms of Labor and Compensation in the Platform Economy*, published in 2020. The portions I have included within this text are drawn from my contributions to the article. Lastly, portions of the conclusion of this dissertation include work I have since published in Brookings Institute, and Slate Magazine

Table of Contents

ABSTRACT OF THE DISSERTATION ii

Acknowledgements iii

Introduction and Method....... 1

Chapter 1: Platform Governance: The Era of New Media Reform....... 23

Chapter 2: Networked Platform Governance....... 51

Chapter 3: The Trust Project: How to Train Your Algorithm 92

Chapter 4: Demonetization, Tiered Governance, and the Perception of Unfairness....... 128

Conclusion: Mediation versus Moderation in Platform Governance....... 159

Appendix A: Panel Presentations at the Content Moderation at Scale Conferences 173

Appendix B: Interview Protocol for Trust Project....... 176

Appendix C: Recruitment Email for The Trust Project Interviews....... 179

References....... 181

Introduction and Method

In the fall of 2015, I had just finished a project I had just completed with Data & Society Research Institute on the impact of data on policing. My contributions — which were two primers on biometric surveillance (Caplan, Ajunwa & Rosenblat, 2015) and the Police Data Initiative (Caplan, Rosenblat & boyd, 2015) — built off work I had done in the years prior looking at the role of data (particularly *open data*) in increasing transparency and accountability in government. I was a PhD student in Journalism and Media Studies but had, thus far, in my PhD work, avoided doing any research on journalism and media (I spent much of my time during my early coursework writing about other issues in information and data policy, such as privacy).

When Data & Society received a small grant titled “Who Controls the Public Sphere in an Era of Algorithms?” I was confident that this would be the first time my personal research interests would not be swayed by a project I was doing for work (thus far, every new project I had taken on had pivoted my scholarship). As my colleagues and I dove into the research, I was even more certain. The question seemed unanswerable; the thought of even trying to define the terms included in the title of the grant — “algorithm,” “control,” and “public sphere” — seemed like it would require years of study. Added to this, the information environment in which we were asking this question, this era of search and social media, was constantly changing, with platform companies like Facebook and Twitter adding on new features, and making alterations to the algorithms they used to prioritize content, constantly. And the broader political environment in which this question was being asked, with the emergence of Donald Trump as the likely Republican candidate for the 2016 election, was, lacking a better word, fraught.

Knowing we did not have the answers ourselves, we did what Data & Society did best: we hosted a workshop. Normally with a workshop like this, our role, as an independent research institute, was to outline the terms of the current debate, so they could be translated across a diverse set of stakeholders (such as academia, government, nonprofits, and the technology industry). The terms of this debate, however, were not known. We had instead a set of questions and concerns stemming from recent controversies that emerged as more communication merged onto social media platforms and search engines. These concerns included the potential implications of Facebook's then-recent "emotional contagion' study (boyd, 2016), anxieties about their experimentation with "nudging" users to the polls (Zittrain, 2014), potential cooperation between Silicon Valley and governments (Yadron, 2016), the increased role of metrics in newsrooms (Hudson & Fink, 2014; Anderson, 2011), and the role algorithms were potentially playing as the *new gatekeepers* of information (Napoli, 2015; Tufekci, 2015). But these concerns also included the potential spread of misinformation over platforms — the journalist Adrien Chen had recently documented the role of the Russian government in the creation of "troll armies" used to spread influence public opinion following the annexation of Crimea (Chen, 2015) — and how divergent conceptions of speech and culture would impact content moderation by platforms (Chen A., 2012; Dewey, 2015; Gillespie, 2015). Instead of just enumerating them, we wrote these up as a set of case studies that, we hoped, were not too dystopic to dissuade productive debate (Caplan & Reed, 2016).

But because the terms of the debate were not yet set, we decided to focus our efforts for the workshop on gaining clarity about those terms. We were hoping that the workshop would be able to make sense of the concerns and issues that we saw mounting as more and more communication — from governments, media organizations, political campaigns, and individuals

— converged onto platforms. My goal to not pivot my interests towards these new questions quickly went awry. We convened a group of academics media, communications, and journalism studies scholars, legal scholars, information scientists, and computer scientists that had been writing about some foundations of these problems over the last decade (Data & Society Research Institute 2016). And I took the once-in-a-life experience of composing my qualifying exams around the academic scholarship of those we invited to convene with us.

The workshop gave us *some* clarity, but it also brought about new questions. For the most part, many of the issues and concerns we had pointed to in our case studies felt like problems of a far-off future. Cracks were beginning to show in the new communication of infrastructure comprised of what we referred to then as "algorithmic media companies" such as Google and Facebook were playing in the production and dissemination of news and information. But oversight also seemed even further off than the potential harms noted by experts. At the time, algorithmic accountability, particularly "reverse engineering" (Diakopoulos, 2014), seemed pivotal in terms of increasing our understanding of how platforms that used algorithms to prioritize, such as Facebook, even worked. But there was also a growing consensus among the workshop participants that analyzing algorithms, and the data they are trained on, would not be enough. This was particularly true if the goal was understanding the impact powerful behemoths like Facebook and Google may be having on the public sphere (Data & Society Research Institute, 2016).

Over the next year, many of the issues we outlined in our case studies, and at our workshop, broke through into the public consciousness. Later that year, and much to the surprise of many (Arkin & Siemaszko, 2016), Donald Trump was elected President of the United States. Questions began to emerge about whether those same "troll armies" documented by Adrien Chen in 2014, had been repurposed as a way to bolster public opinion online in favor of Donald Trump (MacFarquhar, 2018). Immediately following the election, a report by Craig Silverman that purportedly showed how “Viral Fake Election News Stories Outperformed Real News On Facebook” (Silverman, 2016) spurred broad concerns about how changes to the information ecosystem might be facilitating the spread of false information online. In particular, the public’s interest, and their ire, began to turn on online intermediaries like Facebook and Google and the role they could potentially be playing in shaping political life online.

Trying to “Fix Democracy” from a Conference Room

This moment was met with a sense of urgency from academia and civil society. Over the next year, I and many other academics working on these issues, were invited into numerous workshops asking us to consider the role platforms may have on the future of democracy (Yale Law School, 2017; Harmful Speech Online, 2017). Huddled together in ivy-league seminar rooms, a diverse set of stakeholders (at least from the perspective of Yale and Harvard) were tasked with “fighting fake news” and solving the problems of “harmful speech online.” Representatives from the major platform companies — Facebook, Google, Wikipedia, Twitter, and others — were always in attendance, though they mostly remained in the background– listening to the academics tasked with explaining the scope of the problems unfolding over their sites.

My research questions were born inside these rooms, watching these interactions between platform representatives, news media (anxious about changes to their industry), academics with a renewed interest in propaganda studies, and civil society organizations concerned about the impact that both a lack-of-and-too-much regulation of speech online, could have on the participatory internet. The sheer number of these kinds of events[1] gave the impression that we had reached a sort of "critical juncture" that could shape the future of the Internet (McChesney, 2007). Particularly with concerns about the spread of false information over social media, and the impact platforms were having on journalism in general and in their new role as "media companies" (Napoli & Caplan, 2017), there was an impression that the internet itself was standing at a crossroads: How should we reconsider the benefits and downsides of the downsides of a participatory culture that had flourished online, with the need to increase access to reliable, trustworthy, and credible information? And in advocating for a bolstering of more traditional media at the expense of social media, would that mean shutting out the voices of marginalized communities that had been left out of media in the past? Through these ongoing debates about fake news, mis-and-disinformation, "filter bubbles"(Pariser, 2011) and echo chambers, content moderation, the impact of platforms on journalism and the role of algorithmic recommendation systems it was clear that there were competing visions of what the internet *should be* in the future and the role *we* should all have in shaping it.

The research in this dissertation is centered on many of these concerns and the emerging dynamics between platforms as they mediated between different stakeholder groups, each with their vision for the internet and for the development of content standards online. My research

---

[1] The field has begun a crowd-sourced list of similar events from this period that can be seen here: https://docs.google.com/document/d/1xtJ9oQ97hmVyR7dF68OLA9JStPYNWCMCuJwF1UHTTzo/edit?usp=sharing

tries to answer the following questions: How are platforms engaging with stakeholder groups such as academics, users, advertisers, news media and other media organizations, civil society — in how they develop content standards? How does this engagement fit into their organizational goals as companies, versus how might it work to bolster an image of plurality and participation to counter concerns about platform power? How are platform companies acting as *mediators*, not of information in and of itself, but between different stakeholder groups, such as between traditional media or amateur content creators, that are both impacted by the development of content standards by platforms?

Methods

This dissertation uses a variety of different methods and approaches to understand the challenges of content regulation and platform governance. It uses a case study approach, informed by my participation and observation within spaces where debates about content standards were unfolding, to examine the struggles and tensions between competing and collaborating stakeholder groups — platforms and government, traditional media and platforms, and independent creators and traditional media. Using a combination of participant observation and policy and discourse analysis, and explores the values and interests motivating these interest groups setting standards for content online (Aligica, 2006). Though this approach does not provide a holistic or complete picture of the current state of platform governance, it offers a window into the inter-and-intra-organizational relationships that are structuring platform content regulations.

Gaining access to platforms — as companies, and as technologies — has been the main challenge for research in platform governance and regulation. For this reason, there have been a variety of methodological approaches designed to gain access to either the underlying technologies of platform companies, or their legal and economic logics. Just as platforms do, this type of research has spanned across several domains, including research on how platform companies impact areas like criminal justice, education, healthcare, and media. Past research on platforms and their governance has framed this issue of limited access and opacity (Burrell, 2016), to both the algorithms underlying them (Diakopoulos, 2014), and the institutions that produce or use them (Caplan & boyd, 2018). Because of their complexity, and their tendency to be closed down to external investigation, platforms, and their algorithms, have been referred to often as "black boxes" (Pasquale, 2015). This metaphor is drawn from similar analyses within science and technology studies, which has treated established bodies of knowledge, like science, or technological innovations, as "black boxes," whose "contents and behavior may be assumed to be common knowledge," with little understanding as to their inner workings (Pinch & Bijker, 1984).

Researchers have devised many methodologies to gain access to these spaces, however, the methods researchers use depend significantly on whether they consider the *technology*– the algorithm — or the company, as the "black boxed" object of study. Diakopoulos (2014) pioneered a version of "reverse engineering algorithms' input-output relationships" to investigate algorithmic decision-making (p. 2). This type of methodology has been effectively used by journalists, like Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner (2016) to investigate bias in machine learning software used by court systems across the United States to predict potential recidivism rates. Christian Sandvig and colleagues (2014) have proposed a methodology around "auditing algorithms" to detect similar issues like discrimination by internet platforms. These types of approaches towards *investigating algorithms* have been used as the basis for proposed regulations on algorithmic accountability, such as the Algorithmic Accountability Act, which would task the Federal Trade Commission with creating rules for evaluating automated systems (Algorithmic Accountability Act, 2019).

Gaining access to algorithms only, however, offers little information about the context of their creation. For this reason, researchers within this field have looked to other methods to grapple with the social, cultural, and economic contexts in which algorithms and other technologies emerge. These investigations tend to focus on the social production of algorithms and other technologies underlying platforms, their embeddedness in culture, or as Seaver (2017) notes, "as culture." Danah boyd and I (2018) have offered an analysis of algorithms as akin to bureaucratic logics, and suggested that we can examine the private company and their algorithms, through a study of how their algorithms enact dependencies across an industry. Whether algorithms are viewed *as culture* (Seaver, 2017) or *as organizational practice*, as we (Caplan and boyd, 2018) suggest, researchers must use methods that ground the study of algorithms within their broader cultural and political contexts. Angèle Christin (2020) recommends ethnographic methods as a suitable approach for their examination. She offers the technique, "algorithmic refraction" which entails "paying close attention to the changes that take place whenever algorithmic systems unfold in existing social contexts" as a way to grasp the varying ways that algorithms function as "prisms that can reveal existing priorities within groups, organizations, and fields, as well as their changes over time" (p. 907).

These scholars above, particularly Christin (2020) and I, are speaking to a particular challenge in the study of algorithms: the limitations of transparency. As Mike Ananny and Kate Crawford (2018) have argued, transparency — in this case, revealing the algorithm or data used to train it — does little to unveil the "assemblage" of human and non-human actors that make up the algorithmic system (p. 983; Ananny, 2016). Holding these systems accountable, they argue, requires being seeing beyond "any one component of an assemblage but understanding how it works as a system" (p. 983). What Ananny and Crawford are demanding from us is a way to bring the actor-network — the totality of interactions (and non-interactions) which comprise a sociotechnical system — into view.

Given the complexity of algorithmic systems from a sociotechnical perspective, this task feels enormous, if not impossible. To respond to this challenge, my work — which uses a multi-perspectival approach to how content standards are enacted and negotiated by competing stakeholders — is an attempt to ground these systems in their social and institutional complexity. For the last several years, I have been working as a Researcher for the research institute, Data & Society, which places a focus on studying the cultural and social dynamics of data-centric technologies. In this work, a focus is placed on the *sociotechnical,* and the interactions between social and technical factors (Bijker & Pinch, 1987). Using approaches drawn from the social construction of technology (SCOT) (Pinch & Bijker, 1984) and actor-network theory (ANT) (Latour, 2005), the work I have done at Data & Society places an emphasis on examining how different social groups understand the problems associated with sociotechnical systems. The theories I draw from (SCOT and ANT approaches) emphasize a multi-directional view of a concept, or artifact, to understand how multiple relevant social groups embody "a specific interpretation of an artifact, negotiate over its design, with different social groups seeing and constructing quite different objects" (Klein & Kleinman, 2002). This approach places an emphasis on relational struggles between both human and nonhuman actors (Latour, 2005), as well as on the notion of "translation" or the ways in which actors "impose themselves and their definition of the situation on others" at various points in the production of knowledge (Callon, 1984). For this reason, part of the focus of each of the case studies in this dissertation, as well as in the literature review, is on how different social groups mobilize, problematize, and bring different solutions to bear in their use of phrases like *fake news, credibility, trustworthiness,* and related terms (Caplan, Donovan, & Hanson, 2018).

Case Studies Informed by Participant Observation

To understand how multiple, and often interacting, social groups define the problem of credibility in the creation of content standards, this dissertation takes a case study approach, interrogating how the process of standards-making impacts affected communities differently. The choice of these case studies, and the methods within the case studies themselves, were informed by my participation in conferences, workshops, and in working groups about the development of content standards for platforms. While the dissertation itself relies on discourse analysis and interviews with key stakeholders, the choice of subjects, documents, public transcripts and participants were informed by my role as a researcher being asked to take an active role in these conversations as part of my job at Data & Society Research Institute between the years of 2016 and 2020.

I used case studies as a way to bound the analytical fields I examined while taking this unique ethnographically informed approach. There is no single definition of case study research (Cavaye, 1996), however, the case study method generally refers to a way to systematize observation (Weick, 1984). According to Cavaye (1996), the characteristics of case study research include studying a phenomenon within its "natural context," not explicitly manipulating variables, studying the phenomenon at one of a few potential sites, and makes use of qualitative methods and techniques (p. 229). Case studies can be grounded in interpretive/constructivist epistemology (Yazan, 2015). Case study research, like that proposed by Robert Stake (1995), stresses gathering multiple perspectives of those involved in the case, as a way to grapple with collectively agreed upon or diverging understandings of what occurred. This is based on the understanding that there is rarely *one* version of events which can be apprehended by the researcher. For this reason, this dissertation takes a similar set of issues — how platforms develop credibility standards for content produced by a diverse range of users — and gathers perspectives from those across different platforms, the news media industry, and from other content creators.

A constructivist paradigm to case study research works well within the diverse disciplines to which this dissertation speaks. Case studies have been used in policy research and political science since the 1930s, and have been found to be useful for understanding how social groups interact, as an approach to grasping power and influence in policy-making (Lowi, 1964). In information science and information studies, the case study approach–as a qualitative research strategy–has been proposed to counter the heavily quantitative orientation of this field (Benbaset, Goldstein, & Mead, 1987; Orlikowski & Baroudi, 1991). In their article on Critical Approaches to Media Studies, Havens et al. (2009) envision a critical analysis of media studies that uses "grounded institutional case studies that examine the relationships between *strategies*…and tactics," building their use of this method off of a de Certeauian (1984) approach. Case studies, they argue, provide a way to examine institutions through studying micro-level practices, and connecting these to broader discourses within the industry in which they are situated. And lastly, in Science and Technology Studies (STS) and Social Construction of Technology, the case study method has been used by a number of scholars, such as Pinch & Bijker (1984), Edward Hutchins (1995), and Michael Callon (1984), as a way to both situate and study technical systems within social life (and vice versa). This dissertation is no exception in its use of case studies to intervene in a field that has tended to over-emphasize quantitative approaches dependent on data gleaned from platform APIs.

One of the challenges of case study research, however, is that it is difficult both to bound individual case studies, and apply findings to other research if the case study is too specific or narrow (Yazan, 2015). This dissertation attempts to both bound the case study research within the question of *how platforms establish standards for online content*, while viewing this issue as a product of many integrated and interrelated stakeholders and issues. For this reason, this dissertation approaches this concern with *three* case studies all looking at *similar, though not identical* phenomena of standards-development for online platforms, examining the issue from the perspective of platforms (Chapter 2), the media industry (Chapter 3), and amateur content creators (Chapter 4).

Studying Interrelated Systems The Frame of Isomorphism and Relocating the Platform

I chose these three groups — platforms, the media industry, and amateur content creators — because, as this dissertation argues, they are all impacted by (and impact) the same standards, frequently deployed by private platform companies, through content policies applied by human content moderators, or through algorithmic technologies. Though pairing platforms, media, and content creators may seem self-evident currently, the impact of the platform industry on media and content creation, and on other industries, has been a matter of significant debate. This is because technology companies — whether Facebook, Uber, or Airbnb — have used terminology like “platform” to appear neutral (Gillespie, 2010), and to eschew regulatory classification in the past (Cohen, 2016).

This debate has moved forward significantly as regulatory bodies — at the city, state, and federal levels — have sought to understand the impact technology companies are having on a variety of industries (Rosenblat, 2018; Garcia-López, Jofre-Monseny, Martinez-Mazza, & Segú, 2020; Caplan & boyd, 2018). In the media industry, I have argued, with my co-author Philip M. Napoli (2017), that companies like Google, Facebook, and Twitter, have insisted in the past that they should be considered solely as technology companies long after evidence emerged they were engaging in activities akin to media companies, including the production of content, the production of infrastructure for the distribution of content generated by others, and influencing editorial standards.

In past work I have published with danah boyd, (2018) I have also demonstrated that platforms and the media industry can be brought under common view through understanding how common standards unite these industries into a common set of metrics and organizational practices. Drawing from theories from neo-institutionalism and organizational sociology, we show that platforms, the media industry, and amateur content creators can be brought into the same "organizational field" (DiMaggio & Powell, 1983). This can be done through following how changes to rules made by a central organization — enacted through algorithms or policies — can lead to changes or ripple effects on dependent organizations. We use the theory of "institutional isomorphism" introduced by DiMaggio and Powell (1983) to show how Facebook, as a centralized distribution channel for both media creators and everyday users, induces changes in the production of media, essentially serving as a standards-making body for the media industry.

This dissertation extends these theories and dives more deeply into how each stakeholder group — platforms, the media industry, and amateur content creators — engages in this process of standards-making. Through these case studies, it is possible both to see how standards-making in the algorithm era is a collective and interactive process, through which these actors are acting on each other and using methods like multistakeholderism to translate values and needs. This process includes both successes and failures, during which the terms for classification are defined and redefined by the stakeholders involved.

Overview of Case Studies

This dissertation uses a comparative and interactive case study approach, informed by my ethnographic methods, to examine the issue of standards-making within platform governance. It is separated into three different case studies, which all differ in scope and breadth. The method used in this dissertation is unique in that each case study used is conducted at a different level of analysis: institutional, organizational, and lastly, at the level of the user. Each empirical chapter in this dissertation has its own methods section, which goes much deeper into the process through which I gathered the materials and evidence to make my claims. In this section, I will provide a broad overview of the level of analysis, and the actors/stakeholder groups considered in each case study to introduce the multi-perspectival approach I have taken within this work.

Institutional Case Study: Platforms

The first case study (Chapter 2) examines how policy representatives across the platform industry, speak about how they integrate expertise and feedback from external stakeholder groups. This case study takes place at the institutional level, looking at how platforms — particularly in making decisions about how to assess the credibility and authority of content — have distributed the responsibility for content policy to external stakeholders. To conduct this analysis, I used public statements from platform representatives across the industry, corporate blogs, financial reports, trade press, and attended three of the Content Moderation at Scale conference series — attended as well by platform representatives of all the major platforms — which occurred between 2018 and 2019 (number of statements analyzed = 220), and numerous public documents were gathered over the period from 2017-2020. This case study uses a comparative approach, to make the argument that *networked platform governance,* or distributing responsibility for content policymaking across external stakeholders, is a broader trend within the platform industry, and is not isolated to one specific company (or even to one specific team in a platform company). This chapter uses discourse analysis (described below), as well as participant observation, to analyze how platforms are framing their motivations for external outreach, and places these motivations within the context of media policy throughout history, as well as broader cultural discourse regarding the role of expertise.

Organizational Case Study: The Trust Project

The second case study, in Chapter 3, takes a more directed approach, looking at one organization in particular — The Trust Project — and their interactions with platform companies. This case study is distinctly not comparative, and is intended to better highlight what interactions with platform companies may look like over a period of time. This case study relies on primary documents and three semi-structured interviews with The Trust Project's leadership over 2018-2021, conducted over Zoom. In total, I analyzed around 75 news and trade press articles, around 10 corporate blog articles, and around 40 documents made available to me both privately and publicly by The Trust Project as part of this chapter. The case study uses these documents to follow this organization — formed primarily of news media organizations operating in the United States and abroad at various levels — as they worked collectively to develop a set of 'standards' for what they argued could be used to evaluate information credibility. What this case study does not include, however, is the perspective from the platform industry, and focuses instead entirely on the perspective from the external stakeholder group as they worked to navigate platform values and bureaucracies from the outside, and how their relationships between The Trust Project and platforms impacted the effectiveness of their efforts.

Individual-Level Case Study: YouTube and Content Creators

In the final empirical chapter (Chapter 4), I conduct a case study on the demonetization debate on YouTube (Caplan & Gillespie, 2020), focusing on the perceptions and experiences of users as they experience changes to algorithms and content policies that reflect the inter-and-intra-organizational dynamics outlined in the previous two empirical chapters. It uses the debate surrounding "demonetization" on YouTube — the process through which advertising revenue is removed from YouTube videos in response to concerns about advertiser-friendliness — as a strategic entry point to examine how creators are responding to platforms like YouTube as they mediate between different user groups (in this case media organizations, professionalized content creators, and amateur content creators). This chapter relies on publicly available videos from YouTube creators — around 90 in total — to understand how YouTube creators perceive shifts in YouTube's policies as enacted through YouTube's revenue-sharing agreement. It also relies on corporate blog posts, collected using the WayBackMachine, to follow shifts in the YouTube Partner Program as announced and explained by the company, as well as trade press coverage of the demonetization debate. Sections of this chapter were previously published in a journal article co-written with Tarleton Gillespie titled "Tiered Governance and Demonetization: The Shifting

Terms of Labor and Compensation Economy." The goal of this case study was to present the other side of this dynamic — how amateur content creators are impacted by the relationships that platforms are (or are not) making with organizations and individuals. It is oriented around understanding how networked platform governance dynamics are perceived by those who have weaker institutional ties to platform companies, who may be only experiencing these dynamics from the outside.

Methods within Case Studies

Though each case study uses slightly different methodologies, each draw on a common set of qualitative research methods used within areas like public policy, media studies, and information studies. For instance, this work incorporates policy and discourse analysis to understand how different stakeholder groups may be impacting (or not impacting) decisions about how platforms like Facebook and Google prioritize and classify content through algorithmic models. Thomas Streeter (2013) has advocated for the use of discursive approaches to media policy, including discourse analysis (Streeter, 2013, p. 488; Streeter, 1996) In *Selling the Air*, Streeter makes an explicit case for interpretive "rereadings" of legal and policy processes (Streeter, 1996; Streeter, 2013, p. 495). He argues that discursively oriented work on policy contributes to the "understanding of the social construction of modern institutions." Discourse analysis in the realm of policy tends to take as its object struggles over concepts or definitions used by different stakeholders in the governance process, such as that occurring now between different industry actors (platforms and media companies), government-actors, and civil society organizations over the issue of disinformation and credibility of content. Discourse analysis and argumentative policy analysis place an emphasis on language as a key feature, and "necessary component" of policy analysis (Gottweis, 2007, p. 238). I use Critical Discourse Analysis (CDA) in the analysis of policy documents, press releases, mainstream news stories, and official industry and regulatory reports. Critical Discourse Analysis, along with other similar methods in policy analysis, such as the argumentative turn, or Fairclough's Textually Oriented Discourse Analysis (TODA), places an importance on discourse and concept definition as a constitutive process (Epstein, 2012; Stevenson, 2009; Dawes, 2007). This research will use policy and discourse analysis for the study of documents emerging from relevant stakeholder groups, as well as interactions between stakeholders.

Part of the challenge of this type of method of analysis currently is the need to identify these key stakeholders, apart from the independent commissions that have typically governed or proposed such conceptualizations in the past. Documents are therefore also used for purposes of stakeholder mapping, which was used to determine the broader institutional and social field in which organizations are currently involved in discussions of content classification. Mapping is necessary to understand which organizations are currently involved in discussion about news media content classification and limiting the spread of fake news, to determine the roles of stakeholder groups and organizations involved, to identify areas of conflict or collaboration between stakeholders, and for assessing the relative impact of strategies being employed by organizations currently involved in impacting platform decision-making. These techniques are useful for not only telling the story of how policy decisions are made, but for evaluating and determining which interventions or relationships may be most predictive of platform changes in the future. Stakeholder mapping is useful for illustrating both formal and informal ties both within and between organizations, and illustrating the interactional dimensions of organizational life, to describe both the broader environment in which organizations are acting, and constraints, collaborations, or conflicts occurring between organizational and institutional actors.

As this dissertation will demonstrate, platform companies have become mediators for a broad range of interest groups, depending on the industry in which they are operating. For this dissertation, I have chosen to focus on platforms, media, and amateur content creators. These actors are central to the current debate regarding authority and trustworthiness online. However, there are other stakeholder groups — namely, advertisers — that I have not included explicitly within this group of case studies. And yet, the interests of advertisers are present, particularly in the case study on content creators and demonetization on YouTube. As will be shown in that chapter, particularly with the introduction of programmatic advertising, advertisers have frequently pressured platforms to ensure content on their platforms is more "advertiser-friendly," i.e., predictable and uncontroversial. A future iteration of this work would address the role advertisers are playing in the creation of content standards over platforms more explicitly, focusing, for instance, on the role advertisers play in the creation of block lists (Yin & Sankin, 2021).

Overview of Dissertation

The goal of this dissertation is to provide a holistic perspective on a platform governance concern which unfolded over the last several years: How are platforms implementing policies and standards geared around credible information? How are these concepts socially defined and determined through the relationships platforms make with external stakeholder groups? How do power dynamics between platforms and external organizations impact their ability to shape internal platform policies and norms? And lastly, when platforms act as mediators in balancing their own needs and goals, with the perspectives of the stakeholders they engage with in setting policies, how is this process interpreted by those *outside* of this process, those on the outside, i.e., the users and creators who may feel (justifiable or not) that their perspectives are being left out of platform policy.

Trying to grasp a dynamic as complex as this, which involves so many dynamics internal and external to platform companies, and so many stakeholders, is a difficult task to accomplish elegantly. In the second chapter, I conduct a literature review on trust and credibility and media, and where this has begun to intersect with questions and concerns within the field of *platform governance*, a burgeoning discipline that examine show platforms both govern users and are themselves governed (Gorwa, 2019). This literature review provides the broader social and political context and history through which the following empirical chapters should be read, however, each chapter has its own relevant literature review to ground each individual case study. Each chapter also contains its own method section.

In Chapter 2, I conduct an institutional case study of platform companies and their efforts to engage external stakeholder groups. I connect these efforts to scholarship on "networked governance," a concept proposed by Sorensøn and Torfing (2005), which builds off of theories of neo-institutionalism, which works to understand how the state decentralized decision-making, particularly during periods of deregulation. The theory of *networked platform governance* explores how platform companies have used this method to distribute responsibility for policy-making and enforcement to a set of networked actors. This chapter uses statements from platform representatives, as well as an analysis of networked initiatives at platforms, to understand *why* platforms *say* they are using this form of networked governance. It then compares it to how these networked actors — often academics, nonprofits, and media organizations — have themselves spoken about their limited role in forming platform policies. This chapter uses an institutional case study approach to conclude that this approach is not isolated to single platform companies, but rather is the norm across the industry, and should be considered a fundamental part of platform governance. The chapter concludes that in taking a networked platform governance approach, platforms are becoming *mediators* between organizations and individuals in the development of policy, weighing the needs of one group against the needs of others (and, most importantly, against the platform company's own goals).

Chapter 3 builds off of this theory of networked governance, but takes a much different methodological approach, using an in-depth of analysis of one organization — The Trust Project — and their efforts to influence platform policies and how platforms prioritize content through algorithms. Going deeper into one instance of networked platform governance allows not only the power dynamics between platforms and external stakeholder groups to become clear, but it also provides a more thorough understanding of how intra-and-inter-organizational dynamics can impact who can influence platform standards, and under what circumstances.

Chapter 4 addresses another side of mediation by platforms, using the debate over "demonetization" on YouTube (Caplan & Gillespie, 2020) to understand how amateur (and newly professionalized) content creators on YouTube experience policy changes that may be the result of this networked governance. It introduces the term "tiered governance" to underscore how users in different categories of relationships with platforms *are*, or may just perceive they are treated differently, by platforms. The chapter contends that the opacity introduced by these complex negotiations behind-the-scenes, can lead users who are not involved in this policy-making process, to develop alternative explanations for policy decision-making. In this chapter, I introduce the term "tiered governance" to describe how users perceive these networked relationships; far from being more perceived as the result of more horizontal decision-making, they are seen as being a result of unequal treatment by platforms.

Chapter 1: Platform Governance: The Era of New Media Reform

Facebook knew there was a problem with misinformation spreading on its network. On January 20, 2015, Software Engineer, Erich Owens, and Research Scientist, Udi Weinsberg published an article on the Facebook Newsroom — their outward-facing corporate blog — to address the recent rise of hoaxes and "false or misleading news stories" in the Facebook News Feed (Owens & Weinsberg, 2015). Owens and Weinsberg noted Facebook was making some changes to limit the spread of this content, providing an option for users to report a story as "false" similarly to reporting it as "spam." Posts that received many reports, would receive a "warning" tag. Their distribution would also be reduced in the News Feed — the algorithmically generated feed that appears to users as their Facebook home screen, with suggested updates from friends, families, pages, and groups. This reliance on user behaviors — flagging and removing — to define a hoax or false story was intentional. As Owens and Weinsberg noted, Facebook would keep its distance in making these final determinations, "We are not removing stories people report as false, and we are not reviewing content and making a determination on its accuracy."

Despite these efforts, a year later, the problem had only worsened. In November 2016, Craig Silverman from *BuzzFeed News* published a report claiming that in the run-up to the election, a selection of "fake news" stories generated more engagement on Facebook than the top election coverage of 19 major news outlets (Silverman, 2016). Published only a week after the election of Donald Trump in the United States, the article used both trend-level and anecdotal evidence to make the case that the 20 top-performing election stories shared over social media in the run-up to the election, actually came from "hoax sites and hyper-partisan blogs." Despite the fact that hoaxes and false news stories had already been recognized by Facebook the year before, Silverman's analysis noted a rapid shift from "mainstream news" to "fake news" occurring primarily in the period from July, to right before the election (Silverman, 2016).

Silverman's analysis also unintentionally hinted to the challenges in solving the problem of hoaxes and "fake news" online. As part of his report, Silverman released the data he had used to compare news sources across Facebook. Within the list of "fake news" sites, he included sites that began with the clear intention of spreading false news stories, like the *Denver Guardian* (which only existed for a short time online) (Lubbers, 2016), with far-right publications like *Breitbart*.com.

Though the latter is known for publishing many false facts, according to PolitiFact's "truth-o-meter," (PolitiFact, 2018) its close ties to both the Trump campaign, and eventually, the White House, highlight how complicated differentiating between real news, official statements, propaganda, or "fake news" can become. Making determinations on the trustworthiness and credibility of information, can often be politically, socially, and culturally fraught. This was made even more clear as Donald Trump began making use of the phrase "fake news" to criticize other mainstream media sources, such as *CNN* and *The New York Times,* to signal to his followers that *these sources*, which often critique his presidency, are untrustworthy. Since this period, and because of this use of the phrase, the term "fake news" has largely fallen out of favor within broader discussions about the problems of false information spread over social media, being replaced by other terms like "information operations" (Weedon, Nuland, & Stamos, 2017), "false news," (Lyons, 2018), propaganda, mis-and-disinformation, and "junk news" (O'Brien, 2018), among other phrases. However, the term itself has become a stand-in for a variety of concerns expressed with the media ecosystem, which includes the role of platforms, new journalistic actors, *and* legacy media institutions.

Despite Donald Trump's use of the phrase to critique news media, the more recent problem of news-like false information spread or "fake news" has been tied primarily to the growth of social media platforms and search engines. Researchers contend that YouTube, the video-platform, is becoming a source of radicalization, as users follow recommendations down rabbit holes and are driven towards more extreme positions on views they may already hold (Tufekci, 2018). In 2016, Google was criticized for boosting misinformation in its search rankings, through enabling autocompletes like "climate change is a hoax," or that the Sandy Hook shooting never happened (Solon & Levin, 2016). Twitter had repeatedly been associated with the growth of "computational propaganda" and political bots that mimic human users to manipulate public opinion (Woolley & Howard, 2016). Facebook, the largest social media network, has been strongly associated with the rise in "fake news" since they fired the human editors for their Trending Topics module in 2016, subsequently leading to false stories and hoaxes trending over the network (Dewey, 2016). The growth of public interest and concern into the role social media and search engines (and other platforms) are playing in false information spread, has led to various actions on the part of platforms, governments, and other stakeholder groups, to limit the availability of false information over these networks.

This chapter provides an overview of the current debate on platform governance over assessing the trustworthiness or credibility of information online. This issue often seems broad and unwieldy. Not only is the issue of what constitutes "fake" or "real" news and information up for debate, but the actors involved — news organizations, journalists, platforms — have become difficult to bound in the current information era. This chapter thus addresses some broad issues and debates occurring within the area of content governance over platforms. It provides context for the issues in evaluating and assessing content and sources on the internet, and places them in context with changes to the news media industry over the last several decades. It also provides an analysis of *what* we are assessing when we develop standards for content online to guard against false information, focusing on how the issue of trust and "source credibility," (Hovland &Weiss, 1951) has become central in discussions on content governance. Lastly, this chapter will articulate why this problem has been associated with social media platforms and search engines, and will summarize the existing literature on platform and internet governance as it relates to the development of content standards online.

As noted, conducting a literature review in this area is difficult because so much of the terminology used in this field remains undefined and contested. Rather than provide concrete definitions of terms, this chapter will attempt to disentangle *why* this is the case, highlighting the key issues and debates that have become central to terminology like "fake news," "trust and credibility," and "platform governance." In doing so, this chapter presents the issues of "fake news" in particular, as a "boundary object" (Star & Griesemer, 1989), through which we can begin to understand how the concerns and perceived harms of online intermediaries, are being conceptualized and mobilized by different political and social actors (Caplan, Donovan, and Hanson, 2018). Over the last several years, there has been a marked increase in skepticism and scrutiny about the role new media technology companies, like platforms and search engines, have been playing in re-orienting previously defined industries such as hostelry and transportation (Cohen, 2016; Rosenblat, 2018). The impact platforms have had on the news media industry and the information economy has been no different (Napoli & Caplan, 2017). However, though *a car* has had a relatively stable definition for the last century, entities like *news media* and practices like *journalism* have not been as fixed, particularly as more of the world has moved online, and barriers to publishing have been lowered by platforms (DiMaggio, Hargittai, Celeste, & Shafer, 2001). In many ways, the issues within this dissertation and chapter highlight how far these terms must be stretched, as content governance with regard to credible news must increasingly come to contend with other content concerns, such as propaganda, conspiracy theories, misleading content, hate speech, and sensationalism. Though these are already well-worn concerns within the history of news media, what differs now both the potential source (individual versus amateur journalist versus media organization), the scale of the potential reach, and the new *centralized networks,* platforms, where content is produced and distributed.

Always Already Fake News

How to determine the trustworthiness, or credibility, of news and information has always been a concern in communication systems. This was repeatedly noted in the wake of the more recent concerns, to argue that concerns about "fake news" in the social media era, is neither new, nor unique. In an article for *Politico*, Jacob Soll (2016) recounted how false or ideologically motivated news, has played a part in every communication era. Soll makes the case that fake news "has been around since news became a concept 500 years ago with the invention of print," which he argues is "longer, in fact, then verified, 'objective' news, which emerged in force a little more than a century ago." Soll, along with others, have pointed to other periods in which the spread of false stories and propaganda was used by powerful figures, to garner support for their cause. In one such example, Benjamin Franklin, often thought of as the leading printer of the American Republic, sought support for the revolutionary cause, by printing stories about murderous “scalping” Native Americans, who were working with the British (Soll, 2017). Other examples, such as the use of hoaxes like the “Great Moon Hoax” to sell papers in the penny press era (Thornton, 2000), demonstrate how lines between clear falsehoods and conspiracies, sensationalism, hate speech directed towards marginalized communities, and partisan or other ideological content, blurred long before social media.

It is not only the concept itself which is not new. The terminology itself — “fake news” — which has now fallen out of favor, has been used at various points over the last several decades, to describe the use of news media signifiers to spread false, or misleading information. The term was used by the TV Guide in 1992 to refer to the increased use of video news releases (VNRs) sent by public relations firms, within news broadcasts (Lieberman, 1992). Many scholars also used it to refer to the type of political satire and parody used by *The Daily Show* and similar shows that blur the difference between entertainment and news media (Baym G., 2005; Marchi, 2012). Regulatory bodies, like the Federal Communications Commission (FCC) also used the term to refer to the use of news signifiers — such as pretend broadcasts complete with chyrons — by advertisers (Eggerton, 2009). In 2011, the Federal Trade Commission (FTC) charged ten websites with similarly deceptive tactics, using the phrase to describe tactics such as mimicking news sites, and endorsements, for the sale of products (Federal Trade Commission, 2011). In each of these instances, the use of news signifiers, whether they were borrowed from broadcast, print, or digital media, to spread misleading news or satire, were viewed as the primary potential transgression.

When considering the current concerns regarding false information spread over social media, it is also this use of news media signifiers to spread problematic content — such as hate speech, hyper-partisanship, or disinformation — which remains central. However, convergence, "the flow of content across multiple media platforms," central to the digital media era has complicated the use and effectiveness of signifiers even more (Jenkins, 2006). Changes within the information economy altered the structure of news media in significant ways, making it more difficult to differentiate between different content sources. These changes included the co-existence of user-generated content and expert content under the guise of "participatory culture" (Benkler, 2006; Jenkins, 2006), which were aided by technological and social changes to news media distribution, such as news aggregation, blogging, and social media.

The Struggle for Jurisdiction in the Platforms-as-Publishers Era

In many cases, these changes have involved a negotiation of the boundaries between experts and amateurs in content production online (Baym & Burnett, 2009). The production of news media or journalism content has been no different. Professionalized journalists sat alongside a new generation producing content for social media and blogging, occupying a "hybrid user-produser role" or "produser" role (Bruns, 2008). This was accompanied by a broad shift towards what Brooke Erin Duffy calls "aspirational labor" (2018) in the creative industries, where individuals were expected to labor for free, in exchange for network connections that could (maybe) lead to payment in the future. In the news media industry, aspirational labor and the breakdown of professional journalism, was seen both in blogging culture *and in traditional media itself,* as financially struggling publications embraced the unpaid internship model in journalism in the 2000s (Salamon, 2015). These blurred boundaries in the journalism industry has also often mean re-negotiating the boundaries between work and play. Still now, minor acts like clicks and retweets, and more labor-intensive ones, like producing YouTube videos, are still acts of free labor by users that are absorbed into both the platform, and (though less so) news economies (Terranova, 2000; Andrejevic, 2008; Fuchs, 2010).

Some of these changes were welcome counter-forces to what was perceived, particularly within the United States, as what had become a corporate, centralized, and top-down media system, and were part-and-parcel of the media reform movement as it existed in the late twentieth and early-twenty-first centuries (McChesney, 1993; Pickard, 2015). The internet lowered barriers to access for news publishing and distribution, dramatically increasing the number and type of individuals who could produce and distribution information (DiMaggio et al., 2001). This gave rise to new forms of journalism, such as “citizen-journalism” and other forms of user-generated content, and outlets, like political blogs and new players in the form of digital native publications (Gil de Zuniga, Puig-l-abril, & Rojas, 2009; Bruns, Highfield, & Lind, 2012; Couldry, 2010). It was believed by scholars like Yochai Benkler (2008) and Clay Shirky (2008) that this would largely benefit voices that had been previously left out of media, particularly news media, in the past. The rise of independent ethnic or minority media from 2000-onward, was attributed by scholars like Mark Deuze (2006) to be largely a result of “community, alternative, oppositional, participatory, and collaborative media practices” that were, in part, facilitated “by the internet.” (Deuze, 2006, p. 263). Social media, such as Twitter, has also been associated with the further development of “counterpublics,” a term coined by Nancy Fraser (1990) to describe “parallel discusive arenas where members of subordinated social groups invent and circulate discourse to formulate oppositional interpretations of their identities, interests and needs” (p. 123). And yet, these affordances are ideologically agnostic. Though #BlackTwitter has thrived in the age of social media (Graham & Smith, 2016), men’s rights groups (Marwick & Caplan, 2018), and the far-right have also used these same tools to establish “spaces of withdrawal and regroupment…and training grounds for agitational activities directed towards wider publics” (Fraser, 1990, p. 68).

It was not just the emergence of new content producers that problematized the notion of expertise and authority in the information age, but rather new distribution practices. This included the rise of curation and aggregation of original news content, by blogs, alternative publishing outlets, social media, and search engines (King, 2015; Baker, 2012). Curation and aggregation across these media took many forms. Blogs and other digital media outlets, like *The Huffington Post* and the now-defunct *Gawker Media*, became well-known for synthesizing information from a number of third-party sites, adding opinion or commentary, and linking to the original article either within or at the end of a post (Isbell, 2010).

In other instances, social media platforms and search engines merely organized information in new ways that threatened the existing business model of the news media industry. Search engines, like Google News or Yahoo News, came under criticism for aggregating newspaper headlines, Rupert Murdoch argued, and took revenue away from publishers who no longer received advertising revenue from users finding stories through visiting a news publisher’s home page (Jeon & Nasr Esfahani, 2012). Other forms of social aggregators, like Reddit, StumbleUpon, or Digg, relied on users posting content to the site, that could then be upvoted by other users (Schneider, de Souza, & Lucas, 2014).

