Mis- and Disinformation online: a Taxonomy of Solutions

TESIS DOCTORAL

Autora: Anya Stiglitz
Director: Ángel Arrese

Pamplona, 2020
To my parents and my husband
Acknowledgements

A dissertation does not get written without an enormous amount of support and help. There are dozens of colleagues, friends, students, who generously gave their time by reading sections and commenting, recommending readings, arguing with me, sharing their ideas or even giving practical advice about how to get to the finish line. Some who helped are people I’ve known for years and some I met over email and offered to look at sections even though we had never met in person. I am grateful to all.

Despite being busy with her own studies, Chloe Oldham assisted with research and handled the citations even as the Covid 19 pandemic got under way and we were quarantined together for six weeks. In that same period, and before, Michael Schudson managed to read several chapters, point out omissions and give me a list of readings. From the first, Schudson, Susanna Narotzky and Richard John encouraged me to write a dissertation and gave me solid, practical advice about how to do it. Andie Tucher graciously read sections and sent detailed comments even as she was racing to finish her own book. Hawley Johnson sent endless references and ideas and useful articles.

My advisor, Ángel Arrese, lived up to his reputation for being smart, insightful and kind. His commitment to his students is unparalleled, his comments always on point and his care and attention remarkable. I’ve read his work for years but, as his student, I also learned from his example and his diligence. PHD programs can be lonely places but Professor Arrese brought me into a collegial and supportive community. Spending time at the communications department and, going to lectures with the faculty of University of Navarra, was an experience unlike any other and very different from my daily life at Columbia University in New York. I hope to return regularly.

Colleagues around the world who read chapters and sent comments include Mark Hannah, Lili Levy, Jacob Nelson, Phil Napoli, Ryan Whittington, Allison J. Head and
Margie Macmillan. Dozens of people let me interview them and shared their ideas. Most are listed in the appendix in the back. All were knowledgeable, generous and inspiring.

My father, André Schiffrin, died in 2013 so Jonathan Cole took over my father’s role of telling me what to read and sent lists of book recommendations that gave me valuable background for my research on Clyde Miller, McCarthyism and American Universities. Ellen Schrecker talked with me about certainty and interpretation as I read through the archives at Butler library. At times I felt I was back to the period I researched while doing my undergrad thesis on the newspaper PM at Reed College with Richard Fox. This gave me an excuse to re-start some old conversations with Richard and Sarah Berry.

I learned much from conversations with Guy Berger, Gordon Crovitz, Vera Franz, Tom Glaisyer, Ellen Goodman, Ellen Hume, Justin Hendrix, Hawley Johnson, Peter Micek, Joel Simon, Peter Micek, Christoph Deloire, Camille Grenier, Olaf Steenfadt, Maria Kattau, Tim Karr, James Deane, Mira Milosevic, Mark Nelson, Courtney Radsch, Alejandro Romero and Philippe Narval.

Zack Exley, Todd Gitlin, Michael Vachon shared their ideas about politics and the media and over the years I’ve absorbed many of the ideas of Julia Angwin, Emily Bell, Tim Besley, Yochai Benkler, Joan Donovan, Michelle Ferrier, Rana Foroohar, Phil Howard, Tim Karr, Phil Napoli, Jill Nelson, Renée di Resta, Ann Ravel, Amartya Sen, Kara Swisher, Zeynep Tufekci, Heidi Tworek and Ethan Zuckerman.

Some of the PHD students at Columbia were writing their dissertations while I was writing this one and so we helped each other. Efrat Nechushtai and I share an interest in media trust and community engagement. Alex Gonçalves and I have become co-authors and teach together. Javier Sauras and Ava Sirrah and I cheered each other on throughout.

My colleagues at SIPA were particularly supportive including: Ester Fuchs, Claudia Dreifus, Hollie Russon Gilman, Alexander Hertel-Fernandez, Suzanne Hollman, Sarah Holloway, Dan McIntyre, Jenny McGill, Sara Tjossem, Alexis Wichowski, Merit Janow and Colleen Toomey. Special thanks to the people who agreed to serve on my defense committee, as active or backup members: Javier Lesaca, Karolina Koc-Michalska, Susana Narotzky, Laura Moore, Andrea Prat. Conversations with Julia Stiglitz and Vijay Karunamurthy about Silicon Valley and with Jed Stiglitz and Kendra Bischoff, about their research, shaped my thinking too.
The economists in my life including Julia Cagé, Alexander Dyck, Martin Guzman, Antara Halder, Haaris Mateen, Elise Mougin, Andrea Prat, Hamid Rashid and, above all, my husband taught me a different way of looking at media effects and thinking about media systems. Overhearing their long conversations about economics undoubtedly affected my thinking about the world.

I want to acknowledge the researchers who have assisted with my courses at SIPA because, inevitably, some of that work wended its way into this one: Anamaria Lopez, Anna Kordunsky, Hannah Assadi, Matthew Hughes, Susanna de Martino. Ryan Powell came into my life as my RA, became part of my brain and is now a co-author and dear friend.

Other colleagues and helpers make life interesting and fun and carry a big load: Caleb Oldham, Gabriela Plump, Joe Chartier, Karine Lafargue, Monsse Motta and Ofelia Flores.

Many friends including Andrea Gurwitt, Jane Folpe, Jose Antonio Millan, Nicole Pope, Michael Staunton, Ava Seave, Leina Schiffrin, Ita de la Iglesia, Leo, Lara and Katya Sands, Alex, and Tamiko Soros, Robin Stephenson, Jonathan Birchall, Mila Rosenthal and Minky Worden encouraged me over the last three years as I juggled writing this dissertation with working full time.

The marvelous editor Rob Gurwitt read these pages, finding inconsistencies, making suggestions and providing the extra pair of eyes that every writer needs.

Finally: my husband for his endless encouragement, enthusiasm, support and unshakable belief that this subject matters and that I have something interesting to say about it.
# Table of contents

Abstract

Part I – Introduction and trust

Introduction

- Chapter 1: Trust, Journalism Authority and Mis/Disinformation

Part II—Demand-Side Solutions

- Introduction
- Chapter 2: Media Literacy
- Chapter 3: Community Engagement
- Chapter 4: Fact-checking

Part III—Supply-Side Solutions

- Chapter 5: Artificial intelligence, downranking, content moderation
- Chapter 6: intermediary liability, lessons from Europe
- Chapter 7: regulation of online political advertising—getting beyond transparency
- Chapter 8: defamation lawsuits and the marketplace of ideas

Conclusion

Bibliography

Interviews List

Appendixes
Abstract

The premise of this dissertation is that 2016 was the year that societies began to understand the dangers of online/mis/disinformation and decided to fund and implement solutions even though they had not been fully researched or tested. Many of the fixes were, in fact, based on the financial interests or belief systems of the people doing them. Facebook wanted to avoid regulation so it funded fact-checking and put the onus on audiences to become more media savvy/news literate. Journalists thought they could help the problem of online mis/disinformation by putting more effort into getting to know their audiences. Repressive regimes saw an opening to ban speech they didn’t like.

We provide a taxonomy of the variety of initiatives aimed at solving the problem, with the objective of enhancing our understanding of the strengths and limitations of each. Our analysis is set in the context of the academic literature on the problems of propaganda and dis/misinformation and media trust and we provide historical context for many of the solutions. The “solutions” we compare and contrast in this research include fact-checking initiatives and projects by journalists to promote community engagement, media literacy programs, technical fixes such as using as natural language processing/AI to block false and/or inflammatory content and finally government regulation. We discuss each solution, examining the intellectual and research basis for each as well as which groups have supported the different possible fixes and why.

Drawing on dozens of interviews with funders, journalists and regulators as well as archival material and including an exhaustive bibliography, this dissertation looks at the universe of solutions in an organized, structured way and concludes with policy proposals to help promote democracy.
Resumen

La premisa de esta investigación es que 2016 fue el año en que las sociedades comenzaron a comprender los peligros de la desinformación y la misinformación online, y decidieron financiar e implementar soluciones, aunque no se hubieran investigado o probado completamente. Muchas de las soluciones estaban, de hecho, basadas en los intereses financieros o sistemas de creencias de las personas que las hacían. Facebook quería evitar la regulación, por lo que financió la verificación de los hechos y puso la responsabilidad en las audiencias para convertirlas en más conocedoras de las noticias. Los periodistas también pensaron que podían ayudar al problema de la desinformación en línea poniendo más esfuerzo en conocer a sus audiencias. Los regímenes represivos vieron la posibilidad de prohibir los discursos que no les gustaban.

Proporcionamos una taxonomía de la variedad de iniciativas dirigidas a resolver el problema, con el objetivo de mejorar nuestra comprensión de las fortalezas y limitaciones de cada una. Nuestro análisis se enmarca en el contexto de la literatura académica sobre los problemas de la propaganda y la desinformación, y sobre la confianza de los medios de comunicación, y proporcionamos el contexto histórico de muchas de las soluciones. Las "soluciones" que comparamos y contrastamos en esta tesis incluyen iniciativas de comprobación de hechos y proyectos de los periodistas para promover la participación de la comunidad, programas de alfabetización de los medios de comunicación, propuestas técnicas como el uso de procesamiento de lenguaje natural/IL para bloquear el contenido falso y/o inflamatorio y, por último, la regulación gubernamental. Discutimos cada solución, examinando las bases intelectuales y de investigación de cada una de ellas, así como qué grupos han apoyado los diferentes arreglos posibles y por qué.

Basándose en docenas de entrevistas con financiadores, periodistas y reguladores, así como en material de archivo, e incluyendo una bibliografía exhaustiva, esta tesis examina el universo de soluciones de forma organizada y estructurada, y concluye con propuestas de políticas para ayudar a promover la democracia.
“In an ever-changing, incomprehensible world the masses had reached the point where they would, at the same time, believe everything and nothing, think that everything was possible and that nothing was true. ... Mass propaganda discovered that its audience was ready at all times to believe the worst, no matter how absurd, and did not particularly object to being deceived because it held every statement to be a lie anyhow. The totalitarian mass leaders based their propaganda on the correct psychological assumption that, under such conditions, one could make people believe the most fantastic statements one day, and trust that if the next day they were given irrefutable proof of their falsehood, they would take refuge in cynicism; instead of deserting the leaders who had lied to them, they would protest that they had known all along that the statement was a lie and would admire the leaders for their superior tactical cleverness.”

— Hannah Arendt, *The Origins of Totalitarianism*

**Understanding the proposed solutions to the problem of online mis/disinformation**

In 2016, the votes for Brexit and Donald Trump and the later Cambridge Analytica scandal made the public aware of the prevalence of online disinformation (Wardle 2016; Tandoc et al. 2018). Information trickled out about the role of the Russian government, the lies spread by Stephen Bannon and the far-right Breitbart News Network, as well as Fox News, and outrage grew. Attention turned to Facebook and Twitter, which were blamed for spreading lies in the relentless quest for clicks and likes. As journalists began writing about the spread of disinformation, the public and policy makers came to understand that the platforms’ business model was based on generating outrage and anger (Bell 2016; Angwin and Grassegger 2017). The problem was systemic.
It was a shocking wakeup call. The consequences went well beyond interference in democratic elections. Fury at the platforms intensified as it became clear that rumors and hate spread on Facebook and WhatsApp had fueled attacks on Muslims and ethnic minorities in India, Myanmar, and other places (Ingram 2018). By 2019, the Anti-vaxxer movement had grown so large that measles had returned in New York, the Philippines, and Italy and polio had made a comeback in Pakistan (Masood 2019; Shahzad and Ahmad 2019; McNeil Jr 2013).

While it wasn’t clear how much impact online mis/disinformation really had contributed to these shocking events, it was clear that a “weaponization of the digital influence machine” had taken place (Nadler, Crain & Donovan 2018). Policy makers, the platforms, entrepreneurs, journalists and educators immediately galvanized setting up committees, commissions, research groups, searching for new ways—and even new laws and regulations—aimed at tackling the problem of online disinformation. These steps were taken while the academic research was still underway, so the proposed solutions were often not informed by evidence as to what would work or even a deeper analysis of the problem. However, it was a case of needing to do something, so actions were taken before all the needed information was in (Engelke 2019; Nelson 2018).

Writing from the vantage point of 2020, this thesis provides original research (including, but not limited to, extensive interviews with policymakers in different countries) and brings together some of the academic literature on the problems of propaganda and mis/disinformation and media trust along with an understanding of solutions being proposed. We propose a taxonomy of solutions that covers many of the different initiatives aimed at solving the problem and we try to provide historical context for many of the solutions. The solutions discussed in this dissertation include journalistic reporting on the problem of mis/disinformation, promoting quality journalism, promoting media literacy, supporting fact-checking, technical fixes such as using as natural language processing/AI to block false or inflammatory content, and finally, government regulation. We discuss each solution, examining the intellectual and research basis for each, as well as which groups have supported the different possible fixes and why. We do this understanding that the challenges posed by mis/disinformation are so great that there is no “magic bullet”: a whole portfolio of policies will need to be deployed, and even that will not fully address the problems.

Why do so many thoughtful and experienced people come up with such radically different solutions to the problem of online mis/disinformation? One obvious reason is that there are very different financial interests involved. The second reason has to
do with the underlying beliefs of the groups proposing the solutions, including the US aversion to government regulation.

The third reason could be viewed as the exposure effect, as repeated exposure to an idea breeds support for it (Zajonc 1968). Organizations do what they are used to doing and this familiarity makes them think they are doing the right thing. Journalists believe in journalism and so think that more and better journalism is the solution. Wedded to the belief that trust in the media is somehow related to journalism practice, journalists also hope to improve standards and build trust through engagement and fact-checking (Ferrucci 2017; Wenzel 2019; Nelson 2018; Graves 2016). Fact-checkers believe that supporting a culture of truth may save not just journalism but also democracy (Ferrucci 2017; Graves 2016; Wenzel 2019; Cheruiyot and Ferrer-Conill 2018; Amazeen 2018a). Journalists believe they can build trust by engaging with audiences and that this can restore journalism to its rightful role in society (Robinson 2019; Ferrucci 2017). Groups that teach media and promote literacy believe that is the answer (Mihailidis and Viotty 2017). The large platforms and tech entrepreneurs seek to suppress disinformation by doing what they know how to do, i.e. hiring content moderators, changing platform algorithms and blocking certain kinds of false or inciteful content (Dreyfuss and Lapowsky 2019). Similarly, regulators seek regulation. The innate bias towards what is familiar is part of why different actors have backed different solutions.

**The demand for disinformation and the supply of it**

This paper proposes an analytical framework with which we can assess different solutions and which we believe provides some understanding of the limitations of each. For an overall understanding of the different ideas about solutions, we find that the economics terms “supply side” and “demand side” provide a useful framework for understanding the belief systems of the different groups involved in promoting decisions to the mis/disinformation problem. Guy Berger notes that the creation and dissemination of information lies on a continuum, that includes production, transmission, reception and reproduction, and many of the efforts aimed at fixing the problem emphasize one part of the continuum over another (Posetti & Bontcheva 2020; Author interview, Guy Berger 2019).

Those regulators who focus on the supply and transmission, of course, understand that there has always been some mis/disinformation—a point frequently made by those focused on audience consumption patterns. Societies can cope with small amounts that are of limited reach (such as a niche magazine with low circulation) but excessive supply of false information/rumors seeps into mainstream conversations,
overwhelms audiences, results in cognitive fatigue and makes it hard to distinguish true information from false information. Repeated exposure may aggravate the problem as the more audiences see something the more they believe it (Pennycook et al. 2018) even if it’s factually incorrect and later discredited. Corrections may not be seen by the people who originally saw the false information and may not be persuasive when someone’s mind is made up and they want to see their ideas confirmed (Kolbert 2017). Indeed, corrections, rather than having the intended effects, may only enhance distrust (Karlsson et al. 2017).

The regulators who focus on the prevalence of mis/disinformation see the problem as related to an excess supply of mis/disinformation. They focus on the incentives to supply it and the consequences of an excess supply. They ask how changing incentives by putting in regulations, codes of conduct, etc. can lessen the supply of mis/disinformation. The supply siders want Facebook, WhatsApp, and Twitter to limit what they circulate and promote and stop allowing people to make money off producing and disseminating false information. Another way to change the platforms’ incentives would be to make them liable for what appears on their platforms. To the extent that such changes in incentives do not suffice, some regulators believe regulations are necessary, including laws against hate speech or limits on the ability to make certain messages go viral.

By contrast, others focus on improving the ability of consumers to evaluate the information with which they are confronted. They may be relatively unconcerned, arguing that “fake news” and mis/disinformation has always existed, and that accordingly, there is no reason to panic and little evidence that its audiences are persuaded by what they see online (Allcott, Gentzkow and Yu 2019). The tech companies fall in this category, expressing the view that they should not be blamed and that the responsibility lies with society more generally. Some, including Facebook and various foundations (Murgia 2017) fund the teaching of media literacy in schools so that audiences will become more discerning consumers. Others believe in labelling non-verified news in the hope this will get audiences to stop circulating it. Facebook is funding fact-checking efforts throughout the world (Funke 2019). Many free expression groups, particularly in the US, oppose hasty government responses that broaden censorship and liability online for the platforms and could do long-term harm.

**The role of motivated reasoning, financial incentives and ideology**

Incentives and ideology help us understand the position taken by various parties on the desirability of the appropriate measures to deal with mis/disinformation. A term
that originated in social psychology and is used in economics to understand different perspectives is “motivated reasoning” or “reasoning in the service of belief” (Epley and Gilovich 2016; Kunda 1990).

Unsurprisingly, many of the beliefs about solutions to the problem of online mis/disinformation often correspond with the financial incentives particular to each belief-holder. As US muckraking journalist Upton Sinclair is quoted as saying: “It is difficult to get a man to understand something when his salary depends on his not understanding it.”

In the case of the tech companies, there is a vast amount of money at stake. Facebook and Twitter don’t want to be regulated or change their business models, so they would rather off-load responsibility for fixing the problem and donate small amounts of money to help solve it (author interview, anonymous, May 2019). Their ideology is often that of techno-libertarianism, so they reject regulation, or at least regulation that is likely to affect their revenues.

Financial incentives underscore the belief systems of the tech giants but belief in certain solutions over others also results from underlying ideology and belief in what one does. “If you have a hammer then everything looks like a nail” (Maslow 1966). Journalists believe in journalism and so are more likely than others to believe that more and better journalism is the solution. Wedded to the belief that trust in the media is somehow related to journalism practice, journalists also hope to improve standards and build trust through engagement and fact-checking. So too, foundations are accustomed to giving grants, so they see the problem as one that they can help solve by giving grants to organizations trying to research and fix the problem.

Solutions that focus on reducing the supply of false information online are controversial and difficult to implement. Fixes that focus on audience demand may seem more do-able in the short term. It takes years of complicated negotiations to pass a law about online hate speech or transparency of political advertising. Giving a grant to a pre-existing news literacy NGO or a fact-checking organization can be done in a matter of weeks. The appeal of short-term solutions to the tech companies is obvious. Offloading the problem of mis/disinformation takes the onus away from the platforms and puts it on journalists and consumers (Bell 2016). It would be simple and convenient if these ideas worked, but they were implemented at a time when evidence was lacking. Moreover, they are expensive, hard to scale, and slow (Schiffrin et al. 2017).
The role of national bias: US focuses on individual responsibility, Europe is more supportive of regulation. Repressive regimes are repressive

In looking around the world at the different solutions proposed it is clear that national bias and ideology play an important, if unspoken, role. The US is more suspicious of government regulation than Europe and less likely to push for government-led solutions than Germany. Differences within the EU Commission as to how to solve the problem stem in part from the ideologies of Commission officials, with members from former Communist countries less likely to support government regulation and more likely to skew towards voluntary efforts by the platforms (author interviews, Brussels, March 2019).

Governments with less open, or downright repressive, attitudes toward freedom of expression have little compunction in cracking down on the platforms and using the fear of fake news as a reason to practice censorship online. Cuba, China, Singapore, Turkey, and Vietnam are all examples that come to mind. For instance, in Singapore, journalists face potential jail time if they publish stories that are perceived as “falsehoods with malicious intent or going against Singapore's public interest” under the 2019 law intended to combat mis/disinformation (Vaswani 2019).

Many of the US responses highlight the individual responsibility of audience members, exhorting people not to circulate or forward information that is false and to learn how to tell the difference between true and false information. Alan Miller (2019), the founder of the US educational nonprofit News Literacy project, explains, “We need a change in consciousness to counteract this fog of confusion and mistrust. First, we must understand—and take responsibility for—our roles in the 21st-century information ecosystem. Misinformation can’t spread virally unless we infect others with it. We need to slow down before we hit ‘share’ or ‘retweet’ or ‘like,’ and ask ourselves if doing so will mislead, misinform or do harm.” But without regulation this “slowing down” is unlikely to occur. Those spreading information often have reasons for doing so beyond just carelessness. The spreading of political disinformation or non scientific beliefs such as the anti-vaxxer movement are just two examples.

Defining our terms

There are many kinds of mis/disinformation and several attempts have been made to provide typologies. Tandoc, Lim, and Ling (2017) reviewed 34 scholarly articles published between 2003 and 2017 and came up with a typology that included: satire, parody, false images, advertising and public relations, which sometimes overlaps with propaganda. For our purposes we will consider, in a following chapter, the
relationship between propaganda and disinformation and focus too on what Tandoc, Lim, and Ling describe as “news fabrication”. This is often done with the intention to deceive, and the false news is often difficult to identify as it presents as a traditional piece of news with similar format and conventions.

As Tandoc, Lim, and Ling (2017) note:

As with the case of parody, a successful fabricated news item, at least from the perspective of the author, is an item that draws on pre-existing memes or partialities. It weaves these into a narrative, often with a political bias, that the reader accepts as legitimate. The reader faces further difficulty in verification since fabricated news is also published by non-news organizations or individuals under a veneer of authenticity by adhering to news styles and presentations. The items can also be shared on social media and thus further gain legitimacy since the individual is receiving them from people they trust.

The authors also note that “facticity” is another question in the determination of false news, as the false information might be derived from, or rely on, something that is true or partially true: for example, the right-wing website that slaps a false headline on an article from a reputable media outlet. Audience matters as well because under certain conditions, audiences are more receptive to false news.

Another set of discussions around the problem of false news has been the recent interest in disinformation, which is false information spread deliberately to deceive. The English word disinformation resembles the Russian word “dezinformatsiya,” derived from the title of a KGB black propaganda department. The typology created by Claire Wardle, executive director of First Draft, discusses this phenomenon and has been widely used. In her influential papers and reports Wardle said the term “fake news” is misleading and in 2017 released her rubric “Fake News, It’s Complicated,” which is now a standard for the discussion about the problem. In this paper, Wardle describes the types of mis/disinformation as satire and parody, misleading content, imposter content and fabricated content, false connection, false context, and manipulated content (Wardle & Derakhshan 2017). Her paper with Hossain Derakhshan also included a rubric of who the different actors and targets are such as states targeting states, states targeting private actors, corporates targeting consumers (Wardle & Derakhshan 2017).

They further make the point that the intentions of the person creating and/or amplifying the false information are relevant to the definition.

- Misinformation is when false information is shared, but no harm is meant.
Disinformation is when false information is knowingly shared to cause harm.

Mal-information is when genuine information is shared to cause harm, often by moving information designed to stay private into the public sphere.

Sometimes disinformation disguised as parody can spread into conversations. In this dissertation we will focus on mis/disinformation, with occasional forays into propaganda, rather than include satire and parody. Our interest is in news that we believe is affecting social behavior and contributing to the rise of demagogues, violence and anti-science attitudes around the world. This dissertation touches on human psychology and the political and economic incentives of those that disseminate falsehoods online, but our main area of interest is solutions to the problem. We don’t use the term “fake news” because of the problematic usage by President Donald Trump, Brazilian president Jair Bolsonaro, and many others. Post-2016, it quickly became a term used to disparage and undermine journalists and the media.

Why focus on social media?

We are well aware of the fact that in many parts of the world the internet has not reached full penetration and that in Africa, for example, radio, television and mobile are still the most important mediums of communication (Editorial Team 2016). We further understand the power of talk radio in the US, particularly earlier worries about syndicated talk radio hosts such as Rush Limbaugh and cable news (Della Vigna and Kaplan 2007; Hopkins and Ladd 2014; Rosenwald 2019; Nelson 2019; McGirr 2001). However, rather than tackling the general problem of false information, this thesis focuses on the solutions to the problem of mis/disinformation circulated on social media. We take this approach in part because the period since 2016 has underscored the harms caused by the circulation of such information and penetration of social media is growing in many parts of the world. The problem may become worse.

Further, the virality of the content means it can be more influential than a local radio program, with its localized reach and less sophisticated presentation of information. The extensive use of video on YouTube, TikTok and other platforms may make the

---

1 A relatively harmless example is the letter purportedly written by F. Scott Fitzgerald about the Spanish Flu which circulated widely in March 2020.

More serious examples can be found in the works by Peter Pomersantsev and Yochai Benkler, Robert Faris and Hal Roberts which we cite frequently in this dissertation.
 messaging more urgent, attractive, and persuasive (Munger 2020). The social nature of social media means that friends and family as well as communities around the world are reinforcing the messaging and helping groups form and develop (Munger 2020). Often these make use of anonymity and encryption (WhatsApp, 8chan, 4chan) and are therefore harder to monitor. The future technical evolution of these platforms, including the rise of deepfakes, will further contribute to the rapid dissemination of confusing and misleading content. We will delve later into the social science research on media trust, persuasion and attempts to understand whether/how online mis- and disinformation is changing beliefs. Here we will note that although many of the studies are narrowly focused and short-term and don’t provide all the answers, if we take a step back and look at larger social and political trends—for example the rise of the anti-vaxx movement—it is difficult to argue the Null Hypothesis (that nothing is happening).

Since the introduction of massive amounts of false information into the online ecosystem, behavior has changed around the world. This is reason enough to try to understand the phenomenon of online mis/disinformation and possible solutions. There is certainly a school of thought that media effects are not as powerful as imagined (Newton 1999; Schudson 2019; Karpf 2019) but the recent studies of Fox News suggest that there are measurable effects of news consumption. It’s unlikely that researchers will ever be able to quantify with any precision whether exposure to online mis/disinformation or incitement is powerful enough on its own to cause behavior changes or whether it is the interaction of online mis/disinformation with other forms of falsehood that has persuasive effects (Jamieson 2018). To take one example: the killings in Burma, India, and Nigeria that followed inciteful messages spread on WhatsApp. To what extent were people influenced by the messages, their previous prejudices, or inflammatory speech by local political and religious leaders? We’ll never know, and this dissertation cannot tackle every form of media. Rather, we will focus on the most controversial and most rapidly growing media arena of public concern: the social media platforms.

**Why this dissertation?**

This dissertation’s contribution is an argument as to which of the myriad solutions being tried are likely to be effective or what mix of measures might be desirable. We believe that it’s essential to delve into how the solutions were arrived at and the interests and implicit biases involved. By providing a framework for understanding the different types of solutions that exist and a systematic account of each type, this
dissertation will form the basis of a clearer judgment as to what could work and what will not.

**Dissertation structure**

This dissertation has three parts. This introduction lays out the main argument and some key concepts, such as what we mean by mis/disinformation, why it matters, and what we believe the dissertation contribute to the study of the topic.

**Chapter One** looks at different conceptions of trust and how it has been researched. Understanding the idea of trust and its close companion, credibility, is essential to understanding questions of mis/disinformation. Both are part of the larger question of persuasion, and fears of persuasion underscore the anxiety about mis/disinformation. Societies worry that repeated exposure to an argument (whether right or wrong) will influence behavior and, indeed, the rise in hate crimes, the election of demagogues in parts of the world, and the drop in vaccination rates in many countries all suggest that online mis/disinformation is, indeed, a powerful persuader. Parsing the impact of information on the human psyche is extremely difficult, if not impossible, but understanding what we know about trust in media is part of thinking about the impact of information. This dissertation is not the last word on the topic—many have spent their careers trying to understand media persuasion and impact. Rather we will provide an overview of the existing literature and its historical background and context, note some of the research gaps and, most importantly, explain how trust relates to our larger understanding of how to address the problem of online mis/disinformation.

Some of the early discussion about media trust took place in the period around World War II when intellectuals in Europe and the US (particularly those in the Frankfurt School) tried to understand how citizens could be susceptible to Nazi propaganda (Jeffries 2016). In the US there was worry about the influence of demagogues, such as Father Coughlin, who used radio to get their messages across. Fear of new kinds of technology has often contributed to fears of mis/disinformation (Tucher 2013) but we believe that it’s a mistake to dismiss such fears as mere Luddism. Rather, there were objective political catastrophes taking place, which were to have global consequences, and worrying about that and trying to understand the role of propaganda, including how it was created, disseminated, and had influence, was a necessary response to the times.

The rise of Fascism and Communism in the 1920s and 1930s provides a backdrop to **Chapter Two**, in which we discuss media literacy and Clyde Miller’s attempts in the
1930s to teach U.S. schoolchildren how to understand and resist propaganda. Miller was a former journalist who worked at Columbia Teachers College for 10 years. During that time, he raised one million dollars from Boston businessman Edward A. Filene for the Institute for Propaganda Analysis (IPA). Miller’s story has not been fully told and provides important insights for current debates. Our reading of three different archives of Miller’s papers uncovered information about Miller’s own career and the attacks and red-baiting he was subjected to by the right-wing media mogul William Randolph Hearst—in some ways the Rupert Murdoch of the early 20th century. Moreover, Miller’s early media literacy efforts laid the foundations for much of contemporary media literacy work (Hobbs and McGee 2014). His taxonomies of propaganda techniques and his work analyzing pieces of disinformation in the media anticipate many of the techniques used today. Miller and his colleague, Violet Edwards, worked closely with teachers and provided material for them to use, as well as weekly mailings for school children.

Chapters Three and Four discuss efforts to build audience demand for and trust in quality journalism. These efforts try to build trust in journalism by establishing journalism as a force for truth-telling (in the case of fact-checking) or by trying to make media outlets relevant to audiences (community engagement efforts). We argue that after 2016, foundations and the platforms reached for fixes and funded efforts that tackled the demand side for a range of reasons. Facebook wanted to avoid regulation and, further, believing in the importance of free dissemination of ideas and information online, thought that helping audiences become more educated would be a suitable fix. We examine the efforts to combat mis/disinformation online by building trust in journalism and the ability of audiences to distinguish good from bad information. Chapters Three and Four discuss solutions espoused by “demand sides,” who include the media literacy and fact-checking communities and those who hope that building community and engaging with audiences will help restore trust in journalism.

Later chapters will look at solutions related more to the supply, transmission and reproduction of online mis/disinformation: the attempt to use algorithms and machine learning to block and suppress content, and the possibilities for regulation. Chapter Five reviews some of the start-ups that use artificial intelligence and natural language processing and assesses the likelihood of these being effective and able to scale. We conclude chapter five by discussing a few initiatives that rate news outlets and propose standards for journalists to follow and suggest that these may be faster to scale than the tech solutions of the small start-ups.
We conclude this dissertation with Chapters Six, Seven, and Eight on legal remedies. In Chapter Six we discuss the removal of intermediary liability (as has been done in Germany) and see how this could be done in the US. In Chapter Seven we outline some of the problems with political advertising and propose a voluntary provision of political news during elections. In Chapter Eight we explain why the marketplace of ideas is flawed and discuss the use of defamation suits, by journalists, as a way of combatting violent, targeted speech online.

**Personal notes**

Because my family has lived through and been refugees from three authoritarian regimes that made extensive use of propaganda, the topic of political disinformation online is of great personal interest to me. The experiences of the 1930s have particular resonance as we seek to understand the contemporary political economy of mis/disinformation. The rise of right-wing demagogues globally and their creation of messaging to incite hatred, and even violence, is reminiscent of the 1930s. So is the constant attack on migrants and refugees and groups that are considered “other”. But the resemblances are not just between the Far-Right today and in the 1930s. The U.S. has even seen a return to the red-baiting of which Miller was a victim. In short, there are parallels between today’s attacks on truth-seeking institutions such as journalists, universities, the judiciary, and even Congress and the attacks of the 1930s.

This is terrain that is familiar to me as my family were refugees who had to flee several countries in the first half of the 20th century. My Russian grandfather, and his family, left Baku around the time of the 1917 Russian Revolution. My Republican military family had to leave Spain in 1939. My father and his parents, all of whom were French citizens, had to leave Paris in 1941 after my grandfather was fired by Gaston Gallimard from the publishing house my grandfather founded within days of the anti-Jewish laws being imposed by the Vichy government in 1940 (Reichman 2019). My family experienced what happens when dictators come to power. One key question for many of the refugees of the 1930s was timing. Knowing when to leave—or being able to leave—was essential. And yet it’s human nature to assume things aren’t bad or won’t get bad. Part of a robust early warning system is watching the media in order to get a sense of where societies are moving or what ideas government is trying to inculcate. Just as radio in Rwanda incited the killings of the Tutsis, so the propaganda of the Nazis incited hatred of Jews, Roma, and other groups. The Nazi’s Minister of Propaganda, Joseph Goebbels, knew the power of the media when he put a radio in each German household (Meier 2018). They attacked independent journalism and talked about the power of the big lie.
Introduction

It’s not just that propaganda and incitement are unpleasant. We believe there is a connection between mis/disinformation and human behavior. The rise of mis/disinformation online today is affecting political events around the world. Some of what people see online has become a threat to democracy for many reasons: It aggravates feelings of mistrust in institutions, it can incite people to violence, it promotes hatred of migrants and minorities and allows conspiracy theories to circulate. Thus, finding solutions that work is essential to the physical health and safety of citizens as well as the health of our democracy.

Conclusion

We do not believe it’s possible to separate trust in media and susceptibility to disinformation from larger questions about trust in the political system and institutions more generally. We agree that problems of trust in the media cannot be separated from worries about economic inequality, globalization, and larger political anxieties (Schudson 2019). We understand the argument made by US political activist Zack Exley that the problem is the underlying beliefs and not the social media platforms where they manifest (author conversations and messages with Zack Exley 2019 and 2020).

We do believe, however, that online mis/disinformation can provide a toxic ingredient (the match to the dry and brittle house) and that attempts to combat it are part of a larger effort to create norms that uphold important social values like truth, evidence, and the scientific method. Thus, while we criticize aspects of the search for solutions, we think they are critical to the survival of society.

The challenge is to find solutions that work, that do not threaten free expression, and, above all, that cannot be gamed by interest groups such as the tech giants or politicians who are not acting in good faith. Finding laws that cannot be abused by those in power will be difficult. But there is a lack of evidence showing that fixes like fact-checking, media literacy, community engagement, and tweaking algorithms are sufficient. Further, the tech companies have demonstrated that voluntary codes of conduct are not enough. It is the threat of regulation that propels them to act. For this reason, we argue in the final parts of this dissertation that regulations—most likely originating in Europe—will be an essential part of fixing the problem. We further support the initiatives underway to create large funds that will support public-interest media whether by expanding public-service broadcasting or by support small, community-run news outlets.
Online mis/disinformation is difficult to fix, making it all the more important to have an enhanced understanding of the assumptions underlying each solution and an overall framework within which to place the proposed solutions and fixes being implemented globally. We hope the taxonomy and analysis in this dissertation will deepen understanding of the options available to address the problem.
“In an ever-changing, incomprehensible world the masses had reached the point where they would, at the same time, believe everything and nothing, think that everything was possible and that nothing was true. ... Mass propaganda discovered that its audience was ready at all times to believe the worst, no matter how absurd, and did not particularly object to being deceived because it held every statement to be a lie anyhow. The totalitarian mass leaders based their propaganda on the correct psychological assumption that, under such conditions, one could make people believe the most fantastic statements one day, and trust that if the next day they were given irrefutable proof of their falsehood, they would take refuge in cynicism; instead of deserting the leaders who had lied to them, they would protest that they had known all along that the statement was a lie and would admire the leaders for their superior tactical cleverness.”