These aggregators were viewed largely as a financial threat to the news media industry, and as blurring the boundaries between media production and distribution. In 2010, the Federal Communications Commission hosted a workshop on “The Future of Media” which, among other topics, addressed some of the concerns journalists and publishers had with aggregation (Anderson, 2013). According to C.W. Anderson (2013), participants at the workshop quickly became preoccupied with defining the boundaries between “original reporting” and aggregation, with the implication being that the former was not only “essential for democracy,” but was being taken unfairly by aggregators, who were not contributing to the cost of story production (p. 5). This was contrasted by proponents of aggregation, particularly CUNY journalism school professor, and technology advocate, Jeff Jarvis, who made the case that search engines and social news sites increased distribution and audiences, potentially increasing the value of the reporting. As Anderson notes, the workshop spurred important discussions about the definitions of “news media,” and the boundaries between “original reporting” and practices like aggregation (p. 7).

This “struggle for jurisdiction” in the news media, has had even higher stakes as social media and search has come to occupy an even greater role in news media distribution over the last several years (p.13). The current duopoly of digital advertising is now dominated by Facebook and Google, which together form around 60% of the digital ad market (with Amazon, another major platform with an ad business, taking another 8.8%) (Poggi, 2019). This has meant that publishers are not only competing with platforms for ad dollars, but because many of these companies provide for digital advertising marketplaces for key advertising real estate like banners, they are reliant on their infrastructure as well (Kelly, 2018).

This flattening of production and distribution of news media content was potentially hastened by social media companies like Facebook, through technologies like the Open Graph Protocol, an application programming interface (API), that standardized all web content — whether blog, news site, ecommerce, etc. — into the same format when users posted it onto a social media site. Anne Helmond (2015) has referred to this process as the “platformization” of the web, in which platforms, like Facebook, extended “social media platforms into the rest of the web” (p. 1). According to Helmond, APIs play a crucial role in this process by setting up "channels for data flows between social media platforms and third parties," and "function as data channels to make external web data platform ready" for users. The Open Graph Protocol API was an example of this platformization of the web. Used by external sites, it assured that when users would input a URL — regardless of source — the same format of *title*, *image,* and *summary*, would appear in the Facebook News Feed. This worked to reduce and standardize many publications into one feed and format — the inverse of the form of "context collapse" described by Marwick and boyd (2011) in which social media platforms, like Twitter and Facebook, "flatten multiple audiences into one." This standardizing is perhaps how some deceptive sites, which mimicked URLs from major news organizations (such as the ABC spoof, *ABCnews.com.co*) were able to entice users into clicking through from the Facebook News Feed.

Though these changes were occurring alongside broader financial and cultural changes to the news media industry, these shifts present particular challenges to the issue of fake news and disinformation online. In each case, the internet has blurred lines that have previously been established as a way to signal authority and credibility in information. In many cases, this has been a positive shift — the internet has enabled a wider diversity of voices, left out of corporate and top-driven media, to be able to speak to their own experiences, providing new evidence and facts that had been left out of the common narrative. But because of these blurred boundaries, those tasked with defining good versus bad media, real versus fake, expert versus non-expert, have a very difficult, if not impossible, task.

Trust Issues

These blurred lines have coincided with a period of increased distrust in the media within the United States. This rising distrust has been developing for some time. According to a Knight Foundation and Gallup poll, the percentage of U.S. adults who said they “have a great deal or a fair amount of trust in the media” fell from 54% to 32% between 2003 and 2016, increasing again to 41% in 2017 (Gallup/Knight Foundation, 2018). There are a number of potential factors contributing to this decline. Turcotte et al. (2015) have pointed to factors such as increased competition among news outlets, which they argue has heightened “negativity in news, interpretive reporting styles, and attention to partisan news sources” (p. 521). In the United States, distrust in media is growing among both Republicans and Democrats (though Republicans distrust news much more, at a rate of 9 in 10 compared to 4 in 10 for Democrats) (Gallup/Knight Foundation, 2018). Distrust also falls along demographic lines, with men and whites more likely to say they have less trust in media than women and non-whites, despite being most represented in that same media (Abbady, 2017). This is a downward trend that can be traced to the 1990s, during which the media saw a downturn in its popularity as a societal institution, following what could be perceived as its heyday of the 1970s and 1980s (Ladd, 2010). At the same time, research has shown that individuals tend to view their own preferred information sources as *more trustworthy*, as well as their own local news outlets (Arceneux, Johnson, & Murphy, 2012)

Many of the current debates regarding false information spread over platforms entail a discussion of the importance of media to democracies. Citing public sphere theorists, such as Jürgen Habermas (1964), Nancy Fraser (1990), and Michael Warner (2002), these discussions underscore the need to have (multiple and often conflicting) public spaces — including newspapers and other media –to debate and facilitate societal consensus, and provide a counter-power (or many counter-powers) to government power. The internet's potential for consensus-making, in line with Habermasian theory, was touted by early Internet advocates like Nicholas Negroponte (1996), and John Perry Barlow (1996), and Yochai Benkler (2008). These advocates assumed the internet would change the way individuals consumed and accessed news and political information, bypassing the top-down media model that had become so dominant within the 20th century, and interact directly with political representatives (Mickoleit, 2014), with the potential to fundamentally alter civic participation. Though scholars like Zizi Paparachissi (2002) warned against being too optimistic, noting "access to the internet does not guarantee increased political activity or enlightened political discourse," it is the most recent controversies regarding false information and potential propaganda spread over social media, that have underscored how this early optimism may have been displaced. And yet, the extent of the problem, i.e., exactly how the internet has impacted news media production and consumption, and the subsequent impacts on democracy, is not well understood, particularly as news media, amateur content creators, and friends and family continue to compete for trust and attention within the new information landscape.

Though trust in institutional media has reached a new low, there is evidence that trust has been displaced into other individuals and entities. As social media and search engines have become the main ways to discover, share, and consume media content, researchers have begun to re-examine the impact of friends and family, as well as other online influencers, on trust of information (Turcotte et al., 2015). This research builds off of the two-step flow of communication model, introduced by Paul Lazarsfeld and colleagues (Lazarsfeld, Berelson, &

Gaudet, 1948); Katz, 1957; Katz & Lazarsfeld, 1955), which holds that political information diffuses in a unique “two-step process,” where information flows from sources like radio and print, to “opinion leaders,” or those perceived to be influential and trustworthy individuals that others (referred to as “opinion followers”) could turn to for advice on issues (Lazarsfeld, Berelson, & Gaudet, 1948, p. 151; Turcotte, York, Irving, Scholl, & Pingree, 2015). As Turcotte and colleagues show, the move towards social media as a source for political information and news, has meant that news recommendations of trustworthy opinion leaders — friends and family a user perceives as positive — can impact whether a user trusts a news source. *Trust* in this sense can be misleading, though, as research demonstrates that individuals often assess news biased towards their opinion as more credible (Iyengar & Hahn, 2009).

But this has broader implications beyond the opinions of trustworthy friends and family. Online influencers, on Twitter, YouTube, and Facebook — not necessarily tied to an institutionalized media company — are having sway on the political information individuals consume online. This has been accepted within the marketing, public relations, and business literature for some time, with numerous articles on the power the “influencer community” has to wield power “over the perception of brands and companies” (Booth & Matic, 2010). Evidence suggests that online influencers, bloggers, and YouTube stars, are judged according to different criteria when it comes to assessing “source credibility,” i.e., the believability of a communicator, according to their expertise or trustworthiness (Hovland, Janis, & Kelley, 1953; Flanagin, Metzger, Pure, Markov, & Hartsell, 2014). Much of this research has focused on the impact of blogs versus traditional media sources. For instance, Johnson and Kaye (2004) found individuals who rely on blogs for political information tend to judge them as more trustworthy than other sources of information (Johnson & Kaye, 2004). Research by Carroll and Richardson (Carroll &

Richardson, 2011), also suggests that blogs changed the criteria for trustworthiness and credibility, replacing qualities like "expertise, accuracy, and lack of bias" with "interactivity, transparency, and source identification" (Flanagin & Metzger, 2017, p. 421). Other research collected by Flanagin & Metzger (2017) suggest that information gained through social networks is more likely "to be believed more than the same information on Web pages," and that networks like YouTube may worsen the impacts of biased political information (Garrett, 2011) because bloggers and influencers select what information they provide to their followers to be in line with their beliefs and the beliefs of their followers (Wallsten, 2011; Garrett, 2011). This could problematize classic public interest media imperatives, such as promoting information diversity (Napoli, 2001). At the same time, studies have demonstrated that social media have a small, but significant impact on political beliefs (Garett, 2019). Newly democratized media have thus supplanted, in many ways, existing hierarchies and institutional orders, while at the same time introducing new institutions — platforms that host these diverse content streams — which are simultaneously more centralized and opaque, while facilitating this distributed media and trust.

Considering this, there are a number of reasons to be concerned about credibility online. At each stage of the information ecosystem as it flows over platforms, there have been concerns raised about verifiability and reliability. This includes the *source of content* (i.e., is this news or opinion, is the person reliable), as well as the verifiability of the individual and their intention in sharing the content *as a source* (is this account real or a bot, has this person been paid to spread this message). But it also includes the *metrics* used to amplify and determine relevance over algorithmic systems that use user (or bot) behaviors to determine position in a feature like Facebook's News Feed, or a Google search result, which can be easily gamed by "fake clicks, mouse movements, and social network login information, that mimics human users (whiteops, 2016). Fake followers also bolster a lucrative “influencer economy” that has grown online, inflating a celebrity, pundit, or political figure’s popularity, and potentially, their perceived authority and trustworthiness (Confessore, Dance, Harris, & Hansen, 2018). Social media platforms themselves have also been found to inflate *their own metrics*, such as video views, as a way to increase appeal to advertisers (Welch, 2018).

At the center of many of these concerns regarding trustworthiness are online intermediaries; social media platforms and search engines that not only facilitate the production and distribution of content, including news media, from users all over the world, but provide the infrastructure for these networked interactions. Platforms themselves, and the technology industry, are in a period of heightened distrust following decades of cultural and institutional technological utopianism (particularly within the United States). This ebb and flow has mirrored similar waves of optimism and pessimism present in societies with the advent of other new technologies (Marvin, 1988; Winner, 1997; Agre, 1998). With the internet, this unbridled hope took on a similar form. In the 1990s, advocates like John Perry Barlow (1996) declared the internet a new stateless space, free from the “tyrannies” of governments. In the 2000s, the advent of social media, and the promise of Web 2.0 to “Broadcast Yourself” brought new opportunities for the participatory internet (Burgess & Green, 2018). In 2011, protests worldwide were heralded as presenting the pinnacle of potential for new media in social organizing, with platform owners like Mark Zuckerberg, using political events like the Arab Spring as evidence technology companies should be allowed to flourish without regulation (Wintour, 2011).

But for the last several years, particularly since 2016, technology companies, particularly those known in the FAANG (Facebook, Apple, Amazon, Netflix, and Google) have been criticized over the last several years — referred to colloquially as the “techlash” — due to their dominance and monopolies across several industries (Flew, Martin, & Suzor, 2020). In particular, companies like Facebook, Twitter, and Google have been critiqued due to their impact on the news media industry and digital advertising (Vaidhyanathan, 2018), and their potential role in aiding the spread of propaganda and misinformation (Singer & Brooking, 2018). Facebook especially has been under scrutiny in the area of disinformation spread (Banaji, Bhat, Agarwal, Passanha, & Sadhana, 2019). It is due to their centrality, and the degree to which individuals all over the world are coming to *rely* on these networks for their information, that they have become the subject of considerable public interest in individuals and regulators.

The New Media Activism

Similar concerns about content in media systems has been met with public interest media advocacy in the past, and this dissertation holds that this current moment, spurred by issues of trustworthiness and credibility online, is marking another period of media reform (2009). According to Caroll and Hackett (2006), and Napoli (2009), media reform is a rather broad area of social activism. Caroll and Hackett (2006) note it can comprise any "efforts to change any aspect of the media — 'its structure and processes, media employment, the financing of media, content, media ownership, access to media…'" (p. 84). Though Hackett and Caroll's work shows media reform have typically been characterized by a few *frames* — including freedom of the press or expression, media democratization, cultural environment and content concerns, and media justice (Napoli, 2009, p. 389) — the movement has been broad enough that, within these domains, there has been a wide range of ideological positions and agendas.

This new era is equally complex. On the one hand, it has become characterized by concerns about the spread of hate speech and other harmful content — such as disinformation

— over social media platforms, whether it is radical groups here or abroad (Daniels, 2018; Phillips, 2018), or state-sponsored propagandists (Woolley & Howard, 2016). However, there have been equal concerns about privatized governance of speech by platforms (Suzor, 2018; Caplan, 2021) and their power to shape information flows, and the impact of new regulations proposed by governments, to limit speech rights online (Lim, 2020). Other areas of interest and targets for media reform have spanned the ideological spectrum. The issue of fake news and disinformation highlights these cleavages well. Concerns about the dominance of platforms within the digital advertising and media distribution spaces, and the growth of amateur content production and the decline of traditional news (particularly local), has led to a resurgence in support for legacy media, institutional media, and public interest journalism among the left. Conversely, a generation that has grown up online, has remained skeptical of traditional media, and has viewed efforts to re-institutionalize media, as unfair to non-legacy producers (Caplan & Gillespie, 2020).

As Napoli (2009) has shown, even just within the United States, "public interest media advocacy" and "media reform" has covered a broad range of issues and ideological positions over the last two centuries. But research in this area has demonstrated the importance of key moments or "critical junctures," as well as ongoing struggles, in shaping the relationship between media and society (McChesney, 2007). Though the history of media advocacy is too broad and long to cover here, scholars like Robert McChesney (1993), Philip Napoli (2009), and Victor Pickard (2015) have traced the battles and struggles over media reform, particularly across the broadcast era. Victor Pickard has examined competing stakeholder battles in the between public interest advocates in government and commercial powers, across various events and issues, including the New World Communication Order (Pickard, 2007), the Hutchins

Commission (Pickard, 2015), and the failed Blue Book proposals of the 1930s and 1940s (2015). This more recent period of media reform has been characterized by a surge of interest in issues related to platforms and online intermediaries. This emerging field has been referred to as “platform governance” by scholars like Robert Gorwa (2019), to refer to an area of study which has grown over the last year to study the influence of “platforms” over various industries and areas of public life.

As will be demonstrated in more detail below, scholars in this field have varied objects of interest. Partly *because* the term platform is so vague and all-encompassing, there is a desperate need for specification when referring to research in this area. This dissertation specifically examines issues related to platform governance in social media and search, with respect to content concerns such as hate speech and mis-and-disinformation. It therefore falls under the broad umbrella of other work being done in this area which is examining the spread of disinformation online, content moderation (Gillespie, 2018; Klonick, 2017; Caplan, 2018, Roberts, 2019), and the interaction of platforms on the news media industry (Napoli & Caplan, 2017; Caplan & boyd, 2018). This dissertation is thus referring to the governance of *content* and *content platforms.*

Platform Governance: A New Era of Media Reform

This dissertation works to locate these issues of mediation, trust, and the development of content standards, within the realm of media and information policy more generally, and within the burgeoning field of “platform governance” specifically. It considers platforms — their organizational practices and technologies such as algorithms — within a broad theory of “media governance" put forward by Manuel Puppis (2010) which looks beyond just the influence of just the state, to focus on "collective coordination in general" which would include a mix of governing from both public and private actors at various levels, including platforms themselves, industry actors, users, government bodies, and other efforts. Considering the governance of platforms within this broad frame is not only useful but *necessary*, given that the technology industry emerged during a period of *communications deregulation* (particularly within the United States)which has subsequently limited the potential role of government in platform regulation (Flew, Martin, & Suzor, 2020). As will be demonstrated, assessments of what content is credible or trustworthy, will become a matter of debate between stakeholders within platform governance, though heavily leaned towards platforms, with increasing pushback from industry groups and government. The major issue has become *who* should be making standards for content online being distributed over private platforms, and how will these decisions be mediated by existing tensions between stakeholders (for instance, between government and private platforms, between platforms and news media, and between legacy news media and amateur content producers).

The concerns levied against platforms and the algorithms they use, are often not necessarily technical, but cultural. This is due to an increased recognition that platforms are organizations that embed social values (that often go unexamined) in the context of collection, sharing, and distribution of information (boyd & Crawford, 2012; Bowker & Star, 2000). For the last several years, much of the focus has been placed on the *algorithms* structuring information. This work attempts to place a focus on the organizational and inter-organizational, viewing technical processes like algorithmic prioritization as one way to enact standards, alongside others, such as content moderation (Chapter 2), the classification of news content by platforms and by the news media industry itself (Chapter 3), and the categorization of users by platforms (Chapter 4). According to Bowker and Star (2000), standards can be defined as “agreed-upon rules for the production of material or textual objects” that “span more than one community” and unite practices over time and space, through a common set of terminology or metrics (p.13). This dissertation grounds these standards conversations both within the organizational contexts of platforms, and within inter-organizational dynamics.

Within governance, categorization issues impact every level of analysis, with important concerns such as the classification of certain organizations and technologies bringing into view new questions about what standards should apply. For instance, there is still disagreement about whether we should consider *platforms* separate from *media* or *information* within policy. This is important not only to locate the norms and standards with which we hold companies accountable, but to understand *who* the significant actors are in any governance debate. Part of the difficulty stems from the fact that the “platform” industry is very difficult to define and bound. Scholars like Tarleton Gillespie (2018) have made efforts to define the term, pointing to shared features of platforms, such as their capacity to “host, organize, and circulate users’ shared content or social interactions for them,” without producing ‘the bulk of that content” (though companies like Facebook and YouTube frequently produce content alongside what is user-uploaded) (p. 18). He also notes that platforms are primarily built on data and infrastructures for data production and analysis, whether it is used for customer service, advertising or profit. His last criteria — that platforms *moderate* content online — adds an important oversight and mediation role of these companies, and signifies that platforms are often characterized by their power to shape and control what users post and consume. This counters, intentionally, the previous position put forward by platforms that their defining feature was their *neutrality*, their capacity to host content, commerce, and communication, while not directly influencing it (Gillespie, 2010).

It has also been difficult, from a governance perspective, to locate platforms within a clear regulatory domain, due to their tendency to operate in many capacities at once, and their use of language of the technology industry that helps them evade typical classifications built for the industrial era (Cohen, 2016). Using Gillespie's criteria, there is still a wide range of businesses that could be classified as platforms. This has made it difficult to draw contours around businesses operating primarily in the transportation (Uber, Lyft), hostelry (Airbnb), or media (Facebook, Netflix) because though they may impact on industry in particular (Caplan & boyd, 2018), their business is predicated primarily on data, and facilitating the exchange of others they claim are operating in those domains. And yet, there is a much greater understanding currently as to the extent to which platform companies directly influence the activities taking place over their networks, often mediated through technology like algorithms, and other standards built into how these technologies prioritize, classify, and even compensate people, organizations, and actions. This has included, for instance, the management of workers through interfaces like Lyft or Uber (Rosenblat, 2018) and Care.com (Ticona, Mateescu, & Rosenblat, 2018), sentencing guidelines set by algorithms built by private companies used by courts (Christin, Rosenblat, & boyd, 2015), and the prioritization of news and other information over large-scale content platforms like Facebook, Twitter, and Google (Tufekci, 2015). These analyses have embraced existing theories, put forward by scholars like Lawrence Lessig, that algorithms and code serve as a regulating force on actors and behaviors in a system, similar to other forces such as markets, norms, and law (i.e., "code is law") (Lessig, 1999). The field of "platform governance" emerges from the context outlined above, and is a field of study that concerns itself with how platforms both govern users and are governed (Gorwa, 2019; DeNardis & Hackl, 2015).

This work builds on work done in a number of different related disciplines. Gorwa's article, "What is Platform Governance?" (2019), points to a broad area of research which has been focused on the governance of and by platforms, or "the social and political role of platforms divided between platform companies (as architects of online environments), users (as individuals making decisions about their specific behavior in an online group environment), and governments (as the entities setting the overall ground rules for those interactions)" (p. 3). He positions it as an outgrowth of the field of "platform studies," defined by Ian Bogost and Nick Montfort (2009) as a focus of study of digital media which investigates "the underlying computer systems that support creative work." However, it may be more accurately tied to previous work done in internet governance (DeNardis & Hackl, 2015), or platform regulation (Cohen, 2016), both of which have already addressed issues related to the regulatory state of the information age. This dissertation defines platform governance broadly as the norms, rules, and regulations through which platforms are governed, and through which they govern their own users.

Though the field of platform governance is growing and is consequential for broader concerns about platform power, understanding how to locate platforms *within* regulatory domains still matters significantly for the norms, rules, and standards to which we hold these companies to account (Caplan, 2018). The problem remains, how to differentiate behemoth companies like Google, Facebook, and Amazon which are obviously having impacts on industries like media and communications, but are also involved in other domains, as diverse as e-commerce, digital advertising, payment applications, broadband/fiber, and even "solar-powered internet planes" (Coldewey, 2018). Philip Napoli and I (2017) have made the case that certain platform companies, like Facebook and YouTube, should be considered as part of the "media industry." This is not only because they have become inextricably intertwined with the media industry, journalism in particular, but because their arguments *against such classification* (that they do not intervene editorially in content flows, that they do not produce content, and that they are not staffed by journalists) are neither true, nor grounds for immunity from classification as a media company or publisher (Napoli & Caplan, 2017). I have argued with danah boyd (2018) that platform companies (Facebook in particular), have developed strong institutional ties with the news media industry. In our analysis, we took on an organizational approach to understand the extent to which Facebook can influence editorial decision-making, journalist training, and even newsroom dynamics, within the news media industry specifically. Still, the term *platform,* though vague and unhelpful in conversations about governance, remains the most popular way to refer to companies straddling these domains.

Research into governance mechanisms over platforms have focused primarily on the processes put in place by platforms themselves. As platforms or interactive service providers (ISPs) and not "media," platforms were given broad leeway (that has since been limited more) to self-regulate. Part of this, within the United States, is due to Section 230 of the Communications Decency Act, passed as part of the Telecommunications Act of 1996 (47 U.S.C. § 230), which gives providers and users of an "interactive computer service" limited liability for "any information provided by another information content provider" on their network. This law has routinely been cited as the reason platforms cannot be held liable for defamatory content posted on their network by their users, and courts have routinely argued in favor of broad immunity for interactive service providers, including platforms (Glad, 2004). At the same time, this law provides interactive service providers, including platforms, with the capacity to "restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether that content is constitutionally protected" (47 U.S. Code § 230(2)(A)). Combined with state action doctrine (Berman, 2000), and with some exceptions (for instance, with copyrighted or illegal content), platforms are given broad leeway to regulate content on their platforms as they see fit. Though in many cases this has been embraced by legal scholars and technologists (Klonick, 2018; Goldman, 2018), with Jack Balkin writing in 2014 that he considered this rule to be "among the most important protections of free expression in the United States in the digital age" (Balkin, 2014, p. 2313), an increasing number of scholars, like Nicholas Suzor (2018) and me (Caplan, 2021), are skeptical of the form of "private governance" that has become default due to this law.

This dissertation considers platform governance within information policy and "media governance" more broadly, rather than focusing solely on the regulation of media by the state (Braman, 2011; Puppis, 2010). Past frames of analysis within information policy have looked at the role of different stakeholder groups, such as governments, corporations, civil society organizations, and the public have had in enhancing or inhibiting information flows in the past. Scholars like Manuel Puppis (2010) are helpful in orienting discussions of media governance towards broader definitions that encompass the "regulatory structure as a whole, encompassing the entirety of collective rules in society" (Puppis, 2010, p. 137). This broader definition of governance moves it away from discussions weighing the positives or negatives of different governance options emerging from the state, to the entirety of the regulatory system, ranging from the self-organization of industry and civil society to what is thought of as more traditional regulation by government (Puppis, 2010, p. 137). Puppis thus defines "media governance" as "the regulatory structure as a whole, i.e., the entirety of forms of rules that aim to organize media systems" (Puppis, 2010, p. 138). Though the role of the state remains important in this broader conceptualization, Puppis's view of governance stresses the need to understand the numerous and various forms of "management and accountability" occurring *within* and *between* media institutions. This means that relationships between media and society can both be affected by the rules structured within media institutions. Puppis refers to this as "organizational governance" or the "internal rules and control mechanisms" which make up individualized self-regulation, as well as "collective media governance" which encompasses "statutory regulation, coregulation and self-regulation that concern all media organizations in a given industry" (Puppis, 2010, p. 141). These broader frameworks of media governance and information policy stress that governance should be defined instead as "sustaining co-ordination and coherence among a wide variety of actors with different purposes and objectives," between different and competing stakeholders, like political actors and institutions, civil society, governments, corporate interests, and transnational organizations (Puppis, 2010, p. 137).

Lawrence Lessig's (2006) *Code and Other Laws of Cyberspace*, is another way to orient the governance of media beyond just the laws of the state. His work, widely cited by scholars within the area of platform and algorithm studies, stresses that there are actually four major regulators — law, norms, market, and architecture (or code) — all of which constrain, inhibit, or impact the shape of society. His familiar adage, "code is law," which emphasizes the degree to which code and hardware serve as a regulating force on different behaviors and actors within a system, is one way to understand how organizational incentives or values implicit within a private enterprise, such as Facebook, become embedded into the structure of algorithms and serve to structure or rationalize the organizations and individuals that come become dependent on its platform.

As this dissertation will demonstrate, in the development of content standards, particularly through content moderation, platforms have taken a central role in regulating content online, despite claiming they should not be the “arbiters of truth” (Zuckerberg, 2016). Like other regulatory forces, platforms, in both the policies they set and the technologies they use, operate both in terms of setting *positive* sanctions (i.e., determining what is incentivized), and *negative* sanctions for content standards (i.e., determining what is removed) (Durkheim, 1933). Though there has been work on how algorithms like the Facebook News Feed act as new gatekeepers on information, through deciding what user inputs to prioritize and make visible (Tufekci, 2015; Gillespie, 2014) or re-orient editorial norms in newsrooms (Caplan and boyd, 2018; Oremus, 2016), there has been considerably more work done recently on how platforms set standards for what to *remove*, through examinations of content moderation policies and processes (Roberts, 2019; Gillespie, 2018; Klonick, 2018; Caplan, 2018). These scholars have all focused on how platforms both set content standards, and enforce them, though these scholars tend to have a different understanding regarding the capacity of platforms to self-regulate (with Klonick and Gillespie both leaving ample room for self-regulation). And yet, more recently, governments outside the United States are increasingly pushing back on this private power to determine content standards, instituting their own laws regarding content like hate speech or “fake news,” including Germany’s NetzDG Act (Haupt, 2018), New Zealand’s recent prohibition against distributing prohibited content online, (Office of Film & Literature Classification, 2019), and Malaysia’s controversial “Anti-Fake News” bill (Lim, 2020). Industry actors, within the news media industry, including The Trust Project (Chapter 5), are also pushing back against platforms creating standards unilaterally.

This dissertation will explore the complex organizational and stakeholder dynamics currently underpinning the setting of standards regarding information credibility on the major platforms. However, this dissertation goes beyond just content moderation, to examine multiple ways that platform set standards for content in addition to the development and enforcement of community standards regarding what content is or is not allowed to remain on platforms. It uses a case study approach, looking at disagreements and debates between key actors — between governments and platforms, between legacy media and platforms, and between amateur content creators and media — to explore the interdependency of media standards-making within the information era.

--

Chapter 2: Networked Platform Governance

"We do not want to be the arbiters of truth, but instead rely on our community and trusted third parties." – Mark Zuckerberg (2016)

On November 19, 2016, Mark Zuckerberg, founder and CEO of Facebook, took to his personal Facebook page to reassure his users that Facebook was taking misinformation spread seriously. Zuckerberg was responding to recent major events. The United States election had just taken place 10 days prior, and Donald Trump, considered a long-shot by members of his own party, had just been elected President. Craig Silverman from *BuzzFeed*, had recently published stories alleging that "fake election news" had outperformed "real news" on the social media site, heightening fears that Facebook was playing a role in undermining an information ecosystem, news media, considered so important to American democracy (Silverman, 2016). Facebook's role in undermining the news media ecosystem had already been questioned in the years prior, following a study by Pew Research Center that placed Facebook as the new center of news distribution and consumption (Gottfried & Shearer, 2016). The spread of misinformation, disguised as news sites or blogs or ushered in through increasingly partisan media (Caplan, Donovan, & Hanson, 2018), was viewed as a mainly-Facebook problem as well.

While Zuckerberg acknowledged the need to take misinformation concerns "seriously" on his personal page (Zuckerberg, 2016), he had already called claims that it had impacted the election "a pretty crazy idea" (Solon, 2016). Immediately following the election, Zuckerberg was, perhaps understandably, unclear about Facebook's role in the spread, as well as its responsibility for policing misleading content in the future. In his post, he stressed that Facebook, the largest global social media company in the world that also owns the third and sixth most popular social media applications, WhatsApp and Instagram (Statista, 2021), should not be made to be the “arbiters of truth” of what constitutes “accurate content” (Zuckerberg, 2016). He stressed the need to distribute this responsibility to “our community and trusted third parties.” In the months that followed, Facebook took steps to do just that, focusing in particular on establishing partnerships with fact-checking organizations — a well-established field since the 1990s (Graves, 2016) — to make content decisions on misinformation (Ingram, 2018). Users would flag content as potentially “false news” which would then be filtered to these organizations, that were paid a (minimal) sum to review content (Ingram 2018; Ananny, 2018). Users would then also play a role in evaluating the content fact-checkers were able to mark as potentially false (a minimal percentage of content shared on the site (Allcott & Gentzkow, Social Media and Fake News in the 2016 Election, 2017)) — an indication that despite this growing infrastructure, Facebook believes responsibility for judging the truthfulness of content remains with the individual user.

Mike Ananny (2018) has made the case that in this specific case, Facebook was attempting to leverage fact-checking partners’ “different form of cultural power, technological skill, and notions of public service.” But, this has not been the only effort by Facebook and other platforms to distribute responsibility for content on the platform. More recently, Facebook has been working to establish an independent oversight board to review its content decisions (Zuckerberg, 2018), establishing an Oversight Board Trust and LLC to manage Facebook’s funding and oversee its operations (Harris, 2020). The establishment of the Oversight Board was itself the result of a long-process of consultation with external stakeholders, consisting of both a global consultation of experts and organizations that work on issues like “free expression, technology and democracy, procedural fairness and human rights” (Clegg, 2019), as well as a period of public consultation, consisting of a questionnaire and free-form questions that could be submitted by anyone (Harris, Getting Input on an Oversight Board, 2019). Out of this, they have produced draft charters (Facebook, 2019), and bylaws (Oversight Board Bylaws, 2020). They have also repeatedly engaged external actors and third-parties in conducting bias assessments of their technologies and policies. In one case, they commissioned former Republican Senator Jon Kyl to conduct a review on the perceived anti-conservative bias of Facebook (Kyl, 2019). In another case, Facebook engaged an independent non-profit organization with expertise in human rights practices, to conduct a human rights impact assessment of Facebook's response to hate speech spread on their platform against the Rohingya in Myanmar (Warofka, 2018), though not all of their recommendations were adopted (Wong, 2019). This pattern of outreach appears to be strategic, operating not only as a way to leverage "cultural power" as Ananny (2018) has pointed out, but as a way to operate across politics and jurisdictions.

This chapter demonstrates how a broad array of platforms — not just Facebook — are turning towards networked governance to distribute responsibility for the creation and implementation of content standards. It makes the case that platform companies are mitigating issues of scale and potential resource concerns through incorporating feedback, policy ideas, and labor, of media-oriented civil society organizations, academics, and users. In many cases, these efforts may be more symbolic than substantive. Though platforms have been implored by scholars like Tarleton Gillespie (2018) to "share the tools to govern collectively" (p. 212), this chapter will show that efforts to network governance can introduce new opacities and institutionalize inequality between stakeholder groups. This chapter explores the double-edged sword of these attempts to introduce participation into platform governance policy design and implementation, as well as the justifications platforms use when they seek to network platform governance.

Platforms-as-a-Node: Understanding the Current Era of Networked Platform Governance

Platform companies present complex problems for governance. Though they have tended towards monopolies and duopolies within certain domains, current United States-based antitrust law has become difficult to enforce, particularly with tech platforms, which are often comprised of many small businesses (i.e., Amazon began as a book seller, and is now a distribution and logistics business, as well as a cloud server, a producer of retail goods, etc.) (Sagers, 2019), which are often provided to customers for free (for instance with social media and search) leaving it outside a strict antitrust analysis. Platforms also often exist beyond any single jurisdiction; however, though they are subject to the laws in spaces where they operate, they often prefer to build policies and systems to one set of laws (Citron, 2018). And even within single jurisdictions, it is often unclear which laws apply (Tushnet, 2015)]

In response to emerging content regulation, or debates about pulling back Section 230 immunity (Department of Justice Office of Public Affairs, 2020), platforms do seem to be heeding calls to be more "accountable," adding or expanding policy teams (for instance, Reddit added their Trust & Safety team in January 2016) (Ashooh, 2018, 01:12:39), are working with regulators (Rosemain, Rose, & Barzic, 2018), and, are formalizing their current policies and processes (Caplan, 2018). However, platforms are also engaging in another set of policy activities that are notable as a governance strategy; networking and distributing the creation and enforcement of content policies through strategic relationships with nonprofits, academics and other experts, and their users. Platforms are thus creating and making use of the rhetoric and strategies of governance networks, or *networked governance,* in their approach to platform policy-making. And yet, though platforms frequently gesture towards more horizontal forms of governance, making clear they are consulting and learning from other groups, it's unclear whether developing and making use of networked experts and modes of feedback, fundamentally change more hierarchical platform governance processes.

Networked governance, as a governance strategy, works to leverage fields of interdependent (though autonomous) actors in a more horizontal, self-regulating, and informal approach to making governance decisions and achieving organizational goals (Sorenson & Torfing, 2005, p. 203). Governance networks, as described by Sorenson & Torfing (2005) in contrast to more hierarchical forms of government, such as state rule, and as an alternative to market competition (p. 196). As a concept, they are not new. Governance networks have been a subject of interest within political science since the 1990s, emerging at the same time as interest in the "network paradigm" with the rise of the Internet and the "network society" (Limm, 2011, Castells, 2000). They emerged at the same time as the rise of the political scientists began to trace a transition away from theories of "government" towards "governance," which marked a turn away from theories of formal governing by the state, towards the influence of other entities — private corporations, markets, multinational agreements, and other forms of distributed decision-making (Puppis, 2010, p. 135). Their importance in understanding decision-making — particularly the technology industry — grows as the role of the state decreased due to deregulation (Sørenson & Torfing, 2005, p. 202), and the role of other forces, such as markets or private governance by companies such as platforms, have increased (Caplan, 2021).

Networked governance is more conceptual than procedural; it refers to any broadening of politics or governance beyond the single "party" or entity, promising more opportunities for "cooperation, flexible responses, and collective social production" (Bogason & Musso, 2006, p. 6). The theory builds on theories from new institutionalism/neo-institutionalism and organizational sociology (DiMaggio & Powell, 1983), taking as its starting point the "demise of the isolated and sovereign actor or organization" and places an emphasis on "understanding interaction" between interdependent actors and organizations (Bogason & Musso, 2006, p. 4). Networks are generally not legal entities, and are often not bound by formal contracts, but are cooperating towards a collective goal, sharing resources and information (Provan & Kenis, 2007). Though they are often used to refer to the expansion of rule-making beyond the state — to nonprofits, citizens, industry, and other networked actors — firms, particularly the technology industry, have frequently made use of networks as a way to meet "resource and functional needs" (Powell, 1990). As will be shown, content moderation companies frequently refer to democratic ideals such as *participation, diversity,* and *consensus-building,* as well as issues of scale and complexity to explain why they rely on input from external actors in governance decisions (Caplan, 2018).

Networked governance provides opportunities for participatory platform policy-making, but it also presents concerns. It can increase the diversity and expertise of people contributing to decisions about platform policy — a major concern within the technology industry, that has long promised to address its diversity problems, but is still majority white and male (Harrison, 2019). In the case of private governance by platforms (Caplan, 2021; Suzor, 2018), networks can also insert "more negotiated or deliberative models" of decision-making into what was previously done wholly within the company, hierarchically (Bogason & Musso, 2005, p. 5), and can increase the responsiveness of internal policy teams to content issues that are posing problems for local communities. Networked governance also stands in contrast to more hierarchical and bureaucratic forms of governance, which according to work I published with danah boyd (2018), platform companies have been using through centralized decision-making enacted through algorithms. As this chapter will show, this move towards networking governance and incorporating alternative forms of feedback is part of a strategy on the part of platforms to distance themselves from a more centralized system.

But networked governance can itself also present significant problems for governance. It can introduce more ambiguity into how decisions are made, particularly as relationships with external stakeholder groups and actors remain informal and difficult to trace (Bogason & Musso, 2005). In distributing decision-making policies, it can also lead to a situation where "no one is in charge" (Stoker, 2006), mimicking other concerns with distributed responsibility of agency between human engineers and automation that have been noted by scholars like Elish in her concept, the "moral crumple zone"[2] (Elish, 2019). And though consulting external groups, using strategies like "trusted flagger" programs, and integrating user feedback on policy can increase the diversity of those contributing to and enforcing content policies, they can also *increase* power differentials, particularly when they depend on who has *access* to technology companies, which can favor those who already have power (Fischer, 2006). Decisions done through

---

[2] M.C. Elish has argued the "moral crumple zone," using self-driving cars to explore how mistakes made by automation may misattributed to human actors. She makes the case that the "moral crumple zone protects the integrity of the technological system, at the expense of the nearest human operator." See more at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2757236

networks also become more decentralized, placing them even more outside public view (Bogason & Musso, 2005, p. 8), creating more channels of political influence with potentially unevenly distributed access (Sorenson & Torfing, 2005, p. 214). And, though often more horizontal than other forms of governance, hierarchies still exist within networks, with power between stakeholder groups unevenly distributed (Powell, 1990). In the case of platform companies, that are increasingly working to consult outside experts and users in making policy decisions though maintain final control over how policies are made and implemented (Klonick 2020), turning to networks for governance may actually lead to *less* insight into how decisions are made, and more fragmented publics.

Scholars have already noted that platform are playing this regulating role, though the extent to which they locate this as part of a 'network' of interdependent actors tends to differ (Caplan & boyd, 2018). In the area of content moderation, scholars also tend to differ in terms of the power they feel platforms should have *as a regulating force.* Kate Klonick (2017), for instance, pointed out in her work on content moderation that platforms are operating as the "New Governors" of speech, even deploying legal language and frameworks that Klonick argues demonstrates platforms and their policy personnel (including moderators) are operating under not only a First Amendment framework, but in a manner "similar to that of a judge" (p. 49). In his work on the same topic, Tarleton Gillespie (2018) has still called on these companies to give up the myth of the "neutral platform" that acts as their straight-jacket towards policing content with particular societal harms (like disinformation and hate speech), and to own their responsibility. This, in Gillespie's view, would entail these companies change how they are viewing their job as "custodians of the Internet" from *cleaning up bad content,* to becoming active stewards and guiding expectations. He calls on platforms to “take on an expanded sense of responsibility,” and “share the tools to govern collectively” (p. 229).

The goal of this research is to understand the ways in which platforms say they are using relationships — with users, nonprofits, experts, and academics, and even government actors — in the development and enforcement of content standards and policies. In this sense, this research studies the symbolic side of networked governance — why platforms *say* they are networking decision-making. In the following two chapters, I will demonstrate more substantively the ways that platforms use networked governance to mediate between different interest groups as they make content decisions.

When analyzing statements made by platform representatives about the role external stakeholder groups and users play in influencing policy, several key themes emerged. The first was that platforms were often gesturing towards democratic values like participation and deliberation when referring to these relationships, and often used government metaphors or legal structures at the company as a way to structure these forms of participation. Platforms also often highlighted how these relationships helped them bridge resource gaps and other issues with scale, pointing to the variety of different jurisdictions in which they were operating and the global complexity of their operations. Lastly, and relatedly, platforms often pointed to a need to expand their expertise, using relationships with stakeholder groups as a way to address organizational complexity and a global scale, and to insert *context* into platform policy-making.

Method

To study this, I conducted a discourse analysis of public statements made by platform representatives on the issues of scale and moderation. I used statements made by platform representatives during the Content Moderation at Scale Conferences, a series of conferences oriented towards industry, academia, and nonprofits on the challenges of moderating user-generated content. I focused on two of these conferences in particular — held at Santa Clara University Law School in February 2018, and in Washington D.C. on May 7, 2018 (Content Moderation At Scale, 2018) — and I analyzed publicly available transcripts and slides from two sets of panels in which representatives from platform companies provided overviews of their company and team dynamics (see Appendix A for speakers and panels). Though I attended both conferences, I relied on video of the conference for purposes of analysis. This analysis also relied on primary documents from platform companies, including terms of service agreements, SEC filings, community guidelines, and posts from corporate blogs and websites, as well as other public statements made by representatives from platform companies in public interviews and in other research reports. To provide some of the context of how platforms have structured their guidelines about false information in the past, I used the WayBackMachine to gain access to content policies and studied their change over time.

Though the eventual aim of this research is to study examples of networked governance in practice, gaining access to these networked relationships is difficult due to the opacity of platforms (Roberts, 2018), and their tendency to shift their policies frequently. Since the Content Moderation at Scale conference was well attended by the platform industry, with over fifteen representatives from companies serving on panels, it also provided an opportunity to do some comparative analysis across platforms. The companies I focus on were those with a presence at the Content Moderation at Scale Conferences, who took part in two sets of panels focused on policy-making and operations at major platforms, and include Automattic (parent company of WordPress), Dropbox, Facebook, GitHub, Google, Match.com, Medium, NextDoor, Pinterest, Reddit, TripAdvisor, Twitch, Twitter, Vimeo, Wikimedia, and Yelp.

The companies examined here in this chapter go beyond the media/platform divide that forms the central question within this dissertation work, particularly in regard to how platforms, and their interlocutors, structure trustworthy or credible information. All platforms have rules prohibiting certain kinds of speech on their sites. The manner in which they are developed and applied — and what speech is prohibited — has varied over time, and both between and within platforms themselves. Most platform companies have had, since their outsets, rules against illegal content, and copyrighted content (Gillespie, 2018). Over time, platforms have also incorporated rules against public concerns like terrorism and extremist content, revenge porn and harassment, and cyber-bullying. Rules explicitly prohibiting content that could be defined as misinformation or disinformation have been more recent, however, a review of content or "community guidelines" shows that most platforms have had rules regarding the credibility or "authenticity" of content since their outset.