— Hannah Arendt, *The Origins of Totalitarianism*

---

**Introducción**

**Comprendiendo las soluciones propuestas para el problema de la mis- y desinformación en línea**

En 2016, los votos a favor de Brexit y Donald Trump y el posterior escándalo de Cambridge Analytica hicieron que el público tomara conciencia de la prevalencia de la desinformación en línea (Wardle 2016; Tandoc y otros 2018). Mucha información se filtró sobre el papel del gobierno ruso, las mentiras difundidas por Stephen Bannon y la ultraderechista Breitbart News Network, así como Fox News, y la indignación creció. La atención se centró en Facebook y Twitter, a los que se culpó de difundir mentiras en la búsqueda incesante de clics y likes. A medida que los periodistas comenzaron a escribir sobre la propagación de la desinformación, el público y los responsables políticos llegaron a comprender que el modelo de negocio de las plataformas se basaba en generar indignación e ira (Bell 2016; Angwin y Grassegger 2017). El problema era sistémico.
Fue una impactante llamada de atención. Las consecuencias fueron mucho más allá de la interferencia en las elecciones democráticas. La furia contra las plataformas se intensificó a medida que se hizo evidente que los rumores y el odio difundidos en Facebook y WhatsApp habían alimentado los ataques contra los musulmanes y las minorías étnicas en la India, Myanmar y otros lugares (Ingram 2018). En 2019, el movimiento anti-vacunaciones había crecido tanto que el sarampión había regresado a Nueva York, Filipinas e Italia y la poliomielitis había reaparecido en Pakistán (Masood 2019; Shahzad y Ahmad 2019; McNeil Jr 2013).

Si bien no estaba claro cuánto impacto había tenido la desinformación en línea en estos espantosos acontecimientos, sí estaba claro que se había producido una "militarización de la máquina de influencia digital" (Nadler, Crain y Donovan 2018). Los encargados de formular políticas, las plataformas, los empresarios, los periodistas y los educadores impulsaron inmediatamente la creación de comités, comisiones, grupos de investigación, la búsqueda de nuevas formas -e incluso nuevas leyes y reglamentos- destinadas a abordar el problema de la desinformación en línea. Esas medidas se adoptaron mientras se realizaba la investigación académica, por lo que las soluciones propuestas a menudo no se basaban en pruebas de lo que funcionaría o incluso en un análisis más profundo del problema. Sin embargo, se trataba de un caso de necesidad de hacer algo, por lo que se tomaron medidas antes de que se dispusiera de toda la información necesaria (Engelke 2019; Nelson 2018).

Esta tesis, redactada desde la perspectiva de 2020, ofrece una investigación original (que incluye, entre otras cosas, amplias entrevistas con responsables de la formulación de políticas en diferentes países) y reúne parte de la literatura académica sobre los problemas de la propaganda y la desinformación y la confianza de los medios de comunicación, junto con la comprensión de las soluciones que se proponen. Proponemos una taxonomía de soluciones que abarca muchas de las diferentes iniciativas destinadas a resolver el problema e intentamos proporcionar un contexto histórico para muchas de esas soluciones. Las propuestas que se discuten en esta disertación incluyen los reportajes periodísticos sobre el problema de la desinformación, la promoción del periodismo de calidad, la promoción de la alfabetización mediática, el apoyo a la comprobación de hechos, las correcciones técnicas a través del uso del procesamiento del lenguaje natural/AI para bloquear el contenido falso o inflamatorio y, por último, la regulación gubernamental. Discutimos cada solución, examinando las bases intelectuales y de investigación de cada una de ellas, así como qué grupos han apoyado los diferentes arreglos posibles y por qué. Lo hacemos entendiendo que los retos que plantea la desinformación son tan grandes que
no hay una "bala mágica": será necesario desplegar todo un portafolio de políticas, e incluso eso no resolverá completamente los problemas.

¿Por qué tantas personas reflexivas y experimentadas proponen soluciones tan radicalmente diferentes al problema de la desinformación en línea? Una razón obvia es que hay intereses financieros muy diferentes implicados. La segunda razón tiene que ver con las creencias subyacentes de los grupos que proponen las soluciones, incluyendo la aversión de los EE.UU. a la regulación gubernamental.

La tercera razón podría ser vista como el ‘efecto de la exposición’, ya que la exposición repetida a una idea genera apoyo para ella (Zajonc 1968). Las organizaciones hacen lo que están acostumbradas a hacer y esta familiaridad les hace pensar que están haciendo lo correcto. Los periodistas creen en el periodismo y por lo tanto piensan que más y mejor periodismo es la solución. Junto con la creencia de que la confianza en los medios de comunicación está de alguna manera relacionada con la práctica del periodismo, los periodistas también esperan mejorar los estándares y construir la confianza a través del compromiso y la comprobación de los hechos (Ferrucci 2017; Wenzel 2019; Nelson 2018; Graves 2016). Los verificadores de hechos creen que el apoyo a una cultura de la verdad puede salvar no sólo al periodismo sino también a la democracia (Ferrucci 2017; Graves 2016; Wenzel 2019; Cheruiyot y Ferrer-Cardill 2018; Amazeen 2018a). Los periodistas también creen que pueden crear confianza al comprometerse con las audiencias y que esto puede devolver al periodismo el papel que le corresponde en la sociedad (Robinson 2019; Ferrucci 2017). Los grupos que enseñan sobre los medios de comunicación y promueven la alfabetización creen que esa es la respuesta (Mihailidis y Viotty 2017). Las grandes plataformas y los empresarios de la tecnología tratan de suprimir la desinformación haciendo lo que saben hacer, es decir, contratando moderadores de contenido, cambiando los algoritmos de las plataformas y bloqueando ciertos tipos de contenido falso o incitante (Dreyfuss y Lapowsky 2019). Del mismo modo, los reguladores buscan la regulación. El sesgo innato hacia lo que es familiar es parte de la razón por la que los diferentes actores han respaldado diferentes soluciones.

La demanda de desinformación y la oferta de la misma

En este documento se propone un marco analítico con el que podemos evaluar diferentes soluciones y que creemos que proporciona cierta comprensión de las limitaciones de cada una de ellas. Para una comprensión general de las diferentes ideas sobre soluciones, encontramos que los términos económicos "lado de la oferta" y "lado de la demanda" proporcionan un marco útil para comprender los sistemas de creencias de los diferentes grupos involucrados en la promoción de decisiones para
resolver el problema de la desinformación. Guy Berger observa que la creación y difusión de la información se encuentra en un continuo, que incluye la producción, la transmisión, la recepción y la reproducción, y muchos de los esfuerzos encaminados a solucionar el problema ponen de relieve una parte del continuo sobre otra (Posetti & Bontcheva 2020; Entrevista con el autor, Guy Berger 2019).

Los reguladores que se centran en el suministro y la transmisión, por supuesto, entienden que siempre ha habido cierta desinformación, algo que con frecuencia señalan los que se centran en las pautas de consumo de la audiencia. Las sociedades pueden hacer frente a pequeñas cantidades de desinformación con alcance limitado (como una revista de nicho de escasa circulación), pero el suministro excesivo de información/rumores falsos se filtra en las conversaciones principales, abruma al público, provoca fatiga cognitiva y hace difícil distinguir la información verdadera de la falsa. La exposición repetida puede agravar el problema, ya que cuanto más vea el público algo, más lo creerá (Pennycook y otros, 2018), aunque sea incorrecto en cuanto a los hechos y posteriormente desacreditado. Las correcciones pueden no ser vistas por las personas que originalmente vieron la información falsa y pueden no ser persuasivas cuando alguien se decide y quiere ver sus ideas confirmadas (Kolbert 2017). De hecho, las correcciones, en lugar de tener los efectos deseados, sólo pueden aumentar la desconfianza (Karlsson et al. 2017).

Los organismos reguladores que se centran en la prevalencia de la información errónea/desinformada consideran que el problema está relacionado con un exceso de suministro de información errónea/desinformada. Se centran en los incentivos para suministrarla y en las consecuencias de un exceso de oferta. Se preguntan cómo el cambio de incentivos mediante la introducción de reglamentos, códigos de conducta, etc. puede disminuir el suministro de información errónea o falsa. Los proveedores quieren que Facebook, WhatsApp y Twitter limiten lo que circulan, y promuevan y dejen de permitir que la gente gane dinero produciendo y difundiendo información falsa. Otra forma de cambiar los incentivos de las plataformas sería hacerlas responsables de lo que aparece en ellas. En la medida en que estos cambios en los incentivos no sean suficientes, algunos reguladores creen que es necesario regularlas, incluyendo leyes contra la incitación al odio o límites a la capacidad de hacer que ciertos mensajes se vuelvan virales.

En cambio, otros se centran en mejorar la capacidad de los consumidores para evaluar la información a la que se enfrentan. Es posible que se muestren relativamente despreocupados, argumentando que siempre han existido "noticias falsas" y desinformación, y que, en consecuencia, no hay motivos para el pánico y hay pocas
pruebas de que sus audiencias estén persuadidas por lo que ven en línea (Allcott, Gentzkow y Yu 2019). Las empresas tecnológicas entran en esta categoría, expresando la opinión de que no se les debe culpar y que la responsabilidad recae en la sociedad en general. Algunas, entre ellas Facebook y varias fundaciones (Murgia 2017) financian la enseñanza de la alfabetización mediática en las escuelas para que el público se convierta en un consumidor más exigente. Otros creen en el etiquetado de las noticias no verificadas con la esperanza de que esto haga que el público deje de circularlas. Facebook está financiando esfuerzos de verificación de hechos en todo el mundo (Funke 2019). Muchos grupos de en favor de la libre expresión, particularmente en los Estados Unidos, se oponen a las respuestas precipitadas de los gobiernos, que amplían la censura y la responsabilidad en línea de las plataformas y que podrían causar daños a largo plazo.

El papel del razonamiento motivado, los incentivos financieros y la ideología

Los incentivos y la ideología nos ayudan a comprender la posición adoptada por las diversas partes sobre la conveniencia de las medidas adecuadas para hacer frente a la desinformación. Un término que se originó en la psicología social y se utiliza en la economía para entender diferentes perspectivas es el de "razonamiento motivado" o "razonamiento al servicio de la creencia" (Epley y Gilovich 2016; Kunda 1990). No es de extrañar que muchas de las creencias sobre las soluciones al problema de la desinformación en línea a menudo se correspondan con los incentivos financieros propios de cada titular de una creencia. Como se cita al periodista estadounidense Upton Sinclair: "Es difícil hacer que un hombre entienda algo cuando su salario depende de que no lo entienda".

En el caso de las empresas tecnológicas, hay una gran cantidad de dinero en juego. Facebook y Twitter no quieren que se les regule o que se cambien sus modelos de negocio, por lo que prefieren descargar la responsabilidad de arreglar el problema y donar pequeñas cantidades de dinero para ayudar a resolverlo (Entrevista con el autor, anónima, mayo de 2019). Su ideología suele ser la del tecno-libertarianismo, por lo que rechazan la regulación, o al menos la que probablemente afecte a sus ingresos.

Los incentivos financieros subrayan los sistemas de creencias de los gigantes de la tecnología, pero la creencia en ciertas soluciones por encima de otras también resulta de la ideología subyacente y de la creencia en lo que uno hace. "Si tienes un martillo, entonces todo parece un clavo" (Maslow 1966). Los periodistas creen en el periodismo y por lo tanto es más probable que crean que más y mejor periodismo es la solución. Junto con la creencia de que la confianza en los medios de comunicación
está de alguna manera relacionada con la práctica del periodismo, los periodistas también esperan mejorar los estándares y construir la confianza a través del compromiso y la comprobación de los hechos. También las fundaciones están acostumbradas a conceder subvenciones, por lo que consideran que el problema se puede resolver concediendo subvenciones a organizaciones que tratan de investigar y solucionar estos problemas.

Las soluciones que se centran en la reducción del suministro de información falsa en línea son controvertidas y difíciles de aplicar. Las que se centran en la demanda del público pueden parecer más factibles a corto plazo. Lleva años de complicadas negociaciones aprobar una ley sobre la incitación al odio en línea o la transparencia de la publicidad política. La concesión de una subvención a una ONG de alfabetización de noticias preexistente o a una organización de verificación de hechos puede hacerse en cuestión de semanas. El atractivo de las soluciones a corto plazo para las empresas tecnológicas es obvio. Descargar el problema de la desinformación le quita la carga a las plataformas y la pone en los periodistas y consumidores (Bell 2016). Sería simple y conveniente que estas ideas funcionaran, pero se implementaron en un momento en que faltaban pruebas. Además, son costosas, difíciles de escalar y lentas (Schiffrin et al. 2017).

**El papel del sesgo nacional: Los Estados Unidos se centran en la responsabilidad individual, Europa es más partidaria de la reglamentación.**

**Los regímenes represivos son represivos**

Al mirar alrededor del mundo las diferentes soluciones propuestas, está claro que el sesgo y la ideología nacionales juegan un papel importante, aunque no se diga. Los EE.UU. son más desconfiados de la regulación gubernamental que Europa y menos propensos a impulsar soluciones dirigidas por el gobierno que Alemania. Las diferencias dentro de la Comisión de la UE en cuanto a la forma de resolver el problema se derivan en parte de las ideologías de los funcionarios de la Comisión, ya que es menos probable que los miembros de los antiguos países comunistas apoyen la reglamentación gubernamental y es más probable que se inclinen por los esfuerzos voluntarios de las plataformas (entrevistas con los autores, Bruselas, marzo de 2019).

Los gobiernos con actitudes menos abiertas, o francamente represivas, hacia la libertad de expresión tienen pocos reparos en tomar medidas energéticas contra las plataformas y en utilizar el temor a las noticias falsas como motivo para practicar la censura en línea. Cuba, China, Singapur, Turquía y Viet Nam son todos ejemplos que me vienen a la mente. Por ejemplo, en Singapur, los periodistas se enfrentan a posibles penas de cárcel si publican historias que se perciben como "falsedades con
intención maliciosa o que van en contra del interés público de Singapur" en virtud de la ley de 2019 destinada a combatir la desinformación/mala información (Vaswani 2019).

En muchas de las respuestas de los Estados Unidos se destaca la responsabilidad individual de los miembros del público, y se exhorta a las personas a que no hagan circular ni transmitan información falsa y a que aprendan a diferenciar entre la información verdadera y la falsa. Alan Miller (2019), fundador del proyecto educativo sin fines de lucro News Literacy de EE.UU., explica: "Necesitamos un cambio de conciencia para contrarrestar esta niebla de confusión y desconfianza. En primer lugar, debemos entender - y asumir la responsabilidad de - nuestros papeles en el ecosistema de información del siglo XXI. La desinformación no puede propagarse de forma viral a menos que infectemos a otros con ella. Tenemos que ir más despacio antes de llegar a "compartir" o "retweetear" o "gustar", y preguntarnos si al hacerlo nos engañamos, desinformamos o hacemos daño". Pero sin regulación esta "desaceleración" es poco probable que ocurra. Los que difunden información a menudo tienen razones para hacerlo más allá de un simple descuido. La difusión de desinformación política o creencias no científicas como el movimiento anti-vacunas son sólo dos ejemplos.

Definiendo nuestros términos

Hay muchos tipos de desinformación y se han hecho varios intentos de proporcionar tipologías. Tandoc, Lim y Ling (2017) revisaron 34 artículos académicos publicados entre 2003 y 2017 y llegaron a una tipología que incluía: sátira, parodia, imágenes falsas, publicidad y relaciones públicas, que a veces se superpone con la propaganda. Para nuestros propósitos consideraremos, en el siguiente capítulo, la relación entre la propaganda y la desinformación y nos centraremos también en lo que Tandoc, Lim y Ling describen como "fabricación de noticias". Esto se hace a menudo con la intención de engañar, y las noticias falsas son a menudo difíciles de identificar ya que se presentan como una noticia tradicional con formato y convenciones similares.

Como señalan Tandoc, Lim y Ling (2017):

Al igual que en el caso de la parodia, una noticia fabricada con éxito, al menos desde la perspectiva del autor, es una noticia que se basa en memes o parcialidades preexistentes. Los entrelaza en una narrativa, a menudo con un sesgo político, que el lector acepta como legítima. El lector se enfrenta a más dificultades para verificarlo, ya que las noticias fabricadas también son publicadas por organizaciones o personas no relacionadas con las noticias, bajo un barniz de autenticidad al adherirse a los estilos y
y formatos periodísticos. Los artículos también pueden ser compartidos en los medios sociales y así ganar más legitimidad, ya que el individuo los recibe de personas en las que confía.

Los autores también observan que la "facticidad" es otra cuestión en la determinación de las noticias falsas, ya que la información falsa puede derivarse de algo que es verdadero o parcialmente verdadero, o basarse en ello: por ejemplo, el sitio web de derecha que da un titular falso a un artículo de un medio de comunicación de renombre. El público también importa porque, bajo ciertas condiciones, el público es más receptivo a las noticias falsas.

Otro conjunto de debates en torno al problema de las noticias falsas ha sido el reciente interés en la desinformación, que es la información falsa difundida deliberadamente para engañar. La palabra inglesa "disinformation" se parece a la palabra rusa "dezinformatsiya", derivada del título de un departamento de propaganda negra de la KGB1. En sus influentes documentos e informes Wardle dijo que el término "noticias falsas" es engañoso y en 2017 publicó su artículo "Noticias falsas, es complicado", que ahora es un estándar para la discusión del problema. En este documento, Wardle describe los tipos de desinformación como sátira y parodia, contenido engañoso, contenido impostor y contenido fabricado, conexión falsa, contexto falso y contenido manipulado (Wardle & Derakhshan 2017). Su documento con Hossain Derakhshan también incluía un análisis sobre quiénes son los diferentes actores y objetivos, tales como los Estados que apuntan a los Estados, los Estados que apuntan a los actores privados, las empresas que apuntan a los consumidores (Wardle & Derakhshan 2017).

Además, señalan que las intenciones de la persona que crea y/o amplía la información falsa son relevantes para la definición.

- La información falsa (misinformation) es cuando se comparte la información falsa, pero no se pretende hacer daño.

- La desinformación (disinformation) es cuando se comparte información falsa a sabiendas para causar daño.

- La mala información (mala-information) es cuando se comparte información genuina para causar daño, a menudo trasladando a la esfera pública información diseñada para permanecer en privado.

A veces la desinformación disfrazada de parodia puede extenderse en las conversaciones. En esta tesis nos centraremos en la mala/desinformación, con
incursiones ocasionales en la propaganda, en lugar de incluir la sátira y la parodia. Nuestro interés se centra en las noticias que creemos que están afectando el comportamiento social y contribuyendo al aumento de los demagogos, la violencia y las actitudes anti-científicas en todo el mundo. Esta disertación toca la psicología humana y los incentivos políticos y económicos de aquellos que difunden falsedades en línea, pero nuestra principal área de interés son las soluciones al problema. No usamos el término "fake news" por el uso problemático que hacen el presidente Donald Trump, el presidente brasileño Jair Bolsonaro y muchos otros. Después de 2016, se convirtió rápidamente en un término utilizado para menospreciar y socavar a los periodistas y los medios de comunicación.

¿Por qué centrarse en los medios de comunicación social?

Somos muy conscientes del hecho de que en muchas partes del mundo la Internet no ha alcanzado su plena penetración y que en África, por ejemplo, la radio, la televisión y la telefonía móvil siguen siendo los medios de comunicación más importantes (Equipo editorial 2016). Comprendemos además el poder de la radio en los Estados Unidos, en particular las preocupaciones acerca de los presentadores de radio sindicados como Rush Limbaugh y las noticias por cable (Della Vigna y Kaplan 2007; Hopkins y Ladd 2014; Rosenwald 2019; Nelson 2019; McGirr 2001). Sin embargo, en lugar de abordar el problema general de la información falsa, esta tesis se centra en las soluciones al problema de la información errónea/desinformación que circula en los medios de comunicación social. Adoptamos este enfoque en parte porque el período transcurrido desde 2016 ha puesto de relieve los daños causados por la circulación de esa información y la penetración de los medios sociales está aumentando en muchas partes del mundo. El problema puede empeorar.

Además, la viralidad del contenido significa que puede ser más influyente que un programa de radio local, con su alcance localizado y una presentación de la información menos sofisticada. El uso extensivo de vídeo en YouTube, TikTok y otras plataformas puede hacer que el mensaje sea más urgente, atractivo y persuasivo (Munger 2020). La naturaleza social de los medios de comunicación social significa que los amigos y la familia, así como las comunidades de todo el mundo, están reforzando los mensajes y ayudando a los grupos a formarse y desarrollarse (Munger 2020). A menudo estos hacen uso del anonimato y la encriptación (WhatsApp, 8chan, 4chan) y por lo tanto son más difíciles de supervisar. La futura evolución técnica de estas plataformas, que incluye el aumento de las falsificaciones, contribuirá aún más a la rápida difusión de contenidos confusos y engañosos. Más adelante profundizaremos en la investigación de las ciencias sociales sobre la confianza en los
medios, la persuasión y los intentos de comprender si la desinformación y la mala información en línea están cambiando las creencias y cómo lo hacen. Aquí observaremos que, aunque muchos de los estudios están enfocados de forma estrecha y a corto plazo y no proporcionan todas las respuestas, si damos un paso atrás y observamos tendencias sociales y políticas más amplias -por ejemplo, el auge del movimiento anti-vacunaciones- es difícil argumentar la Hipótesis Nula (que no está pasando nada). Desde la introducción de cantidades masivas de información falsa en el ecosistema en línea, los comportamientos han cambiado en todo el mundo. Esta es una razón suficiente para tratar de entender el fenómeno de la desinformación en línea y sus posibles soluciones. Ciertamente hay una escuela de pensamiento que dice que los efectos de los medios no son tan poderosos como se imaginan (Newton 1999; Schudson 2019; Karpf 2019) pero los recientes estudios sobre Fox News sugieren que hay efectos medibles en el consumo de noticias. Es poco probable que los investigadores puedan cuantificar con alguna precisión si la exposición a la desinformación o incitación en línea es lo suficientemente poderosa por sí misma como para causar cambios de comportamiento o si es la interacción de la desinformación en línea con otras formas de falsedad tiene efectos persuasivos (Jamieson 2018). Por ejemplo: los asesinatos en Birmania, India y Nigeria que siguieron a los mensajes incitadores difundidos en WhatsApp. ¿En qué medida se vieron influenciados por los mensajes incitadores, sus prejuicios previos o el discurso incendiario de los líderes políticos y religiosos locales? Nunca lo sabremos, y esta tesis no puede abordar todas las formas de desinformación de los medios de comunicación. Más bien, nos centraremos en el ámbito mediático más controvertido y de más rápido crecimiento de interés público: las plataformas de medios sociales.

¿Por qué esta tesis?

La contribución de esta tesis centra la argumentación sobre cuáles de las innumerables soluciones que se están probando tienen probabilidades de ser eficaces o qué combinación de medidas podría ser deseable. Creemos que es esencial profundizar en la forma en que se llegó a las soluciones y en los intereses y sesgos implícitos involucrados. Al proporcionar un marco para comprender los diferentes tipos de soluciones que existen y un relato sistemático de cada tipo, esta disertación constituirá la base de un juicio más claro sobre lo que podría funcionar y lo que no.

Estructura de la tesis

Esta disertación tiene tres partes. En esta introducción se expone el argumento principal y algunos conceptos clave, como lo que entendemos por desinformación,
por qué es importante y en qué creemos que la investigación puede contribuir al estudio del tema.

En el capítulo uno se examinan las diferentes concepciones de la confianza y cómo se ha investigado. La comprensión de la idea de la confianza y su compañera cercana, la credibilidad, es esencial para comprender las cuestiones de la desinformación. Ambas son parte de la cuestión más amplia de la persuasión, y los miedos a la persuasión subrayan la ansiedad sobre la desinformación. Las sociedades se preocupan de que la exposición repetida a un argumento (ya sea correcto o incorrecto) influya en el comportamiento y, de hecho, el aumento de los crímenes de odio, la elección de demagogos en algunas partes del mundo y la caída de las tasas de vacunación en muchos países sugieren que la desinformación en línea es, de hecho, un poderoso persuasor. Analizar el impacto de la información en la psique humana es extremadamente difícil, si no imposible, pero comprender lo que sabemos sobre la confianza en los medios de comunicación es parte del pensamiento sobre el impacto de la información. Esta disertación no es la última palabra sobre el tema - muchos han pasado sus carreras tratando de entender la persuasión y el impacto de los medios de comunicación. En cambio, ofreceremos una visión general de la literatura existente y de sus antecedentes y contexto históricos, señalaremos algunas de las lagunas de la investigación y, lo que es más importante, explicaremos cómo se relaciona la confianza con nuestra mayor comprensión de cómo abordar el problema de la desinformación en línea.

Algunos de los primeros debates sobre la confianza en los medios de comunicación tuvieron lugar en el período en torno a la Segunda Guerra Mundial, cuando los intelectuales de Europa y los EE.UU. (en particular los de la Escuela de Frankfurt) trataron de comprender cómo los ciudadanos podían ser susceptibles a la propaganda nazi (Jeffries 2016). En los EE.UU. había preocupación por la influencia de los demagogos, como el Padre Coughlin, que usaban la radio para hacer llegar sus mensajes. El miedo a los nuevos tipos de tecnología ha contribuido a menudo a los miedos a la desinformación (Tucher 2013) pero creemos que es un error considerar tales miedos como mero ludismo. Más bien, se estaban produciendo catástrofes políticas objetivas que iban a tener consecuencias mundiales, y preocuparse por ello y tratar de comprender el papel de la propaganda, incluida la forma en que se creaba, difundía y tenía influencia, era una respuesta necesaria a los tiempos que corrían.

El ascenso del fascismo y el comunismo en los años 20 y 30 proporciona un telón de fondo para el capítulo dos, en el que discutimos la alfabetización mediática y los intentos de Clyde Miller en los años 30 de enseñar a los escolares estadounidenses a
entender y resistir la propaganda. Miller fue un ex periodista que trabajó en el Columbia Teachers College durante 10 años. Durante ese tiempo, recaudó un millón de dólares del empresario de Boston Edward A. Filene para el Instituto de Análisis de Propaganda (IPA). La historia de Miller no ha sido contada completamente y proporciona importantes ideas para los debates actuales. Nuestra lectura de tres archivos diferentes de los documentos de Miller reveló información sobre la carrera del propio Miller y los ataques a los que fue sometido por el magnate de los medios de comunicación de derecha William Randolph Hearst, en cierto modo el Rupert Murdoch de principios del siglo XX. Además, los primeros esfuerzos de Miller en la alfabetización mediática sentaron las bases para gran parte del trabajo contemporáneo en ese campo (Hobbs y McGee 2014). Sus taxonomías de técnicas de propaganda y su trabajo de análisis de piezas de desinformación en los medios de comunicación anticipan muchas de las técnicas utilizadas hoy en día. Miller y su colega, Violet Edwards, trabajaron en estrecha colaboración con los profesores y le proporcionaron material para que lo utilizaran, así como envíos semanales para los escolares.

En los capítulos tres y cuatro se analizan los esfuerzos por fomentar la demanda y la confianza del público en el periodismo de calidad. En esos esfuerzos se trata de fomentar la confianza en el periodismo al entenderlo como una fuerza para decir la verdad (en el caso de la verificación de hechos) o al tratar de que los medios de comunicación sean pertinentes para el público (esfuerzos de participación de la comunidad). Sostenemos que después de 2016, las fundaciones y las plataformas llegaron a arreglos y financiaron esfuerzos que abordaron el lado de la demanda por una serie de razones. Facebook quería evitar la reglamentación y, además, creyendo en la importancia de la libre difusión de ideas e información en línea, pensó que ayudar a las audiencias a ser más educadas sería una solución adecuada. Examinamos los esfuerzos para combatir la desinformación en línea mediante el fomento de la confianza en el periodismo y la capacidad de las audiencias para distinguir la información buena de la mala. En los capítulos tercero y cuarto se examinan las soluciones propugnadas por los "demand siders", que incluyen a las comunidades de alfabetización mediática y de comprobación de hechos, y a los que esperan que la creación de una comunidad y el compromiso con el público ayuden a restablecer la confianza en el periodismo.

En capítulos posteriores se examinarán soluciones relacionadas más con el suministro, la transmisión y la reproducción de la desinformación en línea: el intento de utilizar algoritmos y el aprendizaje automático para bloquear y suprimir el contenido, y las posibilidades de reglamentación. En el capítulo cinco se examinan algunas de las empresas de nueva creación que utilizan la inteligencia artificial y el procesamiento del
lenguaje natural y se evalúa la probabilidad de que sean eficaces y escalables. Concluimos el capítulo cinco analizando algunas iniciativas que califican a los medios de comunicación y proponen normas para que las sigan los periodistas y sugieren que éstas pueden ser más rápidas de escalar que las soluciones tecnológicas de las pequeñas empresas de nueva creación.

Concluimos esta investigación con los capítulos seis, siete y ocho sobre los recursos legales. En el capítulo seis discutimos la eliminación de la responsabilidad de los intermediarios (como se ha hecho en Alemania) y vemos cómo podría hacerse en los Estados Unidos. En el capítulo siete esbozamos algunos de los problemas con la publicidad política y proponemos una provisión voluntaria de noticias políticas durante las elecciones. En el capítulo ocho explicamos por qué el mercado de las ideas es defectuoso y discutimos el uso de demandas por difamación, por parte de los periodistas, como una forma de combatir el discurso violento y dirigido en línea.

**Notas personales**

Debido a que mi familia ha vivido y ha sido refugiada de tres regímenes autoritarios que hicieron un uso extensivo de la propaganda, el tema de la desinformación política en línea es de gran interés personal para mí. Las experiencias de los años 30 tienen una resonancia particular al tratar de entender la economía política contemporánea de la desinformación. El surgimiento de demagogos de derecha a nivel mundial y su creación de mensajes para incitar al odio, e incluso a la violencia, es una reminiscencia de los años 30. También lo es el constante ataque a los migrantes y refugiados y a los grupos que se consideran "otros". Pero las semejanzas no son sólo entre la extrema derecha de hoy y la de los años 30. Los EE.UU. incluso han visto un regreso a la "persecución de rojos" de la que Miller fue víctima. En resumen, existen paralelismos entre los ataques actuales a las instituciones que buscan la verdad, como los periodistas, las universidades, el poder judicial e incluso el Congreso, y los ataques del decenio de 1930.

Este es un terreno que me es familiar ya que mi familia fue refugiada que tuvo que huir de varios países en la primera mitad del siglo XX. Mi abuelo ruso y su familia dejaron Bakú alrededor de la época de la Revolución Rusa de 1917. Mi familia militar republicana tuvo que dejar España en 1939. Mi padre y sus padres, todos ellos ciudadanos franceses, tuvieron que abandonar París en 1941 después de que mi abuelo fuera despedido por Gaston Gallimard de la editorial que mi abuelo fundó a los pocos días de las leyes antijudías impuestas por el gobierno de Vichy en 1940 (Reichman 2019). Mi familia experimentó lo que sucede cuando los dictadores llegan al poder. Una pregunta clave para muchos de los refugiados de los años 30 fue la
pregunta sobre el tiempo. Saber cuándo irse, o poder irse, era esencial. Y sin embargo, es la naturaleza humana asumir que las cosas no están mal o no se pondrán peor. Parte de un sólido sistema de alerta temprana es observar los medios de comunicación para tener una idea de hacia dónde se mueven las sociedades o qué ideas el gobierno está tratando de inculcar. Así fue como la radio en Ruanda incitó a la matanza de los tutsis, la propaganda de los nazis incitó al odio hacia los judíos, los romaníes y otros grupos. El Ministro de Propaganda de los nazis, Joseph Goebbels, conocía el poder de los medios de comunicación cuando puso una radio en cada hogar alemán (Meier 2018). Atacaron al periodismo independiente y hablaron del poder de la gran mentira.

No es sólo que la propaganda y la incitación sean desagradables. Creemos que hay una conexión entre la desinformación y el comportamiento humano. El aumento de la desinformación en línea hoy en día está afectando los eventos políticos en todo el mundo. Lo que la gente ve en línea se ha convertido en una amenaza a la democracia por muchas razones: Agrava los sentimientos de desconfianza en las instituciones, puede incitar a la gente a la violencia, promueve el odio a los inmigrantes y a las minorías y permite que circulen teorías de la conspiración. Por lo tanto, encontrar soluciones que funcionen es esencial para la salud física y la seguridad de los ciudadanos, así como para la salud de nuestra democracia.

**Conclusión**

No creemos que sea posible separar la confianza en los medios de comunicación y la susceptibilidad a la desinformación de las cuestiones más amplias sobre la confianza en el sistema político y las instituciones en general. Estamos de acuerdo en que los problemas de confianza en los medios de comunicación no pueden separarse de las preocupaciones sobre la desigualdad económica, la globalización y las grandes ansiedades políticas (Schudson 2019). Entendemos el argumento del activista político estadounidense Zack Exley de que el problema son las creencias subyacentes y no las plataformas de los medios de comunicación social en las que se manifiestan (conversaciones y mensajes del autor con Zack Exley 2019 y 2020).

Sin embargo, creemos que la desinformación en línea puede proporcionar un ingrediente tóxico (una cerilla incendiaria) y que los intentos de combatirlo son parte de un esfuerzo mayor para crear normas que defiendan importantes valores sociales como la verdad, la evidencia y el método científico. Así, mientras criticamos aspectos de la búsqueda de soluciones, pensamos que son críticos para la supervivencia de la sociedad.
El desafío es encontrar soluciones que funcionen, que no amenacen la libre expresión, y, sobre todo, que no puedan ser utilizadas por grupos de interés como los gigantes de la tecnología o los políticos que no actúan de buena fe. Será difícil encontrar leyes que no puedan ser objeto de abuso por los que están en el poder. Pero hay una falta de evidencia que demuestre que arreglos como la verificación de hechos, la alfabetización de los medios, el compromiso de la comunidad y los algoritmos de ajuste son suficientes. Además, las empresas de tecnología han demostrado que los códigos de conducta voluntarios no son efectivos. Es la amenaza de la regulación lo que los impulsa a actuar. Por esta razón, argumentamos en las partes finales de esta disertación que las regulaciones - muy probablemente originadas en Europa - serán una parte esencial para solucionar el problema.

La desinformación en línea es difícil de arreglar, por lo que es aún más importante tener una mejor comprensión de los supuestos que subyacen a cada solución y un marco general en el que situar las medidas y arreglos propuestos que se están aplicando a nivel mundial. Esperamos que la taxonomía y el análisis de este trabajo doctoral profundicen la comprensión de las opciones disponibles para afrontar con éxito el problema.
PART I. LITERARY REVIEW

Chapter 1

Trust, Journalism Authority and Mis/Disinformation

Introduction

Questions of media trust and the credibility of journalism are central to much of the discussion about how to solve the online mis/disinformation problem. Many of the proposed solutions start from the premise that audiences need to trust the information received from journalism instead of the false information found online. Fixes such as fact-checking and community engagement efforts are based on the idea that changing journalism practice will boost trust in, and support for, journalism. Media literacy education is similarly based on helping audiences distinguish between information that can be trusted and that which can’t. Some of the tech fixes, such as using natural language processing to identify disinformation, also assume that when audiences are given prompts to think more analytically they will know what to trust, and that this will change audiences’ behavior, such as becoming more careful about what online information they forward (Pennycook and Rand 2017). Thus, understanding trust in media and how audiences decide what to trust is essential to understanding how different groups address the problem of mis/disinformation online.

There is a policy implication too: As mentioned earlier, many of the measures taken to combat mis/disinformation after the 2016 elections were done in haste and not
informed by the academic scholarship on questions of trust, persuasion, and the media. This is likely to change as more research is done and as the groups involved develop more data. But without a deep understanding of the elements that comprise trust and which factors affect it, all attempts at building trust in the “right” kind of information will fail.

The need for more research is indeed urgent. Our review of the literature on trust and credibility finds that it’s fragmented and inconclusive. Part of the problem is that trust varies enormously from country to country (Van Aelst 2017) and part of the problem comes from the fact that the methodologies used for measuring trust are inconsistent. Even so, it’s clear from the literature that the rise of the internet and the social media platforms upended how audiences receive and share information. It’s also clear that this upending affected trust in media. How to address the problem and how to build trust are far less clear. Above all, to what extent can trust be created or rebuilt after it’s been lost?

Why trust matters

The credibility of journalism and questions about what to trust have been around since the beginning of the circulation of the printed word (Pettegree 2014; Briggs and Burke 2009). The availability of reliable information is generally understood to be central to a functioning democracy, in part because both voters and governments need reliable information in order to make sound decisions about which policies to support, and, in the case of citizens, which policies and politicians to vote for (Besley and Burgess 2001; Prat and Strömberg 2011). Economists use the term “principal-agent” to explain that societies and people need to delegate both decision-making and implementation to others and need to know whether the designated agent is honest and capable of doing a good job (Hamilton 2016). Journalists can provide this information and so assist with the oversight role, but the information provided has to be reliable, credible, and trustworthy (Kohring and Matthes 2007).

Numerous studies show a decline in trust in media, as well as institutions and experts, over the past 30 years (Twenge, Campbell, and Carter 2014; Hanitzch et al. 2018), though it must be noted that trust levels, and what people trust, differ in each country (Moehler and Singh 2001; Ravetti, Quan, and Jin, 2017; Hanitzch 2013). For example, the 2019 Reuters Digital News Report (Nielsen et al. 2019) found that worries about online disinformation vary greatly between nations; 85 percent of Brazilians are
concerned about fake news whereas only 31 percent of people in the Netherlands consider it a problem².