For instance, most of the companies at the Content Moderation at Scale conferences had rules prohibiting misrepresentation — of identity, clicks, or content. These were often included in categories of rules against activities like "impersonation," "spam," or "phishing/spoofing." Automattic (the parent company of WordPress), Dropbox, GitHub, Medium, Google, NextDoor, and Reddit, all have rules against this type of misrepresentation of identity or content. For Automattic, that means a user is not allowed to pretend "to be a person or organization you're not" (Automattic, n.d). This category also often includes prohibitions efforts to misguide other users as to the importance or popularity of your content, including boosting SEO (Automattic), increasing traffic (Google), using false names to deceive others (Medium), or other violations of things "real names policies" (NextDoor, and Facebook). These rules are often comparable to others designed to protect intellectual property, which collapses this notion of "real names" with copyrights and trademarks (a method used by the crowdfunding site, Patreon). In some cases, instances of "fake news" where one site is mimicking or spoofing the trademark of another site (as was the case with ABC.com.co, a spoof on ABC.com), could have been addressed through these types of rules. In many cases, these rules already existed. Using *WayBackMachine*, Automattic had similar language to their current guidelines prohibiting misrepresentation beginning in February 2015, with Dropbox having an even older history of these rules (since 2009), and Google and Match.com including them even earlier (2008 and 2005, respectively).

Some sites, like Facebook and Pinterest, have added rules to explicitly prohibit or deprioritize false content that is mimicking a news website (Facebook, n.d.), or "misinformation" in general (Pinterest, n.d.). Within the platform industry, these policies are rare. Facebook instituted their policy, which does not remove "false news from Facebook" but instead "significantly reduce[s] its distribution by showing it lower in the News Feed," in mid-2018 (Facebook, n. d.). Pinterest addresses "misinformation" in its guidelines, telling users "Don't put harmful misinformation on Pinterest," particularly when it "has immediate and detrimental effects on a pinner's health or public safety" (Pinterest, n.d.). Twitch, the online video gaming streaming site, also has rules explicitly prohibiting "misinformation" such as "feigning distress, posting misleading metadata, or intentional channel miscategorization." Wikimedia has rules against creating "hoaxes," which they consider a subtype of misinformation (Wikipedia, n.d.).

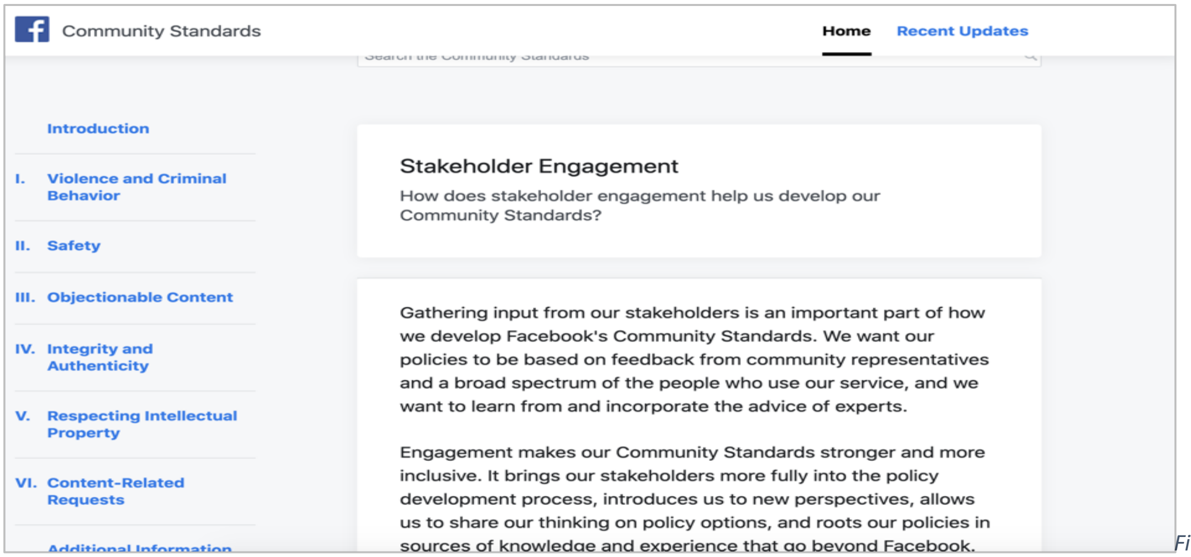
More recently, larger platforms, like Google/YouTube, Facebook, and Twitter, have included specific rules against information designed to encourage voter suppression, and "mislead people about when, where, and how to vote" (Twitter, n.d.). This type of misinformation is often couched within other categories of rules, such as "Spam" for Google, and "Authenticity" in the case of Twitter.

Networked Governance in Practice

At the Content Moderation at Scale conferences, almost all the platform representatives discussed ways they *network governance* to external stakeholder groups, including users, nonprofits, experts/academics, and governments. At the user level, platforms have long made use of users to identity and "flag" potentially offending material, and though the relationships between flagging, moderation, and policy-making have always been opaque (Crawford & Gillespie, 2014) platforms repeatedly noted throughout the conferences that they still rely on flagging by users significantly (Mcgilvray, 2018; Stern, 2018; Sieminski, 2018).

A number of platforms also referred to relationships with external stakeholder groups beyond just these user-level interactions, referring specifically to organizations, such as Lumen (run by the Berkman Klein Center), experts, government agencies, and other specific community members (such as volunteer moderators, or creators). Some platforms have institutionalized this outreach within the company's operations. Twitter has had a "Trust and Safety Council" since 2016, which is composed of nonprofits, academics/researchers, and other grassroots organizations around the world and is still growing (Twitter, n.d.). Facebook also has a team, Content Policy Stakeholder Engagement, that is specifically directed towards doing this kind of work (Stern, 2018, 48:07). They also note within their community standards that "gathering input from our stakeholders is an important part" of how they develop their content policies (Facebook, n.d.). Facebook also has a Safety Advisory Board comprised of "independent online safety organizations and experts" from around the world (Facebook, n.d.) Twitch, a live streaming platform used mostly by gamers, has also joined Facebook and Twitter in establishing a "Safety Advisory Council" comprised of external experts and Twitch streamers who will advise on content policies and procedures (Twitch, 2020). Twitter has a global network of Safety Partners (separate from their Trust and Safety Council) (Twitter, n.d.). In some cases, these partnerships impact how feedback is weighted within the company — for instance, YouTube has a "trusted flagger" program where they provide more "robust tools" to government agencies, individuals, and non-governmental organizations that are "particularly effective at notifying YouTube" of content that violates their guidelines (YouTube, n.d.)

In only very rare cases is the criteria for inclusion into these programs made visible to the public. YouTube's "trusted flagger" program is one such instance, where program eligibility is stated to include "individual users, government agencies, and NGOs" who have "expertise in at least one policy vertical" and who "flag content frequently with a high rate of accuracy and are open to ongoing discussion and feedback with YouTube about various content areas (YouTube).


Community Standards
Home
Recent Updates
Introduction
I. Violence and Criminal Behavior
II. Safety
III. Objectionable Content
IV. Integrity and Authenticity
V. Respecting Intellectual Property
VI. Content-Related Requests
Additional Information

Stakeholder Engagement

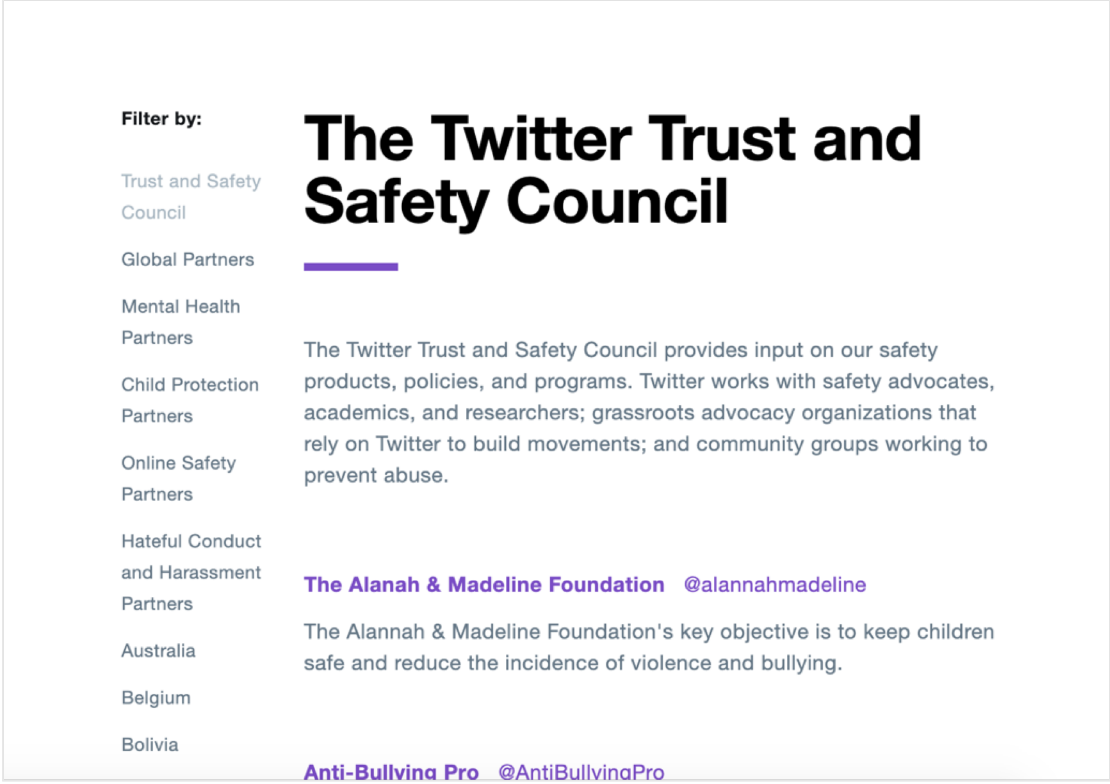
How does stakeholder engagement help us develop our Community Standards?

Gathering input from our stakeholders is an important part of how we develop Facebook's Community Standards. We want our policies to be based on feedback from community representatives and a broad spectrum of the people who use our service, and we want to learn from and incorporate the advice of experts.

Engagement makes our Community Standards stronger and more inclusive. It brings our stakeholders more fully into the policy development process, introduces us to new perspectives, allows us to share our thinking on policy options, and roots our policies in sources of knowledge and experience that go beyond Facebook.

Figure 3-1: Stakeholder Engagement Protocol from Facebook's Community Standards,

accessed May 8, 2020.



The Twitter Trust and Safety Council

The Twitter Trust and Safety Council provides input on our safety products, policies, and programs. Twitter works with safety advocates, academics, and researchers; grassroots advocacy organizations that rely on Twitter to build movements; and community groups working to prevent abuse.

The Alanah & Madeline Foundation @alannahmadeline

The Alannah & Madeline Foundation's key objective is to keep children safe and reduce the incidence of violence and bullying.

Anti-Bullying Pro @AntiBullyingPro

Figure 3-2: Twitter Trust and Safety Council, accessed May 20, 2020

Smaller platform companies have also noted that they rely on relationships with outside organizations and groups, though these relationships may be less formalized than councils and boards, or trusted flagger programs. As noted in Caplan (2018), “artisanal” platform companies often consult an “informal network of experts — mostly academics — to solicit feedback on potential decisions.” For platforms that rely on volunteers to both develop and enforce content rules — sites such as GitHub, Reddit, and Wikimedia — platform policy-making necessarily takes on a networked approach, with users often contributing policies, apart from the hierarchies of the parent organization (Caplan, 2018).

Platform Governmentality

Involving the community, and networking governance, is often framed by companies as part of their governance structure and within their normative ideals. In this sense, they tie outreach and feedback as part of the process of giving a “voice” to their users and the community (Zuckerberg, 2019), as part of their transparency efforts, or as part of their broader governance processes and procedures, which are often designed to mimic or build on existing governmental or legal systems. As was noted with the Zuckerberg quote that opened up this chapter, this often entails gesturing towards democratic ideals of participation and deliberation, either directly, or through absorbing associated values, such as consensus-building, freedom of expression, and transparency.

Both at COMO, and in other interviews and reports (Klonick, 2017; Caplan, 2018), platform representatives often frame themselves within governmental or judicial terms. At the COMO conference in Santa Clara on February 2, 2018, Jessica Ashooh**,** Director of Policy for Reddit, referred to their system of distributing powers between the company’s baseline rules set by Reddit and their policy team, and the rule developed by communities (or Subreddits) as a "federal system" a form of government that divides the powers between national, state, and local governments (Ashooh, 2018, 01:09:56; Caplan, 2021), and at another point referred to their community team as their "diplomatic corps" which is responsible for engaging moderators and identifying and resolving issues before they become a concern. Alex Feerst, Head of Legal for Medium, the publishing platform, referred to their system as a "common-law system of precedence," where "insights you derive from hard cases are up taken into your policies," in conjunction with consultation with "outside experts, and eventually….executives" (Feerst, 2018, 53:11). Paul Sieminski, General Counsel for Automattic agreed with this portrayal, noting "Alex's analogy to the common law is actually a good one," particularly because platforms "want to be clear with our users about what is and is not allowed…but in the application of those rules, there are always going to be edge systems" comparable to "legal systems with statutes and decades of case law."

For instance, the use of terms like "precedent" are also becoming increasingly common as platforms formalize their policy processes. In a post about the structure and governance model for Facebook's Independent Oversight Board, Brent Harris, Director of Governance and Global Affairs for Facebook notes that "precedent," or the expectation that the board will "defer to past decisions" will play an important role in how the Board operates (Harris, 2019). Feerst, then Head of Legal for Medium, referred to their system as a "common-law system of precedence" noting that, when making content decisions one often cites "internal precedence," noting his team used "hard cases up taken into policies" because content moderation often lacks "treatises out there in the world."

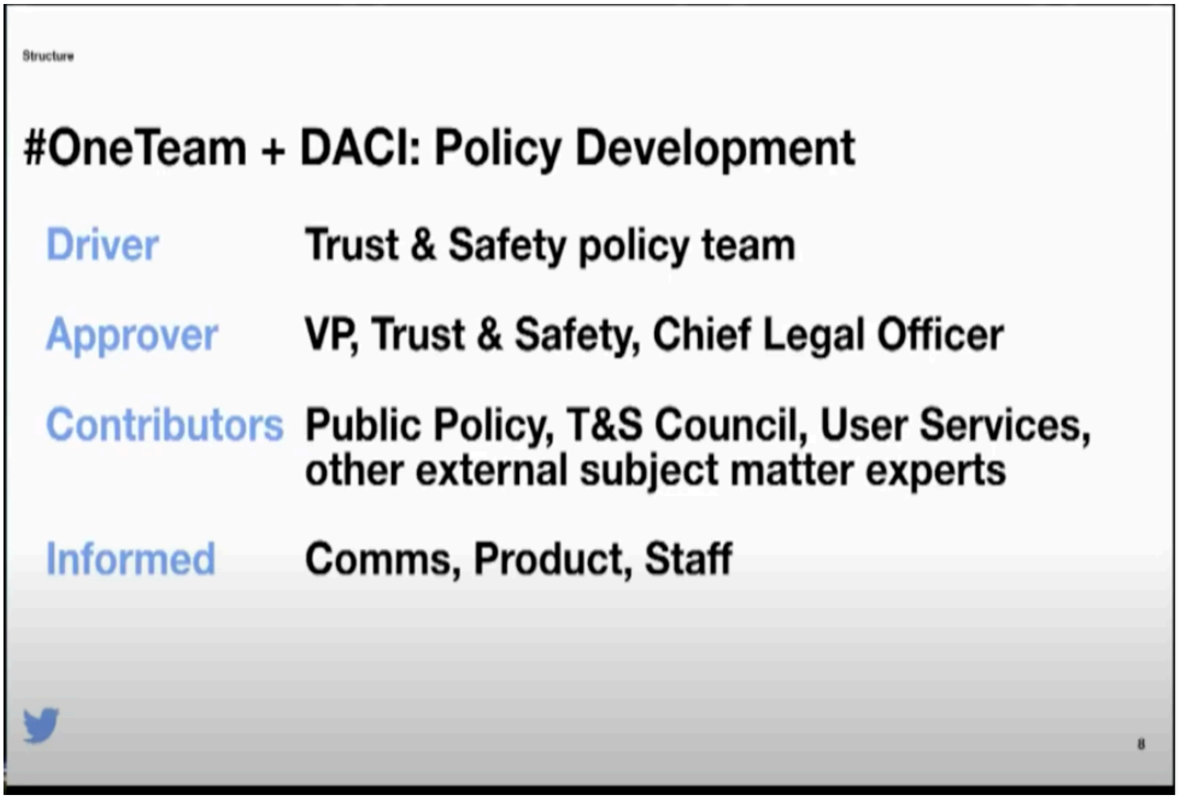
Monika Bickert also used similar terminology in her description of the process used at Facebook, describing the process of input into policies at a large biweekly meeting held at the company, as a “mini-legislative session” where input is considered from teams inside the company, as well as from external experts on issues such as “child safety” (Bickert, 2018, 29:44). Another Facebook employee, Peter Stern, used the same language at the second Content Moderation at Scale conference (noting the process was “sort of like a mini-legislature almost”), making it clear this is a consistent metaphor throughout the content policy team, despite the use of qualifying language — “sort of…almost” — that may signify uncertainty with the comparison. (Stern, 2018, 47:54). These metaphors hint at the ways that platform representatives, often trained as lawyers, or with Trust & Safety departments housed within legal departments, often frame their relationships with users in legal terms, for legal audiences. The boundaries between legal requirements for takedowns and normative or ethical standards (often referred to as “community guidelines”) are blurred by platform representatives, hinting at the ways that networking governance to experts and academics is used to bridge these jurisdictional divides and expertise. Legal and policy teams are often closely related at platform companies, often acting in coordination or with policy teams subsumed within legal departments. For smaller companies, like Medium, they specified that content policies are informed by “the legal risk that provides the underlying backdrop for this, but then on top of that, we all have our various policies” (Feerst, 2018, 49:51). Reddit specified that “policy reports to the General Counsel” with their legal team “as a stakeholder in this process” (Ashooh, 2018, 01:11:29).

Bureaucracy as Democracy

But platforms also often gesture towards other normative values in how they interact with stakeholders as part of the policy-making process, often leaning on democratic and liberal values such as participation, consensus-building, freedom of expression, and transparency when referring to partnerships with external stakeholder groups, or with their users, in forming policy. However, what these terms mean in the sense of rule-making tend to differ. In some cases, the reference to processes like "consensus-building" was very direct, building off of procedures underpinning the structures (such as the "mini-legislature) described above. Both Bickert and Stern used this language when they referred to the "mini-legislative session" where decisions on company policy are made within the company (at a biweekly meeting with internal teams and external stakeholders) (Bickert, 2018, 29:44). Wikimedia also used the term to describe the more distributed process of decision-making that happens in each Wikipedia, each governed by teams of volunteer administrators and editors, who "enforce the rules they come up with based on consensus" (Rogers, 2018a, 01:18:34). Other companies, like Vimeo, also noted that they operate in accordance with "consensus building within the broader moderation team," with Sean Mcgilvray noting that decision-making first exists at the level of the "individual moderator," with difficult-to-decide cases or new issues triggering internal conversations over Slack or in the office ("we will spin our desks around"). In each case, platforms emphasized that consensus-building and participation from internal and external actors, were important for the development of new content rules and policies, and to understand how existing policies should apply in difficult-to-decide cases.

Similarly, participation in policy-making from teams internal and external to the company was also framed as part of the governance process. For instance, GitHub has a policy of "open sourcing" their policies under a cc0 license, and invite their users to "make direct contributions to our policies" (Niv, 2018, 12:13). GitHub's Vice President of Law and Policy framed these efforts as part of a policy of "community governance" and as a way of encouraging openness and transparency in how their legal team "engages with those contributions from users" such as questions or comments on policies. Participation is also framed as a process of "feedback" from a range of stakeholder groups. For instance, according to Adelin Cai, Pinterest's content policies "draw upon feedback from our public relations team, subject-matter experts…industry standards. We also get a lot of feedback from legal" (Cai, 2018, 01:00:47). For companies like Twitch who, like YouTube, rely on user-generated content from "creators" who may receive revenue from the site, there are special mechanisms of feedback for this user group, who "directly" give feedback to Twitch employees who work to "captur[e] their concerns" (Keen, 2018, 44:24). But there are also clear limits to this participation that were noted by some representatives. As Del Harvey, the Vice President of Trust & Safety at Twitter described "contributors," (see fig.3-3) the "public policy, T&S council, User Services, and other external subject-matter experts" as the "lifeblood of this process," but also noted they have a "voice, but not a vote" when it comes to setting content policy. Their role, as she proceeded to describe, is to be a "subject-matter expert" and work both internal and external to the organization, and give input on particular topics (Harvey, 2018, 23:46).

In this sense, bureaucracy– i.e., how teams involve many stakeholders in the decision-making process — was often framed in participatory and democratic terms. Medium explained this process within their procedures of "escalation" when dealing with something "complex in terms of content…[or] legal implications" (Feerst, 2018, 52:27). In his explanation of how escalation works within the company, the Head of Legal for Medium explained that, in one form of escalation, "there is a sort of Socratic thing that happens in which people will argue for or against certain positions….eventually that will come to a vote" (Feerst, 2018, 52:27). However, though democratic process is often integrated into how platform representatives spoke about the decision-making process at their companies, these were often subsumed into larger organizational processes. As Medium's Head of Legal acknowledged, "I as the company's lawyer will have more to say about it," noting that they will "consult outside experts, and eventually, our executives all weigh in on it." This process of feedback can become quite complex.



Structure

#OneTeam + DACI: Policy Development

Driver	Trust & Safety policy team
Approver	VP, Trust & Safety, Chief Legal Officer
Contributors	Public Policy, T&S Council, User Services, other external subject matter experts
Informed	Comms, Product, Staff

Figure 3--3: Presentation by Del Harvey, at the Content Moderation at Scale Conference, Washington, D.C.

As part of participation, "diversity" was emphasized. Keen noted that their process of feedback, and the international makeup of Twitch creators, helped increase "diversity of opinion" on policies. The representative for Medium — then head of legal, Alex Feerst — also stressed an importance on "diversity of opinion" as well as "gender, ethnic, racial, and many other forms of diversity" within their content moderation program, noting that they use a "rotation" program, where other employees at the company that work outside of Trust and Safety, "can spend some of their time doing trust and safety work" (Feerst, 2018, 01:22:29). For Feerst, the diversity *is essential to* content moderation work, that he argues if you do not have it, "you cannot do it…because you don't have enough perspectives to populate the kinds of cultural competencies that people need to really understand this, and even more importantly, you don't know what you don't know."

In a number of cases, companies referenced the importance of transparency in how they made and enforced content policies, highlighting its use as a tool to communicate decision-making with partners and users. In some cases, "transparency" was referred to as a value or a goal which oriented a company's operations. The representative from Dropbox noted that "operating in a transparent way" is part of their "core value" which is "to be worthy of trust" (Dean, 2018, 11:39). Nora Puckett, speaking on behalf of Google, referred to transparency as "a huge goal, whether it is talking about content policies or our legal removals team" (Puckett, 2018, 42:10). Transparency, *as a value,* was also contrasted against other values, like *privacy* or *manipulating rules*. The rep from Medium put this in stark terms, making the case that there is "an understandable value of wanting to be transparent, wanting to say everything, wanting to put it out there, show what you take down, show why you take it down, explain the decisions you're making," but "at the same time, we've all been asked by our users, by the public, by the government, to be extremely mindful of user privacy" (Feerst, 2018, 47:27). Another company, Pinterest, also found transparency to be in contrast with other values of the content moderation team, with Adelin Cai saying, “For my team, we really care about transparency…we know people care about where they stand when they use a platform…but we really do strive to make it very clear as much as we can without giving away so much information that they know how to game the system” (Cai, 2018, 01:52:53). Transparency was also noted as being in conflict with concerns about safety; in a 2018 Proxy Statement filed to the Securities and Exchange Commission, Twitter, Inc., responded to a stockholder proposal for increased transparency on content enforcement with a statement that claimed the proposal “could reduce the effectiveness of our safety efforts by providing a roadmap for those bad actors who are seeking to evade abiding by our terms” (Twitter, Inc., 2018)

Transparency was also cited as a *tool*, with a number of platforms noting their partnership with Lumen, a project from Harvard’s Berkman Klein Center that “collects and studies online content removal requests,” (Berkman Klein Center, n.d.) as part of their transparency efforts (Puckett, 2018, 42:59). Other companies, like Google, Pinterest, Reddit, and Facebook described their own procedures and processes to aid transparency, including the release of “transparency reports” on specific concerns, such as government requests for data or takedowns. For some of these platforms, transparency also means accessible and available content policies, such as Google “making policies very transparent and easily accessible to our users…available online” (Puckett, 2018, 42:10), or Facebook disclosing the details of community standards more publicly, so users can see “the substance of the lines that our reviewers are drawing in day-to-day practice” (Stern, 2018, 27:30). In one case, with Reddit, transparency meant clarifying it for *all* users, not just those affected, when content had been taken down, with Jessica Ashooh noting that “When something is taken down, transparency is very important to us,” with Subreddits that are banned left up with a message explaining why the community was banned.

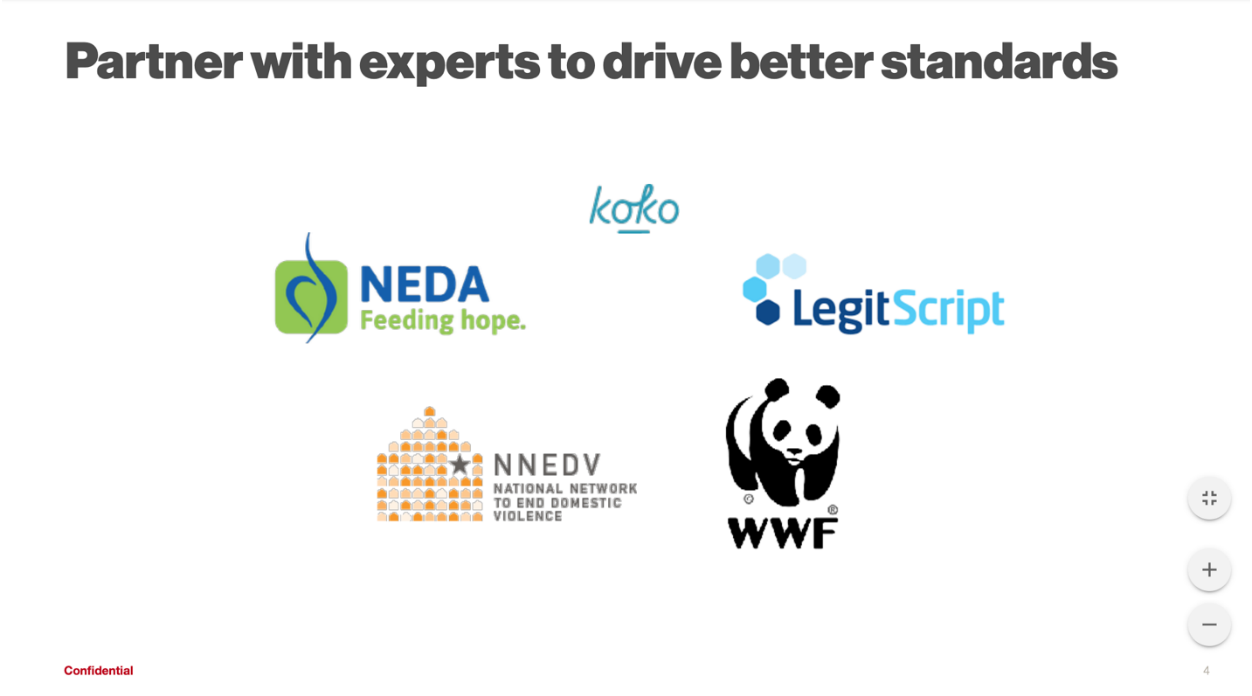
Expertise/Context/Legitimacy

Partnerships with outside experts are often driven by a recognition by the platform that they do not, and cannot, "know everything" (Del Harvey, 2018, 29:58). To address these gaps in expertise and cross-jurisdictional concerns and disputes, platforms have embraced (at least rhetorically) a networked governance approach in their use of partnerships and relationships with civil society and academia. In many cases, these relationships remain *informal*, between individual actors at platforms and people "they know or who have been recommended" (Harvey, 2018, 23:46). In other cases, such as with the increased establishment of trust and safety councils and advisory boards, trusted flagger programs, and other partnerships, these relationships have become more formalized. In all cases, however, the participation of outside actors is structured by the platform, either as the instigator of outreach (and the development of these boards and councils) and as the interpreter of expertise and advice that is then provided to product teams. This outreach seems to be bound with the need for platforms to address issues at scale, but to also address past concerns that platforms acted *too quickly* before considering the effects of the problems they were building, or even *who* the communities were that were impacted. As Del Harvey explained, "by partnering with others, by being really open to the feedback that we get, or to the issues that are raised to us, it makes it a lot easier to not only act quickly and as accurately as possible, but also to build relationships in good faith between us and people who maybe had assumed we didn't care, or that we weren't listening" (2018, 29:58).

In other cases, platform representatives highlighted the ways that partnerships helped them bridge gaps in expertise, and insert context into content decisions. In particular, for bridging gaps in cultural and linguistic expertise, as well as for specific issues like eating disorders, bullying, terrorism, and hate speech, platforms have increasingly looked to outside partners. Platforms often point to partnerships with experts as a way to address how they address content within the contexts of local communities, or online subcommunities, while maintaining their ability to operate at scale. They also tended to stress how these outside partnerships worked to improve standards-setting within particular topic areas or regions. Platform representatives stressed an awareness of the "limits of our expertise," and accordingly, whether they should be making decisions in these areas (Sieminski, 2018, 01:49:15).

Often, the need to bridge gaps was due to issues of scale — most platforms operate globally, but with teams primarily located within the United States, or just with small teams. As the representative from Medium noted, "we are small, but we are global…the internet shows up everywhere" (Feerst, 2018, 50:39). Though Medium tries to hire individuals with additional languages and "cultural capacities," they also address any gaps through partnerships and other "outside resources to deal with all the various countries we are displayed in, the jurisdictions we're working in." The representative for Vimeo, another small platform that nonetheless operates globally, also expressed some difficulties with addressing language gaps. Though he noted that, when faced with a language no one on his team speaks, they look for "context cues…[like] swastikas," but they also "find experts or subject language experts within the company or within third parties that we contract" (Mcgilvray, 2018, 01:03:46). In that sense, the notion of widening the network can both be broader than the individual team, to creating formal (which may include a financial arrangement, like a contract) and informal relationships with outside partners.

Platforms representatives often use partnerships they form with outside organizations as subject-matter experts on particular topics or for particular concerns. Pinterest, for instance, works with groups like the National Eating Disorder Association (as well as the World Wildlife Foundation, Koko, National Network to End Domestic Violence, and LegitScript, see Figure 3-4), which they use when they do not have “specific expertise in certain areas” (Cai, 2018, 01:04:45). For instance, the partnership with the National Eating Disorder Association was used to compile “keywords to identify when a user might be searching for content relating to self-harm” (Cai, 2018, 01:04:45), and to provide Public-Service Announcements that “run against searches for terms like ‘proana’ or ‘suicide’” (Perez, 2013). Facebook also has a “Safety Advisory Board” which Facebook consults on issues related to online safety (Facebook Help Center, n.d.). This global group of nonprofits, which includes (among others) organizations such an India-based women’s empowerment nonprofit called Center for Social Research, the UK-based Childnet International, the National Network to End Domestic Violence, and an Austria-based movement against bullying called PROJECT ROCKIT, provides “expertise, perspective, and insights that inform Facebook’s approach to safety.” Bickert, speaking on behalf of Facebook, compared their own partnerships with expert organizations with those of Pinterest, saying “it’s not as simple as saying ‘we don’t allow bullying.’ You have to have a lot of granular guidance into what bullying is. What we find useful is to engage with the expert groups on a particular topic” (Bickert, 2018, 01:54:18). Twitter’s VP of Trust and Safety, Del Harvey, also stressed that “there are a lot of things we don’t necessarily know. We are not going to be subject-matter experts in everything, try as we might, so it’s really important that we work to identify people who are, and listen to them and make sure we work to understand what they are saying to us is an issue" (Harvey, 2018, 24:32). Harvey clarified this included both issues the company agreed with, and in cases when there is just a "perception" of an issue.



Partner with experts to drive better standards

Confidential

4

Figure 3-0-4: Presentation given by Adelin Cai, Pinterest, at the Content Moderation at Scale conference in Santa Clara, California.

In some cases, this lack of expertise results in a formal relationship with an external organization that becomes embedded into the policies and practices of the company. This has been the case with "trusted flagger" programs, for instance, on YouTube, which are individuals or organization who seem "particularly effective at identifying policy-violating content," which the company gives "more robust tools, like bulk flagging mechanisms," to help identify content (Puckett, 2012, 41:12). Facebook undertook a similar effort with its fact-checking partnership, providing fact-checking organizations that have been certified by a non-partisan fact-checking organization (Ananny, 2018), providing these organizations with tools to identify and review potentially false news over the network (Facebook, n.d.). Facebook has continued to state that this program was built because they “do not believe that a private company like Facebook should be the arbiters of truth” (Facebook Journalism Project, 2021).

Platforms also struggle with jurisdictional concerns when they try to share governance with trusted partners, finding themselves needing to balance complying with the laws and norms of the countries in which they were operating — and the importance of *context*, the specific aspects of social settings (Dourish, 2004). In terms of networked governance, governments become another partner that platforms work with to identify and evaluate takedowns. In some cases, they work through the same procedures that ordinary users must go through. Google noted they provide a “web form” which governments can use to identify the product where potentially violating content has been posted (Puckett, 2018, 33:55), whereas other companies, like Facebook, have more direct relationships with governments used to ensure compliance for requests to restrict access to content a government has deemed illegal (Pearson, 2020). Platforms also work with governments to elevate (rather than restrict) the availability of certain content, particularly during emergencies, using APIs or relationships to facilitate exchanges of information. For instance, Google works with a number of government agencies around the world such as the US National Weather Service, Meteoalarm in Austria, and the India Meteorological Department through the use of a “Common Alerting Protocol” (Google, n.d.). In other cases, platforms can work directly with government agencies to address potential content concerns, as is the case with platforms like Facebook, who have been working closely with the Census Bureau “to help ensure a fair and accurate census,” holding “weekly calls to discuss emerging threats and to coordinate efforts to disrupt attempted census interference” (Murphy, 2020).

Context was not necessarily framed in terms of outside partnerships, but was the underlying thread referenced by companies when explaining why *humans and their social and cultural relationships*, not *machines*, were needed when making content governance decisions, and hinting towards the complexities of doing the work of governance when operating on a global scale. Companies repeatedly referred to various steps they took internal to the company, as a way to introduce this context. For Twitch, human moderation was necessary because "machines really can't get…all the layers of context," expanding to say that the "machine would have to know gaming culture and language, they would need to know the references in that history…the meaning of emotes" (Keen, 2018, 45:29). Shirin Keen, the representative for Twitch, noted that this "level of complexity" was why many of the moderators "come very much from the community, and so they have that deep cultural background and experience." Wikimedia's more distributed governance structure, in which each local language community set and enforce rules that are developed based on consensus, means that the site is quite tailored to "local knowledge and local context" (Rogers, 2018, 01:25:18). For most platforms, the problem of context was used to explain the limitations of automation in content moderation, and the importance of "the human" whether inside the company, or outside (Ashooh, 2018, 01:15:29).

Discussion: Principles of Content Policy or the Construction of the Democratic Platform?

Platforms are increasingly looking to partnerships with external stakeholder groups to inform parts of their content policy-making. Examining *how* these partnerships unfold in practice can, however, be difficult, given the broad range of both formal and informal partnerships (from a side conversation to a formal partnership), difficulty at entering into platform companies to do work, and the use of rampant use of non-disclosure agreements across the tech industry (Roberts, 2019). In rare reporting on the trust and safety councils, or of meetings between platforms and civil society groups, platforms have been criticized by external stakeholders participating in this process. In a letter written to Twitter's CEO Jack Dorsey by members of Twitter's Trust and Safety Council, council members criticized the social media company's approach to outreach and feedback, noting they had gone months without updates and "in some regions," council members were "unable to reach their contacts at the company" (Matsakis, 2019). Though they say the council began strong for its first two years (2016-2018), signatories of the letter said that, in its current version, "The Trust and Safety Council has eroded to practically nothing." Facebook has also been criticized in its exchanges by civil society partners. Organizers of the #StopHateForProfit Facebook advertising boycott, referred to recent statements made by Facebook executives in a meeting as "spin" and as a "powerful PR machine" (Isaac & Hsu, 2020). Some of the leaders of this boycott who have been working with Facebook for years, such as the N.A.A.C.P. took this opportunity to criticize their interactions with the company more generally with Derrick Johnson, the N.A.A.C.P. chief executive saying the meeting was the "same conversation from the past two years," with "no actionable steps" (Isaac & Hsu, 2020).

Despite this frustration, platforms are continuing to expand efforts to partner with outside organizations and individuals, with platforms like Twitch and TikTok recently adding their own Content Advisory Council (Bettadapur, 2020) and Trust and Safety Council for Asia Pacific (Bettadapur, 2020) and Facebook formalizing their advisory process into the independent Facebook Oversight Board. According to how they are represented by platforms, these relationships provide the foundation, the "lifeblood" of the policy-making process at platforms (Harvey, 2018, 22:46). And yet, according to those taking part in this process — the fact-checkers, the civil society organizations and academics giving their time and expertise, the users and community members tapped for their experience — roads into platforms have increasingly been muddied. And yet, it is clear that these relationships provide the underlying structure for the myth making of platform companies. Framing these efforts within the history of networked governance — as a theory and as a way to understand the promise and fallout of more distributed decision-making — provides one way to understand this strategy.

In the frame of networked governance, platforms seem to be appealing to the language of government as a way to locate their participation within this networked process. In this sense, platform representatives are using legal systems, or forms of government, as metaphors. As Lakoff and Johnson (2008) argue, metaphors are powerful tools that make sense of experience, "provide coherent structure, highlight some things and hiding others" (Lakoff & Johnson, 1980). Within that frame, the use of this language seems less of an allusion to a private form of government for speech concerns (Caplan, 2021), than an acknowledgement that, both within the company, and in its dealings with external actors, that they consider feedback and opinions, but ultimately hold more centralized control. But, given that platforms use this language publicly in spaces where external stakeholder groups are often present, platforms may also be using this language as a way to frame these interactions — building off of legal concepts like *precedent* and *common-law,* as well as American values such as *federalism,* to convey to outside actors that decisions are fair and legitimate. In this sense, platform rhetoric should be read through the lens of "procedural justice" and the importance of "perceived procedural fairness" in governance decisions (Sunshine & Tyler, 2003; Bottoms & Tankebe, 2012). In this sense, the use of legal and governmental terms may be used not only to feed off of the legitimacy of these authorities and institutions, but to create a perception of fair procedures that may increase the likelihood that both those internal and external to the company, will defer to the platform's decision-making.

In a more sympathetic reading, however, the use of procedures taken from common law could also be a way to address the *lack of law* elsewhere, particularly in the United States, where speech concerns are largely normative and not legal. As Alex Feerst from Medium noted in his remarks, "there are no treatises out in the world" when it comes to complex content concerns. The use of democratic language by platforms may build x of this lack of legal structure, providing a normative component or pointing to other concepts, such as the marketplace of ideas or freedom of speech, which has been used to ground communications policymaking in the past (Napoli, 2001). The use of terms like *participation, consensus-building, diversity,* and *transparency* also builds off of ideals of the public sphere, and suggests that the process of content policy-making at platforms are open for public debate. The terminology platforms are using to describe their decision-making process does seem to suggest they are at least considering the importance of a deliberative model, integrating discourse from various social groups both internal and external to the company. As Simon Joss (2002) has pointed out, such integration of citizens and interest group representatives is not new in the area of technology assessments, however, this form of participatory policy-making has been limited to government bodies, and most participatory initiatives took place in public. In contrast, many of the outreach activities that platform companies view as participatory take place behind closed doors, often as a matter of direct outreach. Because of this, it's unclear how feedback from users and interest groups is considered relative to other forms of feedback happening at these companies. The framing of this form of input as having a "voice, but not a vote" (Harvey, 2018, 22:46) positions this input within the broader bureaucracy of the company (and the 'voices' inside of it). Without a clear understanding of how feedback is integrated, and from whom, this solicitation remains more symbolic than substantive.

The symbolism of the terms used by platforms, such as “diversity” is important to note particularly *because* these terms have had a long history within communications policy (Napoli, 2001). As Napoli (2001) notes, diversity is a term that has had many meanings — from ownership of media, to increasing minority representation –and there has not been a consensus as to “an adequate definition or measure of this rather ambiguous concept” (p. 126). It has long been considered an essential component of the marketplace of ideas, which Napoli argues, has guided policymakers and courts in communications policymaking. The use of diversity by platforms in their own content policymaking, however, introduces a number of new potential readings of the term. The value of diversity in content moderation is in the *reception and interpretation* of the messages, rather than the messages themselves. Though it is often used to refer to the demographic characteristics of those doing this interpretive work (or “cultural competencies” as was stressed by Medium’s Head of Legal), it is also used to refer to “diversity of opinion,” which may, as Twitch’s representative acknowledged, mean gaining feedback diversity *types* of stakeholder groups (such as “creators” for platforms), but it may also mean acknowledging that, even within groups, there may be different interpretations of content that has been flagged as violating terms of service. A desire for “different viewpoints” was also cited by Twitter in 2017, in their plans to expand their Trust and Safety Council beyond 40 members (Twitter Inc., 2016).

The extensive use of the term transparency, and the creation of documents like “transparency reports” by platform companies also signals that these companies are trying to build on ideals of openness and access (and potentially involvement by the public), as a way to signal their own trustworthiness (Ball, 2009). It also plays an important role in the perceptions of procedural justice that platform companies are trying to communicate to users and the public-at-large; through (Tyler, 2003). Transparency, as a concept within administrative research and governance, is closely tied to accountability. As Ball (2009) illustrates, though increasing the availability of public information became part of a push for greater accountability among state actors, particularly in the 1970s among Watergate and fears of government corruption, "transparency" as a term, has more to do with the influence of nongovernmental organizations and supranational organizations in the 1990s; part of an effort by organizations like Transparency International which worked to advocate for increased access to information for individuals (along with other anti-corruption goals). Transparency, in this sense, also became used by countries as a signifier in negotiations; a way to convey "trustworthiness" and good governance (Ball, 2009, p. 297). Though this term has been constantly reinterpreted and redefined, it is clear that, in some ways, platforms are referring to "transparency" as a goal to be reached within the broader frame of platform accountability (the representative from Google directly said transparency was "a huge goal"). It is also clear that transparency, as a way to convey *good governance*, is weighted by platforms against other values — privacy and potential for manipulation or gaming — that they perceive as contradicting the ideal of transparency as a goal.