Indeed, it’s almost impossible to generalize across societies. In the United States, for example, trust in the media is higher among liberals than among conservatives (Brants 2003; Knight Foundation 2018). The opposite is true in Britain, in part because many liberal readers believe that the right-wing press promotes pro-Brexit views. Conversely, a different paper found that in Israel, people trusted media accounts more when they aligned with their personal experiences (Livio and Cohen 2016). Trust levels seem to depend largely on prior beliefs and biases, although the degree to which this is true is a matter of debate. A study of trust in African media found high levels of trust in state-owned media in relatively undemocratic, illiberal societies where citizens are less likely to think critically about the news they consume (Moehler and Singh 2011). In China, commercial, privately owned media has recently become a viable competitor of party-run print media, but party mouthpieces are still viewed as more credible in the growing world of online journalism (Xie and Zhao 2014; Zhang et al. 2018).

Some research has found that distrust in the media hinders the functioning of democracy (Jones 2004; Ladd 2011), while distrust in government leads to paralysis as elected officials lose their mandate to govern or pursue redistributive policies (Hetherington 2005). Lack of trust in traditional media and in politicians contributed to surprise election victories in 2016 of President Donald Trump, the UK Brexit vote, and 2018 gains in Italy for the Five Star movement and 2017 gains for the far right in Austria (Schultheis 2017; Schultheis 2018).

A short history of trust

Questions about what information to trust go back hundreds of years (Pettegree 2014). As printing expanded in the period between 1450 and 1530, so did the number of publications and availability of ideas, and audiences questioned which sources to trust (Pettegree 2014; Briggs and Burke 2009). As they do today, audiences often turned to first-hand accounts, written by people they knew (Pettegree 2014, p. 2).

Later, technological innovation made more information available and raised questions about credibility as audiences feared that new forms of technology could be manipulated. A debate arose in the 19th century about the ethics of manipulation of photographs. In the US, some argued that embellishing photographs simply made

---

² This section was written before the Covid-19 pandemic. During the pandemic, audiences for quality media rose significantly.

45
them seem more realistic and that faking in the service of providing a larger truth was acceptable (Tucher 2017). This rationale foreshadows the arguments made today that even if Donald Trump is lying about the facts, he is speaking a larger truth that his supporters believe in. However, the defense of “embellishment,” or lying, in order to convey a larger truth may lead to a longer-term undermining of trust and credibility.

While many journalists of the 19th century tried to provide credible information, the media was partisan and false stories appeared regularly. The most infamous include the “Great Moon Hoax” of 1835, in which the New York Sun printed a series of articles and fake pictures of little people on the moon, claiming the images had been picked up by a new powerful telescope (Young 2017). Andie Tucher argues that the moon hoax was understood to be fake and the pleasure lay in the invitation to readers to make up their own minds (Tucher 1994). Another reason for the proliferation of completely false stories was that they helped sell newspapers (Tucher 1994, p. 51). The Moon Hoax helped the circulation of the recently-established Sun soar to almost 20,000 (Young 2017, 12-15). The Morning Journal, owned by press baron William Randolph Hearst, is often credited with generating support for the Spanish-American War of 1898 through its sensational reporting, which included misleading articles and drawings alleging that Spanish officials strip-searched Cuban women on an American ship (Schudson 1981, p. 62). This showed that hoaxes and fraud could be used to further political aims as well as boost circulation (Young, 2017; Tucher, 2013; Carey 2017; Wilkerson 1967). The role of Hearst is just one example of how, by the 18th and 19th centuries, the idea of using media for political purposes was well established.3

This political use of journalism became further entrenched in the early 20th century when the era of mass propaganda through radios and newspapers was born. During World War I, information was used to mobilize people for the war effort and to unify citizens against a common enemy (Burke 2009). Propaganda on a mass scale was spread through an enormous apparatus of state-controlled media and propaganda in the Soviet Union, in Nazi Germany, and later in Communist China (Tong and Sparks 2007; Simons et al. 2017). The U.S. had a massive war office that was also influential in transmitting information (Horten 2002; Knightley 2004).

**The rise of professionalism**

The early examples of irresponsible journalism described above did not give rise to academic research on media trust and credibility but certainly informed discussion about the role of the profession. The significance of political and commercial

---

3 Hearst, of course, continued his career into the 20th century, with his newspapers becoming more right-wing as he aged.
interference was well known and famously portrayed by Upton Sinclair in The Brass Check, published in 1919. In the 1920s, Walter Lippmann argued that journalists should be professional enough to consult trusted experts in government, universities, and think tanks in order to provide the public with necessary information that the journalists didn’t have (Schudson 1981). Scholar Michael Schudson views the professionalization of journalism partly as a response to the partisan reporting of the 19th and early 20th centuries. Professionalization was a way of freeing the journalist from the political and commercial pressures that surrounded them (Schudson 1981; Schudson 2001). Others have argued that the concept of “professionalism” did not take root in newsrooms around the world until later (Hallin and Mancini 2004).

Trust and credibility are often linked and are considered essential to the financial survival of the media (Heflin 2015; Örnebring 2009). Professionalism in journalism—including accuracy and objectivity—helped audiences believe in the credibility of journalism and to trust the media (Schudson 1981). Professionalization of journalism and objective reporting became both a response to propaganda and a viable business model. The idea of professional journalism was built upon the enlightenment model of trust along with verification and replication. Journalists’ responsibility to provide accurate and objective information to help inform the citizenry was at the core of the profession and expected to engender trust in media.

The practice of “objectivity”—in US newsrooms in the 20th century, among other places—was also a way of protecting journalists from publishers and advertisers or other funders who wanted to get involved in editorial decisions by telling journalists what to cover or what to say about a subject (Hume in Schiffrin 2017). Since journalistic objectivity helped make media more credible to the public, sensationalistic reporting was seen as hurting media credibility and contributing to distrust in the media.

**Academic research on media trust began after World War II**

Academic research on media trust began as a response to World War II, as sociologists and political scientists tried to understand how propaganda had apparently been so effective in galvanizing public support. US intellectuals and others became preoccupied with the dangerous effects of propaganda, which they worried had been instrumental in galvanizing citizen support in Germany during the war and in spreading disinformation. One response was to begin promoting media literacy efforts as a way of educating people so they would not believe false information (Turner, 2013). Funded by Edward A. Filene of department store fame, the Institute for Propaganda Analysis published articles about how to spot different kinds of
misinformation and piloted a media literacy project in the Springfield, Mass., school system (Johnson 2006).

Trust in media and the effects of mass media began to be studied in a systematic way in the 1950s, but, as with studies of the media’s impact on foreign policy, the findings are still inconclusive (Hanitzch et al. 2018). Worries about whether news reports could be trusted continued to reemerge every couple of decades, often as a response to more information flows due to new technology. In the 1960s, theories about “video malaise” took hold, which asserted that increasing dependence upon television as a source of news for the American people and the cynicism and negative focus of the news shown on television led to decreasing levels of trust in government and its efficacy (Robinson 1976). In defense of the media, Harvard University political scientist Pippa Norris has argued that there is a “virtuous circle” in which media exposure can create more trust in society (Norris, 2000). Later on, Avery tried to reconcile the two different views and found that trust in the media depends on prior levels of political trust and that exposure to newspapers creates more trust while exposure to television coverage creates less (Avery 2009).

After the birth of the internet, when media outlets began to run native advertising and sponsored content in the early 21st century, this low-quality content and the blurring of the walls between advertising and information hurt media credibility and may have contributed to declines in media trust (Schiffrin et al. 2017; Amazeen and Muddiman 2017; Wojdynski and Evans 2016). The rise of the internet and the collapse of traditional concepts of authority will be discussed further below.

**Practitioners: build trust though media literacy and journalism practice**

Worries about the effects on society of lack of trust have led to repeated attempts to build trust in the media. However, the uncertainty as to the causes of trust and distrust in the media has affected the kinds of efforts made. Although it’s not been established how much trust in media is related to journalism quality or journalism practice, a number of programs trying to build trust in media have focused on either journalism practice or audience education. One example is the civic journalism or public journalism movement of the 1990s, which was aimed at strengthening democracy by promoting audience participation and community involvement, often by soliciting story ideas and priorities from readers (Voakes 2004). Similar efforts are taking place today in dozens of countries (Schiffrin et al. 2017).
The research on media trust and credibility: from post-World War II to the internet

As noted in the introduction, the literature on media trust and credibility spans various fields including political science, psychology, sociology and journalism studies. It is fascinating to see the wide range of perspectives from multitudes of scholars. However, there is a frustrating lack of meta studies and an inconsistency of research methods, making it difficult to compare different studies. While researchers have a general idea about the elements of trust, they have been unable to quantify the importance of each element. For example, it’s not clear how much trust in something or someone derives from prior beliefs or is affected by journalism practice. This uncertainty naturally affects the efforts by practitioners trying to build trust. Nor is there agreement as to how to define trust. For the purposes of this paper we will define trust as a summary judgement about how one assesses the reliability of information being provided. Trust is a relationship in which one person concludes, “I am willing to believe what is said.” Trust is relational, as it involves trusting someone else, and it’s usually conditional and graded—as it’s possible to partially but not wholly trust (Levi and Stoker 2000).

Trust is not the same as credibility, although the terms are often used interchangeably (Kohring and Matthes 2007). Certainly, one can exist without the other. For example, Fox News viewers may believe reports that climate change does not exist while few scientists would find that credible. In fact, by trying to verify or provide credible information, the network could lose its audience’s trust.

1. Measuring trust and understanding what people trust.

Overall, the academic literature takes empirical and conceptual approaches focusing on three areas: defining trust, measuring trust, and understanding what people trust (Kohring & Matthes 2007; Levi and Stoker 2000; Gaziano and McGrath 1986; Meyer 1988). However, a number of scholars have noted a lack of agreed-upon standards and definitions as to what constitutes trust and how it differs from credibility. Indeed, Kohring and Matthes surveyed efforts from the last 50 years and noted the absence of a standardized measure of trust.

In his pioneering research on media trust, Carl I. Hovland and colleagues at the Yale Communication Research Program studied what made audiences find an idea or piece of information persuasive in their 1959 book Communication and Persuasion. They assumed that persuasion was an indicator of trust. Three of the categories used in their research were later used by others trying to understand media trust (Fisher 2018):
- Source credibility, i.e., *who* said something. Hovland and colleagues gave audiences information from sources with varying degrees of credibility (e.g., scientists versus Soviet newspaper *Pravda*) to see whether this had an effect on audience opinions of those sources. Hovland also notes the difference between “expertness” and “trustworthiness.” However, the causality is unclear: It’s not evident if people trust the dentist because he is an expert or if his expertise builds their trust.

- Message content. *What* is said. For example, in trying to understand what kind of appeals are persuasive, Hovland et al. found that invoking fearful feelings got audiences to pay attention, but left them in a state of “emotional tension” which interfered with the “overall effectiveness of a persuasive communication” (p. 271).

- Audience characteristics. This category included group norms and the desire of people to stay in a group and therefore be unpersuaded by ideas that contradict those held by the group despite individual differences in “persuadability.” Hovland explains that highly intelligent people may be more able to absorb new information but also more skeptical. People with low self-esteem who “manifested social inadequacy...showed the greatest opinion change." Eveland and Shah found that perceptions of media bias are also related to conversations with like-minded people and that the effect of such conversations is stronger among Republicans than among Democrats (Eveland and Shah 2003).

Source, message content, and audience characteristics are three of the main elements still used today. Kohring and Matthes later tried to break this down further and proposed what they describe as the “first validated scale of trust in news media in communication research”. Their scale includes trust in the selectivity of topics, trust in the selectivity of facts, trust in the accuracy of depictions, and trust in journalistic assessment (Kohring and Matthes 2007). They found that audiences prioritize selectivity of topics when deciding which outlet to trust.

In 1986, Gaziano and McGrath sought to understand the “multidimensionality” of media credibility, including differences in perception of local and national news, the kinds of media that were perceived to have credibility and which subjects inspired more audience credibility.

## 2. Understanding what people trust and what makes them trust

More recently, researchers have also sought to understand what audiences consider to be trustworthy and the characteristics of what people consider to be trustworthy, e.g., information provided by someone familiar to the reader/viewer (Turcotte et al. 2015; Flanagin and Metzger 2007). There is also debate as to how prior beliefs or
experiences affect audience trust in the media/belief in its credibility or whether journalism practice makes a difference (Livio and Cohen 2016).

Many scholars tried to understand the correlations between trust and audience characteristics or news consumption patterns, but while our literature review yielded a broad spectrum of factors affecting trust, the studies looking at correlations were at times contradictory and inconclusive.

For example, some studies found that people who read newspapers regularly and trust political institutions tend to trust the media (Moy and Scheufele 2000; Tsfati and Ariely 2014; Lee 2010). Others did not find a connection between newspaper consumption and belief that the media is credible (Kioussis 2001). Ardèvol-Abreu and de Zúñiga found that perceptions of media bias dampened use of media (Ardèvol-Abreu and de Zúñiga). There may be a positive linkage between news media use and political trust, but also that for some types of media consumption this relationship weakens across time (Strömbäck et al. 2015). However, a recent study found that audiences with low levels of media trust tend to turn to non-traditional forms of media such as social media blogs and “digital-born providers.” This is particularly true in Northern Europe (Fletcher and Park 2017).

They may also turn to the tabloids, with one study finding that the tabloids are trusted by German audiences who otherwise distrust the media because they view it as being part of a detached elite that neglects the citizens’ interests. However, people with “centrist and anti-outgroup attitudes” have higher trust in tabloid or commercial media. Tabloids that portray themselves as “taboo breakers” and emphasize a “divide between a homogenous people and the “bad elite” portray themselves as the voice of the ordinary people and thus can attract the trust of audiences who reject other media. The authors argue that the determining factor is not whether audience members are on the left or right but whether they subscribe to “anti-elites media populism” (Van Aelst et al. 2014, p. 487).

The relationship between trust in government and media exposure is another key area for researchers. One study of European audiences first looked at how coverage of news and current affairs had evolved since the late 1970s and found that the amount of space devoted to news and current affairs “has risen significantly since the late 1970s” (Van Aelst 2017, p. 100), partly thanks to public broadcasting. The authors then examined the relationship between media use and satisfaction with government and found that “[c]ontrary to what media malaise theories would predict, media use has a consistent and significant positive effect on political support in all nine European democracies in our sample” (Van Aelst 2017, p. 102). Media use mostly
highly correlated with high levels of satisfaction in the Netherlands and Denmark. The relationship was weaker in the UK, Belgium and Ireland. This study covered the period of 1983 to 2014 so does not touch on the period near the 2016 elections. It’s also unclear what kind of media audiences were exposed to, but the authors concluded that the medium, message, and receiver all played a role in levels of trust (Van Aelst 2017).

One study examined the government’s role in affecting audience’s trust. Tsfati and Ariely (2014) studied trust in media across 44 countries, studying state ownership, among other factors, concluding that state ownership of television positively affected media trust in democratic societies, but not in non-democratic societies (Tsfati and Ariely 2014). As noted earlier, regional context is extremely important. The authors of one study of trust in Brazil note that while trust in the media is high there, Brazil’s history has given people reason not to trust the media. For example, the high concentration of media ownership was cited by some focus group participants as a reason not to trust the media. Even so, usage of media and spending on media are both relatively high (Milhorance and Singer 2018).

Understanding the relationship between trust in government and trust in media is important if we are to find ways to rebuild trust. But there is ongoing debate as to what has to do with innate characteristics of audiences. In the studies of source credibility (Gaziano and McGrath 1986) some seek to understand whether feeling strongly about a subject, or being well informed about it, affects trust in media coverage of that subject (Stamm and Dube 1994). One study looked at whether being religious correlated with trusting the media and found that it was not correlated with trust in newspapers and magazines, but that religiosity was correlated with lack of trust in online news (Golan and Day 2010). It has also been argued that prior beliefs matter more than journalism practice. David Jones views the problem as one of journalism practice in that the internet is awash with opinions, so audiences don’t know whom to trust (Jones 2004).

**Trust in the internet age**

The advent of the internet and consumption of information online gave rise to more scholarship on media trust and how it may have changed. Matt Carlson discusses the elements that come together to give authority to the journalistic voice. He argues that there is a contingent relationship which “gives non-coercive authority the right to be listened to” and its many elements include group identity and the belief system that journalists use to demarcate who “may be included as a legitimate knowledge producer,” identifying expert sources for the reader to believe. How information is
presented is another means by which journalists gain authority, e.g., conventions of structure and narrative and the signaling conveyed by where stories are placed on the page—for example, positioning an important story “above the fold,” thus making it more likely to be read first. The narratives that journalists tell themselves and the “meta-discourse” of how journalists are portrayed in society, in books and movies, as well as when they appear on television as talking heads, are also part of journalists’ mantle of authority.

Perhaps most complicated is the relationship with audience. Karlsson et al makes a distinction between “authoritative” and “persuasive” and notes that “credibility” is not synonymous with “authority,” but that credibility is often tied to local news and actors while authority is a broader concept (Karlsson, Clerwall & Nord 2017, p. 162). Karlsson et al. note that the participation of audiences, a hallmark of journalism in the digital age, necessarily undermines the traditional authority of the journalist. Kristen Heflin (2015) explains that journalism is built on expertise, research, and verification and that this is naturally at odds with the idea of mass participation and deliberation that is the hallmark of the internet (Heflin 2015). Journalism practice has also changed as a result of the internet. Online, journalists may cut corners and make less effort to verify information (Nygren and Widholm 2018). However, in Denmark, media outlets explaining their recent plagiarism scandals focused on a few “bad apples” and not the structural changes and time pressures that make such problems more likely (Blach-Ørsten et al. 2018).

Why it’s difficult to create trust online is explained by Katherine Grosser (2016). She says that trust in offline journalism was based on six factors, three of which are self-explanatory (research, selection, and presentation) and three that are less so: proofing (verification such as fact-checking); revision (including the editing process); and coordination (including which editorial desk the piece is vetted by, which section it appears in, and how editorial departments interface).

These traditional factors can be undermined by new elements introduced by online journalism. Grosser defines these new developments as: transparency, reputation transfer (i.e., when well-regarded offline outlets are still trusted online), user generated content, rating cues such as likes and sharing, multimedia, a new definition of “current” due to the constant updating of news, and interactivity, such as comments made by readers.

According to Grosser, online news has yielded greater diversity of information, but the emphasis on being current can give rise to the perception that news reports are incomplete. Comments can undermine trust, particularly when they are negative.
Reputation comes into question when professional journalism that lacks an offline “mother” brand is seen as less trustworthy than the digital offshoots of offline brands. Grosser concludes that “online, the trust mechanism is more complex … the antecedents of trustworthiness are by and large negatively impacted by seven online developments. Consequently, trust in online journalism is more difficult to create than trust in offline journalism” (Grosser 2016; Hanitzch 2018).

**Gauging trust in the internet age**

In the absence of the traditional signals of authority, how do audiences gauge trustworthiness? A 2003 study by JD Greer found that they consider whether a site belongs to a person or a well-known outlet, but not whether the advertisements represent reputable organizations (Greer 2003). Similarly, a 2007 study of user behavior defined two key elements of credibility as being “site” credibility and “sponsor” credibility, and found that respondents trust news sites more than personal sites (Flanagin and Metzger 2007). Lack of transparency (Milhorance and Singer 2018) and use of native advertising are said by consumers to make them less trusting of the media (Amazeen and Muddiman 2018).

Sundar further found that because traditional clues of credibility (bylines, trusted brands) no longer prevail and it’s often not clear on aggregation sites where information originated, online articles with direct quotes from named sources were viewed as more credible than those without (Sundar 1998). Sundar proposes a model of how audiences determine credibility: Certain "affordances," i.e., structural features of the site, lead viewers to employ certain heuristics (mental shortcuts based on preconceived ideas), upon which they make quality judgments and ultimately credibility judgments (Sundar 2008).

**Trust in information from friends and first-hand experience**

As has always been true, audiences trust information that is familiar and/or comes from friends. Coverage of something that people have experienced may also make them more likely to trust media reports (Livio and Cohen 2016). The authors conclude that “perceived correspondence between direct personal experience and news reports was the strongest predictor of trust in journalists when controlling for all other factors…these findings suggest that declining levels of trust in journalists may be associated with actual evaluations of the quality of media performance by individuals thus refocusing the question of trust on journalistic practice rather than on audience attributes.”
A much-publicized paper by Pennycook and Rand found that fake news headlines that were familiar were perceived as substantially more accurate even when they were clearly implausible or contradicted the respondents’ beliefs. Warning labels about the headlines being incorrect had no effect on perceptions of credibility (Pennycook and Rand 2017) or even caused people to share the information more as readers assumed that a Facebook warning meant the story was true (Levin 2017). Pennycook and Rand received more attention in 2018 when they released findings suggesting that motivated reasoning was not the reason people believed “fake news” but rather belief was related to lack of analytical thinking. Prompts that nudged audience members towards cognitive reflection resulted in more accurate understanding of headlines (Pennycook and Rand 2018).

Another survey found that people trusted news more if it was recommended by a Facebook friend whom they viewed as an “opinion leader.” Based on a survey using Facebook data of graduate students about non-partisan news issues, the authors found that recommendations from Facebook friends/opinion leaders caused respondents to trust both the news article and the outlet it came from more and caused respondents to say they would be more likely to read an outlet in the future (Turcotte et al. 2015).

A paper written with Facebook employee Solomon Messing explored whether online social endorsements prompt audiences to view content from places/sources they might normally avoid. The authors argue that since audiences can choose news reports that are circulated by people or select from “most popular” lists, the range of what audiences can choose to see has broadened (Messing & Westwood 2014, p. 1044).

They expect that online, people will be more willing to share controversial items or items that “violate social norms” (Messing & Westwood 2014, p. 1045) “because of the lower levels of social presence in computer-mediated communication compared to in-person discussion” (ibid). In a randomized controlled experiment, the authors recruited 739 subjects from Mechanical Turk to see selection rates of Democrats and Republicans of articles from Fox and MSNBC and found that social endorsements caused higher rates of selection, even for content from ideologically opposed sources. The paper concludes on an optimistic note that social endorsements may help ensure that people see a range of opinions and escape their echo chambers. “Social media may not be a panacea for democracy’s ills, but their technological affordances are a spot of hope in an otherwise dark media landscape” (Messing and Westwood 2014, p. 1057).
After the election of Donald Trump and the Brexit vote in the UK, attention to media trust, disinformation and politics increased substantially (Bennett and Livingstone). As a sign of general interest in the subject of media trust, academic studies were covered extensively in mainstream media (Pennycook and Rand 2019c). Lack of trust in media was understood in the context of social and political polarization and as part of anger at politicians and institutions especially in the aftermath of the 2008 financial crisis. Greece and Spain, which were hard hit by the financial crisis are two countries where trust in the media was affected by the crisis (Köhler and Otto 2018). A 2016 regression analysis of European survey data found that consumption of quality media and public broadcasting strengthens trust and so concluded that all forms of media benefit from high public viewership of quality public broadcasters. Promoting use of media, especially in the young, is of benefit to society, the authors concluded (Schranz et al. 2018). Academics continued to study who consumes and circulates false news and found that in the US, political conservatives are the main generators, disseminators, and consumers of “fake news” (Benkler, Faris and Roberts 2018). A second study found that support of Trump was the best predictor of media distrust in the US (Mourão et al. 2018).

One study found that on Twitter, 80 percent of fake news sources were shared by just 0.1 percent of individuals exposed to fake news and that these people were overwhelmingly older, conservative and politically engaged (Grinberg et al. 2019; Guess, Nagler and Tucker 2019; Loos and Nijenhuis 2020; Lazer et al. 2017). Young people—at least those studied in Germany—seem able to identify misinformation (Heuer and Breiter 2018).

**Ideas for further research: the gaps are urgent**

As questions of media trust are relevant to the functioning of society and democracy, it is essential that scholars find some standardized ways of measuring trust and of conducting research on subjects such as whether trust can be transferred from one institution or outlet to another and the consequences of a lack of trust. Further, more context needs to be provided. Is it, in fact, ever possible to have trust in media in societies where citizens don’t trust government or other institutions? Do we know more about whether trust is correlated with aggregate education levels or individual ones? Is it possible to generalize across countries? What effect does economic anxiety have on trust in institutions and government? In a piece that appeared in *Columbia Journalism Review* in winter 2019, Michael Schudson argued that it’s impossible to separate trust in media from other institutions and that, in any case, journalists should not expect to be liked (Schudson 2019).
Given the 2017 and 2018 revelations about micro-targeting of political advertisements, theories about media trust and persuasion may need to be revised as more data becomes available. Questions that need to be researched include whether micro-targeting of political messaging is persuasive and whether it affects political attitudes, beliefs and voting patterns. It may turn out that a barrage of micro-targeted political messaging has more of an impact on a receptive audience than news that was “liked” by a trusted friend or celebrity. Unfortunately, Facebook and the other platforms have refused to share the data with researchers that could answer lingering questions about the effects of political advertising and micro-targeting in the 2016 elections (Alba 2019).

So, too, given the deliberate attempts by Trump and like-minded politicians in other countries to discredit the media, we have to understand better the processes by which trust is destroyed. Still needed are antidotes to these strategies. Can we learn something from the process of trust destruction about how to rebuild it? How strong are the hysteresis effects, and what can be done to reduce them? The lack of understanding about how to rebuild trust has contributed to the yawning chasm between academic scholarship and theory and what media groups and foundations are actually doing to address problems of trust. As media outlets around the world—often funded by foundations—make time-consuming and expensive attempts to build relationships and credibility with readers, not knowing whether they will have an effect is disheartening. Researchers can help bridge the gap between theory and practice and study these efforts to see if they have an impact.

Scholars looking at recent efforts to rebuild media trust would do well to look at the methodology of a couple of recent studies by economists who ran multiple regressions in order to parse news bias and voting patterns. The work of economists has established some of the causal relationships and direction missing from the media trust literature. Taking into account the position of Fox on the channel dial, economists were able to show the increase in support for Republicans was caused by a rise in viewership of Fox News (Martin and Yurukoglu 2017) and takeovers of local news stations by Sinclair Broadcasting led to a move right in terms of news coverage as well as a decline in local news, with potential for increased polarization (Martin & McCrain 2018).

When trying to understand the impact of practitioner efforts to build media trust, scholars can look at the public journalism programs from the 1990s. They had a strong theoretical framework (Rosen 1999; Örnebring 2009; Heflin 2015; Voakes 2004) and there is a well-developed literature on the effects (or lack thereof) of many
of these projects. Nichols et al. (2006) did a meta-analysis of hundreds of projects and ran regression analyses to understand which story frames, project features, and organizational factors affected which audience responses (Nichols et al. 2006). The authors found that some of the projects analyzed succeeded in promoting civic participation and were able to find correlations between the existence of public journalism projects and increased civic understanding or even participation (Ardèvol-Abreu and Hooker 2017). While the body of literature on public journalism is also inconclusive in some ways, better meta-studies based on this literature could still be useful for those seeking to understand better the determinants of media trust.

Practitioner efforts to address the lack of media trust are predicated on rationality. If given the right signals, audiences will process correctly the information transmitted. If given evidence that the media is trustworthy, it will be trusted more. The demand side of the literature says that that perspective is too simplistic. What people trust is a matter of human psychology. The practitioners of mistrust—Goebbels, Bannon and Trump—figured out pragmatically how to make people trust them and mistrust the standard media. They have grasped that individuals are not rational, and their insights into human psychology have enabled them to manipulate trust.

People who believe in the Enlightenment, in reason based on verifiable evidence, know Fox/Trump are not trustworthy but, to the dismay of the people who believe in Enlightenment traditions, Fox/Trump are trusted, and these media, with the support of the practitioners of mistrust—have succeed in discrediting the conventional media. Unfortunately, research so far has not shown how to reverse the process: it’s not fully known how the conventional media can, for example, show Fox and Trump as the untrustworthy sources of information that they are and re-establish trust in legacy media. Part of this is a lack of deep understanding as to how audience trust relates to perceptions of overall bias and media capture, and how changes in journalism practice can persuade audiences that are suspicious of bias and capture. Without an understanding of these broader questions, attempts at solving the problem of disinformation online—through the prism of building trust—will be unlikely to scale.

There are some places in the world where trust in media is well established and some countries where exposure to news correlates with trust in government and journalism as a whole (Elvestad et al. 2018). This was true in the Netherlands and Denmark in the period 1983-2014, but there was a weaker connection in Belgium, the UK, and Ireland (Van Aelst 2017). But, as in the field of economic development, while it’s understood that good institutions and good governance help societies grow and develop, it’s not clear how to build those institutions and good governance. The
question of trust is similar: It’s not clear how to build trust or how to restore it after it’s gone.
PART II. DEMAND-SIDE SOLUTIONS

Introduction

How media literacy, community engagement, and fact-checking efforts relate to trust

Scholars in different disciplines have put much effort into trying to understand the characteristics of trust in media. Many of them see susceptibility to mis/disinformation online resulting partly from a trust deficit, so efforts to build trust in information provided by journalists are central to efforts to combat mis/disinformation online (Wenzel 2019; Ferrucci 2017). In the post-2016 period, some analysts expanded earlier ideas about improving journalism quality in the hope that this would revive trust in journalism. They thought that trust in quality journalism would help audiences become more skeptical of mis/disinformation and more engaged in civil society, thus leading to a renewal of democracy (Ferrucci 2017).

In the next three chapters, we discuss how media literacy, community engagement, and fact-checking came to be seen as solutions to the problem of online mis/disinformation and how these ideas were interpreted globally. We discuss the post-2016 growth in funding for such efforts. We draw on contemporary scholarship as well as our own research on the donor-funded, nonprofit organizations that carry out such projects. We conclude by looking at the academic research on the effectiveness of fact-checking as well as at public journalism and community engagement efforts. In our examination of the academic research we try to assess whether these efforts are an effective response to the problems of mis/disinformation online. We note that the very premise—that these efforts would build trust—has yet to be studied in any systematic way. This lack of clarity has naturally affected attempts to build trust.
What ties these three efforts (media literacy, community engagement, and fact-checking) together in our taxonomy is that all three are focused on audience demand for accurate information. Media literacy attempts to build discernment skills among audiences so that they can identify which sources to trust. Efforts by journalists to build trust seek to bolster engagement with reliable, relevant material. Fact-checking provides a verification mechanism, establishing what is objectively correct or even “true.”

For journalists, the role of journalism is central and efforts at building trust and thus renewing civic engagement begin with a long, hard look at their own practices. Because journalists believe that their own practices affect audience trust, a great deal of effort has gone into initiatives to improve journalistic practice and thus the quality of journalism. Their hope is that improving the quality of journalism and providing an alternative to online disinformation will lead audiences to turn away from false information, instead trusting credible information produced by credible sources (American Press Institute 2016). This, of course, is optimistic. It’s just as possible that audiences who are annoyed or confused by disinformation will simply turn away from news (Wenzel 2019).

In these chapters we begin with a discussion of the media literacy movement and then move on to efforts to build community engagement as well as fact-checking. In the US during the 1930s, those who wanted to resist fascism and prevent a second world war were uncertain how they might counteract the tide of mis/disinformation without impinging on civil liberties and freedom of expression. They understood that persuasion and misinformation were central to the project of military conquest and totalitarian rule. As discussed earlier, scholars grappled with understanding why some people were susceptible to propaganda, what effect economic conditions had on that susceptibility, how to help audiences treat false claims with skepticism, and how to fight misinformation without resorting to censorship. The search for solutions to the vexing question of how to fight against mis/disinformation without shutting down free speech led to the same answer that is being put forward today: educate people so that they are “inoculated” (McGuire 1961, p. 326) against the “pollution” (Wardle and Derakhshan 2017) of the information ecosystem.

Fact-checking and community engagement efforts are also based on the relationship of journalists to audiences. Without trust there could be no belief in facts or desire for engagement with journalists. Accordingly, these efforts seek to create trust by promoting audience discernment and participation as well as journalistic accuracy, credibility, and transparency of process. All these efforts foreground the importance
of journalism to democracy, the role of journalism practice, and an engaged, active, critical audience. This is very different from the top-down solutions we examine later in this dissertation, such as algorithmic fixes and censorship, which address online mis/disinformation simply by suppressing potentially dangerous information. The media literacy, fact-checking, and community engagement efforts are based on similar philosophies, and in some ways their processes are similar. So, too, all three involve educating and engaging with the public.

The initiatives we discuss in the next three chapters fall into both the supply- and demand-side part of the continuum, as they are based on the hope that by educating the audience, improving the supply of quality journalism, or even changing journalism practice, appreciation of and demand for quality information will be bolstered. After the 2016 elections, support for community engagement efforts grew, as did support for fact-checking. The problem of media capture—which can lead to distrust of media—was and is acknowledged by some media literacy efforts. However, as media capture is a structural problem, it is often amorphous and difficult to solve. Instead, journalists and foundations seeking to build trust in the period after 2016 focused on steps that seemed do-able, particularly if they had already been doing them and had built up expertise. In so doing, these journalism organizations decided that expanding practices and activities they’d already been pursuing could help address the problem of mis/disinformation online (Cheruiyot and Ferrer-Conill 2018). It seemed self-evident that one way of tackling false information online would be to provide audiences with accurate information and an understanding of how that information was produced, and to give them the opportunity to participate in producing the information (RMIT ABC Fact Check 2019; Cheruiyot and Ferrer-Conill 2018). News organizations and foundations also hoped that changes in journalistic practice would encourage trust in media. Building trust was the topic of the day in the post-2016 period, with numerous conferences and discussions about how it could be carried out. These produced a consensus that building trust through fact-checking and changes to journalism practice was preferable to regulatory measures.

An example of this thinking came in the report summarizing discussions that took place at the prestigious February 2017 meeting held at Wilton Park, England. This meeting gathered foundation representatives, academics, and representatives from

---

4 For more information on media capture please see the work of Alina Mungiu-Pippidi, economics papers by Alexander Dyck and Andrea Prat as well as collection edited by the author: https://www.cima.ned.org/publication/media-capture-in-the-service-of-power/
technology companies, NGOs, and human rights groups such as Article 19. The 2017 report declared:

Media organisations should engage in more fact-checking, and if necessary, recruit the help of non-governmental organisations dedicated to news verification to accomplish this. It may also be practical to consider developing a ‘fake news beat’, where ‘fake news’ is covered in-depth by specialists. The media can build credibility by being transparent about organisational structure and ownership, advertising revenues, as well as sources of a story (Babajanian and Wendel 2017).

In 2017 the influential Knight Foundation allocated $1 million in grants to organizations trying to tackle misinformation. Projects that carried out community engagement and fact-checking were among those funded, suggesting that major donors had decided these two fronts were essential in the battle against disinformation (Mullin 2017).

Indeed, in its 2016 report on building trust, the American Press Institute declared that accuracy was the most important element. “For more than a generation, research has tried to identify the qualities that lead people to trust news. The new study employs multiple research methods to drill down into the notion of trust and identifies specific factors that publishers can put into action. Accuracy is the paramount principle...getting the facts right—is most valued regardless of topic” (API 2016).

Community engagement efforts and fact-checking are each based on the idea that journalism practice affects audience views of and trust in journalism. Transparency of process is key to community engagement efforts, opening up both newsrooms and the agenda-setting process to audiences. Fact-checking efforts show how accuracy is determined as well as the sources of information used by the fact-checkers. The idea that focusing on facts and bridging the divide in polarized societies by engaging with communities is viewed as a nonpartisan antidote to political polarization. In many parts of the world, journalists engaged in these kinds of efforts try to avoid partisan politics, although Lucas Graves notes that some organizations feel less strongly about avoiding partisanship (Graves 2018).

It can also be argued that media literacy, fact-checking, and community engagement are about empowering citizens to inform themselves. This empowerment necessarily changes the relationship with journalism, reducing audience dependency on the traditional news media and allowing audiences to bypass the news media (Cheruiyo and Ferrer-Conill 2018, p. 971). While believing in the Walter Lippmann idea of
delegating knowledge-gathering on subjects removed from the experience of audiences to experts (such as fact-checkers), these initiatives also believe in the active participation of audiences in the agenda-setting process (Rosen 1999; Carey in Glasser 1999).