Transparency can also serve as a tool for self-regulation; used as a way to signal the importance of consumer choice (Ball, 2009), or as a way for external actors to evaluate the compliance of the company to their own policies — a form of oversight and enforcement of industry codes by third-parties (Reeve, 2013). In its use by platforms, transparency has been cited by platforms as a tool to demonstrate how platforms are complying with government removal requests (in their partnership with Lumen and the Harvard Berkman Klein Center). Platforms have also noted their use of transparency reports to cite their compliance with their own internal rules (though these transparency reports, and a project by New America, has shown that platforms have greatly expanded their categories of reporting on content rule enforcement, takedowns, and appeals since 2017 (Singh, 2020). But at the same time, these transparency reports can be vague, inconsistent from year to year, and lack standardization despite recommendations for transparency reporting that have been provided by third-party organizations. In an analysis by Singh (2019) in relation to "The Santa Clara Principles on Transparency and Accountability," a set of standards for transparency in content moderation that were established by a coalition of organizations, advocates, and academic experts (notably, at the first Content Moderation at Scale event), Singh found that platforms are falling short of transparency expectations (Sing, 2019). Transparency can be a powerful tool for addressing the extreme information asymmetry that exists between platforms (Tessier, Herzog, & Madzou, 2017), and users, but they have also been criticized on the grounds that they do little to shift these power dynamics, particularly in content moderation decisions. Journalists like Casey Newton (2019) have argued that the reports reveal an appeal process that is "limited and opaque," and how little "recourse people have if they are falsely caught up in a machine-learning dragnet."

Lastly, the need to address context concerns — in terms of language, culture, and expertise — has come to play an important role in content policymaking in the platform era. At the Content Moderation at Scale conferences, the need to address context became a bit of a rallying cry, used in conjunction with an acknowledgement by these technology companies that we need to re-insert the 'human' and guard against the overuse of machine learning and automation when it comes to content concerns.[3] This was stressed time and again by platform representatives, that technology could be used to "flag something that *might be* violating" but that a person had to be there "to make the determination about whether that *is* violating" (Bickert, 2018, 26:56). This was stressed even by representatives from Facebook, who noted that "limitations on the context that we have available means we can't just use technology straight out to do a lot of this work," (Bickert, 2018, 29:44) despite claims by Mark Zuckerberg, Facebook's CEO and Founder, to Congress that AI will solve content moderation concerns. But for people *in* platform policy, using humans to pay attention to context was seen as the only way to grasp the "complexity of human expression," and to treat complaints with "humanity and dignity" (Feerst, 2018, 47:27). In this sense, context is a dynamic feature of the moderation process, a way to overcome the rigidity of computational systems to become more responsive to the different social settings in which they are used (Dourish, 2004).

Specifically, the emphasis on context by platforms is also used as a way to convey the need to pay attention and understand local issues and concerns that emerge from specific social groups. In this sense, it is akin to the "localism principle" in communications policymaking, which focused on how media were elevating local voices through local news and reporting (Napoli, 2001). Context is both in terms of geographic locale (i.e., where moderators are and are not located), but also in terms of language and culture. Platforms often stress how they work to transcend context and legal concerns. In cases where platforms rely on volunteer moderators,

---

[3] This concern, that there are limits to what machines can do when it comes to the interpretation of messages, was additionally the theme of another panel presentation that was not studied here.

such as with Wikimedia and Reddit, they stress the benefit of having volunteers moderating their own communities because of this capacity to apply “local knowledge and local context” (Rogers, 2018, 01:25:18). In other cases, where moderators are U.S. based, the importance of having moderators that can apply “local expertise” is still stressed by platforms like Yelp (Schur, 2018, 01:51:25). For companies like TripAdvisor, who operate globally, being able to have “representation for each of the languages” spoken on the site, as well as “that local perspective” is an important element for deciding whether content should be published or reviewed (Foley, 2018, 13:02). Despite not often having moderators in every space where they are operating, platform representatives often spoke of overcoming these context limitations through activities like “cultural context trainings” (Harvey, 2018, 28:06:00), through hiring people who speak the languages of users being moderated (Stern, 2018, 58:01), or through using “outside resources” such as partnerships with academics and civil society actors (Feerst, 2018, 50:39) For Reddit, a company that relies on volunteer moderators, “moderators are empowered to make their own rules,” which means, in her view, that “cultural customization is built into the system” (Ashooh, 2018, 01:50:54). Notably, however, outside these community-reliant strategies (Caplan, 2018), platforms have continued to push the idea that automation using artificial intelligence and machine learning, is the only way to effectively scale content moderation globally (Gillespie, 2020).

The need to understand the context of user-generated content also extends to subject area and topic; with certain content areas and concerns requiring specific expertise and sensitivities. These knowledge gaps often lead to partnerships between platforms and civil society organizations, or academic experts, that are used by platforms to generate specific policies and recommendations to address content concerns regarding eating disorders, bullying and harassment, hate speech, child sexual exploitation, and animal abuse. Often, when platforms center values like participation and diversity, the need for specific subject expertise or cultural context is necessary. In a sense, this feels almost paradoxical; broadening and widening participation in policy-making is often viewed as at odds with prioritizing the expertise of a few actors (March & Olsen, 1995; Dewey, 1927). And in this sense, the embrace of participation, consensus-building, and transparency do seem to be geared towards opening of platform policy-making to those actors who can offer platforms information, expertise that is lacking at these companies. However, because most of these discussions are happening behind closed doors, and because advice from experts is considered in relation to other forms of feedback happening *within* these companies, it is not clear what role, or to what degree, expert opinions are considered in making content policies. And though there are cases when experts consulted are made available to the public (normally when they are associated with formal bodies, such as the Trust and Safety Councils, Facebook's Safety Advisory Board, and the Facebook Oversight Board), they are often bound by non-disclosure agreements which prevent a more in-depth analysis of the role these experts play in decision-making at these companies (Facebook, n.d.) Additionally, *why* these experts are chosen (and not others), as well as what specific expertise they bring in relation to content policy, is not always easy to assess. This has been the case with the Facebook Oversight Board, whose members range from professors of law, to journalists, to nonprofit executives, to the former Prime Minister of Denmark (Oversight Board, n.d.). Though all of these individuals could be referred to as experts in their respective fields, it is unclear how these forms of expertise will be brought to bear in making decisions about the acceptability of content posted by users around the world.

Conclusion: Appealing to Legitimacy

As of right now, it is unclear what impact these efforts will have on content policy; however, it is clear efforts to engage external actors or distribute responsibility is not a salve for centralized platform power. Facebook has noted that the Oversight Board can *recommend* policies to Facebook, and these recommendations are not binding (Klonick, 2020) and activists in Myanmar have noted that Facebook has implemented policies that have directly countered the recommendations made by the Human Rights Impact it commissioned (Wong, 2019). Though it is impossible, at this point, to trace the many efforts to reach out to external actors, and the variety of ways platforms, like Facebook, integrate recommendations made by different stakeholder groups (such as their users, experts, nonprofits, and others), this chapter attempts to address the role outreach efforts play within the broader context of platform governance.

These networks of outreach in content policy at platforms are an attempt to build on forms of networked governance that have become more popular with governments (Sorenson & Torfing, 2005). In this sense, platforms are using these networks of experts and organizations as a way to both distribute responsibility for policy-making (in cases when decision-making is moving to networked actors, such as with fact-checking partnerships and the Facebook Oversight Board), and to receive feedback from a variety of different actors with different areas of expertise (which is necessary for content concerns, which tend to span a broad array of issues). However, because of the embrace of rhetoric of good governance and by platforms, there has been little attention paid to the challenges of networked governance in general, as well as the specific manner in which platforms are soliciting and integrating feedback from distributed stakeholders.

Firstly, relationships between platforms and networked actors increase the complexity of these decisions, making it impossible to understand or evaluate the relative influence of experts or interest groups in decision-making. Because there is no (or very limited) visibility into how broader conversations are integrated into decision-making, as well as how interest groups involved in these governance networks (including the platforms themselves) work to maximize their own institution's goals. According to Tuebner (2009), this kind of chaos and over-complexity is characteristic of this model, which, after an initial "euphoric phase" tends to lead to failures such as information abundance, coordination concerns, communication issues, asymmetric power relations, and opportunistic behavior (p. 397). Teubner notes that this is because larger networks increase complexity, information overload, and conflicts resulting from different viewpoints.

In the case of platforms, in most cases, networked feedback is absorbed into the function of the organization, even further limiting its effect. Trust and Safety councils do not work publicly, or even *through public channels*; they are bound by non-disclosure agreements and operate behind closed doors. Because of this, criticism that was once done in public by these actors, is absorbed into the operations of these companies. As can be understood by the (rare) public criticisms of this model made by included stakeholders, this form of networked governance can often lead to a communication environment that is both networked and vertical, with networked actors lacking visibility into these more hierarchical decision-making processes (Wilikilagi, 2009). Relative to this, networked actors often have little understanding of their influence relative to the influence of others. This is particularly important as networked relationships continue to be mediated *by platforms* through their own formal channels, *increasing* asymmetry in power relations while gesturing towards horizontal decision-making.

Lastly, governance networks pose new challenges for accountability, particularly as platforms rely on actors and institutions to bolster their own legitimacy and perceived fairness, while not necessarily offering these networked organizations any real power over decisions. As trust in platforms and the technology industry declines, the desire to build on the legitimacy of these other institutions and organizations — and their relationships — is tempting. However, these governance networks tend to complicate classic notions of accountability; that *who should be held accountable can be clearly can be clearly identified and held responsible,* and that pathways for accountability should be direct, with consequences clearly defined (Sorenson & Torfing, 2005). In this sense, networked forms of platform accountability often lead to no accountability at all.

Chapter 3: The Trust Project: How to Train Your Algorithm

In 1997, Sally Lehrman, then President of the Northern California chapter of the Society for Professional Journalists, convened a roundtable to discuss how media ethics were changing in the digital era. Lehrman was concerned that, in the rush to move online, news media was losing some of the same values that had driven traditional journalism in the past. The problem as she saw it was that the "chase" for metrics were undermining journalism. The group, called The Executive Round Table on New Media Ethics and Online Accountability, met for a year. Eventually, Lehrman moved on to an endowed chair at Santa Clara University and a fellowship at the Markkula Center for Applied Ethics (The Trust Project, n.d.).

In 2012, a decade and a half later, Lehrman convened a reunion of the original members of the table. She decided to expand the invitation list beyond journalists in the Bay Area and make it national. What she discovered bringing these journalists together concerned her — many of these journalists were voicing almost the exact same concerns they had voiced in the 1990s: the chase for metrics was still dominating the newsroom, homogenizing content. Publishers were bemoaning the amount of control algorithms (referred to still within the lens of "metrics") had taken control, "and that they were victims to this." Lehrman decided this did not *need to* be the case. She contacted people she knew at platforms — then at Twitter and Google — who both told her the solution was simple: *all you have to do is train the algorithm.*

Research on the impact of platforms on the news media industry generally supports the sentiments expressed by Lehrman and her fellow publishers: platforms, particularly social media, have increasingly taken on a distributive role for news media (Bell, Owen, Brown, Hauka, & Rashidian, 2017). As part of this, algorithms, and data-driven technologies — those rules that underlie the processes that determine things like personalization and prioritization through things like the Facebook News Feed — are playing a part determining what news users receive (Caplan & boyd, 2018). In turn, metrics calculating the traffic of particular stories have come to play an important part in newsrooms and in the journalist's professional life (Petre, 2015). Though there is research on how individual news organizations have addressed the impact of the dominance of metrics in journalists' work practices and individual identities (Christin, 2020), there has been minimal research on how newsrooms have coordinated their response to the challenge. The Trust Project — an international consortium of news organizations working with platforms — provides one opportunity to conduct this type of research. Specifically, The Trust Project is a window for understanding how *networked governance,* between news organizations and platforms, can both facilitate and inhibit an exchange of values between professional groups.

Thus far, this dissertation has explored how relationships — mostly between organizations — can be pivotal for how understanding how certain definitions are socially constructed. In particular, the 'battle' against mis-and-disinformation over social media has led to many stakeholders coming together, particularly because social media companies have been so reluctant to take on the role of defining 'Truth.' In the last chapter, I illustrated many of the ways platforms frame these relationships — drawing in the democratic ideals of media, and the expertise of a broad range of external groups. In this chapter, I dive much deeper into one of those stories — a partnership between platforms and a media association for mutual benefit. What this story shows is how platform efforts to distribute the responsibility for policy-making through collaborating with networked organizations, is impacted by the inter-and-intra-organizational dynamics of these companies. This chapter demonstrates the limitations of networked governance initiatives, particularly in how external organizations bend themselves to the aims of platform companies to increase their legibility, as well as in difficulties in navigating the bureaucracies of platforms from the outside.

Method

This chapter takes a case study approach, using The Trust Project as a starting point of analysis. It relies primarily on publicly available documents made available through The Trust Project's website, trade reporting on The Trust Project between 2013 and 2020, and corporate blogs and articles from Trust Project partners and participants, including from platform partners, such as Facebook and Google, technical partners, such as Schema.org, and news media organizations who had signed on to the project. To understand more about the background of what I was seeing through policy documents, I also conducted semi-structured interviews (n = 3) with leadership of The Trust Project over the course of 2018-2021 (see Appendix B and C). These interviews have been anonymized in accordance with the recommendations from the Rutgers IRB, and to mitigate potential negative impact on the relationship between platforms and The Trust Project. Interviews were sixty to ninety minutes long and were conducted over telephone and zoom. In total, I analyzed around 75 news and trade press articles, around 10 corporate blog articles, and around 40 documents made available to me both privately and publicly by The Trust Project as part of this chapter.

I chose The Trust Project as a case study on networked governance *because* of the extent of interactions the organization has had with a range of networked actors — from media, to nonprofits, to platforms. I selected it because of the uniqueness and depth of their project — an effort to both revisit media ethics for the digital era by publishers around the world, and for their attempt to standardize these ethics and format them within a model legible — culturally and technologically — to platform companies. Their history both predates the current moment of concerns about the trustworthiness of media, and was *accelerated* by the 2016 election and concerns about fake news, and their interactions with platform companies, periods of collaboration followed by periods of limited information, across this period reflect the range of interactions platforms can have with external stakeholder groups in response to the waves of public critique and concerns. Additionally, The Trust Project, in being an organization of primarily journalists, is unique in terms of the expertise they were offering to these technology companies; an expertise those platforms have repeatedly sought (particularly, in the form of fact checkers) as they sought to distance themselves of the responsibility for being the "arbiters of truth" (McCarthy, 2020).

The goal of this case study, as with all case studies, is not to generalize (Thomas, 2016); the experiences of The Trust Project may be quite specific to this organization. This case study also approaches the experiences of The Trust Project *only* from the perspective of the participants of this organization, and not from other networked actors, such as platform companies. The point of this is to both create a boundary, and to also gain a detailed understanding of experiences working *with* platform companies, from the perspective of those external stakeholders. It is also important to note as a limitation and disclaimer to this research that I participated in some early meetings with The Trust Project in my capacity as a Researcher at the Data & Society Research Institute, however, when I realized the organization could be a potential site for research, I removed myself from the organization's activities.

The Context of Metrics in Media

Lerhman and her colleagues first met to discuss the impact of the internet on journalism in the late 1990s. At the time of this first roundtable, the internet was still mostly a walled garden. The majority of Americans were not yet online; Pew Research said that by 1997, only 36% of adults were internet users (Pew Research Center, 1999). In 1997, Yahoo was the leading search engine, followed by Excite, Infoseek, and Lycos (Meeker, 1997); Larry Page and Sergey Brin did not release Google until the next year. Social media platforms, as we know them now, did not yet exist (though there were myriad ways to be social online (boyd, 2015)). Still, as Lehrman describes, news media was already concerned about the impact of the Internet on their business and profession. By 1997, more than 850 commercial newspapers based in the United States offered online services, with 150 of those newspapers appearing between 1996 and 1997 (Singer, 1997). According to Mark Deuze (2000) and J.B. Singer (1997), journalists and editors felt "nervous and concerned" about the new role the Internet was playing in their professional lives. In Deuze's assessment, the Internet was changing journalism in several ways, both in terms of its use as a reporting tool inside newsrooms (referred to as "Computer Assisted Reporting"), and in the growth of "online journalism" produced specifically for the web. The features that Deuze saw as characteristic of online journalism at the time – hypertextuality, multimediality, and interactivity – were the blocks that would slowly build the new future of journalism.

By the 2000s, metrics were coming to play an even more important role in newsrooms, and the lines between journalism and other types of content (like blogs) were beginning to blur with the emergence of social media. This era has now been well documented by scholars examining how newsrooms were re-organized by the use of data and analytics.

Angèle Christin (2020) has done significant work on the role digital metrics came to play in newsrooms throughout the 2000s, embedding herself within two different digital media companies – one in the United States and one in France – to understand how American and French journalists understood and interpreted audience analytics in the production of news. Writing about a similar time period, Caitlin Petre (2015) also examined the role metrics and data analytics came to play in digital media organizations, noting that traffic-based metrics had become highly ranked as form of evaluation, and that metrics were coming to exert "a powerful influence over journalists' emotions and morale." Even earlier, C.W. Anderson (2011) theorized about the increased importance "the algorithm" was beginning to play in mediating between "journalists, audiences, newsrooms, and media products" (p. 530). The anxieties expressed by Lehrman and her colleagues in 1997 predated this period, but were well reflected in this scholarship that posited a major role for metrics and algorithms in shifting the organizational cultures of news, as well as on the emotional life of journalists.

When the table reconvened in 2013, even more had changed in journalism; social media platforms had come to play an even bigger role in how individuals were accessing news. Though Facebook and Twitter were founded in the mid-2000s (2004 and 2006 respectively), subscriber bases for news over social media remained low compared to audience size for print (Ju, Jeong, & Chyi, 2014). In 2011 Twitter was still used most for sharing news links, but this began to slowly change after Facebook introduced the Open Graph Protocol API in 2010, and more news media began implementing it within their content management systems (Overland, 2010). In technical terms, the API standardized the use of metadata used by webpages and how these sites are represented over platforms like Facebook, allowing certain information posted by a news publisher (for instance, title, byline, and lede) to be visible to someone over Facebook (Bodle, 2011). In broader terms, this API had the effect of introducing interoperability across the internet, introducing new "regimes of sharing" which linked "a broad range of platforms, sites, spaces, and people together in a global context" (Bodle, 2011, p. 321).

By 2016, the use of social media platforms like Facebook to read the news had grown markedly; a survey by Pew Research released that year found that a majority of U.S. adults (62%) received their news through social media (Gottfried & Shearer, 2016). Facebook had become *the* major distributor of news content, with 66% of Facebook users accessing news on the site (compared to 59% of Twitter users, which has a much smaller user base across the United States). Perhaps because Pew released the survey in the run-up to a contentious election in the United States, this survey gave rise to a significant amount of scholarship examining the new role platforms like Facebook were playing in the public sphere (for examples see Caplan & boyd, 2016; Van Dijck, Peoll, & De Waal, 2018; Sunstein, 2018), and on the news media industry specifically (Zamith, 2019; Napoli, 2019).

Scholars have payed considerable attention to the role social media platforms are playing in news distribution (Napoli, 2019; Bell & Owen, 2017) and as gatekeepers between newsreaders and news organizations (Tufekci, 2015; Napoli, 2015). Much of this attention has been oriented towards understanding how algorithms and other "computational processes," for instance the Facebook News Feed, are deployed as "gatekeepers" in ways that are similar to the role that a newspaper editor has played in the past (Tufekci, 2015, p. 206). Because of these similarities, and because of the important role they have come to play within the media industry, I, and others, have argued in the past for platforms to be recognized "as media companies" as a way to link the current conversations about platforms and the public interest to the ones that have occurred in past media eras (Napoli, 2015; Napoli & Caplan, 2018). Platforms, however, have resisted this characterization (House Committee On Energy & Commerce, 2018).

In seeking to understand the impact platforms are having on news media organizations, a number of scholars have turned to institutional theory to theorize on the growing dependencies between platforms and the news media industry (Donges, 2007; Katzenbach, 2011; Napoli, 2014; Caplan & boyd, 2018; Ananny, 2016; Meese & Hurcombe, 2020; Christin, 2020). Much of this scholarship has been oriented towards broadening understandings of media regulation to include new sets of interlinked human and non-human actors, like platforms and their algorithms (Ananny, 2016). Other work has sought to use institutional theory and concepts such as “isomorphism” (DiMaggio & Powell, 1983) to trace the linkages between platforms and the media industry, to co-locate these actors within the same “organizational field” (Caplan & boyd, 2018; Meese & Hurcombe, 2020) or to highlight how news media organizations have re-oriented themselves around the goals and incentives of platforms (Christin, 2020).

This chapter also uses neo-institutional theory as its frame of analysis, as a way to move away from theories of media governance (Puppis, 2010) that place an emphasis on “explicit, formulated rules that regulate and coordinate the behavior of actors in a field, towards a theory that embraces the other structures that coordinate behavior across actors in a field, such as norms, shared beliefs, and symbolic systems” (Katzenbach, 2012, p. 6). In W. Richard Scott’s formulation of institutions, interdependent actors are coordinated through not only the “regulative pillar” that Puppis is interested in, but through “normative” and “cognitive” pillars as well (Scott, 2001, p. 35; Katzenbach, 2012, p. 7). As we will see through the interactions between The Trust Project and platform companies, this coordination is not necessarily bi-directional – one group may internalize the symbolic systems and values of another (in this case, The Trust Project adopting the language of technology) in an effort to become legible to a more powerful actor. As Katzenbach notes, a focus on institutions as the "symbolic and behavioral systems containing representational, constitutive, and normative rules together with regulatory mechanisms that define a common meaning system and give rise to distinctive actors and action routines" (Scott & Meyer, 1994, p. 68; Katzenbach, 2012, p. 7) broadens our understanding of institutions to the informal mechanisms that constrain and shape the behavior of interdependent actors. As Katzenbach notes, it is these structures of coordination – the social interactions and institutional frameworks – that should (and can be) studied through research, rather than the structures of regulation themselves. Though rules are often more visible, they tell us nothing about when and to whom they are applied; the next chapter on YouTube and tiered governance demonstrates the extent to which relationships – between platforms, organizations, and specific creators – have come to mediate the applicability of content moderation rules on the platform.

Concepts that have emerged from institutional theory, such as "institutional isomorphism" also provides a mechanism to begin tracing the impact of interdependencies across an organizational field (DiMaggio & Powell, 1983, p. 147). As danah boyd (2018) and I have demonstrated in past work, this theory can be used to study the impact of platforms – and their algorithms – on the now dependent news media industry. According to DiMaggio and Powell (1983), there are three forms of isomorphism (defined as the process through which mechanisms through which a field becomes more similar or homogenized): *coercive isomorphism* which emerges from explicit rules, regulations, and accreditation, *mimetic isomorphism* which emerges as organizations copy each other's actions as a way to deal with uncertainty, and *normative isomorphism*, which is associated with professional values and ethics. In the article we wrote (Caplan & boyd, 2018), we demonstrated that Facebook has not only exerted a coercive force across the news media industry through setting and resetting standards for how the company calculates “high-quality” news media in their algorithmic News Feed (p. 6), but they also led to significant mimetic isomorphism across the news media industry. This is because the company made constant changes to the News Feed, leading to a degree of uncertainty across the field (Caplan & boyd, 2018; Christin, 2020). This paper makes the case that algorithms are extensions of bureaucratic processes, and should be analyzed as such. As this chapter will demonstrate, however, the non-algorithmic forms of bureaucracy at platform companies also impact how policies and products take shape; in this instance, how the impact of the ethos of Silicon Valley where goals and bureaucracy are constantly shifting (“pivoting”) is experienced by external stakeholders of these organizations (Christin, 2020, p. 63)

This chapter takes on the perspective of a media association, comprised of over 200 news media organizations around the world, that sought to influence the development of content standards at the major platform companies, like Facebook, Google, and Twitter. It examines the effort of this media association to impact how these platforms set standards, including how they categorize and classify content produced by news media organizations. It looks at how these interdependent actors were impacted by inter-and-intra-organizational dynamics as they worked to coordinate their efforts, what structures helped and impeded, and how values were exchanged in this process (in this case, often uni-directionally). What The Trust Project demonstrates is that the process of networked governance described in Chapter 2 can lead external stakeholder groups to make significant investments in the development and creation of content standards, while remaining outside the company, and thus impacted, but not able to influence, internal company dynamics that impact their work.

One of the shortcomings of institutional theory is that it is often criticized for being too descriptive – too caught within context and a particular time and place (Scott, 2001, p. 5). This chapter is no exception to that rule – working to tell the story of one set of relationships between the largest platform companies in the world, and a member of an industry that had found itself suddenly displaced within the new information order.

The Trust Project

For the public, the 2016 election in the United States brought to light many issues and concerns regarding *trust in media*, and the growing role platforms *as intermediaries* were playing in the public sphere. *BuzzFeed News* epitomized these issues in an article published one week after the election, which claimed that "viral fake election news stories" had generated more engagement on Facebook than coverage by 19 major news publishers (Silverman, 2016). Though Mark Zuckerberg, CEO of Facebook, called claims that "fake news" had influenced the election "a pretty crazy idea" (Solon, 2016), the phrase "fake news" had suddenly reached public consciousness.

Though this period brought on a new sense of urgency and visibility, the work Lehrman and her colleagues had begun with The Trust Project was already well underway. At the reconvened 2012 Roundtable on Journalism Ethics, participants were already expressing concerns about the ways that metrics ("clicks") were impacting news ethics and quality. Trust in media in the United States was also already in a steady decline – dipping significantly in 1997 and 2004 and remaining below 50% (Brenan, 2020; Anonymous Interview-2, Personal Communication, December 3, 2020). For members of Trust Project, that this decline coincided with the rise of the mass Internet was no coincidence: “there seemed to be this intersection between the decline in trust in the news, and the arrival of news into the digital space.” In a blog Lehrman wrote with Richard Gringas (then-head of Google News) in 2014, Lehrman and Gringas proposed The Trust Project as *the* solution to this problem of trust, placing the onus on news media companies to include more transparency about how they operate, such as clearly stating their objectives, better labeling their content, and including better citations (Gringas & Lehrman, 2014). In this view, the decline of trust was a result of a failure to adapt older models of media gatekeeping to the digital cultural ethos which, perhaps just rhetorically, emphasized transparency and choice.

At the time, this position was thoroughly critiqued by media scholars Jeff Jarvis, C.W. Anderson, and Emily Bell in a Twitter exchange that the journalist Mathew Ingram captured in a post on *GigaOm.com* (Ingram 2014). For Jarvis, “trust” was still the right frame through which to think about these concerns, but he suggested Google, Gingras’s employer, should play a greater role in “favoring news organizations, journalists, and other sources that follow standards,” integrating these characteristics into search rankings (Jarvis, 2014). But both Bell and Anderson questioned whether “trust” was the right way to think about the value of media, arguing that “trustworthiness” did not necessarily correspond with whether someone found a news source, or story, valuable (Ingram, M., 2014). C.W. Anderson noted that journalism that was “trusted” often had very little correlation with quality, pointing to the news media environment of the 1950s, which centered on trusted news anchors but was notoriously homogenous and narrow in how it presented the news. Still, for publishers like the BBC and The Guardian, the trust they had cultivated with readers was still seen as central to editors. Though their academic peers were not necessarily bought into the idea, Lehrman and Gringas were beginning to find partners across the news media and platform industries, as well as from funders.

Early Partnerships: Control the Algorithm or be Controlled

Though The Trust Project began with a journalist, she included perspectives from Silicon Valley early on. At the 1997 roundtable, Richard Gringas was working in content at a broadband company; by the time the table reconvened, he was the Vice President of News at Google. Though Gringas was not there in an "official" capacity for the search engine, members of The Trust Project acknowledged that his presence, and other connections at platforms "helped open doors." However, his presence, and the presence of other platform representatives, also brought in a way of thinking oriented around the goals of platforms. When The Trust Project brought their concerns about the impact digital platforms were having on quality journalism to these platform employees, they told them the solution was simple: "all you have to do is train the algorithm" (Anonymous Interview-2, Personal Communication, December 3, 2020).

Inspired by the user-centered design efforts she had encountered throughout her experience in Silicon Valley, The Trust Project's founder, Sally Lehrman, quickly adapted to this way of thinking. She began this work in early 2016, working with a human-centered design consultant to conduct "one-on-one user interviews" with Americans across age, gender, and ethnicity, to better understand their "daily news journey" and how readers establish "trust in a source" (Kurjan, 2016). From these interviews, she gathered together executives from 20 news organizations to marry the information gained through these interviews with "journalism values" (The Trust Project, n.d.). There were no platform representatives present at this May 2016 workshop; rather participants came from major news publishers, such as The New York Times, Washington Post, The Guardian, News Corp. the BBC, and El Universal. At the same time, though journalistic values, such as localism, were present, workshop participants had the goal of making these legible to platforms, to "producing useful signals for distribution platforms." This included thinking through how features such as "author bios" could be used within a platform's recommendation system (p. 9).

These signals or "trust indicators" (the preferred nomenclature of The Trust Project) were thought of very early on by Lehrman and other participants as a potential "industry standard and set of tools" to enhance the viability of news in the platform era (The Trust Project, 2016). Though they were developed with platforms-as-adopters in mind, participants also considered how the trust indicators could, *or should*, reshape newsrooms, and how news organizations communicated ethical commitments in the digital era. In that sense, The Trust Project was seeking to establish a set of "standards" for news media ethics in the Bowker and Star (2000) of the term, to serve as a "set of agreed-upon rules for the production of (textual or material) objects," which "spans more than one community of practice (or site of activity)" extended over space and time (p. 13). For The Trust Project, these standards, or indicators, would provide the link necessary to connect a more traditional approach to media ethics, with the needs and norms of platform companies.

The process of developing the trust indicators" was lengthy, spanning across several years through a mixture of in-person summits and design sprints hosted at Hearst Tower (The Trust Project, 2016) and the Washington Post (The Trust Project, 2017). Volunteers separated into several groups to take on certain parts of the project – working groups like Trust Project Branding, UX, and Development, with an additional set of working groups for some of the indicators, such as "Best Practices," "Journalist Expertise," "Citations/References," "Article

Type," "Methodology," and "Local" (The Trust Project, n.d.). At the heart of it stood The News Leadership Council, which advised on the Trust Indicators and core issues related to "information literacy and rebuilding trust in journalism within a fractious, so-called post-fact environment."

Nearly all the volunteers for the Council and the working groups came from the news media industry, with only two volunteers from Google (one of which was Gringas), and one volunteer who had formerly worked with Intel. And yet, there was considerable diversity across news media, an industry known for considerable competition and its resistance to industry cooperation (Anonymous Interview-1, Personal Communication, July 1, 2020). Volunteers came from public, private, and nonprofit media. There were publishers who operated at the national, international, and local levels. Some volunteers came from digital native companies, such as Mic. And though they were largely Western, there were a number of non-English publications involved as well, such as La Stampa/La Repubblica, Deutsche Presse-Agentur, and Zeit Online (The Trust Project, n.d.).

Out of this process came the eight "Trust Indicators" The Trust Project holds as "a widely accepted standard for assessing the integrity behind a news site" (The Trust Project, n.d.). These indicators, which were inspired by the 1947 Hutchins Commission (The Trust Project, n.d.), spread across a broad range of issues and concerns both familiar within the history of media ethics and policy as well as some new additions, which could be viewed as adaptations of media ethics to the platform era (Pickard, 2015; Napoli, 2001). Indicators like "Best Practices," "methods," and "references" are largely aimed at increasing transparency, asking news publishers to be more visible in their source of funding, their mission, and their own standards and ethics, or how long they took to research a story (and their methods). Indicators like "localism" and "diversity" are intended to make newsrooms be more visible (and perhaps rethink internally) their commitments to hiring from the communities they are covering, and their commitments to bringing in diverse perspectives, across race, class, generation, gender, sexual orientation, and region (The Trust Project, n.d.). Lastly, "labels" is aimed more at addressing the convergence (Jenkins, 2006) and context collapse (Marwick & boyd, 2011) concerns introduced by platforms, and involves publishers making clear when a story is opinion, advertisement, or news.

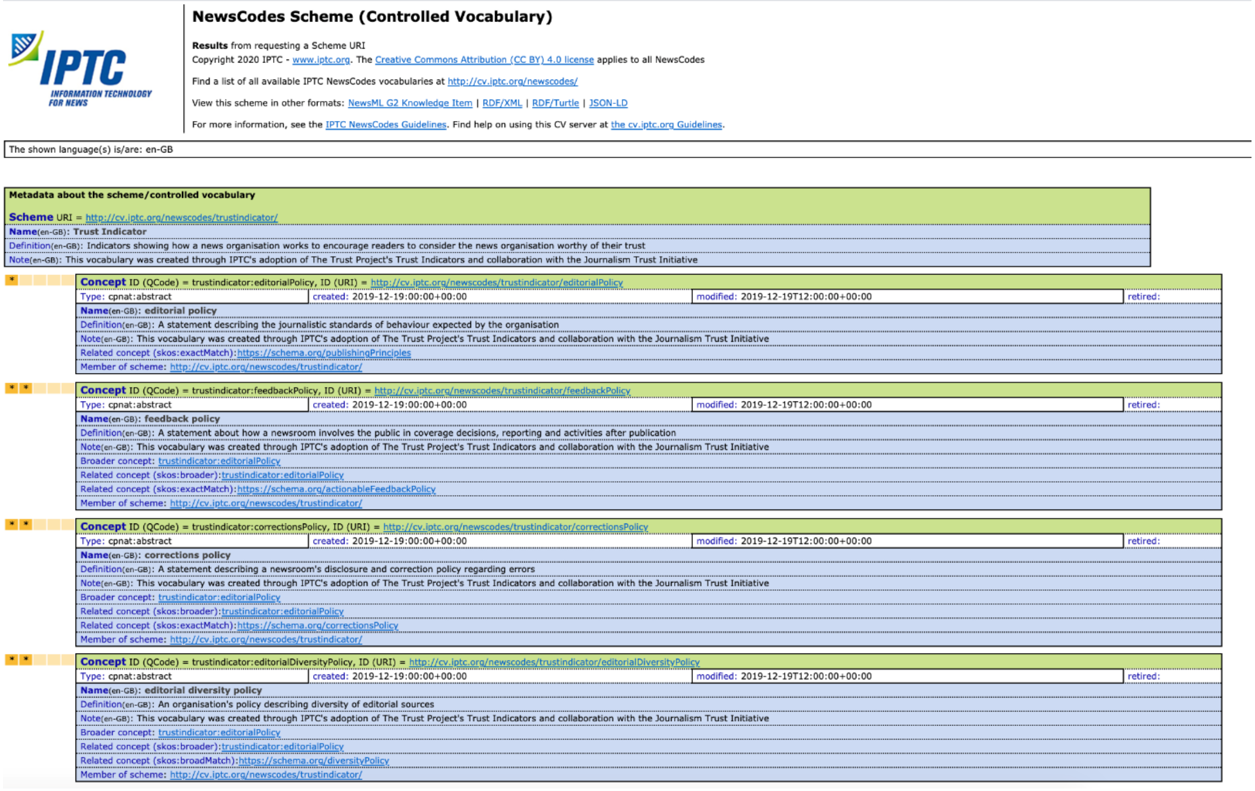
The goals of The Trust Project could be viewed as twofold: (1) To guide publishers on how to make their commitments and standards more transparent and visible to potential readers; and (2) To make publisher commitments "machine-readable" in the hopes that platforms would use this information to differentiate them from their competitors (Smith, 2017).

In pursuit of the first goal, The Trust Project partners would, more-or-less, make visible information about ethics policies and processes like their approach to verification and fact-checking, their ownership/funding, and their approach to diverse voices.[4] In practice, much of this information was housed outside the main publisher's site, within the larger website for the larger corporation or brand. For example, The Economist, one of the publishers that was more vocal about their commitments to The Trust Project features information on mission, ownership, guiding principles, and governance (Smith, 2017)., alongside their general press releases and information about job opportunities on The Economist Group website (The Economist Group,

---

[4] PEN America partnered with The Trust Project to conduct an analysis of newsrooms transparency efforts. Many of the publishers that receive, such as The Guardian, and The Economist, are Trust Project partners.]

n.d.). For the second goal, The Trust Project decided to work with Schema.org, a collaborative effort from the search engines Google, Bing, and Yahoo, to create and support a "common vocabulary for structured data markup on web pages" (Guha, 2011). This markup language is what makes a document *machine-readable* by software, in both presenting information on the page, and indexing. Though Schema.org's NewsArticle markup structure was developed with IPTC, the global standards body for metadata for news media, both Schema.org and IPTC amended their standards to include the "trust indicators" (Figure 4-1) (IPTC News Architecture Working Group, 2020; Schema.org, n.d.). This includes metadata standards for having policies on fact-checking, bylines, unnamed sources, and other broader ethics concerns. These standards thus become part of the structured data layer which is made available to external platforms through a JSON object in a website's HTML (Smith, 2017). In practice, it looks like Figure 4-3 below.



NewsCodes Scheme (Controlled Vocabulary)

Results from requesting a Scheme URI
Copyright 2020 IPTC - www.iptc.org. The Creative Commons Attribution (CC BY) 4.0 license applies to all NewsCodes

Find a list of all available IPTC NewsCodes vocabularies at http://cv.iptc.org/newscodes/

View this scheme in other formats: NewsML G2 Knowledge Item | RDF/XML | RDF/Turtle | JSON-LD

For more information, see the IPTC NewsCodes Guidelines. Find help on using this CV server at the cv.iptc.org Guidelines.

The shown language(s) is/are: en-GB

Metadata about the scheme/controlled vocabulary

Scheme URI = http://cv.iptc.org/newscodes/trustindicator/
Name(en-GB): **Trust Indicator**
Definition(en-GB): Indicators showing how a news organisation works to encourage readers to consider the news organisation worthy of their trust
Note(en-GB): This vocabulary was created through IPTC's adoption of The Trust Project's Trust Indicators and collaboration with the Journalism Trust Initiative

*

Concept ID (QCode) = trustindicator:editorialPolicy, ID (URI) = http://cv.iptc.org/newscodes/trustindicator/editorialPolicy
Type: cpnat:abstract | created: 2019-12-19:00:00+00:00 | modified: 2019-12-19T12:00:00+00:00 | retired:
Name(en-GB): **editorial policy**
Definition(en-GB): A statement describing the journalistic standards of behaviour expected by the organisation
Note(en-GB): This vocabulary was created through IPTC's adoption of The Trust Project's Trust Indicators and collaboration with the Journalism Trust Initiative
Related concept (skos:exactMatch):https://schema.org/publishingPrinciples
Member of scheme: http://cv.iptc.org/newscodes/trustindicator/

* *

Concept ID (QCode) = trustindicator:feedbackPolicy, ID (URI) = http://cv.iptc.org/newscodes/trustindicator/feedbackPolicy
Type: cpnat:abstract | created: 2019-12-19:00:00+00:00 | modified: 2019-12-19T12:00:00+00:00 | retired:
Name(en-GB): **feedback policy**
Definition(en-GB): A statement about how a newsroom involves the public in coverage decisions, reporting and activities after publication
Note(en-GB): This vocabulary was created through IPTC's adoption of The Trust Project's Trust Indicators and collaboration with the Journalism Trust Initiative
Broader concept: trustindicator:editorialPolicy
Related concept (skos:broader):trustindicator:editorialPolicy
Related concept (skos:exactMatch):https://schema.org/actionableFeedbackPolicy
Member of scheme: http://cv.iptc.org/newscodes/trustindicator/

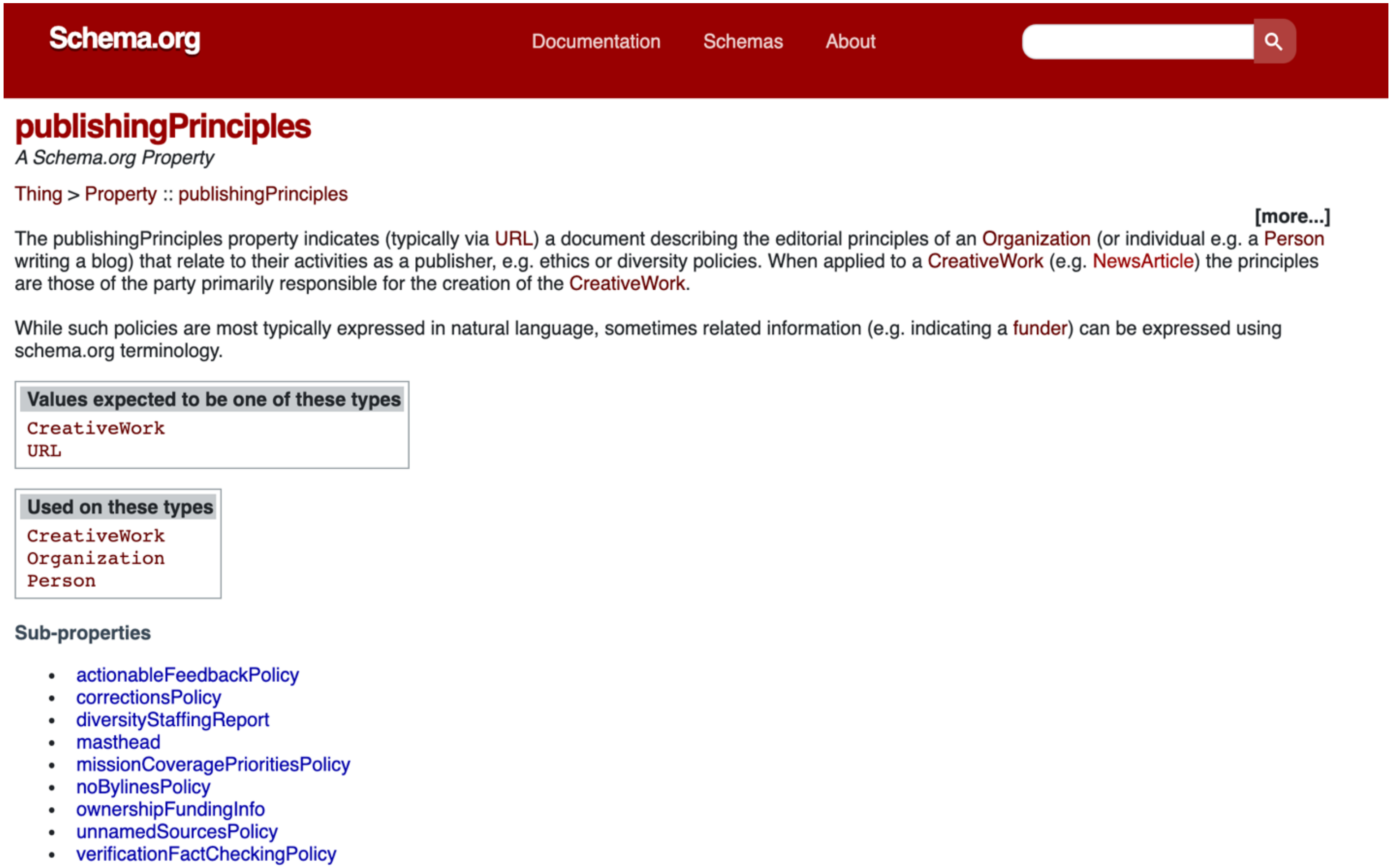
* *

Concept ID (QCode) = trustindicator:correctionsPolicy, ID (URI) = http://cv.iptc.org/newscodes/trustindicator/correctionsPolicy
Type: cpnat:abstract | created: 2019-12-19:00:00+00:00 | modified: 2019-12-19T12:00:00+00:00 | retired:
Name(en-GB): **corrections policy**
Definition(en-GB): A statement describing a newsroom's disclosure and correction policy regarding errors
Note(en-GB): This vocabulary was created through IPTC's adoption of The Trust Project's Trust Indicators and collaboration with the Journalism Trust Initiative
Broader concept: trustindicator:editorialPolicy
Related concept (skos:broader):trustindicator:editorialPolicy
Related concept (skos:exactMatch):https://schema.org/correctionsPolicy
Member of scheme: http://cv.iptc.org/newscodes/trustindicator/

* *

Concept ID (QCode) = trustindicator:editorialDiversityPolicy, ID (URI) = http://cv.iptc.org/newscodes/trustindicator/editorialDiversityPolicy
Type: cpnat:abstract | created: 2019-12-19:00:00+00:00 | modified: 2019-12-19T12:00:00+00:00 | retired:
Name(en-GB): **editorial diversity policy**
Definition(en-GB): An organisation's policy describing diversity of editorial sources
Note(en-GB): This vocabulary was created through IPTC's adoption of The Trust Project's Trust Indicators and collaboration with the Journalism Trust Initiative
Broader concept: trustindicator:editorialPolicy
Related concept (skos:broader):trustindicator:editorialPolicy
Related concept (skos:broadMatch):https://schema.org/diversityPolicy
Member of scheme: http://cv.iptc.org/newscodes/trustindicator/

Figure 4-0-1: The NewsCode Scheme, developed by IPTC. Accessed in February 2021.



publishingPrinciples

A Schema.org Property

Thing > Property :: publishingPrinciples

[more...]

The publishingPrinciples property indicates (typically via URL) a document describing the editorial principles of an Organization (or individual e.g. a Person writing a blog) that relate to their activities as a publisher, e.g. ethics or diversity policies. When applied to a CreativeWork (e.g. NewsArticle) the principles are those of the party primarily responsible for the creation of the CreativeWork.

While such policies are most typically expressed in natural language, sometimes related information (e.g. indicating a funder) can be expressed using schema.org terminology.

Values expected to be one of these types
`CreativeWork` `URL`

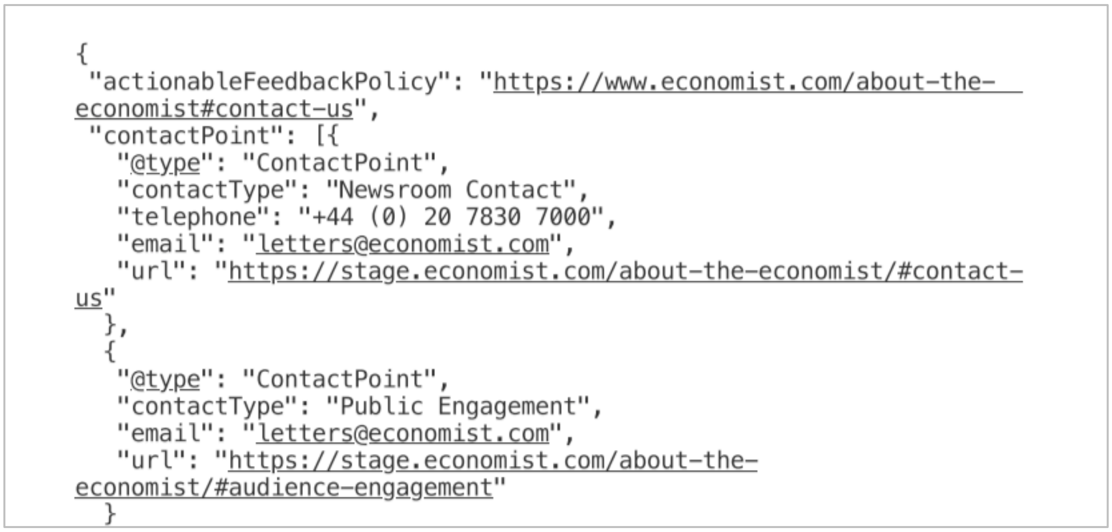
Used on these types

`CreativeWork`
`Organization`
`Person`

Sub-properties

- actionableFeedbackPolicy
- correctionsPolicy
- diversityStaffingReport
- masthead
- missionCoveragePrioritiesPolicy
- noBylinesPolicy
- ownershipFundingInfo
- unnamedSourcesPolicy
- verificationFactCheckingPolicy

Figure 4-2: publishingPrinciples from Schema.org, developed using The Trust Project categories. Accessed March 2021.



```
{
 "actionableFeedbackPolicy": "https://www.economist.com/about-the-
economist#contact-us",
 "contactPoint": [{
   "@type": "ContactPoint",
   "contactType": "Newsroom Contact",
   "telephone": "+44 (0) 20 7830 7000",
   "email": "letters@economist.com",
   "url": "https://stage.economist.com/about-the-economist/#contact-
us"
  },
  {
   "@type": "ContactPoint",
   "contactType": "Public Engagement",
   "email": "letters@economist.com",
   "url": "https://stage.economist.com/about-the-
economist/#audience-engagement"
  }
```

Figure 4-3: JSON example of the Trust Indicators, as described by Smith (2017) in a Medium post for The Economist.

Concerns About Fake News Push Platforms to Sign On

Though the Trust Indicators have now been embedded in markup language and metadata standards, this meant little in terms of broader adoption if the major platforms did not sign on. Though Google, through Gringas, had been an early “unofficial” supporter and, by this point, a funder (Chang, 2017), they had not made a commitment to using the indicators in the ways Lehrman and the other publisher had hoped: to differentiate publishers who had *used the indicators* from those that did not.

By November 2017, many of the major platforms were making public commitments to The Trust Project, even if these commitments did not clearly outline how they would be using the indicators, and under what circumstances. By October 2017, Facebook had tested a new feature which would give users more context into the articles they saw on the News Feed. By

November, Facebook had begun to display the Trust Indicators as part of that module, allowing a small number of publishers to “upload links to additional information through their Brand Asset Library under their Page Publishing Tools – including information on their ethics policy, corrections policy, fact-checking policy, ownership structure, and masthead” (Facebook, n.d.). Around this same period, Google also announced they would be working with The Trust Project towards a “labeling effort,” however, they did not indicate how they would be using the trust indicators, nor how they would be displayed next to articles appearing on Google News or Google Search (Chang, 2017). The Trust Project also claimed in a press release that Bing and Twitter also agreed to use the indicators (Santa Clara University, 2017), however, these companies have not made a public statement about their commitment to the project independently.

Platforms and third-party services have used The Trust Project to inform ratings and indexes on the “reliability” of news outlets (Caldwell, 2019). NewsGuard, a browser tool that provides “trust ratings” for news sites founded by journalists Steven Brill and Gordon Crovitz, came to Lehrman and asked if they could use The Trust Project as a foundation for their own rating system (Anonymous Interview-1, Personal Communication, July 1, 2020). The tool is used by Microsoft in their Edge mobile browser to warn users of “untrustworthy news sites” (Warren, 2019). NuzzelRank, a site which bills itself as the “authority ranking of news sources” has also partnered with The Trust Project to calculate a publisher’s “NuzzelRank” score (NuzzelRank, 2018). The project has played a small role in helping to revise categories platforms are making to classify news media versus other content. Facebook said The Trust Project, along with feedback from “dozens of publishers” such as Axios, The Economist, and Bloomberg, were involved in the creation and testing of their Ad Archive, a system the company introduced in September

2018 which, controversially, decided to include promoted stories by news publishers as part of a public political ad archive (Moses, 2018). Feedback from these publishers led to Facebook tweaking this policy, separating news publishers from traditional political and issue ads (Moses, 2018).

Though the platforms have been quite muted in how they say they are using the trust indicators, members of The Trust Project have noted their involvement. Often the involvement expressed by publisher partners seems overstated in light of the limited ways platforms have publicly declared they are using the trust indicators. For example, when announcing their involvement with The Trust Project, the Liverpool ECHO noted it was "backed by Google, Facebook, and Twitter who have all agreed to use the indicators to identify and supply you with quality journalism from media organizations who are open and honest" (Machray, 2017). When the site Sci.Dev.Net joined the association, they noted that the machine readability of the trust indicators means "Google News and Facebook can use algorithms to identify and clearly label quality journalism on their platforms" (Deighton, 2018). The East Bay Times framed the potential similarly, saying "the indicators will also be used by Google, Facebook, and other platforms to help identify legitimate news sources" (Chase, 2018). In rare cases, a publisher would announce their involvement without the mention of technology platforms. For instance, CBC News, Canada's public broadcasting network did not mention Facebook, Google, Twitter, or Bing in their announcement, instead focusing on how transparency standards "will help audiences to assess whether news 'comes from a credible source,'" and pointing to other similar news agencies that had already signed on, such as the BBC, dpa news agency in Germany, and La Repubblica in Italy (CBC News, 2018). However, the use of the trust indicators as a "standardized technical language" has consistently been front and center in terms of how the project presents itself (Lohse, 2017).

Tracing the Use of the Trust Indicators by Platforms

And yet, though platforms were, at certain points in The Trust Project's history, eager to sign onto the project publicly and through funding, exactly how these platforms *are using* the trust indicators remains incredibly opaque. When asked about how platforms are continuing to use the indicators, one member of The Trust Project was unclear, saying "I can't tell you exactly how they're using it." Though they know they *are*, partly because they have gone back and forth with the companies in how she communicates their involvement in The Trust Project's press releases, and that the companies have noted they use the trust indicators in their "systems" (Interview 2). According to members of The Trust Project, news partners are also reporting they are seeing differences in performance in search after they implement the indicators (Anonymous Interview-1, Personal Communication, July 1, 2020), however, The Trust Projects did not provide evidence to support these reports.

In interviews with one member of The Trust Project, this lack of clarity in how platforms are using the indicators is a constant concern. This opacity was an undercurrent in our conversations, and frequently, when I asked *how exactly* the trust indicators were being used, it felt almost like an untouchable subject. In some cases, like with Twitter, though they had committed at one time or another to using the indicators, and had engaged in several conversations with The Trust Project, the interviewee was not sure the extent of their use (Anonymous Interview-1, Personal Communication, July 1, 2020). In other cases, previous uses of the trust indicators, such as in the “context” button for news on Facebook, were no longer actively being worked on by the company (Anonymous Interview-2, Personal Communication, December 3, 2020). In other cases, platform companies have said they are using the indicators to guide internal processes at the company, such as for human raters being used to train algorithms (Anonymous Interview-2, Personal Communication, December 3, 2020).[5] When pressed for how exactly these companies are continuing to use the indicators, the interview subject noted they had “lost track” of how they were being used.

This uncertainty was certainly at odds with the public commitments that stakeholders made; beyond The Trust Project, publishers, *and* platforms had all, at various points, noted their commitment to the project. Behind-the-scenes, however, platforms had expressed some hesitation with the project, and members of The Trust Project began to perceive a clear difference in values between how publishers and platforms were approaching the issue of trustworthiness and authority in information shared online. The Trust Project found that relationships with platforms were strained by different expectations regarding the timing and scale of the project (Anonymous Interview-1, Personal Communication, July 1, 2020) and there was a distinct perception of those in The Trust Project that platform companies did not understand that the “community organizing approach” The Trust Project was taking, which

---

[5] Google uses human raters to evaluate its search engine quality. Though the trust indicators are not mentioned by name in their public guidelines, the interview subject did note that the trust indicators seem to be inherent in some of the guidelines given by Google. See more here: https://static.googleusercontent.com/media/guidelines.raterhub.com/en//searchqualityevaluatorguidelines.pdf

required "lots of different voices, diverse voices" to come to "collective decisions" takes time (Anonymous Interview-1, Personal Communication, July 1, 2020). They also felt some resistance from the platforms because The Trust Project was unable to scale quickly (Interview 2). Though The Trust Project has grown from a list of 50 partners to now over 200 publishers (as of April 2021), they felt pressure from the tech companies to grow faster. This need to scale faster was at odds with the values of The Trust Project, which wanted to be able to be able to assess how closely publishers were hewing to the standards they had developed (Anonymous Interview-2, Personal Communication, December 3, 2020). Though The Trust Project had already embedded what they perceived as the terminology and processes of the technology industry within their method (focusing on the need for "machine-readable" standards), they were unable to operate according to the technology industry's ever-present goal of scale.

However, there were other organizational barriers that prevented the collaboration from moving forward, or at least, added to the uncertainty The Trust Project perceived about how their work was used by the company. These barriers were grounded within the operations and organizational dynamics of the platform companies (Caplan, 2018). Though personal relationships with employees at platforms helped organizers get through the door, and served as important "ambassadors" for the project (Anonymous Interview-1, Personal Communication, July 1, 2020), The Trust Project found that shifts within the company impacted their ability to move forward. These shifts often took the form of constant personnel changes (The Trust Project member referred to it as "rapid turnover"), which impacted *who* at the company was responsible with maintaining momentum for the partnership (Anonymous Interview-2, Personal Communication, December 3, 2020). Why these shifts occurred, and even when, was often opaque to the external Trust Project. They noted that one platform company did not adopt the trust indicators because of “turmoil within” the company (Anonymous Interview-2, Personal Communication, December 3, 2020). In other cases, the relationship was hindered by changes in staff and reprioritization within the company; however, they were not often provided with much detail as to what happened (for instance, they were told the project had just been “pushed down 6 months”). In those cases, The Trust Project has to work to re-establish a relationship with the new person responsible for the partnership, and do additional outreach to get on “their calendar” (Anonymous Interview-2, Personal Communication, December 3, 2020). To overcome these barriers, The Trust Project noted they need “strong internal advocates” within the companies (Anonymous Interview-2, Personal Communication, December 3, 2020).

Still, though publishers have noted they see benefits from their involvement within The Trust Project, it is often unclear how the indicators are used. Constantly shifting relationships with platform companies often leaves this external stakeholder *reading into* public statements and commitments made by the companies where they find their ideas may inform a policy, but are not cited (Anonymous Interview-1, Personal Communication, July 1, 2020). In this particular relationship, platforms have shown a willingness to engage in a years-long partnership with the external group, however, seem less willing to credit the nonprofit publicly. In other cases, The Trust Project is mentioned as part of a long list of organizations that informed a particular policy or standard; this was the case in a 2020 announcement by Facebook that they were “Prioritizing Original News Reporting on Facebook” (Brown & Levin, 2020). This one-of-many strategy is in line with theories of networked governance that have been outlined in the previous chapter, which can contribute to a lack of understanding and clarity about the respective contributions of networked organizations.

Benefits for Publishers: “This is not an Industry Association”

For publishers, however, the use of trust indicators for publishers may not be the primary goal in joining the consortium. As part of their public commitments to The Trust Project, publishers mentioned a range of benefits for joining the coordinated effort. As previously mentioned, some mentioned to leaders of The Trust Project they saw performance changes after implementing the trust indicators (Anonymous Interview-1, Personal Communication, July 1, 2020). In a survey conducted by one of the news partners (with the sample recruited via a third party), Reach Plc (formerly Trinity Mirror) they found that implementing the trust indicators “significantly improved readers’ perception of The Mirror across a number of trust related marks” including a nine percent increase in user perceptions of Mirror journalists as trustworthy (Tenzer, 2018). In a similar study conducted by the University of Texas at Austin’s Center for Media Engagement, researchers found that the trust indicators increased positive evaluations of news articles compared to articles without indicators (Curry & Stroud, 2017). In both of these studies, however, the benefits emerge between the relationships between publishers *and their readers*, with an emphasis placed by platforms on how participation in The Trust Project signifies *trustworthiness* to their audiences (Tenzer, 2018) . The potential impact between readers and publishers was also present in other announcements made by entities joining on. For instance, The Mercury News and the East Bay Times framed their commitment to The Trust Project as part of “what we’re doing to preserve your trust” (Chase, 2018). The Canadian Broadcasting Corporation (CBC) also noted the benefits of their involvement in similar terms, saying the project’s aim is to “establish transparency standards that will help audiences to assess whether news ‘comes from a credible source’” (CBC News, 2018).

And yet, there may be other benefits to participation that emerge from the coordination from numerous publishers working in tandem. As one of the leaders of The Trust Project noted, such industry-wide coordination is not common in the news media industry, which tends to be marked by competition and independence (Anonymous Interview-1, Personal Communication, July 1, 2020). For smaller sites, like Sci.Dev.Net, affiliating themselves with The Trust Project helps them address many of the issues of convergence and context collapse that occur over social media. As they note “While Sci.Dev.Net has worked hard to build a reputation for fair and balanced reporting, not everyone knows who we are. It means that for people who come across our articles on social media, we need to distinguish ourselves from the misleading news that is being used around the word to reinforce prejudices and spread misinformation” (Deighton, 2018).

There are other potential benefits of being involved with The Trust Project stemmed from being in coalition with a broad range of national and international publishers working in concert through a single channel to platforms. This broad coalition has, in rare occasions, been used to the benefit of smaller local publications. For instance, when Facebook instituted their policy to include publishers within their Ad Archive publishers found they were suddenly not able to advertise their own stories without being classified as political advertising (Moses, 2018).The Trust Project was able to bring those concerns to Facebook, representing the consortium as a whole (Anonymous Interview-2, Personal Communication, December 3, 2020). Though it’s unclear whether this action in particular led to Facebook changing the policy, The Trust Project had already been working with Facebook on their “news publisher index” which was used by the platform to determine which publishers would be exempt from the rule (Constine, 2018). However, these examples are rare, and though The Trust Project listens to the concerns of publishers, they maintain they are definitively *not* an industry association and are not willing to "lobby" in this way. It was clear from our discussions that The Trust Project did not want to take any actions that could potentially alienate their group from their working relationship with platform companies.

Discussion

This chapter provided an in-depth look into one set of networked relationships between platforms and an external stakeholder group. The aim was to understand the ways that the organizational dynamics *between* and *within* these groups impacted the transfer of values and incentives from platforms to these publishers (as well as the reverse), as a way to explore and understand how *networked platform governance*, the strategic networking and distributing of policy responsibilities to nonprofits, academics, and other experts, works in practice. In the case of The Trust Project, this particular instance of networked governance was used to fill a gap in *subject-matter expertise*, specifically of journalistic practices.

What this research shows is that, in instances like The Trust Project, the exchange of values *between organizations* may not begin on an equal footing. The publishers that were part of The Trust Project early on had already found themselves having to navigate and conform to the incentives and goals of platforms; they began this work by noting their lack of control in relation to personalized recommendation systems, or "the algorithm," which was, in their view, homogenizing content and newsrooms (this is supported by work done by Caplan & boyd (2018) and Christin (2020). In an effort to attract participation from platforms, the group adopted the language and goals of platforms early on, heeding advice from advisors from platform companies that the solution was to “train the algorithm” *in favor of* journalistic values and practices that might be more in line with the *quality journalism* as Lehrman and her group defined it. The Trust Project spent a significant amount of effort conforming to these expectations *before they had official buy-in from platform companies*, working with organizations like Schema.org to translate their coalition-built media ethics and standards into a machine-readable format.

Though platforms were willing to be public with their participation with The Trust Project, particularly in the year following the United States election (during which many academics, policymakers, and journalists criticized platforms for their role in enabling ‘fake news’ to rise), they rarely provided specifics about their involvement. In one sense because initiatives like Schema.org and IPTC are already a collaboration between many of the major platforms, they have already bought into the project implicitly. However, platform companies like Google, Facebook, Twitter, and Bing were consistently opaque about the exact ways they were using the indicators, and how related initiatives that involved The Trust Project changed over time. In a number of cases, such as with Facebook’s “context” button, a project that had been announced with fanfare by the company was allowed to slowly drop off over time, raising questions about the motivations of platform companies to sign on so publicly. For Trust Project, a nonprofit organization, finding evidence of their successes meant reading into past uses of the indicators, as well as relying on constantly changing messengers from platform companies or *reading into* public documents to find language and recommendations they perceived as being drawn from their work. As theories of networked governance discussed in the previous chapter show, though platforms were willing to borrow from the legitimacy of this network in their public communication (in this case, particularly with a network of 200 publishers, many of which had an already adversarial relationship with the company (AdAge, 2018). This did not necessarily translate into The Trust Project's importance within the platform companies, and they found they needed to rely on strong internal advocates to move their work forward.

The Trust Project was also aware that the platform companies were working on a number of different initiatives, in concert with many external stakeholder groups, however, it was often unclear which organization had the ear of the company at any given time. When they were referred to publicly in corporate blogs and statements, they were frequently named as among a group of organizations the company had consulted throughout their process. In one example they were included in a list of four organizations, which included similar organizations such as the SOS Support Public Broadcasting Coalition, the Global Forum for Media Development, and Reporters Without Borders' Journalism Trust Initiative, as well as 20 other global media experts (Brown & Levin, 2020). They were also listed by Facebook as part of a group of organizations involved in Facebook's "journalism project" alongside Arizona State University, the International Center for Journalists, the News Literacy Project, and the public relations group Weber Shandwick (Rob, Lever, & Chapman, 2017). In countless other cases, they were not mentioned (though leaders of The Trust Project perceived the use of their language/standards within press releases and corporate documents). Facebook and Google in particular seemed to be in a constant effort to engage with or coordinate networked organizations on this topic of trust in journalism; there was a broad range of different groups they were supporting, such as the "CrossCheck" fact-checking platform, an initiative with Poynter (Google and Facebook had both signed on) (FirstDraft News, 2017), the Credibility Coalition (Credibility Coalition, n.d.),[6] the Certified Content Coalition which has said they are working with Twitter (RAND Corporation, n.d.),[7] and the Trusting News project out of the Reynolds Journalism Institute out of the University of Missouri (News, 2018), as well as their own fact-checking initiatives (Ananny, 2018). There were so many of these initiatives– particularly ones that included the phrase "trust" in their name – that Nieman Lab offered wrote an article in 2018 titled "So what is that, er, Trusted News Integrity Trust Project all about? A guide to the (many, similarly named) new efforts fighting for journalism" (Schmidt, 2018).

What The Trust Project demonstrates is that, though platforms appear to be opening up their policy-making process through consistent outreach with subject-matter experts and related organizations, these relationships do little to mitigate the opacity that has come to define platforms and their policies. Members of The Trust Project have had working relationships with these companies over several years, and though direct channels to platforms have certainly meant that The Trust Project has been involved in several initiatives, these efforts are challenged by changes within the companies. Internal dynamics to which external stakeholders are not privy – such as changes in positions and role, and reprioritization of goals – can leave groups like The Trust Project scrambling to reorient. This reorientation requires that external stakeholders constantly to not only re-assert their importance and value to platform companies, but also

---

[6] Credibility Coalition has not posted updates on their website since 2019 (See more at https://credibilitycoalition.org/working-groups/).

[7] As of May 2021, the CertifiedContentCoalition.org website is no longer available.

requires a navigation of the organizational dynamics of these companies (finding out who the new employee is responsible for these partnerships, establishing a new working relationship, etc.).

When platforms implement these strategies of networked governance, they can broaden the scope of expertise informing feedback into products and policies. For The Trust Project, the group was frequently engaged by platforms on topics related to journalism ethic and polices, how to define and categorize publishers within platform ecosystems, and how to identify ‘quality’ journalism. And yet, it’s unclear even to this particular group how this feedback is used, and where their work fits into the development of policies (particularly in relation to other organizations that platforms are also engaged with concurrently). As noted by Bogason and Musso (2005), decisions done through networks become more decentralized, placing them even more outside public view. And though more horizontal than if platforms were making these decisions alone,, intermittent engagement of stakeholder groups like The Trust Project means external stakeholders are often putting significant amounts of work in, without any guarantees of buy-in from platforms.

Conclusion

In the previous chapter, I demonstrated how platforms are increasingly making use of relationships with outside organizations as a way to distribute the responsibility for creating and enforcing policies. Companies use this at many levels of these companies and at different stages of product development – it seems as if everywhere you turn there is a new “advisory council” (Medzini, 2021) or “Oversight Board” (Klonick, 2020). This form of self-regulation has now been well documented by scholars like Rotem Medzini (2021), who demonstrate how platforms repeatedly use this strategy as a way to “reallocate[e] content responsibilities to intermediaries.” This phenomenon is comparable to other strategies of networked governance platform companies are using in their operations, as a way to externalize content policy-making to other acts. What this example shows is that these efforts do not make platform governance more horizontal. Rather, external stakeholders doing this content standards and policy work are kept very external to these companies, elevated only when it suits their needs.

When I first encountered The Trust Project in 2016, it felt like finding the grown-ups in the room during a period of chaos and turmoil. It was a collaboration between several well-established national and international publications and I watched as some of the most powerful editors, executives, and journalists at these publications, took the time and care to think through difficult media ethics questions for the global digital media era together. Even before they went to the platforms, it was impressive to see how these publications reached across funding models (public, private, nonprofit), scope (national, international, and local), and across borders, to come to an agreement on a set of global media standards. This process was not without its difficulties; there were frequent disagreements where values would come into conflict – for instance, in requiring bylines, a nonstarter for publications like the Economist – that was thought through (and occasionally resolved) through this process.

Platforms, for their part, were eager to be in the room, listening in, and even funding this and similar efforts (Fischer, 2019), particularly as criticisms mounted regarding their role in facilitating the spread of false information. As a researcher with Data & Society, I was also invited into many of these rooms.[8] I saw representatives from platform companies like Google, Facebook, and Twitter move from event to event. Their presence at these events was viewed as a win in-and-of-itself, though they frequently held back, listening more than contributing, and often resisted any firm commitment to recommendations made by the group. With The Trust Project, early buy-in from a VP at Google meant that this effort seemed to begin with a foot already in-the-door. And their approach, standardizing media ethics across publishers with an eye towards machine readability, felt as if it was designed to increase the *legibility of journalistic practices to platforms*, even if in practice there was sure to be something lost in that translation as media ethics were contorted and flattened into metadata and markup language.

Studying networked governance in practice presents significant difficulties, similar to those challenges experienced by these networked actors. It is almost impossible to move past one's own position within the network, to understand how a powerful node, like the platform, is interacting with other organizations and individuals. Unless you gain access to these platform companies, and can engage with their teams over time (which has been notably difficult for qualitative researchers), it is also very hard to determine how product and policy integrate feedback, and how this feedback is weighed against that of other experts and related organizations. Studying The Trust Project in-depth provided one way to look at these interactions over time; to view a cycle of engagement and disengagement in response to concerns and critiques emerging from the public and press.

---

[8] Insert many of the fake news events you went to between 2016-2020

This approach has limitations. Taking on *one* perspective when trying to understand the behavior of networked actors gives you a very narrow perspective of how events may have unfolded. Ideally, this work would take a more holistic approach, using one event or issue and engaging a range of stakeholders to understand their role in how it unfolded or was addressed. And yet, that approach threatens to flatten networked actors within one horizontal field. This approach, which focused on one specific organization seeking to influence policy at platforms, provides a view into how power is distributed (unevenly) among networked actors.

Chapter 4: Demonetization, Tiered Governance, and the Perception of Unfairness

"What's up, you beautiful bastards?" This is how popular YouTube creator Philip DeFranco greeted his fans in nearly every one of his hundreds of videos—until 31 August 2016, when he warned his followers that his greeting, and other elements of his show, might have to change. New "advertiser-friendly" rules enacted by YouTube would, he warned, impact his capacity to earn advertising revenue on the video-sharing platform (DeFranco, 2016).[1] Some of his videos had already been "demonetized" for containing "graphic content or excessive strong language." The advertising revenue DeFranco earned from views of those videos would cease.

DeFranco's video was a harbinger of nearly 3 years of complaints from YouTube creators, including the "adpocalypse," when some advertisers threatened to pull out of the site after discovering their ads paired with videos encouraging terrorism and hate (Kumar, 2019). In response, YouTube adjusted their rules for creators and demonetized videos across the site, leaving creators frustrated and grasping for some kind of explanation. His video spawned the creation of thousands of similar videos from creators expressing their discontent with what they felt was a formal set of policies designed to de-prioritize user-generated content and prioritize traditional media clips that were more predictably advertiser-friendly.

These demonetization videos provide a strategic entry point for examining the current economics and governance of the YouTube Partner Program (YPP). As this chapter will demonstrate, the monetization of user-generated content is far more complex than even YouTube creators may be aware; even so, their concerns, criticisms, and lay explanations offer a powerful diagnostic of the fundamental tensions YouTube has been attempting to navigate for more than a decade. These tensions have emerged due to several competing forces related to YouTube's "formalization" (Lobato & Thomas, 2015); the monetization of amateur content, the production of YouTube's own content, and the simultaneous embrace of legacy media on the video platform. As Stuart Cunningham and David Craig (2019) have framed it, the history of YouTube should be seen as one of a Northern California company coming to terms with the Southern California "fundamentals of entertainment, and content and talent development" (p.41).

This chapter provides an overview of how multiple contrasting commercial media logics (Van Dijck & Poell, 2013) are simultaneously reconciled and fractured through YouTube's advertiser-friendly guidelines. Though YouTube began with a participatory call to action – to "Broadcast Yourself"– the company quickly worked to structure the underlying labor relations between creators and platform, financially, technologically, and contractually. Revenue sharing, though wrapped in a rhetoric of giving back and "reward[ing]" the "most dedicated community members" (YouTube, 2007b) was used to entice prolific users to generate even more content for the site, increasing the overall value of the YouTube platform (YouTube, 2010). While YouTube partners have a variety of motivations for producing content for the site, some now treat the promised financial compensation as a form of entrepreneurial employment, as part of the gig economy (Burgess & Green, 2018). What may have begun as a "partner" revenue sharing arrangement, a bonus offered to already motivated and prolific creators, has in practice set the terms of the labor of media production at YouTube, imposing specific expectations for users who count on that revenue.

In this chapter, I will examine the complex landscape of these arrangements and what happens when it shifts beneath creators' feet. This work – which specifically examines how users are impacted as platforms work to both reconcile competing media logics, and in their networking of governance – is intended to complement the previous two empirical chapters in a number of ways. In one sense, it uses the issue of demonetization as a way to highlight how platforms use the policies and processes of content moderation (and advertiser-friendly guidelines) to manage the expectations and needs of multiple competing stakeholders. In another sense, it highlights how the vast majority of users producing content for social media platforms like YouTube, experience these shifts in policy as platforms seek to make sense of their position in the management and monetization of content. It demonstrates the ways in which the interorganizational relationships inherent in networked approaches to governance are experienced by the vast majority of users who exist outside of governance infrastructures as defined by platforms (and the myriad ways they try and re-insert themselves back into the policy debate).

Amid the many concerns expressed by creators in “demonetization” videos, three stand out as a diagnostic of the underlying tensions YouTube is attempting to manage:

(1) YouTube Partners articulated a contradiction within the social imaginary of YouTube – the impossibility of squaring YouTube’s stated values as an open platform of expression, with the increasingly cautious rules regarding acceptable content and constantly shifting financial and algorithmic incentive structure.

(2) YouTube’s *tiered governance* strategy, offering different users different sets of rules, different material resources and opportunities, and different procedural protections when content is demonetized, powerfully structures production – in ways substantially different from the “open platform” promise that YouTube and other social media make to their entire user base, and different from the expectations of fairness creators bring to the table.

(3) When rules are ambiguous or poorly conveyed, creators are more likely to develop their own theories for why their content has been demonetized. This opacity provides some creators a tactical opportunity to advance politically motivated accusations about why their content has been demonetized.

Demonetization, the Partner Program, and Advertiser-Friendliness

What distinguishes YouTube from most other platforms is the YouTube Partner Program (YPP), their long-standing practice of sharing advertising revenue with some of their video creators. A few other platforms have developed similar programs, though few are funded through programmatic advertising like YouTube. Medium, the online publishing platform, has a partner program (though "partners" pay a small fee to join) (Grinberg, 2018) funded through subscription revenue rather than advertising (Medium, n.d.). Twitch, the live streaming video platform owned by Amazon, shares with some creators revenue from viewer subscriptions, the sale of virtual goods, and advertising. Some smaller sites including DLive, DTube, Mixer, Portal, Stream, Tsu, and 8 also enable partnerships, and are often floated as alternatives in the debates that follow YouTube's policy changes (Twitch, n.d.; Alexander, 2018b).[9]

If advertiser revenue can be shared, it can also be withheld. "Demonetization" refers to YouTube excluding a specific video from the ad-revenue sharing arrangement, or excluding a creator from the Partner program altogether.[10] Demonetization is usually imposed as a penalty, for videos that violate YouTube's "advertiser friendly" content guidelines, specific to videos in the Partner program; videos might also be demonetized when the terms for participating in the Partner Program change. This is not the same as a video being taken down. A demonetized video remains on the platform, and can still be seen; only the ad revenue is halted. Nor does demonetization prevent the YouTube creator from earning revenue entirely; other means remain, including selling merchandise or taking donations directly from users (Hall, 2018).

Demonetization is just one element of a broader suite of governance mechanisms (Gorwa, 2019) available to YouTube: content moderation, which includes the removal of individual videos or the suspension of entire accounts for violating guidelines around sexual content, violence, harassment, hate speech, or misinformation; the placement of videos behind age barriers or other interstitial warnings indicating graphic content; the removal of videos

---

[9] According to Alexander (2018), YouTubers have also been losing their accounts over promoting alternative platforms on their channels (though we did not independently verify this fact).

[10] This meaning of "demonetization" should not be confused with a second, contemporary meaning, where the same word has been used to describe the efforts of a community or nation to forego material currency in exchange for digital or some other informational form of monetary exchange.

deemed to be copyright infringement, privacy violations, or spam. The Partner Program, then, is only a part of this larger platform governance, but one that is specific to YouTube and platforms that share revenue with their creators.

Over time, YouTube has adjusted the rules and parameters of the YouTube Partner Program – sometimes to the benefit, sometimes to the detriment, of the now hundreds of thousands of creators who receive revenue from it. Though these arrangements are largely hidden from public view, recent changes that tethered that revenue to follower counts and imposed stricter content guidelines, pushed the YPP toward a public reckoning (Alexander, 2018a)– driven in part by YouTube creators themselves, like DeFranco, raising their concerns with their audiences. Frustration has become so widespread that videos about demonetization have effectively become a YouTube genre in their own right. A search for the terms "demonetization" or "adpocalypse" returns hundreds of thousands of videos featuring YouTubers offering their own experiences or commenting on the policy changes in general. Most recently, conservative YouTubers and Republican politicians have pointed to specific cases of demonetization as evidence that YouTube and other platform companies are targeting conservatives for their political views (Birnbaum, 2019).

Competing Media Logics and the Perception of Fairness

YouTube is, by an order of magnitude, the largest video hosting platform in the world, and has been for some time. The site began in 2005 primarily to host the amateur videos scattered across the web – though from the start, its founders already imagined partnering with the biggest producers and broadcasters (Burgess & Green, 2018) . Today, YouTube hosts an immense amount of content from established media networks, multi-channel networks (MCNs), and third-party developers, as well as producing their own content in-house (Burgess, 2012; Lobato R. , 2016; Vondreau, 2016).[11] But they remain, culturally and financially, bound to their predominant role as host for user-generated video. In this sense, they have navigated a similar path as many other major social media platforms: growing wildly by providing an open space for amateur participation, then struggling to fit that participation with viable revenue streams – usually advertising.

Though YouTube began as a site for user-generated content, it has also worked to attract content from more traditional media sources. Though this began as an intentional move to "mitigate the risk of litigation from the major rights holders" (Cunningham & Craig, 2019, p. 44), YouTube has continued to partner with the traditional media industry, at the expense of their creators. Demonetization and the "adpocalypse" is only one such instance of how YouTube negotiates the demands (and risks) of the various users producing content for the platform. These competing users and logics have been defined and theorized by various scholars in different ways – each of which has their own conceptual limitations. Van Dijk and Poell (2013) have articulated the two as "social media logic" versus "mass media logic" (2013), highlighting the ways in which the former operates according to "programmability, popularity, connectivity, and datafication," which is contrasted with "mass media logic" where content creation is more controlled and hierarchical, with roles (between producers and audiences) more strictly

---

[11] According to documents from a copyright-related court case between Google and Viacom, YouTube's founders Chad Hurley, Steve Chen, and Jawed Karim, initially pitched the company to venture capitalists as containing both amateur content, that would "eventually sit alongside legitimately uploaded, professionally produced media content." From Burgess & Green, 2018.

delineated. Conversely, Cunningham and Craig (2019) use the competing business cultures of Northern California and Southern California as a way to distinguish between the “two very distinct, world-leading industrial cultures that are increasingly clashing, converging, and becoming interdependent.” This analogy is useful *because* of the way the theory re-orients us to the distinct business cultures – and the rivalry – of media production associated with old and new media; NoCal, associated with Silicon Valley, emphasizes “aggressive disruption,” iteration, and measurement, whereas SoCal, the location of Hollywood, is more hierarchical, less amenable to change, and has, in the past, prioritized relationships (often at the expense of diversity).[12]
This chapter does not adopt these framings wholesale, but rather introduces them as a way to highlight the deep cultural tensions at work – between traditional media, creators, and platforms – in the demonetization debate and in the governance of content over platforms generally, particularly as platforms seek to converge the NoCal and SoCal business cultures, using demonetization as a way to discourage the creation of content that may not appeal to mass audiences. These tensions, and how they are mediated by platforms, are the thread connecting the chapters of this dissertation, in terms of the values driving this kind of mediation by platforms (Chapter 3), and the advocacy work done by the media association to change their position within the platform ecosystem (Chapter 4). This chapter works to understand the other side of this dynamic – namely, the experience of content creators– as platforms continually adapt their policies to reflect shifting institutional arrangements. The return to the beginning of the

---

YouTube Partner Program is also a return to a different moment in the sociological examination of platforms and digital culture – one in which the early calls for "participatory culture" look hopelessly naïve.

The dynamics exposed through the demonetization debate reflect underlying labor concerns in the creator economy (Caplan & Gillespie, 2020) and concerns about potential exploitation (Terranova, 2000; Scholz, 2013; Bruns A. , 2006; Andrejevic, et al., 2014). However, they also reflect other concerns regarding the governance of content by platforms that, as Philip Napoli (Napoli, 2021) recently argued, reflect an "ongoing discussion about the notion of fairness and how fairness should guide the content curation and moderation practices of social media companies." For Napoli, this focus is specifically oriented towards calls to re-introduce a new Fairness Doctrine for social media, amid broader (currently unfounded) concerns that platforms are privileging one set of views over another. However, the demonetization debate also shows how fairness as a normative ideal has been extended in discussions about social media policies to *different types of media producers* – individuals, media organizations, politicians and public figures – and which sources should be viewed as authoritative and trustworthy. In this sense, platforms are increasingly acting as gatekeeper between media sources and audiences (Napoli, 2015; Caplan & boyd, 2018; Tufekci, 2015). Increasingly, this also has come to mean that platforms are making important distinctions about the source of media content and its trustworthiness and authority (Caplan, 2020).

In the case of demonetization, questions about bias should also be connected to ongoing discussions "algorithmic fairness" that have unfolded in academic and civil rights circles throughout the last decade (Sweeney, 2013; Borocas & Selbst, 2016; Noble, 2018; Kleinberg, Ludwig, Mullainathan, & Rambachan, 2018). This is because, for the most part, policies about advertiser-friendliness are enacted. In the case of YouTube, it is unclear the extent to which algorithmic harms are perceived or actual – YouTube has not always been open about their use in demonetization decisions (Romano, 2019). Additionally, though there are some clear ways in which the demonetization debate intersects with broader civil rights concerns – a number of YouTubers claim demonetization algorithms are targeting them due to their sexual orientation or race – most of the concerns expressed about demonetization are centered on other concerns such as the use of profanity, and political views.

For YouTube creators impacted by demonetization, algorithmic harms are both perceived and actual. YouTube does not provide a significant amount of information about how demonetization algorithms function to their creators; though they provide a public list of terms, as well as examples, that can result in lost advertiser revenue, how the algorithm determines the presence of these terms is not explained (YouTube, n.d.). This information asymmetry between the platform and its users (Rosenblat & Stark, 2016) means that creators on the platform are left to theorize about how the algorithms function, so they can regain their monetization status. In this sense, this work contributes to a broad range of work that has been done on folk theories of algorithms and "algorithmic gossip," a phrase coined by Sophie Bishop (Bishop, 2019) to describe the ways that users create "communally and socially informed theories and strategies pertaining to recommender systems" as a way to increase the potential for visibility on these networks. Though demonetization is not a recommender system, creators also share information about when and whether their content has been demonetized, as a way to make sense of and avoid algorithmic penalization. This work thus contributes to broader work on folk theories of algorithms, and how users make sense of and share knowledge about their labor and compensation (Bechky, 2003; Bucher, 2017) . However, it should be noted that demonetization on YouTube is not determined through algorithm alone; not only has YouTube made individual determinations in the past about organizations and individuals (Rosenberg E. , 2019), humans are inserted at many stages of the demonetization process, including in appeals (YouTube, n.d.).

There are many other reasons why fairness has come to play such an important role in content moderation decisions. For Nicolas Suzor (2018) fairness should just be considered a principle of good governance. Suzor has suggested that, as private companies take on a governing role of the public sphere, that we should "care deeply about the extent to which private governance is consensual, transparent, equally applied, and relatively stable, and fairly enforced" (p. 1). A lack of fairness or "perceived arbitrariness," in Suzor's view, is a "key source of anxiety around governance" (p. 7). This work builds off of work done on "procedural fairness" (also referred to as "procedural justice") by Tom Tyler (Tyler, 2003; Taylor, 1997). Tyler argues that the legitimacy of institutions is often tied to "the fairness of the procedures" through which "authorities make decisions"). Though Tyler's work has examined procedural fairness through the lens of criminal justice in the past, he has recently worked with Facebook to examine how "perceived procedural justice" or "perceived fairness" impacts how users experience content moderation decisions (specifically, on their likelihood to re-offend) (Tyler, Katsaros, Meares, & Venkatesh, 2018). They found perceived justice/fairness significantly shaped whether users would violate the rules following a removal of content.

But as we will see, though perceived fairness plays an important role in content moderation, it is not necessarily a diagnostic on the treatment between individuals (though in some cases it is). Rather, complaints about fairness or a *lack of fairness* unveil underlying tensions about how platforms are creating the categories of advertiser-friendly (and not advertiser-friendly) content and, as an extension, their process of *classification* (Bowker & Star, 2000, p. 11) of different user groups creating content. As the previous chapters have shown, this mediation by platforms – the way they are differentiating between data sources – is part of an ongoing process of standards development by platforms as they work to draw boundaries between trustworthy and untrustworthy sources. This chapter specifically examines how this ongoing development of standards impacts non-traditional media, i.e., YouTube creators or "partners," and how this ongoing process of distinguishing between sources is experienced by creators, who are working to reconcile their beliefs about YouTube as a participatory platform, with the reality of a platform company adjusting to the demands of its other user groups (namely, advertisers).