Community engagement efforts, such as Hearken, solicit ideas from audiences as to what should be covered (Schmidt, Nelson, and Lawrence 2019). Fact-checkers rely on audiences to send in questions and circulate corrected information on social media. In some ways they are taking on the “boundary work” described by Matt Carlson which helps delineate the traditional role of journalists as professionals creating, safeguarding and transmitting knowledge. (Carlson 2017). Or we may think of fact-checkers as helping Michael Schudson’s “Monitorial Citizen” become more engaged and informed, and thus better able to carry out his/her monitorial function.

There is an inherent paradox in the idea of fact-checking. On one hand it purports to establish the truth. On the other, it needs public support and recognition of the truth (Graves 2016, p. 193) while acknowledging that public opinion is often unpredictable, illogical, and unpersuaded by facts.

No longer able to play the role of deciding what’s news, reporters must take up the trickier task of deciding what’s true. But it is important to recognize that their ambivalence towards the internet is really ambivalence toward the public that animates the network, an often hostile and unreasoning audience that reporters know more intimately than ever. This makes it all the more striking that, in editorial discussions and in their published work, fact-checkers hew so closely to the idealized audience of democratic citizens envisioned by traditional journalistic values (Graves 2016, p. 193).

This may be true more of fact-checking than of community engagement efforts. When people are in a room together they are more likely to reach a compromise (Moyer 2012; Sunstein 1999). When they are online, they are more likely to disagree and become polarized (Suler 2004; Barnidge 2017). It stands to reason, therefore, that community engagement efforts which bring community members into the same room are more likely to bring out the best in their audiences. By contrast, fact-checking elicits attacks from members of the public who disagree with the fact-checkers’ verdict on the veracity of a statement made by a politician they support. “Even more than their peers, these journalists are inundated in hostile and sometimes unhinged communication from their readers,” Graves notes (Graves 2016, p. 189).
While both fact-checking and community engagement initiatives rely on the “imagined audience” (Nelson 2018; Graves 2018; Ferrucci 2017) the reality of that audience turns out to be somewhat different than imagined. The audience is hard to find and it’s hard to reach. Fact-checkers, too, “have difficulty in building and maintaining diverse audiences” (Amazeen 2018b, 15). Fact-checking and building relationships with audiences are time-consuming, expensive, hard to scale, and laborious. The nature of this work means it’s almost impossible to measure impact, a challenge compounded by the fact that each group involved with community engagement has different goals and different ideas as to how it would measure success (Moffitt 2013; Nechushtai 2020). Further, it’s impossible to determine the counter-factual. What would public discussion have been like without the community engagement efforts? In this section we discuss the “demand-side” fixes and we begin by looking at Clyde Miller and the media literacy movement of the 1930s. His views on the role of citizens as active participants in democracy and the importance of information foreshadow not just the media reform movements of the 1940s but in some way resemble the public journalism movement of the 1990s (Pickard 2018).
Chapter 2

Demand-side solution: teaching media literacy in order to fight disinformation—in 1939

The political catastrophes of the 1930s gave studies of trust and how to combat propaganda new urgency. Faced with fascist regimes taking over much of the world and beaming propaganda globally, US demagogues spouting rhetoric against the government and world Jewry, the rise of Stalinism, and the beginning of the red-baiting that foreshadowed McCarthyism, scholars and journalists struggled to understand how people could fall for lies and overblown rhetoric.

In the 1930s propaganda and people’s belief in mis/disinformation had real and dangerous consequences. Like the post-2016 world, the world of the 1930s was politically polarized and authoritarian leaders with territorial ambitions threatened democracy globally. These leaders used propaganda to stoke fear and to transmit their messages.

The problem of online mis/disinformation may seem unique to the early 21st century, but technological changes to the production and distribution of news have often been accompanied by worries about declining information quality (Tucher 2013). The rise

---

5 Thanks to Chloe Oldham for her research, Andrea Gurwitt for her editing, and Professor Andie Tucher for her comments. Thanks to Thai Jones and the librarians working with the archives at the New York Public Library, Nicholas M. Butler, and Hannah Ontiveros and Nancy Mclean for sending me hundreds of pages of Duke University’s archives on Clyde Miller.
of radio, television, and the internet all provoked fears that societies as a whole, and young people especially, would be inundated with overwhelming amounts of false or sensationalistic news.

The fact that fears about “fake news” have surfaced periodically is often cited by those who accuse the people who raise these fears of being alarmists or unwilling to adapt to technological change. Like the politician who dismisses accusations of wrongdoing as “old news,” Silicon Valley executives like to say that societies have always resisted new forms of technology. They trot out well-worn oversimplifications, such as “medieval monks hated the printing press” or “there was no privacy in pre-industrial society.” But just because a problem is old or has appeared before does not mean it shouldn’t be addressed. To the contrary.

**Propaganda and Disinformation**

Historians who have analyzed media effects and propaganda through the ages have, among other topics, examined propaganda during the Boer War, in the Soviet Union under Lenin and Stalin, and in China after 1948, as well as the use of radio before the 1994 Rwanda genocide, and film and photography during the Vietnam War. They have studied propaganda techniques, the role that propaganda played in the course of events, and the phenomenon of mass persuasion. There is much to be learned from this illuminating and disturbing work. In this chapter, however, we will look at the US response to disinformation and propaganda in the 1930s, as Hitler took power just before World War II. We will consider one group’s launch of a media literacy program and an anti-racism curriculum as a way of inoculating children against propaganda.

It can be argued that propaganda is not the same as mis/disinformation, but we believe that in many ways it was the 1930s equivalent of today’s mis/disinformation (Pomerantzev 2015). Propaganda often crossed the line into falsehoods and truth was often obscured just as it is today. Another point in common is the intentionality of the person producing the disinformation (Bok 1978). One key difference may be the dynamic nature of mis/disinformation today compared to the 1930s. Today’s version is less predictable, as messages are disseminated by multiple actors and the technology shifts constantly (deep fakes, etc). Sowing distrust is the main goal (Persily 2017). By contrast, in the 1930s the propaganda seemed more predictable and had more clear-cut political or strategic ends.

Propaganda was central to totalitarian societies of the 20th century. Lenin created the largest apparatus for propaganda that the world had seen (Simons et al. 2017; Tong
and much has been written about Communist control of the media, both its structure and how Party press came to dominate after 1922, replacing the newspapers that had been in place before the revolution. So, too, in Nazi Germany where the Nazi propaganda machine included anti-Semitic children’s books, new newspapers such as *Der Stürmer*, edited by Julius Streicher, and the use of radio, loudspeakers, and film to transmit messages. As head of the propaganda ministry created by the Nazis, Joseph Goebbels oversaw an extensive apparatus of control over arts, culture, and information (Fest 1970). As well as making extensive use of relatively new technology (film and radio) the regime also controlled previously existing outlets, converting them to pro-party organs. Indeed, there is extensive literature on the collaboration of publishing houses, journalists, and writers during World War II (Reichman 2019; Guéhenno 1947). This quote from the memoirs of German author Raimund Pretzel, who wrote under the pseudonym of Sebastian Haffner during his exile in England, describes the atmosphere for news in Germany during the mid to late 1930s:

> Many journals and newspapers disappeared from the kiosks—but what happened to those continued in circulation was much more disturbing. You could not quite recognize them anymore. In a way a newspaper is like an old acquaintance: you instinctively know how it will react to certain events, what it will say about them and how it will express its views. If it suddenly says the opposite of what it said yesterday, denies its own past, distorting its features, you cannot avoid feeling that you are in a madhouse. That happened. Old established democratic broadsheets such as the *Berliner Tageblatt* or the *Vössische Zeitung* changed into Nazi organs from one day to the next. In their customary, measured educated style they said exactly the same things that had been spewed out by the *Angriff* or the *Völkischer Beobachter*, newspapers that had always supported the Nazis (Haffner 2000, p. 197)

Because the persuasive effects of propaganda seemed to affect mass attitudes towards authoritarian governments, trying to understand the mechanisms became important. At various times, Hannah Arendt, Walter Lippman, and, later, Jacques Ellul became preoccupied with the effects of propaganda on societies, and there was no better case study than World War II. The subject was to absorb Arendt for her entire adult life, and she is often quoted about the importance of trust and what happens to societies once they become engulfed in mistrust.

> Factual truth, if it happens to oppose a given group’s profit or pleasure, is greeted today with greater hostility than ever before…. More disturbing is that to
the extent that unwelcome truths are tolerated in free countries they are often
consciously or unconsciously transformed into opinions, as though the fact of
Germany’s support of Hitler or France’s collapse before the German armies in
1940 or of Vatican policies during the Second World War were not a matter of
historical record but of opinion…. What is at stake here is this common and
factual reality itself, and this is indeed a political problem of the first order

Long before these intellectuals were writing about propaganda effects, Clyde Miller, a
former journalist based at Columbia Teachers College in New York launched the first
media literacy program in the US. He did this because he believed the American way
to combat propaganda was to understand and explain it, not emulate or suppress it.
His efforts lasted no more than a decade, but his legacy lived on even after World
War II. Media trust and media credibility were further studied by academics after the
war, and schools in the UK, Germany, Canada, and Australia, among other countries,
launched their own media literacy programs. A number of these were influenced by
the work of Miller’s Institute for Propaganda Analysis (IPA).

This chapter will discuss Miller’s career and his influence on the media literacy
programs that followed him. This is not a text about pedagogical teaching methods,
but rather a parable of what happens when Enlightenment values of truth and
reasoning come under attack. Mis/disinformation is a tool of anti-democratic forces,
and the far-right of the 1930s used propaganda to play on fears of refugees, world
Jewry, Communists, and trade-union activists. As well as being used domestically
within Germany, propaganda emanated from Germany as part of its attempts to
support appeasement and forestall war. In the US, German propagandists even wrote
speeches for US Sen. Ernest Lundeen of Minnesota, foreshadowing the way Russian
disinformation is circulated today by far-right politicians in the US and Europe (Hart
2018a, p.104; Benkler, Faris, and Roberts 2018).

At a time when disinformation and extreme propaganda were considered the problem
of the day, journalists were well aware of the overt partisanship and capture by
business tycoons with political agendas that had marked their profession in preceding
decades. By the 1930s values of “professionalism” had been instilled in much of the
US press (Schudson 2001). It was therefore not surprising that, when the western
democracies faced a global disinformation problem, journalistic values of education,
critical thinking, and verification seemed like a solution, in the interwar period, just as
they do today.
Miller and the IPA studied and tried to counteract all kinds of propaganda during that era—whether from Goebbels, the CPUSA, or PR agencies. They called for an articulation of the techniques used by propagandists and developed an analytical framework for understanding and combatting that era’s propaganda. Their example can inspire us to do the same as we consider our contemporary information ecosystem and think about how we can combat the mis/disinformation we see on social media.

It is striking how many of the arguments made today appeared in the IPA Bulletin’s pages. Countering false information with true information is difficult when the people spreading false information refuse to change their ideas or, when told something is not true, simply double down on their prejudices. Just as Trump supporters say that factual errors are unimportant because Trump is speaking a larger truth, Henry Ford and Father Coughlin said that even if some information in the Protocols of the Elders of Zion was not authentic it could still be “factual” (Jan 1 1939 issue of the Bulletin, p. 5).

Anticipating today’s critique (Boyd 2018) by Data & Society founder Danah Boyd, the IPA faced criticism that its techniques would make students cynical. The IPA responded by saying that its teachings were needed to equip students for their future as engaged citizens: “The teacher who acts as a guide to maturity helps her pupils to think critically and to act intelligently on the everyday problems they are meeting [...] by its very nature [the] process will not build attitudes of cynicism and defeatism.”

Not only were the techniques and patterns of prejudice similar to those we see in the early 21st century, but so was much of the content of the mis/disinformation in the 1930s and 1940s: Hysteria about a subversive conspiracy at a Baltimore restaurant6, and dishonest advertisements paid for by doctors warning against socialized medicine. Far-right groups whipped up by weekly broadcasts and taking to the streets, where they were protected by sympathetic police officers. There were accusations of treason in the US government. When progressive candidate, Upton Sinclair, ran for governor in California he was defeated because of false accusations on film. A sociology professor who had authored a definitive textbook was accused by the business community of being too liberal. Accusations that refugees were getting jobs ahead of Americans. These false claims sound contemporary, and many similar incidents took place in the pre- and post-2016 period (Boak 2019), but the examples were in this paragraph all happened in the 1930s and 1940s and all were analyzed in the IPA’s monthly bulletin for teachers.

---

6 Reminding the author of the Comet Ping Pong conspiracy theory.
Miller, a reporter from the Cleveland Plain Dealer, had been appointed associate professor of education at Columbia Teachers College in 1935 at an annual salary of $6,700\(^7\) and was soon made its head of communications. With funding from department-store magnate Edward A. Filene, Miller founded (Nicotra 2009) the Institute for Propaganda Analysis in 1937. He brought a flair for publicity to this new cause, which, ultimately, rested on the belief that education could be used as a tool to fight prejudice and propaganda. Filene was a liberal philanthropist who had earlier invited the legendary muckraker Lincoln Steffens to Boston to help bring about reform of Boston politics. The so-called Boston 1915 project failed, but Filene remained committed to funding programs that helped support liberal democratic ideals. He set up the IPA with a $10,000 grant and ended up spending $1 million before it was disbanded at the start of World War II.

Many of the IPA’s ideas—especially about the need for personal reflection and understanding of how personal experience shapes one’s ideas—form the foundation (Hobbs and McGee 2014) of media literacy programs taught today. So do the techniques for understanding propaganda pioneered by Miller. Understanding Miller and his wide-ranging critique of the politics of his time, and the role of the media, provides context to the demand-side efforts of today.

While it was clear that Germany’s defeat in World War I and dire economic conditions, including widespread unemployment, paved the way for the rise of Adolf Hitler, academics and journalists tried to parse how Nazi propaganda had been so effective in galvanizing public support for the regime.

IPA’s position was that the interplay of economic conditions and individual tendencies created susceptibility to propaganda. But it saw only one possible solution that was compatible with the First Amendment.

“There are three ways to deal with propaganda—first, to suppress it; second to try to answer it by counter-propaganda; third, to analyze it. Suppression of propaganda is contrary to democratic principles, specifically contrary to the provisions of the United States Constitution. Counter-propaganda is legitimate but often intensifies cleavages. Analysis of propaganda, on the other hand, cannot hurt propaganda for a cause that we consider ‘good’,” Miller said in a speech in New York in 1939.

One of the first things the IPA did was to describe the techniques used by people spreading propaganda, creating a taxonomy of the different kinds in use, because IPA staffers thought that understanding the techniques of persuasion would arm people

\(^7\) Equal to about $124,000 today.
against them. Others have done this since, including Claire Wardle (2016) and Edson Tandoc (Tandoc et al. 2017). IPA also gave a checklist of questions the public could ask when confronted with information so as to understand whether it was true or not. More recently the Hapgood website, run by Mike Caulfield’s Digital Polarization Initiative, did the same with his well respected SIFT Model (Caulfield 2019).

The Group Leader’s Guide to Propaganda

As well as the taxonomy, the IPA published a foundational document advancing its view that both personal proclivity and social forces induced susceptibility to propaganda. In the lengthy “Group Leader’s Guide to Propaganda Analysis” (Edwards 1938), IPA’s educational director, Violet Edwards, argued that the “common man” was “tragically confused” by overwhelming amounts of information and having to make decisions without first-hand information.

Instead of the town hall or the cracker barrel of yore, where citizens could meet to discuss the topics that affected them personally, citizens had to rely on information from others about how society should be organized and which policies should be pursued far from home, Edwards wrote. Many others, including Walter Lippmann and later Jacques Ellul, argued that the common man should rely on journalism to help sift through and distill the excessive information available.

The IPA’s 250-page report/position paper is a detailed and thoughtful book on propaganda and media literacy. It pulled together much of the then-current thinking on the topic and anticipated much of our contemporary discussion. The IPA discussed confirmation bias and the role of advertising in paving the way for propaganda, and it called for journalists to go into communities and build relationships with their communities to explain the importance of journalism.

The IPA wanted its readers to understand that propaganda didn’t appear out of nowhere, that instead it arose from overarching historical, economic, and political patterns and trends that were important to identify. It summarized much of the thinking about media literacy and explained the importance of critical and scientific method.

Edwards argued that in order to understand the second-hand information on which citizens depend, readers must adopt critical thinking and scientific methods. The IPA hoped that people using its analytical techniques would be able to think rationally about the information they encountered.

---

8 https://hapgood.us/2019/06/19/sift-the-four-moves/ The moves are “Stop, Investigate the source, Find Better Coverage, Trace the claims, quotes and media to their original context.”
Seven Propaganda Devices described by the Institute for Propaganda Analysis

1. **Name-Calling.** Based on “hate and fear,” the propagandist gives “bad names” “to those individuals, groups, nations, races, policies, practices, beliefs and ideals which he would have us condemn and reject.”

2. **Glittering generalities.** Words are used to “stir up our emotions and befog our thinking....[T]he propagandist identifies his program with virtue by use of ‘virtue words’. Glittering Generalities is a device to make us accept and approve, without examining the evidence.”

3. **Transfer:** “The propagandist carries over the authority, sanction, and prestige of something we respect and revere to something he would have us accept.”

4. **Testimonial:** For instance, “When I feel tired, I smoke a Camel and get the grandest lift.”

5. **Plain Folks:** They “win our confidence by appearing to be people like ourselves.” For example, a commercial says, “It’s our family’s whiskey, neighbor; and neighbor, it’s your price.”

6. **Card Stacking:** “The propagandist deploys all the arts of deception to win our support for himself [...] by means of this device propagandists would convince us that a ruthless war of aggression is a crusade for righteousness.”

   “He stacks the cards against the truth. He uses under-emphasis and over-emphasis to dodge issues and evade facts. He resorts to lies, censorship and distortion. He omits facts. He offers false testimony.”

7. **Bandwagon:** “A device to make us follow the crowd to accept the propagandists’ program en masse.” … “Propagandists will appeal to us as Catholics, Protestants, or Jews” and use biases common to a group. “The theme of this type of propaganda may be summed up in the statement, ‘Everybody’s doing it: come along and follow the great majority, for it can’t be wrong.””

--------

Some ABC’s of Propaganda Analysis (December, 1937)
“ASCERTAIN the conflict element in the propaganda you are analyzing. All propaganda contains a conflict element in some form or other—either as cause, or as effect, or as both cause and effect.”

“BEHOLD your own reaction to this conflict element. It is always necessary to know and to take into consideration our own opinions with regard to a conflict situation about which we feel strongly, on which we are prone to take sides. This information permits us to become more objective in our analysis.”