Method and Analysis

To answer these questions, I collected YouTube videos, beginning with keywords like "demonetization," "adpocalypse," "adsensegate," and "youtubeisoverparty." I identified ninety videos between August 2016 and June 2018. To ensure these videos were diverse in terms of politics and genre, I also paired these keywords with others, like "lgtbq+," "alt-right," "prank," "horror," and "beauty" to identify videos that may not have been popular enough to surface in the broader search, but may have nevertheless been significant in their own circles. I found that ninety videos were sufficient to establish clear patterns in how creators spoke about the experiences of being demonetized, with significant redundancy and overlap as to how these creators perceived how policies were being made and enforced at YouTube.[13]

To situate these experiences within the broader history of the company's practice, I also conducted a discourse analysis of YouTube's corporate communications about the Partner Program over more than a decade, drawn from YouTube's official blog, and the YouTube Creator Blog. I used the WayBackMachine to track when Partner Program policies had been instituted or adjusted, and how they changed over time. I drew on additional secondary sources, coverage in the technology press, and trade publications whenever possible.

Broadcast Yourself…But Keep it "Advertiser-Friendly"

Frustration about demonetization has grown over time, particularly as the policy, and how it has been implemented, shifts frequently. There have been at least seven periods of demonetization that have fueled and structured the public debate. Each revolved around a controversial policy change in which YouTube re-negotiated and re-defined who counts as a partner, what content was acceptable for advertisers, or how the rules and decisions were conveyed. In response to each, creators took to the platform to express their discontent and disillusionment with the increasingly attenuated promise of shared revenues. It should also be

---

[13] In a previous co-authored version of this paper (Caplan & Gillespie, 2020), we also integrated six additional interviews with YouTube partners ranging from smaller subscriber counts (below 1000) to larger ones (above 100,000). These interviews will be referenced (as needed) as part of the original paper, and not as a primary source.

noted that this should not be considered a definitive list and different writers on the topic (Alexander, The Yellow $: A comprehensive history of demonetization and YouTube's war with creators, 2018b) tend to conflate one or two of these periods together. However, these are topics of debate, and time periods, that informed the collection of videos for analysis.

- August – December 2016: The #YouTubeIsOverParty, which began after DeFranco and others received notifications that their content had been demonetized. In the context of this dissertation is important to note that this predated the U.S.-election and subsequent concerns about political content online;
- March – May 2017: "The First Adpocalypse,"[14] after advertisers expressed concern about their ads being paired with terrorist content (during this time, Disney also cut ties with YouTuber PewDiePie for sharing anti-Semitic imagery (Alexander, 2018a);
- August – November 2017: the "Yellow Dollar Sign" controversy, after YouTube added icons to creators' dashboards indicating which videos have been demonetized, fueling suspicion about YouTube's algorithmic methods of policing videos;
- November – December 2017: "The Second Adpocalypse," following another wave of advertiser concerns, this time about lewd comments on videos featuring children (Handly, 2017);

---

[14] The First, Second, and Third Adpocalypse are the terms given by the YouTube community for theses controversies, as indicated through the YouTube fandom wiki, available here: https://youtube.fandom.com/wiki/YouTube_Adpocalypse

- December 2017-January 2019: The “Logan Paul incident” after the prominent YouTuber uploaded a video depicting the body of a suicide victim, and the public outcry led to changes regarding creator access, privileges, and benefits (YouTube, “Creator influence,” n.d.);
- January-February 2018: The spam change, after YouTube announced a significant policy change that affected who had access to the Partner Program, excluding small channels under 1000 subscribers (YouTube, 2018, December 13);
- February 2019: “The Third Adpocalypse,” following the release of a video by Matt Watson demonstrating how YouTube family channels were being used by bad actors and pedophiles, leading to more advertiser pressure (Alexander, 2019).

DeFranco was not the first to complain about how the Partner Program had deviated from its initial promise (Alexander, 2018b) but he was one of the first high-profile YouTubers to do so. His videos helped spur a public discussion that soon consolidated around the hashtag #YouTubeIsOverParty (Ingram, 2016; Kircher, 2017a; Maheshwari & Wakabayashi, 2017); Maheshwari & Wakabayashi, 2017). Creators across the site were taking issue with the vague “advertiser-friendly” guidelines that YouTube claimed it used to determine whether a video was acceptable for revenue sharing. Because these guidelines included broad prohibitions against “controversial or sensitive subjects and events, including subjects related to war, political conflicts, natural disasters and tragedies,” YouTubers like DeFranco challenged them as tantamount to the “censorship” of political and cultural expression (DeFranco, 2016). Curiously, YouTube had not actually changed its “advertiser-friendly” policies; only how users were *alerted* to violations (Kain, 2016).[15] But what this new clarity made clear was that, YouTubers earning a share of ad revenue must meet different content standards for what is acceptable to advertisers, and their videos can lose monetization for violating those rules.

These experiences of creators were frequently at odds with the claims being made publicly by YouTube, fostering suspicion between the platform and creators, and between creators themselves. Among YouTubers, suspicion that advertisers were setting YouTube's agenda became persistent. This was because many of YouTube's policy changes appeared to have been triggered by major news coverage about potential harmful or predatory content on the network (Wakabayashi & Maheshwari, 2019), which often included a very public outcry from major advertisers. For example, in the first adpocalypse, the companies Johnson & Johnson and AT&T threatened to pull their advertisements following reports about terrorist content and antisemitism on the site (Mostrous & Bridge, 2017). During the second adpocalypse, major advertisers like Adidas, Deutsche Bank, Cadbury, and Hewlett-Packard froze advertising on the platform after finding their advertisements paired alongside content that included material that featured child-exploitation (Ong, 2017). In response, YouTube improved enforcement of its

---

[15] A search through YouTube's past policies confirms that the same five categories of "sexually suggestive content," "violence," "inappropriate language," "promotion of drugs," and "controversial or sensitive subjects and events, including natural disasters and deaths," were present as early as March 11, 2015, with this last category expanded to include "subjects related to war, political conflicts, natural disasters, and tragedies" later that year (YouTube, 2015, September 6).

guidelines for videos aimed at children, removing ads from over 3.5 million videos from June to November 2017 (Dave, 2017).

Creators had mixed reactions to YouTube's responsiveness to advertisers. Many were sanguine about the reality of an advertising supported model, sympathetic to YouTube's financial bid. One individual noted that "most creators won't admit it, YouTube is a business." Another commended YouTube for actually sharing advertising revenue with creators at all, something few other platforms had done. Some felt that "creators have to keep advertisers happy" as an entrepreneurial strategy because advertisers are not bound to the platform. [16] But others rejected these guidelines as contrary to YouTube's ethos and the central promise of the Partner program. Some felt duped; one lamented that YouTube had been a space where he felt free to speak unhindered, that was not "regulated and censored in the way that sterile, safe, sanitized, boring, and brain-dead old television had become."[17] PewDiePie, the site's most popular user throughout this period, echoed this sentiment, warning that YouTube was being "forced" by advertiser pressure to turn "into television."[18] Another noted YouTube's hypocrisy, discouraging non-advertiser-friendly content while also depending on a recommendation algorithm that "rewards you for being an edge lord," i.e., being as 'edgy' or risqué as possible.[19]

---

[16] From YouTube video ID #8.
[17] From YouTube video ID #16.
[18] From YouTube video ID #19.
[19] From YouTube video ID #80.

Beyond the sense of indignation, YouTube creators also confessed to their audiences that demonetization meant significant material and social losses, which they felt were also at odds with YouTube's narrative. Exact figures are difficult to come by and easy to misrepresent. Some YouTubers reported losing as much as "97% of ad revenue,"[20] while others described drops not quite so drastic, but nevertheless significant. Others enumerated the significant investments amateur production actually required, whether material resources like camera equipment (Caplan & Gillespie, 2020),[21] "shoes and clothing," software like Final Cut Pro,[22] and even just the costs of charging cameras.[23] Others noted the massive commitments of time necessary; one YouTuber noted she "puts more work into her YouTube channel than when she did a 9-5."[24]

To counter the loss of revenue, many implored their audiences to support them financially in other ways (Caplan & Gillespie, 2020). Many urged fans to donate through Patreon – a crowdfunding service widely used by artists and creators – or to buy related merchandise, both now common tactics for YouTubers needing to diversify their revenue streams[25] and limit the risk of, as one creator put it, "putting all our eggs into the YouTube basket."[26] In a few cases, YouTubers fretted that the lost revenue was reason enough to leave the platform. One noted that

---

[20] From YouTube video ID #8.
[21] From YouTube video ID #5.
[22] From YouTube video ID #74.
[23] From YouTube video ID #64.
[24] From YouTube video ID #5.
[25] From YouTube video ID #60.
[26] From YouTube video ID #27.

she had stopped “relying on YouTube,” noting the loss of revenue had affected her “motivations to make videos here.”[27] For some this meant a return to full-time employment and a “normal job,”[28] or a migration to alternative platforms like Twitch,[29] or DLive, or to cryptocurrency-based alternatives like DTube and Steemit. One creator described how one alternative platform had used the demonetization crisis as a way to recruit him to create exclusive content (Caplan & Gillespie, 2020). Still, network effects kept most creators tied to the platform, with one YouTuber lamenting, “it’s still the biggest audience.”

Tiered Governance and the Perception of Fairness

The #YouTubeIsOverParty was only the first public interrogation of YouTube’s revenue sharing arrangements. Over the next three years, advertiser concerns led to additional changes in how YouTube implemented its policies across the YouTube platform. In many cases, these changes highlighted tensions around YouTube’s *tiered governance* approach, in which different users – amateurs, professionalized amateurs, legacy media organizations, and YouTube’s contracted producers of original content – are held to different standards in different ways. A frequent complaint in the demonetization videos, was that YouTube was prioritizing the interests of advertisers over the needs of creators: that established media personalities were seen as “advertiser-friendly” and were thus being treated differently by the platform (Dwoskin, 2019),

---

[27] From YouTube video ID #88.
[28] From YouTube video ID #31.
[29] From YouTube video ID #87.

that user-generated content was policed separately, more strictly, and through different mechanisms – including an over-reliance on flawed automation techniques.

Jimmy Kimmel, host of *Jimmy Kimmel Live!* was frequently held up as evidence that the "advertiser-friendly guidelines" imposed on amateur creators were not applied equally to media partners. One series of videos pointed out that YouTube demonetized videos by Casey Neistat – a well-established YouTuber with millions of followers – that discussed the 2017 Las Vegas shooting, while Kimmel's monologues discussing the same tragedy remained, with advertisements.[30] Others pointed to Vevo as indicative of YouTube's hypocrisy. Though creators were demonetized for "bikini try-ons" that the platform argued violated their terms of service, Vevo music videos with sexual suggestive content continued to collect advertising revenue.[31]

Some felt that YouTube was "picking and choosing who they want to have in their Premium" (Caplan & Gillespie, 2020). Neistat himself argued that YouTube may treat Kimmel differently because the partnership with ABC is structured differently, allowing ABC to sell *its own ads* alongside the content it posted.[32] Like many, Neistat believed that selective enforcement of the rules, uneven partnerships, and sudden losses of revenue were hurting all those creators who did not enjoy such deals, noting the community had "lost faith in YouTube."

But focusing on this one distinction, between amateur creators and established media partners, belies the complex landscape of distinctions YouTube makes between many types of

---

[30] From YouTube videos ID #41, #42, #48.

[31] From YouTube video ID #2, #67.

[32] From YouTube video ID #48.

creators: “whitelisted” media partners that sell their own ads (Alexander, 2017), the NonProfit Partner Program (YouTube, n.d.-d), channels supported by third-party multi-channel networks (MCNs), YouTube’s own content, and YouTube’s Partner Program. YouTube also maintains an elite set of the most popular and acceptable channels that advertisers can pay a premium to be paired with; these “Google Preferred” channels are selected according to oblique, algorithmic criteria including their popularity, engagement, and propriety (YouTube, n.d.-b). But even within the Partner Program, YouTube has instituted a tiered structure, offering benefits and material resources to some more than others, based on factors such as subscriber count, engaged time, and other measures of popularity (Popper, 2017).

There have always been constraints about who gets to be a YouTube Partner. When the program was first launched in 2007, YouTube hand-selected who could join, prioritizing “the most popular and prolific content creators” (YouTube, 2007a). Creators like Lonelygir15, smosh, and HappySlip were chosen because they had, according to YouTube, “built and sustained large, persistent audiences through the creation of engaging videos,” content that “has become attractive for advertisers.” Even as the YPP was extended to the broader YouTube community, the offer was limited to specific geographic regions: first to those within the United States and Canada, with Japan, Australia, Ireland, Brazil, and Spain added the next year (YouTube, 2007a ; YouTube, 2018b).

The broadened YPP came with parameters for who could be included. Partners needed to have a large subscriber base and regular engagement; they also had to remain in good standing with the platform, ensuring their content not only follow the site-wide “Community Guidelines” imposed on all users, but also that it be “advertiser friendly” – a second set of more restrictive guidelines Partners must meet, with “signals like community strike, spam, and other abuse flags” used to determine inclusion (YouTube, 2018b).

Eventually, the Partner Program was systematized into an explicitly tiered structure, distinguishing not only between Partners and not, but between Partners of different prominence, a “ladder” (Arnstein, 1969) that rewarded creators for increased engagement. In 2011, YouTube announced a “Medals” program (which built off of their existing “Honors” and “Goodies” programs, though we could find no information about those programs from official YouTube sources). These medals – Gold, Silver, and Bronze (Fratella, 2016)[33] – rewarded users who had the most all-time engagement (Wilms, 2011). At first, these were merely awards, symbolic commendations of a YouTuber’s success. Eventually, the medals became “benefit levels” (YouTube have since added lower categories such as Graphite and Opal), each indicating a higher stratum of subscribers and more “watch-time.”

These are more than mere commendations. Though the specific benefits have changed, each tier offers material resources and additional opportunities that could be used to further a creator’s career on the platform. For instance, while all partners have access to YouTube’s Creator Hub website (YouTube, 2012a), Opal Partners receive special invitations to onsite events and workshops that train creators in production techniques and audience management. Bronze

---

[33] Graphite for under 1000 subscribers, Opal for 1000-10,000 subscribers and more than 1000 hours of watch time in the past year, Bronze for 10,000-100,000 subscribers and 10,000 watch hours, and Silver for attaining more than 100,000 subscribers and 100,000 watch hours. In addition, much like the music industry, awards are given to those who cross the million and ten million marks.

partners get access to “creator spaces,” decked out studio spaces owned by YouTube that include equipment like “DSLRs and cinema cameras” (YouTube, 2012b)– privileging creators near enough to London, Los Angeles, Berlin, Mumbai, New York, Dubai, Rio, and Toronto, or with the sufficient funds to get there. Silver partners, with more than 100,000 subscribers, are each assigned a YouTube Partner Manager who promises one-on-one support, as well as insider access to new products and features (YouTube, n.d.-c).

These tiers thus offer not just different material benefits, but also differentiated access to the company itself. A direct line to a YouTube representative means unique expertise and resources, but it also offers an opportunity to build up trust with a real human point of contact within the company – who might then be a source of much-needed clarification of internal policies, an avenue for more directly appealing moderation or demonetization decisions against the creator, or a source of advice on how to carefully avoid such decisions (Klonick, 2018, p. 1654). Caplan and Gillespie (2020) found that creators with smaller audiences had very limited means to contest a demonetization, while more popular creators used their Partner Manager or other people they knew at the platform as a backdoor. Kumar notes that YouTube quietly changed its appeals process to benefit established stars: only videos with at least a thousand views in a week could be re-evaluated by a human, but this restriction did not apply to channels with over ten thousand subscribers (Kumar, 2019, p. 6).

This tiered governance is wholly defined by YouTube. The terms of participation can be changed by the platform arbitrarily, unilaterally, even capriciously – changes that could have an immediate impact on a creator’s audience size and reach. This was the issue in the “small channel” purge in February 2018. YouTube tightened who qualified for the Partner Program, raising the number of watch hours and subscribers required (Welch, 2018). Smaller YouTubers who had already been receiving ad revenue were suddenly dropped from the program altogether. Though some acknowledged that this change might benefit creators overall by easing competition in the YouTube attention economy,[34] others again took to the platform to complain about the new policy and how YouTube had implemented it. As one trans YouTuber worried, smaller channels from more marginalized communities might feel compelled to become more like "what mainstream America wants to see," or use "clickbaity titles" to bring in more traffic, all to ensure their sustained revenue.[35]

Policies have to occasionally change. But this degree of unevenness, as a feature of the "private governance by platforms" (Gillespie, 2018; Suzor, 2019; Klonick, 2017; Caplan, 2020), can violate one of the "most commonly agreed upon principles of the rule of law… that rules are applied equally and predictably" (Suzor, 2018, p. 6). Suzor notes that the "perceived arbitrariness" of rules being applied differently can be a "key source of anxiety around governance" (p.7). This anxiety, propelled by information asymmetries and a lack of communication from YouTube about their policy changes, leads to a plethora of sometimes conspiratorial explanations from YouTubers about why content was demonetized, or why the rules were changing.

---

[34] From YouTube video ID #31.
[35] From YouTube video ID #78.

Information Asymmetries and Alternative Explanations

On August 7, 2017, YouTube implemented another change, not the *rules* themselves but rather, *how* Partners would be alerted to the fact they had been demonetized (YouTube, 2017a). It may not be immediately obvious to a user when they've had a video demonetized: users who earn significant revenue typically produce many videos; and, those revenue streams vary week to week based on viewership. In 2017 YouTube added three new icons to users' dashboards, to indicate the revenue-generating status of each video: Green meant "videos that can earn money from the broadest set of advertisers and from YouTube Red;" Yellow for videos that have "limited or no ads because the video has been classified as either not suitable for all advertisers, or has been fully demonetized;" A dollar sign with a strike through it meant the video was fully demonetized due to a copyright infringement, Content ID claim, or Community Guidelines violation. YouTube made additional changes to how users could appeal a decision and what information about the "status" of their appeal was provided. The icons, meant to offer greater transparency, were not well received.

The "Yellow Dollar Sign" controversy highlighted broader tensions around the information asymmetry or "information flux" (Bossewitch & Sinnreich, 2012) between creators and the platform company. YouTube creators have long complained not just about being demonetized, or the capricious, ad hoc, and unpredictable changes in the rules they so often made, but about YouTube failing (or avoiding) to communicate the rules and expectations clearly. Even as communication was being improved, opacity in how the rules were applied, along with the lack of access to personnel at YouTube who might explain, led users to develop their own theories about how and why videos were selected for demonetization.

Around this time, YouTube was relying more on machine learning algorithms to help determine whether a video was "advertiser-friendly" (YouTube, 2017b). With the dashboard icons, it was easier for creators to scrutinize which videos had been demonetized and which had not. Creators began noticing that both current[36] *and past* videos had been demonetized.[37] By matching this with drops in revenue, creators surmised that the detection algorithm had actually been in place for several months before it was announced, and that its criteria were changing.[38] With so little information to go on, YouTubers struggled to understand what exactly in their video was being judged unacceptable, leading to "an anxiety laden environment of second-guessing, self-surveillance and continuous tweaking" (Kumar, 2019, p. 7). One YouTuber criticized the system as "pure AI logic at work," providing no information about "why the video is flagged, just that it is."[39] This left Creators with guesses rather than guidance as to how to stay within the rules in future videos: one beauty vlogger noted that without an explanation, "I can sit here and speculate all day long….we can't even curate our content to fit their algorithm."[40]

Of course, "being in algorithmically controlled spaces entails trying to figure out how the algorithmic mechanisms of the platforms they use work" (Bucher, 2018, p. 139). In the absence of clear communication from YouTube, creators can draw their own conclusions (Morris, 2018;

---

[36] From YouTube video ID #50.
[37] From YouTube video ID #44.
[38] From YouTube video ID #37.
[39] From YouTube video ID #37.
[40] From YouTube video ID #39.

Kumar, 2019) about how and why YouTube governs the Partner Program in the way that it does. Some YouTubers theorized that the demonetization algorithm was parsing keywords, metadata, and text, and might unfairly penalize words with multiple meanings – all by using a form of reverse engineering (Diakopoulos, 2014; Kitchin, 2017), changing tags or titles to see if the algorithmic system remonetized their content. A creator who consistently posted about LGBTQ+ issues reported running an experiment, in which she gave the same video two titles – "gay couple" and "straight couple." The "gay couple" video was demonetized, the other was not.[41] What may have been a detection error, the YouTuber took as a "form of censorship," with the system treating the word "gay" "as if it's a dirty word." Another YouTuber asserted that every time he used the word "ISIS," he was demonetized, leading him to believe the algorithm equated all discussion about ISIS with pro-ISIS content.[42]

For YouTubers who discuss controversial or contested subjects, it was easy to suspect they were demonetized *because of* their beliefs. Far-right YouTubers have gained particular traction with their audiences by asserting that YouTube is targeting them for their political beliefs, ending in their "mass demonetization."[43] For example, some critics, like podcaster Joe Rogan, have accused YouTube of demonetizing videos in order to "stop the spread of conservative ideology and spread liberal ideology" (JRE Clips, 2018). More recently, YouTuber Steven Crowder claimed YouTube had revealed its bias against conservatives by demonetizing several of Crowder's videos for using homophobic language and selling homophobic

---

[41] From YouTube video ID #56.
[42] From YouTube video ID #3.
[43] From YouTube video ID #30.

merchandise (a clear violation of YouTube's policies) (Rosenberg, 2019). Such concerns about political bias in content moderation and platform governance more broadly have been gaining traction, even brought to the U.S. Senate by Republican politicians like Ted Cruz and Josh Hawley (Shepardson, 2019).

In fact, YouTube's efforts to police creators through the rules of revenue sharing are not limited to one ideology. YouTubers across the political spectrum and across issues have asserted that their demonetizations were politically motivated. Among the YouTubers who reported being demonetized, political content was common, but this included creators from right, left, and center. Members of marginalized communities have also felt targeted. One user recounted how every video she makes about police violence, "My Alton Sterling video, my Philando Castile video…demonetized. Anything where I'm talking about Black issues, or sensitive issues, it gets demonetized." YouTubers speaking out against environmental issues, mental health, and LGBTQ issues have also had videos demonetized (Kircher, 2017a). One YouTuber used evidence from his demonetization history to argue that YouTube was targeting him because of his sexual orientation: "I'm sure it's not a gay thing, but it feels a bit like a gay thing. Feels a bit suspect, homophobic, and like YouTube is trying to say to me 'you and your kind do not belong here.'"[44]

---

[44] From YouTube video ID #66.

YouTube does maintain an appeals process, but significant lag times mean that even videos demonetized in error may lose revenue at the worst time, particularly in the first 24-48 hours when the most views are likely.[45] Because of this, Partners have developed a variety of tactics (de Certeau, 1984) to test whether a video will be demonetized: Borrowing a tactic used to determine potential copyright strikes (Caplan & Gillespie, 2020), some Partners first upload their video set to private,[46] giving the algorithm time to assess the video before it is made public. Of course, implicit in this tactic is the widespread assumption that YouTube's judgments should be consistent, though many complained that they are not: "it got demonetized twice, the third time it got through …I'm not sure how the YouTube algorithm works, but I'm sure it works a bit randomly. If it's just machine learning, it should flag the same video every time."

Conclusion

On the one hand, demonetization reveals how YouTube and other platforms are continuing to shape creative production, asserting the terms under which creators, and even legacy media, labor; an assertion of the terms of the "contract" familiar within the media industry (Caplan & boyd, 2018; Caplan & Gillespie, 2020). However, demonetization, and the YouTube Partner Program also reveals a shift in YouTube's broader orientation towards creators: an explicit turning away from the participatory rhetoric that characterized its early history, towards

---

[45] From YouTube video ID #65.
[46] From YouTube video ID #58.

a more structured governance of media production and labor. As this chapter shows, this shift has actually been quite slow, with YouTube tiering creators early on in the Partner Program, creating the initial grooves that would separate users into their own streams. For YouTube, this has been framed as a process of cultivation – of giving resources to creators to help them grow – but it has also been one of management and mediation between creators, advertisers, traditional media, and a concerned public.

What is at stake with demonetization is how platforms continue to mediate *between* information sources – between amateurs and traditional media and between old and new media models. As their approach to Covid-19 shows, platforms are viewing monetization as akin to a stamp of approval; one solution to the broader problem of trustworthiness and credibility online. Though monetization seems to give platforms some leverage to enforce platform rules, the tiered governance of how these rules are implemented will have broad consequences for the participatory internet (Caplan, 2020), re-institutionalizing forms of gatekeeping YouTube (claims) it was initially trying to disrupt. YouTube's stratification of its users is matched by its stratification of its policies. Accusations of censorship and bias, and a multitude of other reactions by creators, show how such stratification leave users suspicious of platforms and their policies, and searching for more satisfying answers than they are willing to provide.

The demonetization debate shows the tensions YouTube is attempting to manage as their new and old media models begin to clash. YouTube still wants to structure, fund, and govern the amateur production of media, and sustain the promise of participatory culture. At the same time, the need to secure stable revenue streams, combined with pressures from advertisers to improve the quality and predictability of content, has driven them to install a system of tiered governance, hinged on rewarding audience size and celebrity, that is more akin to the contractual arrangements of traditional media. The YouTube Partner Program, and the tensions revealed by the demonetization debate, attest to the difficulty of holding these models in suspension.

Conclusion: Mediation versus Moderation in Platform Governance[47]

As the Coronavirus spread across the world, many of the major platforms began to take more aggressive action to prioritize information from trustworthy sources (Caplan, 2020). For YouTube, algorithmic demonetization became not the *last resort* in their governance arsenal, but the first. Framed as a measure for public safety, the company demonetized all content that spoke about coronavirus (YouTube Creators, 2020). Creators could be demonetized, but the process would be gradual; YouTube would begin with creators who "accurately self-certify their videos and a select number of news partners." The process of self-certification – which had been rolled out in 2019 – consists of creators self-evaluating whether their content complies with YouTube's advertiser-friendly guidelines (YouTube, n.d.). As YouTube notes, the process is geared towards building *trust*, with YouTube as the judge of trustworthiness. As they explain "when you're able to rate your content accurately, it lets us know that we can start to trust your ratings for future uploads, and not have to rely on our automated systems."

Where once platforms were reluctant to make decisions about content, we are increasingly seeing companies like Facebook, Google, and Twitter take a firmer stand about what users are able to access on their platforms. This has been true especially for COVID-19

---

[47] Portions of this chapter have been taken from previously published work that was authored solely by the author of this dissertation. See Caplan, R. (2020). "COVID-19 is a Crisis of Content Mediation" *Brookings Institute*. https://www.brookings.edu/techstream/covid-19-misinformation-is-a-crisis-of-content-mediation/ and Caplan, R. (2020). "Pornhub is Just the Latest Move Toward a Verified Internet." *Slate Magazine*. *https://slate.com/technology/2020/12/pornhub-verified-users-twitter.html*.

(Caplan, 2020) and the United States election (Election Integrity Partnership, 2020). In these cases, platforms seem to be willing to make decision about what constitutes authoritative or trustworthy content, often framing these efforts within concerns about public safety or public health. According to Mark Zuckerberg, the company has taken quicker action against COVID-19 misinformation because it was "easier to set policies that are a little more black-and-white and take a much harder line" (Smith B. , 2020). And yet, again and again, we have seen that medical knowledge of COVID-19 is limited, changing, and uncertain. This can mean that what was accepted as fact about the disease has varied over time and place, and as the diseases continues to morph. Young people *are affected* (contrary to earlier beliefs) (Maragakis, 2020). Masks *do* limit spread (contrary to earlier American public health guidelines) (Dwyer & Aubrey, 2020). And as new symptoms have emerged, more people realize they may have been infected.

A shifting understanding of the virus and changes in guidance from public health authorities represents a challenge for platforms that must distinguish between expert and non-experts, and official and non-official sources. They are doing this through a two-pronged strategy that combines the *mediation* of sources (i.e., deciding what content is authoritative), with the removal of content that might contradict it, otherwise known as *content moderation*. Though significant attention has been paid recently to content moderation (Klonick, 2017; Roberts S. T., 2019; Gillespie, Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media, 2018; Caplan, 2018), less attention has been paid to how platforms are *mediating* between sources in choosing what to prioritize and why.

Platforms have, to some extent, always placed themselves in the position of mediating between information sources. This has been done through algorithmic gatekeeping (Napoli, 2014; Tufekci, 2015) and content policies (Gillespie, 2018), through financial partnerships (Song

& Wildman, 2012) and relationships with governments (Pearson, 2020), or through prioritizing their own goods and services at the expense of others (Commission, 2015). However, little attention has been paid to intra-and-inter-organizational dynamics contributing to these decisions, and the consequences these decisions can have on networked individuals and organizations that are also vying for influence over platform standards-making.

This dissertation offered three perspectives on the impacts of this kind of standards-making from platforms, a media association, and from content creators over YouTube. It began with a context-setting chapter which presented problems of mis-and-disinfomration, specifically the impossible to pin-down "fake news," as a boundary object through which we can understand how the concerns and perceived harms of online intermediaries – platforms – are being conceptualized and mobilized by different political and social actors right now (Star & Griesemer, 1989). It noted that what constitutes trustworthy or credible information has always been a concern in communication systems, but that recent concerns about the spread of false information over platforms stem from a renegotiation of the boundaries between amateurs and experts in knowledge production that has been unfolding over the last several decades with the rise of platforms and social media (Baym & Burnett, 2009). This chapter placed this process of renegotiation within the broader context of a new area for media reform, referred to as "platform governance" (Gorwa, 2019), an emerging field of study which focuses on governance of and by platforms, and the "social and political role of platforms divided between platform companies (as architects of online environments)" and those networked entities that impact or are impacted by their rules. This dissertation also sought to understand how platforms are navigating these tensions – competing visions of what the internet should be – and how they are, themselves, appealing to pluralism, diversity, and inclusivity in designing their content policies, while maintaining strict control for making and changing those same content policies and standards.

This dissertation contends that, in addition to content policies and algorithms, there are other dynamics which must be paid attention to when considering how and why platforms make content rules and standards. Namely, it sought to understand, in particular, in how inter-organizational relationships – between platforms and academics, nonprofits, media organizations, users – can impact the formation and application of content rules. In the first case study of this dissertation, Chapter 2, I examined how platform companies are increasing the ways through which they engage external stakeholder groups in making and enforcing content policy. I placed these efforts within the frame of "networked governance," an existing theory within political science written about by Sørenson & Torfing (2005) and others that looks at how the neoliberal state similarly has distributed policy-making to connected groups and individuals over the last several decades. This chapter places the efforts of platforms within that same political history, as both benefitting from forms of networked governance by the State, while building their own governance arrangements on a similar set of principles. This chapter concludes by making the case that *networked platform governance* is one of the mechanisms through which platforms are coming to *mediate* between the individuals and organizations that seek to influence platform policy.

In the second case study, Chapter 3, I looked more deeply into one example of networked governance, looking at an international consortium of news publishers that worked together over several years to create content standards for platforms they felt better reflected the history of media ethics. This chapter provides an example of how some information producers are working to influence the development of content standards, particularly through appealing to their own domain expertise. However, it also shows the other side of networked platform governance, for instance, in the limited influence external stakeholders can have over platform policies. It demonstrates how the complex internal dynamics of platforms, particularly in how they define their own company priorities and how they allocate personnel to these relationships external to the company, can impact the effectiveness and influence of those being tapped by platforms to give feedback and advice. The uncertain future of The Trust Project also points to how opacities within networked platform governance can leave organizations working with platforms unclear about the future of their work.

The final case study (Chapter 4) showed the other side of this dynamic, and how users (in this case, YouTube content creators) contend with constantly shifting content policies that reflect the needs and demands of external stakeholders (in this case, advertisers). It uses the demonetization debate, which had seven different phases, as a strategic entry point to understand how users make sense of constantly changing policies. As I demonstrated in this chapter, users will employ multiple strategies to try to make sense of these changes, and to predict how policies may impact their work, including trying to reverse engineer rule-making, and creating alternative explanations for why their content may or may not have been demonetized. Some of those narratives, including a prevalent perception that platforms *tier governance*, implementing different rules for different user groups and give priority to those with stronger institutional ties to platforms, reflect the ongoing tensions between amateurs and experts in the platform economy. These tensions are bound up with broader cultural concerns about the production of knowledge and information, particularly in claims made by YouTubers that traditional media are being given special treatment.

This dissertation does not present a clear picture of networked platform governance dynamics. As we can see with The Trust Project, even those who are working more closely with platforms and are assumed to be privileged in those interactions, are left in the dark about how and when platform companies choose to make changes to their content policies. These publishers have, at times, felt like the YouTube content creators; impacted by policies (or "the algorithm") with no way to push back. In both cases, these groups tried to use the collective power of their peers and their industry to create a stronger counter-weight to platform power. Whereas YouTube creators asked their audiences to go to YouTube's social media and make their case heard, news media publishers worked together over several years to create standards they felt better reflected their goals. In each case, however, there is little indication this sort of collective work made a significant impact in platform policies.

What we see instead is platforms continuing to respond to external pressures, particularly pressure from the media or potential government regulation (Caplan, 2020). Platforms continue to seek the feedback and advice of external stakeholder groups who give input on platform policies, and yet, it is unclear to what extent those individuals and organizations are impacting those rules. For proponents of governance networks, some argue they can increase the diversity and expertise of people contributing to decisions about policy – a major concern in the technology industry, and can insert "more negotiated or deliberative models" and responsiveness to local needs (Bogason & Musso, 2005, p. 5). However, they also present significant problems– they can introduce ambiguity into how decisions are made, and decisions become more decentralized, placing them even more outside public view, creating more channels of political influence with potentially unevenly distributed access. Networked governance arrangements may be more horizontal, but it does not mean power is evenly distributed.

Limitations of Case Studies

This dissertation took a multi-perspectival approach towards understanding the complex intra-and-inter-organizational dynamics underlying the creation of platform content standards within the emerging frame of networked platform governance. It bounded this analysis within a particular frame and agenda that became a major focus of platforms and policymakers over the last several years – namely, how to distinguish between credible or trustworthy information online from mis-and-disinformation. This work did not attempt to offer an answer to that question. Instead, it demonstrated how platform companies are negotiating concerns about credibility and drawing distinctions between user groups – professionals from amateurs, experts from non-experts, and partners versus non-partners – in making these assessments. It offers the perspective that it is through these *relationships* between platforms and their interlocutors, that the concepts driving standards around *trust, credibility,* and *authority* are built.

Though the case studies included within this work are bound within a particular time period and frame, they are limited from a point of comparison because while they address overlapping and interrelated concerns, they are not grounded within a single event or case study. A potentially more powerful approach would have been to take one single event or initiative, such as the use of fact-checking organizations by platforms to address concerns about credibility in information, and then using conducting interviews with the various competing and collaborating stakeholder groups involved. Around 2017 I began this research taking that approach, conducting interviews with fact-checking organizations, news media, platforms, and digital advertisers as part of a broader project on networked governance for Data & Society. What I found, however, was that as public concerns about mis-and-disinformation and the impact of platforms on news media began to unfold, the common ground between the interviews – how to define the problem of credibility online – was not a stable enough base on which to build this approach. It was also too complex of a problem – too many overlapping positions and stakeholders – to be able to write about without considering each group separately. This is why each chapter of this dissertation takes on the perspective of each stakeholder group and grounds it within the assumptions of each stakeholder group in a manner that can both stand alone as a study, but also be considered in terms of the whole.

Within each case study, I have presented there are limitations, however, their complementarity is an attempt to address the limitations of the corresponding studies. For the first institutional case study on platforms, it relied on public statements made by platform representatives as well as primary documents such as corporate blogs, SEC filings, and other statements by company executives. These case study relied on public documents because, in most cases, often those working with platform companies are required to sign non-disclosure agreements (Roberts S. T., 2019). This case study therefore does not offer an analysis of the organizational dynamics of networking governance. Rather it is an analysis of the many ways platforms are using this form of input in content policy decisions, their stated aims, and how it fits into their broader goals.

The second case study I conducted attempts to address some of the limitations with relying on public statements instead of observing the intra-and-inter-organizational dynamics between platforms and external organizations. It does this through an in-depth analysis of one consortium continually tapped by platforms to give input into content policy (The Trust Project). This study also has several limitations, particularly because I was limited in terms of how many interviews I could do with The Trust Project members (I attempted through multiple forms of outreach between 2018-2021 with limited success). Future extensions of the research conducted for this chapter will include additional interviews with The Trust Project members. This chapter also leaves out the perspectives of platform representatives who have worked with The Trust Project, and is only told through The Trust Project's perspective.

My last case study which focuses on how amateur creators respond to changes in YouTube's monetization policy attempts to address another side of this debate that the previous two case studies does not address. In particular, it considers how users experience changes in policy that reflect these inter-and-intra-organizational dynamics and negotiations. This research relied on videos about demonetization that were posted by users on the YouTube platform. This case study is limited in terms of the arc of the overall dissertation in that it does not directly address how these changes reflected networked governance arrangements (though it does allude to various financial partnerships YouTube has made with established media organizations).

Lastly, this dissertation is notable in terms of the stakeholder groups it leaves out. Firstly, I did not conduct a case study of digital advertisers, which was one of the stakeholder groups I identified being of importance in my dissertation proposal. This was primarily due to time constraints. Future extensions of this work would include a case study on digital advertisers and how they influence content decisions. As was demonstrated in the chapter on YouTube, advertiser pressure is an important factor in the making of content rules or "advertiser-friendly guidelines" (Caplan & Gillespie, 2020). Advertisers have always been an important stakeholder group in the media industry in their ability to exert influence over content (An & Bergen, 2013). With platforms, the influence of new techniques for automating and optimizing advertising (known as "programmatic advertising") (McGuigan, 2019) has also increased unpredictability in terms of advertising placement. Future research will explore to what extent this unpredictability creates some pressure for advertisers to control content across a platform, and the impact it could have on amateur content production.

This dissertation also does not include a case study on how governments are trying to influence content standards over platforms. This is for several reasons. Firstly, there is ample work currently being done by scholars investigating the current debate about revising Section 230 immunity (Hawkins & Stanford, 2020; Goldman, 2020), as well as work investigating other non-US such as the German NetzDG (Gorwa, Elections, institutions, and the regulatory politics of platform governance: The case of the German NetzDG, 2021), and work examining specifically "anti-fake news" legislation such as rise and fall of legislation proposed in Malaysia (Lim, 2020). Secondly, within the United States, and during the Trump administration, debates regarding amending Section 230 became much more complex and divisive, and during the course of the writing of this dissertation, the terms of the debate were very much unclear. That made it too difficult to identify the relevant stakeholders that would be needed to be interviewed for this study.

Future Research: Platform Mediation and The Verified Internet

As platforms come to embrace their role as *mediators* of the Internet, they are increasingly using tools like verification (i.e., the blue check mark) as a way to distinguish between official and unofficial sources. Recently, the online adult video platform, Pornhub, announced they had removed all unverified videos, limiting uploads to verified users only (Cole, 2020). The move followed an investigative opinion piece by *The New York Times's* Nicholas Kristof that followed the lives of sexual abuse victims whose videos were uploaded to the site. Kristof alleged that rape videos, including child rape videos, were allowed to remain and spread on the site unchecked (Kristof, 2020). In response, major payment companies like Mastercard and Visa began their own investigations, eventually announcing they would stop processing payments with Pornhub. (Robertson, 2020). Pornhub's move to "verified users only" means that uploads can only come from official content partners and members of their "Model Program" (Pornhub, 2020).

Pornhub, however, is not alone in this move towards prioritizing verified users and content as a way to mitigate content concerns. As I demonstrated throughout this dissertation, platforms have begun embracing more publicly their role as *mediators* of information, and between interest groups vying for status online. What is happening on Pornhub and many other platforms is part of this broader shift online: Many, even most, platforms are using "verification" as a way to distinguish between sources, often framing these efforts within concerns about safety and trustworthiness. For instance, Airbnb announced in 2019 that it would verify all of its listings (Yaffe-Bellany, 2019), including the accuracy of photographs, addresses, and other information posted by hosts about themselves and their properties. Tinder has rolled out a blue checkmark verification system to deter catfishing, asking users to take selfies in real time and match poses in sample images (Carman, 2020). Social media platforms like Twitter and Instagram have long included blue verification checkmarks. Perhaps in recognition of the importance verification will play in the future of the internet, Twitter has opened a draft of their new verification system to public comment (Twitter, n.d.; Twitter Inc., 2020). This dissertation has also shown that other platforms where legacy media and amateur content creators converge, such as YouTube, have different content moderation rules and processes for different user groups (Caplan & Gillespie, 2020).

Verification is being viewed as one solution to the broader problem of trustworthiness or credibility online, often framed within the lens of mis-and-disinformation. In some cases, it is being used as a way to mediate and highlight credible and authoritative information – as was the case with Twitter during the early stages of COVID-19 (Lunden, 2020)–or content from platform-approved sources (the case with Pornhub). In other cases, it is serving primarily as an external badge, oriented more towards users as they navigate the internet–part of Silicon Valley's long-term tendency to emphasize users' individual responsibility for evaluating content (Ananny, 2020). In both senses, verification signals a broader shift in content moderation away from *content* and toward *sources.*

At this stage, it is not clear whether this move towards the verified internet is bad or good. Performers on Pornhub have, in the past, advocated for this move towards verification, as a way to curb piracy and prevent the spread of nonconsensual porn (Dickson, 2020). Journalism groups, such as The Trust Project, have been working for years to convince platforms to address differences in how information is produced by news media (specifically organizations that operate according to a standard set of media ethics) versus other information sources. And housing activists have been calling out platforms like Airbnb since at least 2015 for being used as a front for professional management companies posing as individual homeowners (Samaan, 2015).

Verification will have important consequences for the participatory internet, particularly for the large swaths of users and creators who do not get a checkmark. Nikki Usher, Jesse Holcomb, and Justin Littman (2018) have found that verification is not conferred evenly, with male journalists more likely to be verified than female journalists. There is also a dearth of publicly available information about the demographics of verification in general–for instance, whether Black users are verified at the same rates as white users. Criteria based on notability, which is a common feature of verification, has been found, in the case of Wikipedia, to reproduce existing inequalities (Harrison S. , 2019). Similarly, news organizations are able to apply for verification on behalf of their journalists, which could hint towards a bias towards legacy media, and away from amateur content creators (Usher, Holcomb, & Littman, 2018). The move towards the verified internet will thus have important consequences for communities and movements, such as Black Lives Matter, that have used social media to circulate their own narratives without relying on mainstream media (Freelon, McIlwain, & Clark, 2016). Depending on how broadly it is implemented on various platforms, it will also have important consequences for privacy, anonymity, online identity, and freedom from stigmatization (van der Nagel, 2020).

And yet, as this dissertation has also shown, it is not clear platforms are listening to these groups when forming their policies and processes. This dissertation has provided the groundwork on which to build additional research to document this next phase of platform governance–*the verified internet*. As platforms come to embrace their roles as mediators, in addition to their role as moderators, how platforms approach verification, and how they work to institutionalize or even automate these decisions, will have important consequences for how we evaluate information online. Thus far, we have seen how platforms gesture towards pluralism and democratic principles, particularly in how they use their role as intermediary, to position themselves as only one actor within a set of interlinked actors; as an entity that circulates power rather than controls it. In the next phase of platform governance research–with the verified internet–we must consider how these strategies are working to consolidate platform power. This includes understanding how the struggle for verification can reveal the logic of *platforms-as-*

institutions (Thornton & Ocasio, 2013) that are organizing and structuring individuals and organizations online and off.

Appendix A: Panel Presentations at the Content Moderation at Scale Conferences

Ashooh, Jessica. “Overview of Each Company’s Operations: Reddit.” Transcript of speech given at the Content Moderation at Scale Conference, Santa Clara University, February 2, 2018 https://santaclarauniversity.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=6e2bf22d-52cd-4e3f-9324-a8810187bad7.

Bickert, Monika. “Overview of Each Company’s Operations: Facebook.” Transcript of speech given at the Content Moderation at Scale Conference, Santa Clara University, February 2, 2018 https://santaclarauniversity.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=6e2bf22d-52cd-4e3f-9324-a8810187bad7.

Burton, Casey. “Under the Hood: UGC Moderation (Part 1): Match.com” Transcript of speech given at the Content Moderation at Scale Conference, Washington D.C. May 7, 2018. https://www.youtube.com/watch?v=stB23tNBl2o&index=3&list=PL4buVHalBRoMgSatKZoIj0vy4LjNP-iaz&t=0s

Cai, Adelin. “Overview of Each Company’s Operations: Pinterest.” Transcript of speech given at the Content Moderation at Scale Conference, Santa Clara University, February 2, 2018 https://santaclarauniversity.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=6e2bf22d-52cd-4e3f-9324-a8810187bad7.

Dean, Ted. “Overview of Each Company’s Operations: Dropbox.” Transcript of speech given at the Content Moderation at Scale Conference, Santa Clara University, February 2, 2018 https://santaclarauniversity.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=6e2bf22d-52cd-4e3f-9324-a8810187bad7.

Feerst, Alex. “Overview of Each Company’s Operations: Medium.” Transcript of speech given at the Content Moderation at Scale Conference, Santa Clara University, February 2, 2018 https://santaclarauniversity.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=6e2bf22d-52cd-4e3f-9324-a8810187bad7.

Foley, Becky. “Under the Hood: UGC Moderation (Part 1): TripAdvisor.” Transcript of speech given at the Content Moderation at Scale Conference, Washington D.C. May 7, 2018. https://www.youtube.com/watch?v=stB23tNBl2o&index=3&list=PL4buVHalBRoMgSatKZoIj0vy4LjNP-iaz&t=0s

Harvey, Del. “Under the Hood: UGC Moderation (Part 1): Twitter.” Transcript of speech given at the Content Moderation at Scale Conference, Washington D.C. May 7, 2018. https://www.youtube.com/watch?v=stB23tNBl2o&index=3&list=PL4buVHalBRoMgSatKZoIj0vy4LjNP-iaz&t=0s

Keen, Shirin. “Under the Hood: UGC Moderation (Part 1): Twitch.” Transcript of speech given at the Content Moderation at Scale Conference, Washington D.C. May 7, 2018. https://www.youtube.com/watch?v=stB23tNBl2o&index=3&list=PL4buVHalBRoMgSatKZoIj0vy4LjNP-iaz&t=0s

Mcgilvray, Sean. “Under the Hood: UGC Moderation (Part 1): Vimeo.” Transcript of speech given at the Content Moderation at Scale Conference, Washington D.C. May 7, 2018. https://www.youtube.com/watch?v=stB23tNBl2o&index=3&list=PL4buVHalBRoMgSatKZoIj0vy4LjNP-iaz&t=0s

Niv, Tal. “Under the Hood: UGC Moderation (Part 2): GitHub.” Transcript of speech given at the Content Moderation at Scale Conference, Washington D.C. May 7, 2018. https://www.youtube.com/watch?v=SRRxJAp0j0&list=PL4buVHalBRoMgSatKZoIj0vy4LjNP-iaz&index=4

Puckett, Nora. “Overview of Each Company’s Operations: YouTube.” Transcript of speech given at the Content Moderation at Scale Conference, Santa Clara University, February 2, 2018 https://santaclarauniversity.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=6e2bf22d-52cd-4e3f-9324-a8810187bad7.

Puckett, Nora. “Under the Hood: UGC Moderation (Part 2): YouTube.” Transcript of speech given at the Content Moderation at Scale Conference, Washington D.C. May 7, 2018. https://www.youtube.com/watch?v=SRRxJAp0j0&list=PL4buVHalBRoMgSatKZoIj0vy4LjNP-iaz&index=4

Rogers, Jacob. “Overview of Each Company’s Operations: Wikimedia” Transcript of speech given at the Content Moderation at Scale Conference, Santa Clara University, February 2, 2018 https://santaclarauniversity.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=6e2bf22d-52cd-4e3f-9324-a8810187bad7.

Rogers, Jacob. “Under the Hood: UGC Moderation (Part 2): Wikimedia.” Transcript of speech given at the Content Moderation at Scale Conference, Washington D.C. May 7, 2018. https://www.youtube.com/watch?v=SRRxJAp0j0&list=PL4buVHalBRoMgSatKZoIj0vy4LjNP-iaz&index=4

Schur, Aaron. “Overview of Each Company’s Operations: Yelp.” Transcript of speech given at the Content Moderation at Scale Conference, Santa Clara University, February 2, 2018 https://santaclarauniversity.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=6e2bf22d-52cd-4e3f-9324-a8810187bad7.

Sieminksi, Paul. “Overview of Each Company’s Operations: Pinterest.” Transcript of speech given at the Content Moderation at Scale Conference, Santa Clara University, February 2, 2018 https://santaclarauniversity.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=6e2bf22d-52cd-4e3f-9324-a8810187bad7.

Stern, Peter. “Under the Hood: UGC Moderation (Part 2): Facebook.” Transcript of speech given at the Content Moderation at Scale Conference, Washington D.C. May 7, 2018. https://www.youtube.com/watch?v=SRRxJAp0j0&list=PL4buVHalBRoMgSatKZoIj0vy4LjNP-iaz&index=4

Appendix B: Interview Protocol for Trust Project

Interview Protocol for Computational Systems and Public Interest Frameworks

Thank you so much for taking the time to speak with me. The goal of this project is to better understand the impact of nonprofit and industry coordination on platform policy-making, specifically content policy. You have been chosen due to your participation in a group which has collaborated on content standards to influence platforms.

As the consent form has outlined, your interview will be anonymized. This is not because I expect any risks from this research (there are no foreseeable risks given that you have already taken part in this collaborative effort), but rather to enable you to speak as freely as possible. I may refer to you in broad terms related to your industry, but will not use your name, affiliation, or place of employment in any publication out of this research. You may skip questions at any point in this interview.

I will be recording this interview, and the interview recording will be encrypted.

Do you have any questions about the consent procedure or what I'm trying to do?

1. What is your position at The Trust Project you work for? How long have you been involved in this organization?
2. At what point did platforms get involved?
3. How have platforms taken the standards? How are they using them now?
4. What would you like to change about your relationships with platforms? Have your connections to them changed?
5. What has it been like trying to navigate these tech companies as these positions change?

6. How has this relationship changed over time?

7. Have you seen power in bringing them together? Does it make publishers feel like they have some power in relation to platforms?
8. How have your relationships with platforms changed over time?
9. How do you think politics have undermined or helped these discussions?
10. Can you give me a bit of history about how this project first began in the 1990s? How was this project changed by broader discussions about misinformation, false information, and the need for more credible content?
11. In your position, have you had to work directly with platforms? I.e. have you had contact with someone who works at a company like Facebook or Google?
12. How has this relationship changed over time?
13. You once told me that one of the real changes were that they didn't used to have a person that deals with these issues, and now they do. Is that still true? Has that shifted?
14. Are you satisfied with the relationship you currently have with platforms?
15. How did this process of standards-making alter the way you defined credibility or trustworthiness?
16. What was the process of making these standards machine-readable? What organizations are you working with, and how has it been implemented by platforms? How would you *like* it to be implemented?
17. How does it challenge or change your business model for news to be alongside user-generated content?

18. Do you think it causes issues for credibility?
19. Do you consider this a global organization?
20. Does The Trust Project have any insight into when platforms change their algorithms to define high-quality news?
21. Have they made decisions, particularly in using your standards, that you disagree with?
22. Have you ever changed your business model as a result? Was this successful?
23. Do you have a relationship with someone at the platform? Prior to joining The Trust Project? Afterwords?
24. How do you think The Trust Project changed your relationships with platforms?
25. How has it changed your relationship with your newsrooms? With how you report stories?
26. Would you recommend The Trust Project to other publications?
27. What do you see the benefit of The Trust Project being?
28. Is there a possibility to do a follow-up interview if I need clarifications?

Thank you so much for taking part in this study. You may obtain a copy of your transcript within one-month time of this interview, and a copy of the full results upon completion of the study (around 6 months from interview). You can retain a copy of your transcript or the results through requesting it through the investigator, Robyn Caplan, at Robyn.Caplan@rutgers.edu.

Appendix C: Recruitment Email for The Trust Project Interviews

Robyn Caplan (PhD Candidate at Rutgers University, Researcher at Data & Society) is conducting a study on the impact of industry and non-profit coordination on platform and technology policy-making (Title: Influence in Platform Governance: A Case Study of Industry/Non-Profit Coordination). She is seeking to interview between 5 and 10 members of The Trust Project for this study.

Caplan's work seeks to investigate the relationships between platforms and publishers as one of ongoing exchange and influence. It is well understood that platforms, and their algorithms, have been having an impact on the news media industry (Caplan & boyd, 2018). Though platforms have expressed the need to integrate the expertise of external stakeholder groups – particularly news media – there is little research done, yet, as to the role this external expertise plays within policy decisions at platforms. This work seeks to understand how external stakeholder groups, such as The Trust Project, are speaking *back* to platforms collectively (for instance, as a way to translate the values and expertise of the industry into platforms), to mitigate the influence these companies can have on dependent industries.

The interview will take 60 minutes of your time. It will cover the basics of your participation in The Trust Project, your experience working with platforms before and after participation, and how your organization was/is now impacted by platform policies.

Though Caplan will collect some basic information about your organization (such as whether you are a local, national, or international publication), the interview will be anonymized to allow you to speak freely. It will be part of her dissertation work and could be published as part of journal articles and book chapters.

The faculty advisor for this study is Susan Keith, and she can be reached at susank@comminfo.rutgers.edu.

References

47 U.S.C. § 230.

A post-truth era of fake-alternative facts? (2017, March 22). Rutgers University School of Communication and Information .

Abbady, T. (2017, May 1). *The Modern Newsroom is Stuck Behind the Gender and Color Line.* Retrieved from NPR Code Switch: https://www.npr.org/sections/codeswitch/2017/05/01/492982066/the-modern-newsroom-is-stuck-behind-the-gender-and-color-line

AdAge. (2018, November 29). *Facebook exempts news publishers from its archive on political ads.* Retrieved from AdAge: https://adage.com/article/media/facebook-exempts-news-publishers-archive-political-ads/315766

Alexander, J. (2017, October 10). *YouTube fixing monetization issues with 'premium tier' partners after complaints.* Retrieved from Polygon: https://www.polygon.com/2017/10/10/16453306/youtube-monetization-ads-casey-neistat-philip-defranco

Alexander, J. (2018a, July 12). *YouTube creators reportedly losing accounts over Twitch stream promotions.* Retrieved from Polygon: https://www.polygon.com/2018/7/12/17564060/youtube-accounts-twitch-spam-deceptive-surny-linus-tech-tips-astrosizt

Alexander, J. (2018b, May 10). *The Yellow $: A comprehensive history of demonetization and YouTube's war with creators.* Retrieved from Polygon: http://www.polygon.com/2018/5/10/17268102/youtube-demonetization-pewdiepie-logan-paul-casey-neistat-philip-defranco

Alexander, J. (2019, February 19). *YouTube still can't stop child predators in its comments.* Retrieved from The Verge: https://www.theverge.com/2019/2/19/18229938/youtube-child-exploitation-recommendation-algorithm-predators

Algorithmic Accountability Act, S. 1108 (116th) (2019).

Aligica, P. (2006). Institutional and Stakeholder Mapping: Frameworks for Policy Analysis and Institutional Change. *Public Organization Review, 6*, 79-90.

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives, 31*(2), 211-36.

Allcott, H., & Gentzkow, M. (2017). *Social Media and Fake News in the 2016 Election.* Stanford University, Stanford Institute for Economic Policy Research.

An, S., & Bergen, L. (2013). Advertiser pressure on daily newspapers: A survey of advertising sales executives. *Journal of Advertising, 36*(2).

Ananny, M. (2016). Towards an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness. *Science, Technology, & Human Values, 41*(1), 93-117.

Ananny, M. (2018). *The partnership press: Lessons for platform-publisher collaborations as Facebook and news outlets team to fight misinformation.* New York: Tow Center.

Ananny, M. (2020). Making up political people: how social media create the idea, definitions, and probabilities of political speech. *Georgetown Law Technology Review, 4*(2), 351-366. Retrieved from Georgia Law and Technology Review.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society, 20*(3), 973-989.

Anderson, C. W. (2011). Deliberative, Agonistic, And Algorithmic Audiences: Journalism's Vision of its Public in an Age of Audience Transparency. *International Journal of Communication, 5*, 529-547.

Andrejevic, M. (2008). Watching Television Without Pity: The Productivity of Online Fans. *Television & New Media, 9*(1), 24-46.

Andrejevic, M., Banks, J., Campbell, J. E., Couldry, N., Fish, A., Hearn, A., & Oullette, L. (2014). Participations: Dialogues on the participatory promise of contemporary culture and politics. *International Journal of Communication, 8*, 1089-1106.

Angwin, J. (2017, June 28). *Facebook's Secret Censorship Rules Protect White Men from Hate Speech But Not Black Children.* Retrieved from ProPublica: https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms

Arceneux, K., Johnson, M., & Murphy, C. (2012). Polarized political communication, oppositional media hospitality, and selective exposure. *The Journal of Politics, 74*(1), 174-186.

Arkin, D., & Siemaszko, C. (2016, November 8). *2016 Election: Donald Trump Wins the White House in Upset.* Retrieved from NBC News: https://www.nbcnews.com/storyline/2016-election-day/2016-election-donald-trump-wins-white-house-upset-n679936

Arnstein, S. (1969). A Ladder of Citizen Participation. *JAIP, 35*(4), 216-224.

Baker, S. E. (2012). Retailing retro: Class, cultural capital and the material practices of the (re)valuation of style. *European Journal of Cultural Studies, 15*(5), 621-641.

Balkin, J. M. (2014). Old-School/New-School Speech Regulation. *Harvard Law Review, 127*(2296), 2296-2342.

Ball, C. (2009). What is Transparency? *Public Integrity*, 293-307.

Banaji, S., Bhat, R., Agarwal, A., Passanha, N., & Sadhana, P. M. (2019). *"WhatsApp vigilantes: An exploration of citizen reception and circulation of WhatsApp misinformation linked to mob violence in India.* London: London School of Economics.

Barlow, J. P. (1996). *A Declaration of the Independence of Cyberspace.* Davos: Electronic Frontier Foundation.

Baym, G. (2005). The Daily Show: Discursive Integration and the Reinvention of Political Journalism. *Political Communication, 22*(3), 259-276. Retrieved from Political Communication.

Baym, N., & Burnett, R. (2009). Amateur experts: International fan labor in Swedish independent music. *International Journal of Cultural Studies, 12*(5), 433-449.

Bechky, B. (2003). Sharing Meaning Across Occupational Communities: The Transformation of Understanding on a Production Floor. *Organization Science, 14*(3), 312-330.

Bell, E. J., Owen, T., Brown, P. D., Hauka, C., & Rashidian, N. (2017). *The Platform Press: How Silicon Valley Reengineered Journalism.* New York: Tow Center for Digital Journalism.

Bell, E., & Owen, T. (2017). *The Platform Press: How Silicon Valley Reengineered Journalism.* New York: Tow Center for Digital Journalism.

Benbaset, I., Goldstein, D., & Mead, M. (1987). The Case Research Strategy in Studies of Information Systems. *MIS Quarterly, 11*, 369-386.

Benkler, Y. (2008). *The Wealth of Networks.* New Haven: Yale University Press.

Berkman Center for Internet and Society. (2017). Harmful Speech Online. *Workshop convened by Berkman Center for Internet and Society, Institute for Strategic Dialogue, and the Shorenstein Center for Media, Politics, and Public Policy.* Cambridge: Harvard University .

Berkman Klein Center. (n.d.). *Lumen.* Retrieved June 2020, from Berkman Klein Center: https://cyber.harvard.edu/research/lumen

Berman, P. S. (2000). Cyberspace and the State Action Debate: The Cultural Value of Applying Constitutional Norms to Private Regulation. *University of Colorado Law Review, 71*, 1263.

Bettadapur, A. N. (2020, March 18). *Introducing the TikTok Content Advisory Council.* Retrieved from TikTok: https://newsroom.tiktok.com/en-us/introducing-the-tiktok-content-advisory-council

Bettadapur, A. N. (2020, September 21). *Introducing the TikTok Asia Pacific Safety Advisory Council.* Retrieved from TikTok Newsroom: https://newsroom.tiktok.com/en-sg/tiktok-apac-safety-advisory-council

Bijker, W. E., & Pinch, T. (1987). The Social Construction of Facts and Artifacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other. In W. E. Bijker, T. P. Hughes, & T. Pinch, *The Social Construction of Technological Systems.* Cambridge: MIT Press.

Birnbaum, E. (2019, June 5). *YouTube blocks controversial conservative from making money off ads.* Retrieved from The Hill: https://thehill.com/policy/technology/447145-youtube-demonetizes-conservative-commentator-after-saying-he-didnt-violate

Bishop, S. (2019, June 15). Managing visibility on YouTube through algorithmic gossip. *New Media & Society, 21*(11-12), 2589-2606.

Bodle, R. (2011). Regimes of Sharing: Open APIs, Interoperability, and Facebook. *Information, Communication, and Society, 14*(3), 320-337.

Bogason, P., & Musso, J. A. (2006). The Democratic Prospects of Network Governance. *American Review of Public Administration, 361*(1), 3-18.

Bogost, I., & Montfort, N. (2009). Platform Studies: Frequently Questioned Answers. *Proceedings of the Digital Arts and Culture Conference*.

Booth, N., & Matic, J. A. (2010). Mapping and Leveraging Influencers in Social Media to Shape Corporate brand Perceptions. *Proceedings of the Conference on Corporate Communication .*

Borocas, S., & Selbst, A. (2016). Big data's disparate impact. *California Law Review, 104*, 671.

Bossewitch, J., & Sinnreich, A. (2012, July 23). The end of forgetting: Strategic agency beyond the panopticon . *New Media & Society.*

Bottoms, A., & Tankebe, J. (2012). Beyond Procedural Justice: A Dialogic Approach to Legitimacy in Criminal Justice. *The Journal of Criminal Law & Criminology, 102*(1), 119-170.

Bowker, G. C., & Star, S. L. (2000). *Sorting Things Out: Classification and its Consequences.* Cambridge: MIT Press.

boyd , d., & Crawford, K. (2012). Critical Questions for Big Data. *15*(5), 662-679.

boyd, d. (2015). Social Media: A Phenomenon to be Analyzed. *Social Media + Society*, 1-2.

boyd, d. (2016). Untangling research and practice: What Facebook's "emotional contagion" study teaches us. *Research Ethics, 12*(1), 4-13.

Braman, S. (2011). Defining Information Policy. *Journal of Information Policy, 1*, 1-5.

Brenan, M. (2020, September 30). *Americans Remain Distrustful of Mass Media.* Retrieved from Gallup: https://news.gallup.com/poll/321116/americans-remain-distrustful-mass-media.aspx

Brown, C., & Levin, J. (2020, June 30). *Prioritizing Original News Reporting on Facebook.* Retrieved from Facebook Newsroom: https://about.fb.com/news/2020/06/prioritizing-original-news-reporting-on-facebook/

Bruns, A. (2006). Towards produsage: Futures for user-led content production. In F. Sudweeks, H. Hrachovec, & C. Eds, *Proceedings cultural attitudes towards communication and technology* (pp. 275-284). Murdoch University.

Bruns, A. (2008). *Blogs, Wikipedia, Second Life, and beyond: From production to produsage* (Vol. 45). Peter Lang.

Bruns, A., Highfield, T., & Lind, R. A. (2012). Blogs, Twitter, and breaking news: The produsage of citizen journalism. *Producing theory in a digital world: The intersection audiences and production in contemporary theory, 80*(2012), 15-32.

Bucher, T. (2017). The algorithmic imaginary: Exploring the ordinary effects of Facebook algorithms. *Information, Communication & Society, 20*(1), 30-44.

Bucher, T. (2018). Cleavage-Control: Stories of Algorithmic Culture and Power in the Case of the YouTube 'Reply All'. In Z. Paparachissi, *A Networked Self: Platforms, Stories, Connections.*

Burgess, J. (2012). YouTube and the formalization of amateur media. In D. Hunter, R. Lobato, M. Richardson, & J. Thomas, *Amateur media: Social, cultural, and legal perspectives* (pp. 53-57). Routledge.

Burgess, J., & Green, J. (2018). *YouTube: Online Video and Participatory Culture.* John Wiley & Sons.

Burgess, J., & Green, J. (2018). *YouTube: Online video and participatory culture (2nd ed.).* Cambridge: Polity Press.

Burrell, J. (2016, January 6). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 1-12.

Caldwell, J. (2019, January 23). *Microsoft builds Fake News spotter NewsGuard into Edge mobile browser.* Retrieved from ONMSFT.com: https://www.onmsft.com/news/microsoft-builds-fake-news-spotter-newsguard-into-edge-mobile-browser

Callon, M. (1984). Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fisherman of St Brieuc Bay. *The Sociological Review, 32*(1), 196-233.

Caplan, R. (2018). *Content or Context Moderation: Artisanal, Community-Reliant, and Industrial Approaches.* Data & Society Research Institute, New York. Retrieved from Data & Society Research Institute: https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf

Caplan, R. (2020, May 7). *COVID-19 misinformation is a crisis of content mediation.* Retrieved from Brookings: https://www.brookings.edu/techstream/covid-19-misinformation-is-a-crisis-of-content-mediation/

Caplan, R. (2020, December 18). *Pornhub is just the latest example of the move toward a verified internet.* Retrieved from Slate: https://slate.com/technology/2020/12/pornhub-verified-users-twitter.html

Caplan, R. (2021). The Artisan and the Decision Factory: The Organizational Dynamics of Private Speech Governance. In H. Landemore, R. Reich, & L. Bernholz, *Digital Technology and Democracy Theory* (pp. 167-190). Chicago: University of Chicago Press.

Caplan, R., & boyd, d. (2016). *Mediation, Automation, power.* Retrieved from Data & Society Research Institute: https://datasociety.net/library/mediation-automation-power/

Caplan, R., & boyd, d. (2018). Isomorphism through algorithms: Institutional dependencies in the case of Facebook. *Big Data & Society, 5*(1), 1-12.

Caplan, R., & Gillespie, T. (2020). Tiered Governance and Demonetization: The Shifting Terms of Labor and Compensation in the Platform Economy. *Social Media + Society, 6*(2).

Caplan, R., & Reed, L. (2016). *Who controls the public sphere in an era of algorithms: Case studies.* Data & Society Research Institute.

Caplan, R., Ajunwa, I., & Rosenblat, A. (2015). *Data & Civil Rights: Biometric Technologies in Policing.* New York: Data & Society.

Caplan, R., Donovan, J., & Hansen, L. (2018). *Dead Reckoning: Navigating Content Moderation After Fake News.* New York: Data & Society Research Institute. Retrieved from Data & Society Research Institute: https://datasociety.net/library/dead-reckoning/

Caplan, R., Rosenblat, A., & boyd, d. (2015). *Open data, the criminal justice system, and the Police Data Initiative .* New York: Data & Society Research Institute .

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

Carman, A. (2020, January 23). *Tinder will give you a verified blue check mark if you pass its catfishing test* . Retrieved from The Verge: https://www.theverge.com/2020/1/23/21077423/tinder-photo-verification-blue-checkmark-safety-center-launch-noonlight

Carroll, B., & Richardson, R. R. (2011). Identification, transparency, interactivity: Towards a new paradigm for credibility for single-voice blogs. *International Journal of Interactive Communication Systems and Technologies, 1*(1), 19-35.

Carroll, W. K., & Hackett, R. A. (2006). Democratic media activism through the lens of social movement theory. *Media, culture & society, 28*(1), 83-104.

Cavaye, A. (1996). Case study research: a multi-faceted research approach for IS. *Information Systems Journal, 6*, 227-242.

CBC News. (2018, October 9). *CBC joins international initiative to boost transparency in news* . Retrieved from CBC : https://www.cbc.ca/news/cbc-news-trust-project-1.4855367

de Certeau, M. (1984). *The Practice of Everyday Life.* Berkeley: University of California Press.

Chang, J. (2017, November 16). *Identifying credible content online, with help from The Trust Project.* Retrieved from Google - The Keyword: https://blog.google/outreach-initiatives/google-news-initiative/sorting-through-information-help-trust-project/

Chase, N. (2018, October 9). *What we're doing to preserve your trust.* Retrieved from East Bay Times: https://www.eastbaytimes.com/2018/10/09/what-were-doing-to-preserve-your-trust/

Chen, A. (2012, February 16). Inside Facebook's outsourced anti-porn and gore brigade, where 'camel toes' are more offensive than 'crushed heads'. *Gawker*.

Chen, A. (2015, June 2). The Agency. *New York Times Magazine*.

Christin, A. (2020). *Metrics at Work: Journalism and the Contested Meaning of Algorithms.* Princeton: Princeton University Press.

Christin, A. (2020). The ethnography and the algorithm: beyond the black box. *Theory and Society, 49*, 897-918.

Christin, A., Rosenblat, A., & boyd, d. (2015). *Courts and Predictive Algorithms.* New York: Criminal Justice Program 38.

Citron, D. K. (2018). Extremist Speech, Compelled Conformity, and Censorship Creep. *Notre Dame Law Review, 93*(3), 1035-1072.

Clegg, N. (2019, January 28). *Charting a Course for an Oversight Board for Content Decisions.* Retrieved from Facebook Newsroom: https://about.fb.com/news/2019/01/oversight-board/

Cohen, J. (2016). The Regulatory State in the Information Age. *Theoretical Inquiries in Law, 17*(2).

Cohen, J. (2016). The Regulatory State in the Information Age. *Theoretical Inquiries in Law, 17*(2), 1-36.

Coldewey, D. (2018, June 26). *Facebook Permanently Grounds its Aquila Solar-Powered Internet Plane*. Retrieved from TechCrunch.com: https://techcrunch.com/2018/06/26/facebook-permanently-grounds-its-aquila-solar-powered-internet-plane/

Cole, S. (2020, December 14). *Pornhub just purged all unverified content from the platform.* Retrieved from Vice: https://www.vice.com/en/article/jgqjjy/pornhub-suspended-all-unverified-videos-content

Commission, E. (2015). *Antitrust: Commission sends Statement of Objections to Google on comparison shopping service .* Brussels: European Commission.

Confessore, N., Dance, G. J., Harris, R., & Hansen, M. (2018, January 27). *The Follower Factory.* Retrieved from The New York Times: https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html

Constine, J. (2018, November 19). *Facebook exempts news outlets from political ads transparency labels.* Retrieved from TechCrunch: https://techcrunch.com/2018/11/29/facebook-news-ads-transparency/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAABymZO2vkr1uliDa62-vU0HIi_ZraM-GJCLrTQyJWS_A9BS94pjx2EMQHbgVMnNd0ASEQ5OETozhZfGwyTmyrfxNbvaQ8f8bVS-7hNR-pteq

Content Moderation At Scale. (2018). *COMO.* Retrieved from Comoatscale.com.

Couldry, N. (2010). *Why voice matters: Culture and politics after neoliberalism.* Sage Publications.

Crawford, K., & Gillespie, T. (2014). What is a flag for? Social Media Tools and the Vocabulary of Constraint. *Social Media & Society, 18*(3), 410-428.

Credibility Coalition. (n.d.). *Our goal: to understand the veracity, quality, and credibility of online information.* Retrieved May 2021, from Credibility Coalition: https://credibilitycoalition.org/

Cunningham, S., & Craig, D. (2019). *Social media entertainment: The new intersection of Hollywood and Silicon Valley.* New York: New York University Press.

Curry, A., & Stroud, N. J. (2017). *Trust in Online News.* Austin: The University of Texas at Austin Center for Media Engagement.

Daniels, J. (2018). The algorithmic rise of the "alt-right". *Contexts, 17*(1), 60-65.

Data & Society Research Institute . (2016, May 16). Who Controls the Public Sphere in an Era of Algorithms. *Data & Society.*

Dave, P. (2017, November 22). *YouTube steps up takedowns as concerns about kids' videos grow.* Retrieved from Reuters: https://www.reuters.com/article/us-youtube-content/youtube-steps-up-takedowns-as-concerns-about-kids-videos-grow-idUSKBN1DM2YP

Dawes, E. T. (2007). Constructing reading: Building conceptions of literacy in a volunteer read-aloud program. *Language Arts, 85*(1), 10-19.

de Certeau, M. (1984). *The practice of everyday life.* (S. T. Rendall, Ed.) University of California Press .

DeFranco, P. (2016, August 31). *YouTube is shutting down my channel and I'm not sure what to do.* Retrieved from YouTube: https://www.youtube.com/watch?v=Gbph5or0NuM&t=592s

Deighton, B. (2018, August 10). *SciDev.net joins The Trust Project.* Retrieved from SciDev.net: https://www.scidev.net/global/editorials/scidev-net-joins-the-trust-project/

DeNardis, L., & Hackl, A. (2015). Internet governance by social media platforms. *Telecommunications Policy.*

Department of Justice Office of Public Affairs. (2020). *Justice Department Issues Recommendations for Section 230 Reform.* Washington D.C.: The United States Department of Justice.

Deuze, M. (2000). Online Journalism: Modelling the First Generation of News Media on the World Wide Web. *First Monday.*

Deuze, M. (2006). Ethnic media, community media and participatory culture. *Journalism, 7*(3), 262-280.

Dewey, C. (2015, January 27). Two weeks after Zuckerberg said 'je suis Charlie,' Facebook begins censoring the prophet Muhammad. *The Washington Post* .

Dewey, C. (2016, October 12). Facebook has Repeatedly Trended Fake News Since Firing Its Human Editors. *The Washington Post.*

Dewey, J. (1927). *The Public and its Problems.* New York: Henry Holt.

Diakopoulos, N. (2014). *Algorithmic accountability reporting: On the investigation of black boxes.* Tow Center for Digital Journalism.

Dickson, E. (2020, December 11). *Pornhub upended the porn industry. How new changes could destroy sex workers' livelihoods.* Retrieved from Rolling Stone: https://www.rollingstone.com/culture/culture-news/pornhub-visa-mastercard-nicholas-kristof-sex-work-1102150/

DiMaggio, P. J., & Powell, W. W. (1983). The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizations. *American Sociological Review, 48*(2), 147-160.

DiMaggio, P., Hargittai, E., Celeste, C., & Shafer, S. (2001). *From Unequal Access to Differentiated Use: A Literature Review and Agenda for Research on Digital Inequality.* Cambridge: Russell Sage Foundation.

Donges, P. (2007). The new institutionalism as a theoretical foundation of media governance. *Communications: European Journal of Communication Research, 32*(3), 325-330.

Dourish, P. (2004). What we talk about when we talk about context. *Pers Ubiquit Comput, 8*, 19-30.

Duffy, B. E. (2018). *(Not) getting paid to do what you love.* New Haven: Yale University Press.

Durkheim, É. (1933). *The Division of Labor in Society.* New York: MacMillan.

Dwoskin, E. (2019, August 9). *YouTube's arbitrary standards: Stars keep making money even after breaking the rules.* Retrieved from The Washington Post: https://www.washingtonpost.com/technology/2019/08/09/youtubes-arbitrary-standards-stars-keep-making-money-even-after-breaking-rules/

Dwyer, C., & Aubrey, A. (2020, April 3). *CDC now recommends Americans consider wearing cloth face coverings in public.* Retrieved from NPR: https://www.npr.org/sections/coronavirus-live-updates/2020/04/03/826219824/president-trump-says-cdc-now-recommends-americans-wear-cloth-masks-in-public

Eggerton, J. (2009, July 22). Senator Wants FCC to Look Harder Into 'Fake News'. *Broadcasting + Cable*.

Election Integrity Partnership. (2020, October 28). *Evaluating Platform Election-Related Speech Policies .* Retrieved from EI Partnership : https://www.eipartnership.net/policy-analysis/platform-policies

Elish, M. C. (2019). Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science, Technology, and Society, 5*, 40-60.

Epstein, D. (2012). Manufacturing Internet Policy Language: The Inner Workings of the Discourse Construction at the Internet Governance Forum. *TPRC 2011.*

Facebook. (2019). *Draft Charter: An Oversight Board for Content Decisions.* Retrieved from Facebook Newsroom: https://about.fb.com/wp-content/uploads/2019/01/draft-charter-oversight-board-for-content-decisions-2-1.pdf

Facebook Help Center. (n.d.). *What is the Facebook Safety Advisory Board and what does the board do?* Retrieved from Privacy and Safety : https://www.facebook.com/help/222332597793306

Facebook. (n.d.). *Launching New Trust Indicators from The Trust Project for News on Facebook.* Retrieved May 2021, from Facebook for Media: https://www.facebook.com/formedia/blog/launching-new-trust-indicators-from-the-trust-project-for-news-on-facebook

Facebook. (n.d.). *Stakeholder Engagement: How does stakeholder engagement help us develop our Community Standards.* Retrieved from Facebook.com: https://www.facebook.com/communitystandards/stakeholder_engagement

Facebook. (n.d.). *What is the Facebook Safety Advisory Board and what does this board do?* Retrieved from Facebook Help Center: https://www.facebook.com/help/222332597793306

Federal Trade Commission. (2011, April 19). *FTC Seeks to Halt 10 Operators of Fake News Sites Making Deceptive Claims About Acai Weight Loss Products.* Retrieved from FTC.gov: https://www.ftc.gov/news-events/press-releases/2011/04/ftc-seeks-halt-10-operators-fake-news-sites-making-deceptive

FirstDraft News. (2017). *FAQ.* Retrieved May 2021, from CrossCheck: A collaborative journalism project: https://crosscheck.firstdraftnews.org/france-en/faq/

Fischer, S. (2019, May 30). *Exclusive: The Trust Project raises $2.25 million, becomes non-profit.* Retrieved from Axios.com: https://www.axios.com/the-trust-project-funding-craig-newmark-facebook-6b23535d-2e55-4621-b091-de9478e0b39c.html

Flanagin, A. J., Metzger, M. J., Pure, R., Markov, A., & Hartsell, E. (2014). Mitigating risk in ecommerce transactions: perceptions of information credibility and the role of user-generated ratings in product quality and purchase intention. *Electronic Commerce Research, 14*(1), 1-23.

Flanagin, A., & Metzger, M. J. (2017). Digital media and perceptions of source credibility in political communication. In *The Oxford handbook of political communication* (p. 417).

Flew, T., Martin, F., & Suzor, N. (2020, May 21). Internet regulation as media policy: Rethinking the question of digital communication platform governance. *Journal of Digital Media and Policy*.

Fowler, G. A., & Esteban, C. (2017, April 9). *14 years of Mark Zuckerberg saying sorry, not sorry.* Retrieved from The Washington Post: https://www.washingtonpost.com/graphics/2018/business/facebook-zuckerberg-apologies/https://www.washingtonpost.com/graphics/2018/business/facebook-zuckerberg-apologies/

Fraser, N. (1990). Rethinking the Public Sphere: A contribution to the Critique of Actually Existing Democracy. *Social text, 25/26*, 56-80.

Fratella, D. (2016, July 14). *YouTube debuts Creator Hub to organize creator resources.* Retrieved from SocialBlade: http://socialblade.com/blog/youtube-debuts-creator-hub-organize-creator-resources/

Freelon, D., McIlwain, C. D., & Clark, M. D. (2016, February 29). *Beyond the hashtags: #Ferguson, Blacklivesmatter, and the online struggle for offline justice.* Retrieved from Center for Media & Social Impact: https://cmsimpact.org/resource/beyond-hashtags-ferguson-blacklivesmatter-online-struggle-offline-justice/

Friedman, B., Kahn Jr. , P. H., Hagman, J., Severson, R. L., & Gill, B. (2006). The Watcher and the Watched: Social Judgments About Privacy in a Public Place. *Human-Computer Interaction, 21*, 235-272.

Fuchs, C. (2010). Labor in Informational Capitalism and on the Internet. *The Information Society, 26*(3), 179-196.

Gallup/Knight Foundation. (2018). *Indicators of News Media Trust.* Gallup Inc.

Garcia-López, M.-À., Jofre-Monseny, J., Martinez-Mazza, R., & Segú, M. (2020). Do short-term rental platforms affect housing markets? Evidence from Airbnb in Barcelona. *Journal of Urban Economics, 119*.

Garett, R. K. (2019). Social media's contribution to political misperceptions in U.S. Presidential elections. *PIoS one, 14*(3).

Garrett, R. (2011). Troubling consequences of online political rumoring. *Human Communication Research, 37*(2), 255-274.

Gil de Zuniga, H., Puig-l-abril, E., & Rojas, H. (2009). Weblogs, traditional sources online, and political participation: An assessment of how the Internet is changing the political environment. *New Media & Society, 11*(4), 553-574.

Gillespie, T. (2010). The Politics of Platforms. *New Media & Society, 12*(3), 347-364.

Gillespie, T. (2014). The relevance of algorithms. *Essays on Communication, Materiality, and Society, 167*, 167-199.

Gillespie, T. (2015). Platforms Intervene. *Social Media + Society*.

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media.* New Haven: Yale University Press.

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*.

Glad, B. J. (2004). Determining What Constitutes Creation or Development of Content Under the Communications Decency Act. *34 Sw. U. L. Rev. 247*, 266.

Goldman, E. (2018). The Complicated Story of FOSTA and Section 230. *First Amendment Law Review, 17*, 279-293.

Goldman, E. (2020). Section 230 Letter from 46 Academics.

Gorwa, R. (2019). What is Platform Governance? *Information, Communication & Society, 22*(6), 854-871.

Gorwa, R. (2021). Elections, institutions, and the regulatory politics of platform governance: The case of the German NetzDG. *Telecommunications Policy, 45*(6).

Gottfried, J., & Shearer, E. (2016). News Use Across Social Media Platforms 2016. *Pew Research Center* .

Gottweis, H. (2007). Rhetoric in Policy Making: Between Logos, Ethos, and Pathos. In F. Fischer, G. Miller, & M. Sidney, *Handbook of Public Policy Analysis: Theory, Politics, and Methods.* Boca Raton, FL: CRC Pres.

Graham, R., & Smith, S. (2016). The Content of Our #Characters: Black Twitter as Counterpublic. *Sociology of Race and Ethnicity, 2*(4), 433-449.

Graves, L. (2016). *Deciding What's True: The Rise of Political Fact-Checking in American Journalism.* New York: Columbia University Press.

Grinberg, D. (2018, June 20). *Digital dilemma: Is Medium a pay scam for writers.* Retrieved from Linkedin.com: https://www.linkedin.com/pulse/writing-mediums-partner-program-payment-scam-david-b-grinberg/

Gringas, R., & Lehrman, S. (2014, October 16). *Online Chaos Demands Radical Action by Journalism to Earn Trust.* Retrieved from Medium: https://medium.com/@GingrasLehrman/online-chaos-demands-radical-action-by-journalism-to-earn-trust-ea94b06cbccb

Guha, R. (2011, June 2). *Introducing schema.org: Search engines come together for a richer web.* Retrieved from Google Official Blog: https://googleblog.blogspot.com/2011/06/introducing-schemaorg-search-engines.html

Habermas, J. (1964). The public sphere: an encyclopedia article. *The idea of the public sphere: A reader*, 114-120.

Hall, J. (2018). Japan's right-wing YouTube: Finding a niche in an environment of increased censorship. *Asia Review, 8*(1), 315-47.

Harmful Speech Online. (2017, June 29). Retrieved from Berkman Klein Center, Institute for Strategic Dialogue and the Shorenstein Center on Media, Politics, and Public Policy at Harvard University: https://medium.com/berkman-klein-center/exploring-the-role-of-algorithms-in-online-harmful-speech-1b804936f279.

Harris, B. (2019, September 17). *Establishing Structure and Governance for an Independent Oversight Board.* Retrieved from Facebook Newsroom: https://about.fb.com/news/2019/09/oversight-board-structure/

Harris, B. (2019, April 1). *Getting Input on an Oversight Board.* Retrieved from Facebook Newsroom: https://about.fb.com/news/2019/04/input-on-an-oversight-board/

Harris, B. (2020, January 28). *Preparing the Way Forward for Facebook's Oversight Board.* Retrieved from Facebook Newsroom: https://about.fb.com/news/2020/01/facebooks-oversight-board/

Harrison, S. (2019, October 1). *Five Years of Tech Diversity Reports – and Little Progress.* Retrieved from Wired: https://www.wired.com/story/five-years-tech-diversity-reports-little-progress/

Harrison, S. (2019, March 26). *The Notability Blues.* Retrieved from Slate Magazine: https://slate.com/technology/2019/03/wikipedia-women-history-notability-gender-gap.html

Haupt, C. E. (2018, June 13). *Regulating Speech Online: A Comparative Constitutional Perspective.* Retrieved from Data & Society Research Institute: https://datasociety.net/library/online-speech-regulation-a-comparative-perspective/

Havens, T., Lotz, A. D., & Tinic, S. (2009). Critical Media Industry Studies: A Research Approach. *Communication, Culture & Critique, 2*(2), 234-253.

Hawkins, M. D., & Stanford, M. J. (2020). Uproot or upgrade? Revisiting Section 230 immunity in the digital age. *University of Chicago Law Review Online, 1.*

Helmond, A. (2015). The platformization of the web: Making web data platform ready. *Social Media + Society, 1*(2).

House Committee On Energy & Commerce. (2018, April 11). Facebook: Transparency and Use of Consumer Data . Washington , D.C. , USA.

Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and persuasion; psychological studies of opinion change.* New Haven: Yale University Press.

Hudson, M., & Fink, K. (2014). The algorithm method: Making news decisions in a clickocracy. *Columbia Journalism Review, 50*(5).

Hutchins, E. (1995). *Cognition in the Wild.* Boston: MIT Press.

Information Society Project Yale Law School. (2017). Fighting Fake News Workshop Report. *Fighting Fake News.* New Haven: The Information Society Project and The Floyd Abrams Institute for Freedom of Expression.

Ingram, M. (2014, October 27). *Trust is a crucial aspect of journalism, but it's a slippery concept and hard to measure.* Retrieved from GigaOm: https://gigaom.com/2014/10/27/trust-is-a-crucial-aspect-of-journalism-but-its-a-slippery-concept-and-hard-to-measure/

Ingram, M. (2016, September 1). *Here's what you need dot know about the #YouTubeIsOverParty uproar.* Retrieved from Fortune: http://fortune.com/2016/09/01/youtube-advertising/

Ingram, M. (2018). *The Weekly Standard and the flaws in Facebook's fact-checking program.* New York: Columbia Journalism Review.

IPTC News Architecture Working Group. (2020, September 24). *Expressing Trust and Credibility Information in IPTC Standards.* Retrieved from IPTC.org: https://iptc.org/std-dev/NewsML-G2/documentation/trustindicators.html

Isaac, M., & Hsu, T. (2020, July 7). *Facebook Fails to Appease Organizers of Ad Boycott.* Retrieved from The New York Times : https://www.nytimes.com/2020/07/07/technology/facebook-ad-boycott-civil-rights.html

Isbell, K. (2010). *The Rise of the News Aggregator: Legal Implications and Best Practices.* Cambridge: The Berkman Center for Internet & Society.

Iyengar, S., & Hahn, K. S. (2009). Red media, blue media: Evidence of ideological sensitivity in media use. *Journal of Communication, 59*(1), 19-39.

Jarvis, J. (2014, October 25). *Building Trust in News.* Retrieved from Medium.com: https://medium.com/whither-news/building-trust-in-news-e85cfc07321f

Jenkins, H. (2006). *Convergence Culture: Where Old and New Media Collide.* New York: New York University Press.

Jenkins, H. (2006). *Convergence Culture: Where Old and New Media Collide.* New York: New York University.

Jeon, D.-S., & Nasr Esfahani, N. (2012). News Aggregators and Competition Among Newspapers in the Internet (Preliminary and Incomplete). *The Economics of the Postal Sector in the Digital World.* Porto: Toulouse ICT Workshop .

Johnson, T. J., & Kaye, B. K. (2004). Wag the blog: How reliance on traditional media and the Internet influences credibility perceptions of weblogs among blog users. *Journalism and Mass Communication Quarterly, 81*, 622-642.

Joss, S. (2002). Toward the Public Sphere – Reflections on the Development of Participatory Technology Assessment. *Bulletin of Science, Technology & Society, 22*(3), 220-231.

JRE Clips. (2018, Mar 20). *Joe Rogan – Is YouTube Demonetization Censorship?* Retrieved from YouTube : https://www.youtube.com/watch?v=pJzfy0g7Y3E

Ju, A., Jeong, S. H., & Chyi, H. I. (2014). Will Social Media Save Newspapers? Examining the Effectiveness of Facebook and Twitter as News Platforms. *Journalism Practice, 8*(1), 1-17.

Kain, E. (2016, September 1). *New "advertiser friendly" YouTube isn't actually new, isn't censorship.* Retrieved from Forbes : https://www.forbes.com/sites/erikkain/2016/09/01/new-advertiser-friendly-youtube-policy-isnt-actually-new-isnt-censorship/

Katz, E. (1957). The two-step flow of communication: An up-to-date report on a hypothesis. *Public opinion quarterly, 12*(1), 61-78.

Katz, E., & Lazarsfeld, P. F. (1955). *Personal Influence. The Part Played by People in the Flow of Mass Communication.* New York: Free Press.

Katzenbach, C. (2011). Technologies as Institutions: Rethinking the Role of Technology in Media Governance Constellations. In N. Just, & M. Puppis, *Trends in Communication Policy Research.* Intellect.

Kelly, R. (2018, June 21). *How Publishers can Compete With the Facebook-Google Duopoly.* Retrieved from AdWeek: https://www.adweek.com/tv-video/how-publishers-can-compete-with-the-facebook-google-duopoly/

Kircher, M. M. (2017a, March 20). *Lesbian-wedding video among LGBTQ content blocked by YouTube restricted mode.* Retrieved from New York Magazine: https://nymag.com/intelligencer/2017/03/youtube-restricted-mode-blocks-lgbtq-videos.html

Kircher, M. M. (2017a, March 20). *Lesbian-wedding video among LGBTQ content blocked by YouTube restricted mode* . Retrieved from New York Magazine:

https://nymag.com/intelligencer/2017/03/youtube-restricted-mode-blocks-lgbtq-videos.html

Kircher, M. M. (2017a, March 20). *Vloggers are crying censorship, but YouTube says it's nothing new.* Retrieved from New York Magazine : https://nymag.com/intelligencer/2016/09/what-is-the-youtubeisoverparty-censorship-issue.html

Kirkpatrick, D. (2010). *The Facebook Effect* . New York: Simon & Schuster Paperbacks.

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society, 20*(1), 14-29.

Klein, H., & Kleinman, D. L. (2002). y, & Human Values The Social Construction of Technology: Structural Considerations. *Science, Technology, & Human Values, 27* (1), 28-52.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic Fairness. *AEA papers and proceedings, 108*, 22-27.

Klonick, K. (2018, April 10). The New Governors: The People, Rules, and Processes Governing Online Speech. *Harvard Law Review.*

Klonick, K. (2020). The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression. *The Yale Law Journal.*

Kristof, N. (2020, December 4). *The Children of Pornhub.* Retrieved from The New York Times: https://www.nytimes.com/2020/12/04/opinion/sunday/pornhub-rape-trafficking.html

Kumar, S. (2019). The algorithmic dance: YouTube's Adpocalypse and the gatekeeping of cultural content on digital platforms. *Internet Policy Review, 8*(2), 1-21.

Kurjan, C. (2016). *Engaging with News: Indicators of Trustworthiness.* Santa Clara: Quiver Consulting for Sally Lehrman, Director of TheTrustProject.org.

Kyl, J. (2019). *Covington Interim Report.* Retrieved from Facebook Newsroom: https://fbnewsroomus.files.wordpress.com/2019/08/covington-interim-report-1.pdf?mod=article_inline

Ladd, J. M. (2010). The role of media distrust in partisan voting. *Political Behavior, 32*(4), 567-585.

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live by.* Chicago: University of Chicago Press.

Latour, B. (2005). *Reassembling the Social: An Introduction to Actor-Network Theory* . New York: Oxford University Press.

Lazarsfeld, P. (1944). The election is over. *Public Opinion Quarterly*, 317-330.

Lazarsfeld, P. F., Berelson, B., & Gaudet, H. (1948). *The People's Choice: how the Voter Makes Up His Mind in a Presidential Campaign.* New York: Columbia University Press .

Lessig, L. (2006). *Code: Version 2.0.*

Lieberman, D. (1992, Feb 22-28). Fake News: A Special Report: What We See Isn't Always News – It's Public Relations. *TV Guide*, pp. 13-14.

Lim, G. (2020). *Securitize/Counter-Securitize.* New York: Data & Society Research Institute.

Limm, J. (2011, January 6). *Networked Governance: Why it is Different and How it Can Work.* Retrieved from Civil Service College: Singapore: https://www.csc.gov.sg/articles/networked-governance-why-it-is-different-and-how-it-can-work

Lobato, R. (2016). The cultural logic of digital intermediaries: YouTube multichannel networks. *Convergence: The International Journal of Research into New Media Technologies, 22*(4), 348-360.

Lobato, R., & Thomas, J. (2015). *The Informal Media Economy.* Cambridge: Polity Press.

Lohse, D. (2017, November 16). *Leading News Outlets Establish Transparency Standards To Help Readers Identify Trustworthy News Sources.* Retrieved from The Trust Project: https://thetrustproject.org/2017/11/16/launch/

Lowi, T. J. (1964, July). American Business, Public Policy, Case-Studies, and Political theory. *World Politics, 16*(4), 677-693.

Lubbers, E. (2016, November 5). *There is no such thing as the Denver Guardian, despite that Facebook post you saw.* Retrieved from The Denver Post: https://www.denverpost.com/2016/11/05/there-is-no-such-thing-as-the-denver-guardian/

Lunden, I. (2020, March 21). *Twitter prioritizes blue-check verification to confirm experts on COVID-19 and the novel coronavirus.* Retrieved from TechCrunch: https://techcrunch.com/2020/03/21/twitter-prioritizes-blue-check-verifications-to-confirm-experts-on-covid-19-and-the-novel-coronavirus/?guccounter=1&guce_referrer=aHR0cHM6Ly9zbGF0ZS5jb20v&guce_referrer_sig=AQAAAChti2ExJFPEimvo5EsSm4GQHkuwOETKRo62F-0HJJCd

Lyons, T. (2018, May 23). *Hard Questions: What's Facebook's Strategy for Stopping False News?* Retrieved from Facebook Newsroom: https://newsroom.fb.com/news/2018/05/hard-questions-false-news/

MacFarquhar, N. (2018, February 18). *Inside the Russian Troll Factory: Zombies and a Breakneck Pace.* Retrieved from The New York Times: https://www.nytimes.com/2018/02/18/world/europe/russia-troll-factory.html

Mcgilvray, S. (2018). Comments made at the Content Moderation at Scale Conference. Washington, D.C.

Machray, A. (2017, November 17). *ECHO signs up to The Trust Project as launch partner – What that means for readers.* Retrieved from Liverpool ECHO: https://www.liverpoolecho.co.uk/news/liverpool-news/echo-signs-up-trust-project-13915974

Maheshwari, S., & Wakabayashi, D. (2017, March 22). *AT&T and Johnson & Johnson pull ads from YouTube.* Retrieved from The New York Times : https://www.nytimes.com/2017/03/22/business/atampt-and-johnson-amp-johnson-pull-ads-from-youtube-amid-hate-speech-concerns.html

Maragakis, L. L. (2020). *Coronavirus and COVID-19: Younger adults are at risk, too.* Retrieved from John Hopkins Medicine: https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/coronavirus-and-covid-19-younger-adults-are-at-risk-too

March, J., & Olsen, J. (1995). *Democratic Governance.* New York: Free Press.

Marchi, R. (2012). With Facebook, Blogs, and Fake News, Teens Reject Journalistic "Objectivity". *Journal of Communication Inquiry, 36*(3), 246-262.

Marwick, A., & boyd, d. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society, 13*(1), 114-133.

Marwick, A., & Caplan, R. (2018). Drinking male tears: Language, the manosphere, and networked harassment. *Feminist Media Studies , 18*(4), 543-559.

Matsakis, L. (2019, August 23). *Twitter Trust and Safety Advisers Say They Are Being Ignored.* Retrieved from Wired: https://www.wired.com/story/twitter-trust-and-safety-council-letter/

McCarthy, T. (2020, May 28). *Zuckerberg says Facebook won't be 'arbiters of truth' after Trump threat.* Retrieved from The Guardian: https://www.theguardian.com/technology/2020/may/28/zuckerberg-facebook-police-online-speech-trump

McChesney, R. (1993). *Telecommunications, Mass Media, & Democracy: The Battle for the Control of U.S. Broadcasting, 1928-1935.* New York: Oxford University Press.

McChesney, R. W. (2007). *Communication revolution: critical junctures and the future of media* . New York: New Press.

McGuigan, L. (2019). Automating the audience commodity: The unacknowledged ancestry of programmatic advertising. *New Media & Society, 21*(11-12), 2366-2385.

Medium. (n.d.). *Join the Partner Program.* Retrieved from help.medium.com: https://help.medium.com/hc/en-us/articles/115011694187-Join-the-Partner-Program

Medzini, R. (2021). Enhanced self-regulation: The case of Facebook's content governance. *New media & Society*.

Meeker, M. (1997). *The Internet Advertising Report.* Morgan Stanley.

Meese, J., & Hurcombe, E. (2020). Facebook, news media, and platform dependency: The institutional impacts of news distribution on social platforms. *New Media & Society*(June).

Mickoleit, A. (2014). *Social media use by governments: A policy primer to discuss trends, identify policy opportunities and guide decision maker.* OECD.

Mitchell, A., Kiley, J., Gottfried, J., & Matsa, K. E. (2014, October 21). *Political Polarization & Media Habits.* Retrieved from Pew Research Center : https://www.journalism.org/2014/10/21/political-polarization-media-habits/

Morris, E. (2018). *Desperately seeking the producer: Audiences, identity, and the margins of the internet.* University of Pennsylvania.

Moses, L. (2018, September 18). *Facebook rolls out system to include news publishers in its controversial Ad Archive.* Retrieved from DigiDay: https://digiday.com/media/facebook-rolls-system-include-news-publishers-controversial-ad-archive/

Moses, L. (2018, June 28). *Facebook tweaks political ads policy, but not enough to satisfy irate publishers.* Retrieved from Digiday: https://digiday.com/media/facebook-tweaks-political-ads-policy-not-enough-satisfy-irate-publishers/

Mostrous, A., & Bridge, M. (2017, November 24). *YouTube adverts fund pedophile habits.* Retrieved from The Times of London : https://www.thetimes.co.uk/article/youtube-adverts-fund-paedophile-habits-fdzfmqlr5

Murphy, L.W. (2020). *Facebook's Civil Rights Audit – Final Report.* Facebook. Retrieved from https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf

Napoli, P. M. (2001). *Foundations of Communications Policy: Principles and Process in the Regulation of Electronic Media.* Cresskill, New Jersey: Hampton Press, Inc.

Napoli, P. M. (2001). The Localism Principle in Communications Policymaking and Policy Analysis: Ambiguity, Inconsistency, and Empirical Neglect. *Policy Studies Journal, 29*(3), 372-387.

Napoli, P. M. (2009). Public Interest Media Advocacy and Activism as a Social Movement. *Annals of the International Communication Association, 33*(1), 385-429.

Napoli, P. M. (2009). Public Interest Media Advocacy and Activism as a Social Movement. *Annals of the International Communication Association*, 385-429.

Napoli, P. M. (2014). Automated Media: An Institutional Theory Perspective on Algorithmic Media Production and Consumption. *Communication Theory, 24*(3), 340-360.

Napoli, P. M. (2015). Social media and the public interest: Governance of news platforms in the realm of individual and algorithmic gatekeepers. *Telecommunications Policy, 39*(9), 751-760.

Napoli, P. M. (2021). Back from the Dead (Again): The Specter of the Fairness Doctrine and its Lessons for Social Media Regulation. *Policy & Internet*.

Napoli, P. M., & Caplan, R. (2017). Why media companies insist they're mot media companies, why they're wrong, and why it matters. *First Monday, 22*(5).

Negroponte, N. (1996). *Being Digital.* New York: Vintage Books.

News, T. (2018, October 11). *Trusting News project expands research and training through University of Georgia partnership.* Retrieved from RJI Online : https://www.rjionline.org/stories/trusting-news-project-expands-research-and-training-through-university-of-g

Newton, C. (2019, November 14). *Why tech companies owe us more than a quarterly transparency report.* Retrieved from The Verge: https://www.theverge.com/interface/2019/11/14/20963124/facebook-transparency-report-limits-justice-oversight-board

Noble, S. (2018). *Algorithms of Oppression.* New York: New York University Press.

NuzzelRank. (2018, June 18). *Announcing NuzzelRank – Authority Ranking of News Sources Using Signals From Top Business Influencers.* Retrieved from Web Archive - Blog.Nuzzel.com: https://web.archive.org/web/20210409111256/https://blog.nuzzel.com/nuzzelrank/

O'Brien, M. (2018, May 16). *Inside Facebook's race to separate news from junk.* Retrieved from PBS News Hour : https://www.pbs.org/newshour/show/inside-facebooks-race-to-separate-news-from-junk

Office of Film & Literature Classification. (2019). *Christchurch attacks classification information.* Classification Office New Zealand.

Orlikowski, W. J., & Baroudi, J. J. (1991). Studying Information Technology in Organizations: Research Approaches. *Information Systems Research, 2*(1), 1-28.

Ong, T. (2017, November 28). *YouTube pulls ads on 2 million inappropriate children's videos.* Retrieved from The Verge: https://www.theverge.com/2017/11/28/16709214/youtube-autocomplete-search-function-child-exploitation-ads

Overland, H. M. (2010, June 17). *What is Facebook Open Graph.* Retrieved April 2020, from Search Engine People: https://www.searchenginepeople.com/blog/what-is-facebook-open-graph.html

Oversight Board Bylaws. (2020). Retrieved from https://about.fb.com/wp-content/uploads/2020/01/Bylaws_v6.pdf

Oversight Board. (n.d.). *Meet the Board.* Retrieved from https://www.oversightboard.com/meet-the-board/

Owens, E., & Weinsberg, U. (2015, January 20). *Showing Fewer Hoaxes.* Retrieved from Facebook Newsroom: https://newsroom.fb.com/news/2015/01/news-feed-fyi-showing-fewer-hoaxes/

Papacharissi, Z. (2002). The virtual sphere: The internet as a public sphere. *New media & society, 4*(1), 9-27.

Pariser, E. (2011). *The Filter Bubble: What The Internet Is Hiding From You.* Penguin Books.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information.* Cambridge, MA: Harvard University Press.

Pearson, J. (2020, April 21). *Exclusive: Facebook agreed to censor posts after Vietnam slowed traffic.* Retrieved from Reuters : https://www.reuters.com/article/us-vietnam-facebook-exclusive/exclusive-facebook-agreed-to-censor-posts-after-vietnam-slowed-traffic-sources-idUSKCN2232JX

Pearson, J. (2020, April 21). *Exclusive: Facebook Agreed to Censor Posts After Vietnam Slowed Traffic - Sources.* Retrieved from Reuters: https://www.reuters.com/article/us-vietnam-facebook-exclusive/exclusive-facebook-agreed-to-censor-posts-after-vietnam-slowed-traffic-sources-idUSKCN2232JX

Perez, S. (2013, June 20). *Over a year after new content policies, "self-harm social media" still thrives.* Retrieved from TechCrunch: https://techcrunch.com/2013/06/20/over-a-year-after-new-content-policies-self-harm-social-media-still-thrives/

Petre, C. (2015). *The Traffic Factories: Metrics, Chartbeat, Gawker Media, and The New York Times .* New York: Tow Center for Digital Journalism.

Pew Research Center. (1999, January 14). *The Internet News Audience Goes Ordinary.* Retrieved from Pew Research Center: https://www.pewresearch.org/politics/1999/01/14/the-internet-news-audience-goes-ordinary/

Phillips, W. (2018). *The oxygen of amplification.* New York: Data & Society Research Institute.

Pickard, V. (2007). Neoliberal Visions and Revisions in Global Communications Policy: From WCICO to WSIS. *Journal of Communication Inquiry, 31*, 118.

Pickard, V. (2015). *America's Battle for Media Democracy: The Triumph of Corporate Libertarianism and the Future of Media Reform.* New York: Cambridge University press.

Pinch, T. J., & Bijker, W. E. (1984). The Social Construction of Facts and Artefacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other. *Social Studies of Science, 14*(3), 399-441.

Poggi, J. (2019, February 20). *Google-Facebook Duopoly Set to Lose Some of Its Share of Ad Spend.* Retrieved from AdAge: https://adage.com/article/digital/duopoly-loses-share-ad-spend/316692/

Popper, B. (2017, April 6). *YouTube will no longer allow creators to make money until they reach 10,000 views.* Retrieved from The Verge: https://www.theverge.com/2017/4/6/15209220/youtube-partner-program-rule-change-monetize-ads-10000-views

Pornhub. (2020, December). *The latest on our commitment to trust and safety.* Retrieved from Pornhub: https://www.pornhub.com/blog/11422

Powell, W. W. (1990). Neither Market Nor Hierarchy: Network Forms of Organization. *Research in Organizational Behavior, 12*, 295-336.

Provan, K. G., & Kenis, P. (2007). Modes of Network Governance: Structure, Management, and Effectiveness. *Journal of Public Administration Research and Theory.*

Puppis, M. (2010). Media Governance: A New Concept for the Analysis of Media Policy and Regulation. *Communication, Culture & Critique, 3*(2), 134-149.

RAND Corporation. (n.d.). *Certified Content Coalition.* Retrieved May 2021, from Fighting Disinformation Home: https://www.rand.org/research/projects/truth-decay/fighting-disinformation/search/items/certified-content-coalition.html

Reeve, B. (2013). Private Governance, Public Purpose? Assessing Transparency and Accountability in Self-Regulation of Food Advertising to Children. *Bioethical Inquiry*, 149-163.

Rob, Lever, R., & Chapman, G. (2017, April 3). *Industry, academic partners team up to fight fake news.* Retrieved from Yahoo News: https://www.yahoo.com/news/industry-academic-partners-team-fight-fake-news-042225050.html

Roberts, J. J. (2019, April 29). *Why You Should be Worried About Tech's Love Affair With NDAs.* Retrieved from Fortune: https://fortune.com/2019/04/29/silicon-valley-nda/

Roberts, S. T. (2019). *Behind the Screen: Content Moderation in the Shadows of Social Media.* New Haven: Yale University Press.

Robertson, A. (2020, December 10). *Visa and Mastercard cut off Pornhub after report of unlawful videos.* Retrieved from The Verge : https://www.theverge.com/2020/12/10/22168240/mastercard-ending-pornhub-payment-processing-unlawful-videos

Romano, A. (2019, October 10). *A group of YouTubers is trying to prove the site systematically demonetizes queer content.* Retrieved from Vox: https://www.vox.com/culture/2019/10/10/20893258/youtube-lgbtq-censorship-demonetization-nerd-city-algorithm-report

Rosemain, M., Rose, M., & Barzic, G. (2018, November 12). *France to 'embed' regulators at Facebook to combat hate speech.* Retrieved from Reuters: https://www.reuters.com/article/us-france-facebook-macron/france-to-embed-regulators-at-facebook-to-combat-hate-speech-idUSKCN1NH1UK

Rosenberg, E. (2019). *A right-wing Youtuber hurled racist, homophobic taunts at a gay reporter. The company did nothing.* Retrieved from The Washington Post: https://www.washingtonpost.com/technology/2019/06/05/right-wing-youtuber-hurled-racist-homophobic-taunts-gay-reporter-company-did-nothing/

Rosenberg, E. (2019, June 6). *A right-wing YouTuber hurled racist, homophobic taunts at a gay reporter. The company did nothing.* Retrieved from The Washington Post : https://www.washingtonpost.com/technology/2019/06/05/right-wing-youtuber-hurled-racist-homophobic-taunts-gay-reporter-company-did-nothing/

Rosenblat, A. (2018). *Uberland: How Algorithms Are Rewriting the Rules of Work.* Oakland: University of California Press.

Rosenblat, A., & Stark, L. (2016). Algorithmic labor and information asymmetries: A case study of Uber's drivers. *International Journal of Communication, 10.*

Sagers, C. (2019, October 17). Antitrust and Tech Monopoly: A General Introduction to Competition Problems in Big Data Platforms. *Testimony Before the Committee on the Judiciary of the Ohio Senate.*

Salamon, E. (2015). (De)valuing Intern Labor: Journalism Internship Pay Rates and Collective Representation in Canada. *TripleC*, 438-458.

Samaan, R. (2015). *Airbnb, rising rent, and the housing crisis in Los Angeles.* Los Angeles: Laane.

Santa Clara University. (2017 , November). *The Trust Project Helps Readers Identify Reliable News.* Retrieved from SCU.edu: https://www.scu.edu/news-and-events/press-releases/2017/nov-2017/the-trust-project-helps-readers-identify-reliable-news.html#:~:text=Google%2C%20Facebook%2C%20Bing%20and%20Twitter,Indicators%E2%80%9D%20to%20highlight%20credible%20journalism.

Schema.org. (n.d.). *NewsArticle.* Retrieved May 2021, from Schema.org: https://schema.org/NewsArticle

Schmidt, C. (2018, April 5). *So what is that, er, Trusted News Integrity Trust Project all about? A guide to the (many, similarly named) new efforts fighting for journalism.* Retrieved from Nieman Lab: https://www.niemanlab.org/2018/04/so-what-is-that-er-trusted-news-integrity-trust-project-all-about-a-guide-to-the-many-similarly-named-new-efforts-fighting-for-journalism/

Schneider, D., de Souza, J., & Lucas, E. M. (2014). Towards a typology of social news apps from a Crowd Computing perspective. *IEEE International Conference* . San Diego: IEEE.

Scholz, T. (2013). *Digital labor: The internet as playground and factory.* Routledge.

Scott, W. R. (2001). *Institutions and Organizations.* Thousand Oaks: Sage Publications .

Scott, W. R., & Meyer, J. W. (1994). *Institutional environments and organizations: structural complexity and individualism.* Thousand Oaks Sage.

Seaver, N. (2017). Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society, 4*(2).

Shepardson, D. (2019, April 10). *Facebook, Google accused of anti-conservative bias at U.S. Senate hearing.* Retrieved from Reuters: https://www.reuters.com/article/us-usa-congress-socialmedia/facebook-google-accused-of-anti-conservative-bias-at-u-s-senate-hearing-idUSKCN1RM2SJ

Shirky, C. (2008). *Here comes everybody: The power of organizing without organizations.* Penguin.

Sieminski, P. (2018). Overview of Each Company's Operations. *Content Moderation at Scale.* Santa Clara.

Silverman, C. (2016, November 16). *This Analysis Shows How Viral Fake Election News Stories Outperformed Real news on Facebook.* Retrieved from BuzzFeed News: https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook

Silverman, C. (2016, November). *This Analysis Shows How Viral Fake Election News Stories Outperformed Real News on Facebook.* Retrieved from BuzzFeed news: https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook

Singer, J. B. (1997). Changes and consistencies: Newspaper journalists contemplate online future. *Newspaper Research Journal, 18*(1-2).

Singer, P., & Brooking, E. (2018). *LikeWar: The Weaponization of Social Media.* New York, NY: Houghton Miflon Harcourt.

Singh, S. (2019, May 7). *Assessing YouTube, Facebook, and Twitter's Content Takedown Policies.* Retrieved from New America, Open Technology Institute: https://www.newamerica.org/oti/reports/assessing-youtube-facebook-and-twitters-content-takedown-policies/?utm_medium=email&utm_campaign=OTI%20-%20One-year%20anniversary%20Santa%20Clara%20Principles&utm_content=OTI%20-%20One-year%20anniversary%20Santa%20Cl

Singh, S. (2020). *Transparency Report Tracking Tool: How Internet Platforms are Reporting on the Enforcement of their Content Rules.* New York, NY : New America, Open Technology Institute.

Smith, A. (2017, November 16). *Joining The Trust Project: The Economist takes part in an initiative to help readers spot good journalism.* Retrieved May 2021, from Medium.com: https://medium.com/severe-contest/joining-the-trust-project-1f7cec4e6038

Smith, B. (2020, March 15). *When Facebook is more trustworthy than the president* . Retrieved from The New York Times: https://www.nytimes.com/2020/03/15/business/media/coronavirus-facebook-twitter-social-media.html

Soll, J. (2016, December 18). *The Long and Brutal History of Fake News.* Retrieved from Politico Magazine: https://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535

Solon, O. (2016, November 10). "Facebook's failure: Did fake news and polarized politics get Trump elected? *The Guardian*.

Solon, O. (2016, November 10). *Facebook's Fake News: Mark Zuckerberg rejects 'crazy idea' that it swayed voters.* Retrieved from The Guardian.

Solon, O., & Levin, S. (2016, December 16). How Google's Search Algorithm Spreads False Information with a Right-Wing Bias. *The Guardian*.

Song , S. Y., & Wildman, S. S. (2012). Evolution of strategy and commercial relationships for social media platforms: The case of YouTube. In *Handbook of Social Media Management* (pp. 619-632).

Sorenson, E., & Torfing, J. (2005). The Democratic Anchorage of Governance Networks. *Scandinavian Political Studies, 28*(5).

Stake, R. (1995). *The Art of Case Study Research.* Thousand Oaks: SAGE Publications.

Stamos, A. (2017, September 6). *An Update on Information Operations on Facebook.* Retrieved from Facebook Newsroom: https://about.fb.com/news/2017/09/information-operations-update/

Star, S. L., & Griesemer, J. (1989). Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology. *Social Studies of Science, 19*(3), 387-420.

Statista. (2021, January). *Most popular social networks worldwide as of January 2021, ranked by most active users.* Retrieved May 2021, from Statista: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Stern, P. (2018). Comments made at the Content Moderation at Scale Conference. Washington D.C.

Stevenson, S. (2009). Digital Divide: A Discursive Move Away from the Real Inequities. *The Information Society, 25*(1), 1-22.

Stoker, G. (2006). Public value management: a new narrative for networked governance? *The American review of public administration, 36*(1), 41-57.

Streeter, T. (1996). *Selling the Air: A Critique of the Policy of Commercial Broadcasting in the United States.* Chicago: University of Chicago Press .

Streeter, T. (2013). Policy, Politics, and Discourse. *Communication, Culture & Critique, 6*(4), 488-501.

Sunshine, J., & Tyler, T. (2003). The Role of Procedural Justice and Legitimacy in Shaping Public Support for Policing. *Law & Society Review, 37*(3), 513-548.

Sunstein, C. (2018). *#Republic: Divided Democracy in the Age of Social Media.* Princeton University Press.

Suzor, N. (2018). Digital constitutionalism: Using the rule of law to evaluate the legitimacy of governance by platforms. *4*(3).

Sweeney, L. (2013). Discrimination in online ad delivery. *Communications of the ACM, 56*(5), 44-54.

Taylor, T. R. (1997). Procedural fairness and compliance with the law. *Swiss Society of Economics and Statistics (SSES), 133*, 219-240.

Tenzer, A. (2018). *Measuring the Impact of The Trust Project.* The Trust Project, Reach Plc.

Terranova, T. (2000). Free labor: Producing culture for the digital economy. *Social Text, 18*(2), 33-58.

Tessier, M., Herzog, M., & Madzou, L. (2017). Regulation at the Age of Online Platform-Based Economy, Accountability, User Engagement, and Responsiveness. In L. Bell, & N. Zingales, *Platform regulations: how platforms are regulated and how they regulate us* (pp. 175-188).

Teubner, G. (2009). And if I by Beelzebub cast out Devils: An Essay on the Diabolics of Network Failure. *German Law Journal, 10*(4), 395-416.

The Economist Group. (n.d.). *Results and governance.* Retrieved May 2021, from EconomistGroup.com: https://www.economistgroup.com/

The Trust Project. (2016). *20 May Trust Summit Workshop.* New York: The Trust Project.

The Trust Project. (2017). *The Trust Project: Citations and References design sprint.* Washington, D.C.: The Trust Project.

The Trust Project. (n.d.). *#TrustedJournalism.* Retrieved May 2021, from TheTrustProject.org: https://thetrustproject.org/trusted-journalism/

The Trust Project. (n.d., May 18). *About.* Retrieved 2021, from The Trust Project: https://thetrustproject.org/about/

The Trust Project. (n.d.). *Frequently Asked Questions.* Retrieved May 2021, from TheTrustProject.org: https://thetrustproject.org/faq/

The Trust Project. (n.d.). *Trust Project Working Groups.* Retrieved 2021, from TheTrustProject.org: https://thetrustproject.org/trust-project-working-groups/

Thomas, G. (2016). *How to do Your Case Study.* London: Sage Publications Ltd. .

Thornton, B. (2000). The Moon Hoax: Debates About Ethics in 1835 New York Newspapers. *Journal of Mass Media Ethics, 15*, 89-100.

Thornton, P. H., & Ocasio, W. (2013). Institutional Logics. In R. Greenwood, C. Oliver, K. Sahlin, & R. Suddaby, *The SAGE Handbook of Organizational Institutionalism.* SAGE Publications .

Tufekci, Z. (2015). Algorithmic Harms Beyond Facebook and Google: Emergent Challenges of Computational Agency. *Colorado Tech Law Journal*, 203.

Tufekci, Z. (2018, March 10). YouTube, the Great Radicalizer. *The New York Times*.

Turcotte, J., York, C., Irving, J., Scholl, R. M., & Pingree, R. J. (2015). News recommendations from social media opinion leaders: effects on media trust and information seeking. *Journal of Computer-Mediated Communication, 20*(5), 520-535.

Tushnet, M. (2015). Internet Exceptionalism: An Overview of General Constitutional Law. *56 Wm. & Mary L. Rev. 1637, 1638.*

Twitch. (2020, May 14). *Introducing the Twitch Safety Advisory Council.* Retrieved from Twitch: https://blog.twitch.tv/en/2020/05/14/introducing-the-twitch-safety-advisory-council/

Twitch. (n.d.). *Twitch partner program.* Retrieved from https://www.twitch.tv/p/partners/

Twitter Inc. . (2016). *Letter Q1 2016 Shareholder Letter.* San Francisco, California : SEC.

Twitter Inc. (2020, November 24). *Help us shape our new approach to verification.* Retrieved from Twitter Blog: https://blog.twitter.com/en_us/topics/company/2020/help-us-shape-our-new-approach-to-verification.html

Twitter. (n.d.). *About verified accounts.* Retrieved from Twitter Help Center: https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts

Twitter. (n.d.). *The Twitter Trust and Safety Council.* Retrieved May 2020, from About.Twitter.com: https://about.twitter.com/en_us/safety/safety-partners.html

Twitter. (n.d.). *Working with partners to improve the health of the public conversation*. Retrieved from Safety at Twitter: https://about.twitter.com/en_us/safety/safety-partners.html#online-safety-partners

Twitter, Inc. . (2018). *Proxy Statement: Notice of 2018 Annual Meeting of Stockholders.* San Francisco: Twitter, Inc.

Tyler, T. R. (2003). Procedural justice, legitimacy, and the effective rule of law. *Crime and Justice, 30*, 283-357.

Tyler, T., Katsaros, M., Meares, T., & Venkatesh, S. (2018). *Social media governance: Can companies motivate voluntary rule following behavior among their users.* Facebook Research.

Usher, N., Holcomb, J., & Littman, J. (2018). Twitter makes it worse: Political journalists, gendered echo chambers, and the amplification of gender bias. *The International Journal of Press/Politics, 23*(3), 324-344.

van der Nagel, E. (2020). Embodied verification: Linking identities and bodies on NSFW Reddit. In K. Warfield, C. Abidin, & C. Cambre, *Mediated Interfaces: The Body on Social Media* . New York: Bloomsbury.

Van Dijck, J., & Poell, T. (2013). Understanding social media logic. *Media and communication, 1*(1), 2-14.

Van Dijck, J., Poell, T., & De Waal, M. (2018). *The platform society: Public values in a connective world.* Oxford University Press.

Vondreau, P. (2016). The video bubble: Multichannel networks and the transformation of YouTube. *Convergence: The International Journal of Research into New Media Technologies, 22*(4), 361-375.

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

Wakabayashi, D., & Maheshwari, S. (2019, February 20). Advertisers Boycott YouTube After Pedophiles Swarm comments on Videos of Children. *The New York Times* .

Wallsten, K. (2011). Many sources, one message: Political blog links to online videos during the 2008 campaign . *Journal of Political Marketing, 10*(1), 88-114.

Warner, M. (2002). Publics and Counterpublics. *Public culture, 14*(1), 49-90.

Warofka, A. (2018, November 5). *An Independent Assessment of the Human Rights Impact of Facebook in Myanmar.* Retrieved from Facebook Newsroom: https://about.fb.com/news/2018/11/myanmar-hria/

Warren, T. (2019, January 23). *Microsoft is trying to fight fake news with its Edge mobile browser.* Retrieved from The Verge: https://www.theverge.com/2019/1/23/18194078/microsoft-newsguard-edge-mobile-partnership

Weedon, J., Nuland, W., & Stamos, A. (2017, April 27). *Information Operations and Facebook.* Retrieved from Facebook Newsroom: https://fbnewsroomus.files.wordpress.com/2017/04/facebook-and-information-operations-v1.pdf

Weick, K. E. (1984). Theoretical Assumptions and Research Methodology Selection. In F. W. McFarlan, *The Information System Research Challenge.* Boston: Harvard Business School.

Welch, C. (2018, October 17). *Facebook may have knowingly inflated metrics for over a year* . Retrieved from The Verge: https://www.theverge.com/2018/10/17/17989712/facebook-inaccurate-video-metrics-inflation-lawsuit

Welch, C. (2018, January 16). *YouTube tightens rules around what channels can be monetized.* Retrieved from The Verge: https://www.theverge.com/2018/1/16/16899068/youtube-new-monetization-rules-announced-4000-hours

whiteops. (2016). *The Methbot Operation.* Human Security.

Wikipedia. (n.d.). *Wikipedia: Do not create hoaxes.* Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Wikipedia:Do_not_create_hoaxes

Wilikilagi, V. (2009, October 26). *What is Network Governance and Its Impact on Public Policy Formulation.* Retrieved from SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1494757

Wilms, K. (2011, November 9). *Now announcing Medals, a new way to celebrate great videos.* Retrieved from YouTube Creator blog: https://youtube-creators.googleblog.com/2011/11/now-announcing-medals-new-way-to.html

Wintour, P. (2011, May 26). *Facebook founder Zuckerberg tells G8 summit: Don't regulate the web.* Retrieved from The Guardian: https://www.theguardian.com/technology/2011/may/26/facebook-google-internet-regulation-g8

Wong, J. C. (2019, February 7). *Overreacting to failure': Facebook's Myanmar strategy baffles local activists.* Retrieved from The Guardian: https://www.theguardian.com/technology/2019/feb/07/facebook-myanmar-genocide-violence-hate-speech

Wong, J. C. (2019, February 7). *'Overreacting to failure': Facebook's new Myanmar strategy baffles local activists.* Retrieved from The Guardian: https://www.theguardian.com/technology/2019/feb/07/facebook-myanmar-genocide-violence-hate-speech

Woolley, S. C., & Howard, P. N. (2016). Political Communication, Computational Propaganda, and Autonomous Agents. *International Journal of Communication, 10*, 4882-4890.

Woolley, S. C., & Howard, P. N. (2016). Political Communication, Computational Propaganda, and Autonomous Agents. *International Journal of Communication, 10*, 4882-1890.

Yadron, D. (2016, June 8). Silicon Valley appears open to helping U.S. spy agencies after terrorism summit. *The Guardian*.

Yaffe-Bellany, D. (2019, November 6). *Airbnb to verify all listings, C.E.O. Chesky says.* Retrieved from The New York Times: https://www.nytimes.com/2019/11/06/business/airbnb-verify-listings.html

Yale Law School. (2017). *Fighting Fake News.* Retrieved from Yale Law School: https://law.yale.edu/sites/default/files/area/center/isp/documents/fighting_fake_news_-_workshop_report.pdf

Yazan, B. (2015). Three Approaches to Case Study methods in Education: Yin, Merriam, and Stake. *The Qualitative Report, 20*(2), 134-152.

Yin, L., & Sankin, A. (2021, April 9). *Google Blocks Advertisers from Targeting Black Lives Matter YouTube Videos.* Retrieved from The Markup: https://themarkup.org/google-the-giant/2021/04/09/google-blocks-advertisers-from-targeting-black-lives-matter-youtube-videos

YouTube . (n.d.). *YouTube Trusted Flagger Program*. Retrieved from YouTube Help: https://support.google.com/youtube/answer/7554338?hl=en

YouTube . . (n.d.-b). *Google preferred lineups* . Retrieved from https://www.youtube.com/google-preferred/.

YouTube. (2007a , December 10). *Partner program expands.* Retrieved from Google Blog: https://youtube.googleblog.com/2007/12/partner-program-expands.html

YouTube. (2007b, May 3). *YouTube elevates most popular users to partners.* Retrieved from http://youtube.googleblog.com/2007/05/youtube-elevates-most-popular-users-to.html: YouTube Google Blog

YouTube. (2010, July 9). *Investing in the future of video: YouTube announces partner grant program.* Retrieved from YouTube Official Blog: https://blog.youtube/news-and-events/investing-in-future-of-video-youtube

YouTube. (2012a, November 28). *YouTube Creator Playbook and Creator Hub: Fresher, bolder, and live on the web.* Retrieved from YouTube Creator Blog, authored by Lauren Vilders: https://youtube-creators.googleblog.com/2012/11/youtube-creator-playbook-and-creator.html

YouTube. (2012b, July 25). *Introducing the YouTube creator space.* Retrieved from YouTube Creator Blog: https://youtube-creators.googleblog.com/2012/07/introducing-youtube-creator-space.html

YouTube. (2017a, August 7). *Expanding the ability to appeal more videos.* Retrieved from YouTube Creator Blog: https://youtube-creators.googleblog.com/2017/08/expanding-ability-to-appeal-more-videos.html

YouTube. (2017b, October 30). *Update on monetization appeal and the appeals process* . Retrieved from YouTube Help, authored by Marissa, Community Manager: https://support.google.com/youtube/forum/AAAAiuErobUG-hcnZ1x0RM/?hl=en&gpf=%23!msg%2Fyoutube%2FG-hcnZ1x0RM%2Fjw0ohBYCAQAJ&msgid=jw0ohBYCAQAJ

YouTube. (2018b, December 13). *Removing spam subscriptions from YouTube.* Retrieved from YouTube Help, authored by Jordan: https://support.google.com/youtube/forum/AAAAiuErobUAWHJfWAsqVk/?hl=en&gpf=%23!topic%2Fyoutube%2FAWHJfWAsqVk

YouTube Creators. (2020, March 20). *Coronavirus and YouTube: Answering Creator Questions.* Retrieved from YouTube: https://www.youtube.com/watch?v=i352PxWf_3M

YouTube. (n.d.). *Advertiser-friendly content guidelines.* Retrieved May 2021, from YouTube Help: https://support.google.com/youtube/answer/6162278#zippy=%2Cguide-to-self-certification

YouTube. (n.d.). *Request human review of videos marked "Not suitable for most advertisers".* Retrieved May 2021, from YouTube Help: https://support.google.com/youtube/answer/7083671?hl=en#:~:text=Go%20to%20the%20video%20you,video%20is%20eligible%20for%20appeal.

YouTube. (n.d.). *Self-Certify videos for ads.* Retrieved May 2021, from Creator Academy: https://creatoracademy.youtube.com/page/lesson/ad-friendly_self-certification_video

YouTube. (2018b, December 13). *Removing spam subscriptions from YouTube* . Retrieved from YouTube Help, authored by Jordan: https://support.google.com/youtube/forum/AAAAiuErobUAWHJfWAsqVk/?hl=en&gpf=%23!topic%2Fyoutube%2FAWHJfWAsqVk

YouTube. (n.d.-c). *Partner managers.* Retrieved from https://www.youtube.com/creators/partner-managers/

YouTube. (n.d.-d). *YouTube Social Impact.* Retrieved from https://socialimpact.youtube.com/.

Zamith, R. (2019). Algorithms and Journalism. In *Oxford Research Encyclopedia of Communication.*

Zittrain, J. (2014, June 1). Facebook could decide an election without anyone ever finding out . *New Republic*.

Zuckerberg, M. (2019, October 17). *Standing for Voice and Free Expression.* Retrieved from Facebook : https://about.fb.com/news/2019/10/mark-zuckerberg-stands-for-voice-and-free-expression/

Zuckerberg, M. (2016, November 19). *A lot of you have asked what we're doing about misinformation.* Retrieved from Mark Zuckerberg Facebook Page: https://www.facebook.com/zuck/posts/10103269806149061

Zuckerberg, M. (2018, November 15). *A Blueprint for Content Governance and Enforcement.* Retrieved from Facebook.com: https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/

Zuckerberg, M. (2019, March 6). *A Privacy-Focused Vision for Social Networking.* Retrieved from Facebook.com: https://www.facebook.com/notes/mark-zuckerberg/a-privacy-focused-vision-for-social-networking/10156700570096634/