“CONCERN yourself with today’s propagandas associated with today’s conflicts. These are the ones that affect directly our income, business, working conditions, health, education, and religious, political, and social responsibilities. It is all too easy to analyze some old example of propaganda, now having little relation to vital issues.”

“DOUBT that your opinions are ‘your very own.’ They usually aren’t. Our opinions, even with respect to today’s propagandas, have been largely determined for us by inheritance and environment….We resemble others with similar inheritance and environment and are bound to them by ties of common experience….We tend to distrust the opinions of those who differ from us in inheritance and environment. Only drastic changes in our life conditions, with new and different experiences, associations, and influences, can offset or cancel out the effect of inheritance and long years of environment.”

“EVALUATE, therefore, with the greatest care, your own propagandas [beliefs]. We must learn clearly why we act and believe as we do with respect to various conflicts and issues—political, economic, social, and religious…. This is very important.”

“FIND THE FACTS before you come to any conclusion. There is usually plenty of time to form a conclusion and believe in it later on. Once we learn how to recognize propaganda, we can most effectively deal with it by suspending our judgment until we have time to learn the facts and the logic or trickery involved in the propaganda in question. We must ask:
‘Who is this propagandist?
‘How is he trying to influence our thoughts and actions?
‘For what purpose does he use the common propaganda devices?
‘Do we like his purposes?
‘How does he use words and symbols?
‘What are the exact meanings of his words and symbols?
‘What does the propagandist try to make these words and symbols appear to mean?
‘What are the basic interests of this propagandist?”
‘Do his interests coincide with the interests of most citizens, of our society as we see it?’

“GUARD always, finally, against omnibus words. They are the words that make us the easy dupes of propagandists. Omnibus or carryall words are words that are extraordinarily difficult to define. They carry all sorts of meanings to the various sorts of men. Therefore, the best test for the truth or falsity of propaganda lies in specific and concrete definitions of the words and symbols used by the propagandist. Moreover, sharp definition is the best antidote against words and symbols that carry a high charge of emotion.”

IPA’ other books and publications

The IPA published analyses of political speeches using little icons—the emojis of the day—after each phrase to explain which technique the speaker was using. One such book (Lee 1939) analyzed the anti-Semitic radio broadcasts of the infamous Father Coughlin, a Catholic priest in Detroit who was estimated to have 30 million listeners for his broadcasts that included conspiracy theories about so-called “international Jewry” and President Roosevelt.

Bringing what the IPA described as “the newspaper man’s passion for simplifying complicated subjects,” Miller published monthly Bulletins describing an important topic in the news, analyzing the propaganda techniques used by all sides, and recommending further reading and discussion questions for classrooms. The IPA struck a nerve: Some 10,000 people subscribed to the Bulletin, which cost $2.00 a year, and 18,000 bought the bound volume of back issues published at the end of each year.

The IPA’s monthly Bulletins also contained teaching guides, including discussion questions and topics. These usually related back to the seven propaganda techniques described by the IPA. Like many groups today (Atkins 2017), the IPA believed that media literacy efforts in schools were essential to spreading their message and techniques. They sought to put students on guard against propaganda and to make them into sophisticated news consumers. The IPA formed a relationship with Scholastic magazine, in 1939 and 1940, producing a series called, “What Makes You Think So? Expert Guidance to Help You Think Clearly and Detect Propaganda in Any Form,” which was distributed in schools. By the late 1930s, one million school children were using its methods to analyze propaganda, and the IPA corresponded with some 2,500 teachers (Sproule 1996).
The IPA stressed critical thinking and understanding the broader political economy. It outlined techniques for audiences to look for when reading. But it went further than that. In the IPA’s monthly bulletins, Miller and colleagues provided detailed analyses of world politics and displayed a broad understanding of the context in which the media operated and how propaganda worked. Rather than simply detailing propaganda techniques, Miller exposed the strategies revealed by news coverage of a variety of topics and analyzed the interests of the different players involved. He tackled the growing public relations industry, debates over the rise of chain stores, anti-union press coverage, the quality of evidence presented to the Dies Committee (later known as the House Un-American Activities Committee), as well as tactics used by white supremacists, domestic fascists, and Nazi sympathizers. The IPA analyzed their methods in detail but also provided accurate information in order to do what are now referred to as “debunkings.” Looking at the IPA’s teaching materials from today’s vantage point it’s clear they belong to a pre-internet era when attention spans were longer and there was an appetite for detailed critiques and analysis.

In the first issue of his monthly bulletin, dated October 1937, Miller drew a historical parallel by comparing the use of propaganda in the Pullman strike of 1894 with that used by Henry Ford and the Johnstown Citizens. In an affidavit to the Dies Committee, Hill & Knowlton later disputed the claims made by the IPA about Hill & Knowlton’s role in the infamous Youngstown Steel Strike.

In another bulletin, the IPA dissected disinformation about refugees stealing jobs from Americans:

Persistently since the influx of refugees from the war areas began, a story has bobbed up in numerous American cities about the alleged heartless—and actually unreal—discharging of regular employees [sic] by stores to make places for ‘foreigners.’ The story usually is anti-Semitic; the store with which it is connected has Jewish owners, and Jews are said to get the jobs.

One large store in New York City which has been a victim of the story has spent considerable sums trying to trace the source and find some way of stopping it. The efforts have been fruitless. The story keeps reappearing, and mimeographed leaflets have even been circulated picturing the Jewish manager welcoming a long line of Jewish refugees while turning away another line of fine Nordic types [May 26, 1941].

What made Miller’s work different from many current news literacy efforts is the breadth of his vision. Many current efforts focus on helping students decipher what is
true and what is not, teach students techniques used by fact-checking groups, and offer useful skills such as how to do reverse image searches and identify a photo that has been mislabeled or taken out of context. There is a debate in the world of media literacy about whether it’s necessary to discuss topics such as media ownership and advertiser power (Maskl et al. 2017). Renee Hobbs, one of the best-known scholars in the US media literacy field, has argued that questions of ownership need to be part of the media literacy curriculum (Hobbs 2010). This is part of a broader discussion in the field regarding differences between news literacy and media literacy. It’s also the position of many that news literacy and media literacy skills are simply part of the broader critical thinking skills that everyone needs.

For the IPA, this wasn’t even a question. In May of 1938, *The Bulletin* delved into the techniques of Nazi propaganda and discussed how authoritarian regimes lack the “competing organizations” that create a marketplace of ideas. The IPA argued in part that it’s not just propaganda that is a problem, but the state’s domination of information flows.

“Political, economic, educational and religious spokesmen are able to and, actually do, disseminate rival propagandas. This gives those at whom the rival propagandas are directed some freedom of choice among the alternatives offered them” (Volume 1, number 8, May 1938).

Miller used his ideas about propaganda techniques such as “testimonials” and “glittering generalities” and explained how they applied to Nazi propaganda. He discussed the neglect of the German middle class after World War I, arguing that this made them more susceptible to Nazi propaganda, and he singled out such German industrialists as Fritz Thyssen, who funded the NSDAP, for wanting to crush labor and Communism (p. 3, Volume 1, number 8, May 1938). He also parsed the use of Nazi symbols like swastikas and the othering of Jewish people, and dissected how these tropes surfaced in the publications of domestic US Nazi groups like the American Nationalist Confederation, detailing their resemblance to the 19th century white supremacist organization the White Camelia. Many of the leaders were later put on trial but not imprisoned (Hart 2018a).

Miller also understood the use of soft power such as ownership and advertising to capture the media (Volume 1, number 5, February 1938). The *Bulletin* devoted two issues to “Newspaper Analysis,” returning to the point that in democracies there are “many voices, many opinions and many propagandas.” Quoting University of Missouri journalism professor Roscoe Ellard, the *Bulletin* made the point that editors of small-town newspapers can’t afford to lose advertising revenue and so can
gradually succumb to pressure from the business community as well as from readers. The result can be a softening of the newspaper’s stance and the adoption of a captured mind-set and self-censorship. Large newspapers that have the pressure to resist are the most “reliable” and the Bulletin provided a list of credible newspapers, which included The New York Times and The Baltimore Sun. Based on a survey of working journalists in Washington DC, the list of newspapers found to be unreliable included (in first place) The Chicago Tribune, which at that time was owned by William Randolph Hearst. The CP’s Daily Worker also made the list.

**IPA community programs and backlash**

This broader context is essential to understanding the response to the IPA and to Miller’s work, including the attacks by Hearst and the investigation of the IPA by the Dies Committee. The committee had been set up to analyze both Communist and fascist activities in the US; it later became notorious as the House Committee on Un-American Activities, used by Joseph McCarthy after World War II and in the 1950s before finally being disbanded. McCarthyism was a feature of the 1950s, so it’s not as well known that hearings against Communists began in the 1930s. In January 1940, the IPA devoted an issue of the bulletin to analyzing the Dies Committee.

The committee was often criticized for not going after the domestic far right and for focusing too much on Communism (Heale 1986), but this was not the focus of Miller’s analysis in the Bulletin. Rather, he discussed the testimony given to the committee by many of the witnesses and argued that there was far too little evidence presented and far too much gossip and scandal-mongering. Miller noted that committee chairman Martin Dies and his supporters took the view that “some of his witnesses might be unreliable but insisted, nevertheless, that valuable evidence of Communist activities was being uncovered where none had been suspected.”

After the Bulletin was published, IPA president Professor Kirtley F. Mather of Harvard signed a letter to Congress calling for the cessation of the committee’s activities (Special to The New York Times, February 23, 1941). This was not the first such letter (Heineman 1992). A few weeks after the bulletin was published, Dies’ deputy, J.B Matthews, announced that an investigation into the IPA had been underway for two years, saying that Committee members wanted to understand who was influencing American students (Special to the New York Times, February 23, 1941).

Meanwhile, the IPA further ruffled feathers by helping design a curriculum aimed at promoting civic engagement and racial and religious tolerance that was piloted (Johnson 2006) in the Springfield, Mass. school district, which had a sympathetic
superintendent. The program petered out after a few years partly due to criticism by the Catholic Church and lack of local support as religious tensions rose in Springfield after World War II (Bresnahan 1971). By the early ‘50s, as McCarthyism was under way, there were murmurings that the plan contained “subversive” elements.

However, although it was phased out in Springfield, the plan’s ideas lived on. According to education professor Lauri Johnson, “the Springfield Plan became the most well-publicized intercultural educational curriculum in the 1940s, talked about and emulated by school districts across the country and into Canada” (Johnson 2006, p. 302).

The IPA ceased publishing the weekly Bulletin in 1942, and it disbanded after the US entered World War II and needed to start producing its own propaganda to galvanize support for the fight against Hitler. In its farewell issue of Jan. 9, 1942, headlined “We Say Au Revoir,” the IPA explained that the board of directors had voted to suspend operations.

“The publication of its dispassionate analysis of all kinds of propaganda ‘good’ and ‘bad’, is easily misunderstood during a war emergency, and more important, the analyses could be misused for undesirable purposes by persons opposing the government’s efforts. On the other hand, for the Institute, as an Institute, to propagandize would cast doubt on its integrity as a scientific body” (Bulletin Jan 9, 1942, volume IV).

This final Bulletin expressed satisfaction with the work achieved by the IPA, warned that wartime is usually accompanied by a rise in intolerance, and expressed the hope that the IPA’s techniques for analyzing propaganda would be used in the future, which indeed they were. But the ending was not so simple. The US entered World War II, then transitioned into the Cold War and hurried down the road of McCarthyism—all of which required vast amounts of propaganda and sometimes disinformation.

There was no room in wartime America for the IPA. Miller’s sophistication and nuance led to attacks from the right in the US. He was smeared by Hearst, investigated by the Dies Committee, and lost his job. Miller’s time at Columbia Teachers College came to a sad end as he apparently fell victim to the intolerance he had warned against. Along with other faculty, he was placed on leave in 1944 when the college was facing a financial crisis and was never brought back to work. In 1948, Miller was officially let go and told it was because of departmental restructuring, but Miller argued that it was because his work had offended William Randolph Hearst.
Hearst was known for attacking “Reds” in the universities and schools and even sending in reporters disguised as students to try to uncover left-wing professors in the classroom (Alwood 2007).9

His newspaper, The World Telegram, had criticized the church group that Miller was involved in. The red-baiting Hearst columnist, Frederik Woltman, had written nasty articles about the Methodist Federation for Social Action, “accusing the federation of being sympathetic toward the policies of the Soviet Union” (New York Times, May 8, 1948). Miller protested as soon as they appeared and called on Columbia Journalism School to rescind the Pulitzer Prize they had bestowed on Woltman in 1947 (IPA files at Butler Library, Columbia University). Speaking to the press after he was let go, Miller said it was possible that the Teachers College board had been influenced by his campaign against Woltman (New York Times 1948; New York Herald Tribune 1948).

Many professors during the McCarthy years were let go quietly (Schrecker 1986) and Miller was apparently one of them. He lost his Columbia housing and salary and wrote repeatedly to Columbia’s president decrying the “violation of tenure and academic freedom.” After that, the trail seemingly runs cold. Miller died on a trip to Australia in 1977 and is buried there. We’ve not been able to find his obituary.

We may never know for sure whether Miller was a victim of the political climate of the 1930s and an early victim of the repressive climate that took hold in the 1950s. The documents in the archives at Columbia suggest that the president of Teachers College had long disliked Miller’s work and had begun suggesting to Miller as early as 1939 that Miller seek work elsewhere. After Miller was dismissed he spoke to the New York Times saying it was a violation of academic freedom, got a lawyer to write and ask Columbia for $100,000 in damages, and asked the American Association of University Professors to write to the provost. The dean of Columbia Teachers College William Fletcher Russell, responding to the questions about the circumstances of Miller’s firing, on June 29, 1948 sent a 13-page letter to provost Albert G. Jacobs outlining the case for firing Miller. Russell’s lengthy defense of his decision to fire Miller was that his staff didn’t get along with him, he’d never done much teaching and, the IPA and Springfield Plan had no relation to Teachers College. Russell concluded by saying, “There is no justification for any charge that might be made by Miller that academic freedom or liberty of teaching has been in anyway infringed.” The letter seemed to satisfy the administration, which replied thanking Russell for his

---

9 Unfortunately similar to what happened during the Covid-19 pandemic of 2020.
“thorough account” (Columbia University Central Files Series 1: 1895-1971 of the archives at Butler, Boxes 352, 217, 412).

On Dec 7, 1948 Russell wrote another memo to the provost, this time asking that Russell’s job title be changed from “Dean” of Teachers College to “President” of Teachers College. In March 1949, Russell submitted another lengthy memo, proposing “A Substantial Program by Columbia University and Teachers College for Education for Americanism” aimed at strengthening “loyalty to American ideals and action.” Saying that Americans have “weak defenses,” Russell outlined the threat from Communists and said that schools needed to inculcate an appreciation for liberty so that citizens could resist.

“Soviet agents, for thirty years, have conducted a skillful campaign to capture posts of leadership in government service, labor unions, agencies of communications and in colleges and schools…the Communist Professors at the University of Washington—and I fear there are others—those Americans who follow blindly the Soviets; those who let themselves be drawn into the organization of the hammer and sickle; maybe editors of great ability, labor leaders of power, government servants whose superiors may commend them highly, professors who know and contribute to their subject; but they do not know enough about liberty, and its diseases and dangers, to keep them from being agents of a foreign power or from giving unconscious aid and comfort to the enemy in a cold war.”

Subsequent correspondence suggests that the proposal was adapted, becoming less explicitly anti-Communist (Columbia University Central Files Series 1: 1895-1971 of the archives at Butler, Boxes 352, 217, 412) and was funded by the Carnegie Corporation (Griffin and Felix 1951).

We can’t say with certainty that Russell’s multi-year campaign to oust Miller from Teachers’ College was over political disagreements, but it is interesting that after the Institute of Propaganda Analysis was dismantled, Teachers College launched a program focused on promoting citizenship rather than critical thinking about propaganda.

The attacks on Miller and the IPA by the right show that media literacy, resisting propaganda, and supporting truth threaten those who try to gain power through propaganda and mis/disinformation. These same tendencies are alive today as the new demagogues attack “fake news” and journalism. While full-blown McCarthyism did not get under way until after the Second World War, there were people later
known as “premature Anti-Fascists” who opposed anti-Communist witch-hunting in the 1930s. Miller and his colleagues were perhaps premature anti-disinformationists.

Much of what Miller feared came to pass. Propaganda and lies were used widely during World War II. As well as suppressing information, governments created and disseminated propaganda as a way of boosting the war effort (Tucher, forthcoming 2021).

**Measuring success in media literacy education**

Even so, the IPA’s legacy lived on. The IPA’s attempts to inject scientific method and rational discourse as a way of countering disinformation are being replicated today (Pennycook and Rand 2017). The IPA’s approach of working with teachers and students to train the next generation to think critically about information is also being tried again.

Media or news literacy has become part of the curriculum in many countries, passing through a number of phases. It’s generally argued that the first phase is “inoculation,” in which “attitudinal resistance is conferred by pre-emptively highlighting false claims and refuting potential counterarguments” and that critical thinking is a later phase (Van der Linden et al. 2017). Much of this work is predicated on the idea that warning people about the presence of false information may help prevent them from believing it.

“The rate of cultural transmission, or infection, may be slowed through a process known as attitudinal inoculation. In medicine, resistance to a virus can be conferred by exposing someone to a weakened version of the virus (a vaccine)—strong enough to trigger a response (i.e., the production of antibodies), but not so strong as to overwhelm the body’s immune system. The social–psychological theory of attitudinal inoculation follows a similar logic: A threat is introduced by forewarning people that they may be exposed to information that challenges their existing beliefs or behaviors. Then, one or more (weakened) examples of that information are presented and directly refuted in a process called “refutational pre-emption” or “prebunking.” In short, attitudinal resistance is conferred by pre-emptively highlighting false claims and refuting potential counterarguments (Van der Linden et al. 2017).

---

10 Today governments still get involved in counter speech as a way of counteracting the supply of mis/disinformation online. We’ll discuss the “counter speech” arguments in Chapter Seven.
Because US education is governed by states, there is a lack of comprehensive federal legislation mandating media literacy education in schools, even if the different groups could agree on what should be taught. Most media literacy programs fall into three categories: 1) those that look at the “economic motivations that undergird popular media and information streams,” 2) education that relates to democracy and civic life; and, most commonly, 3) efforts to “evaluate the quality of information” (Huguet et al. 2019). Even so, media literacy efforts are often fragmented and inconsistent, which contributes to the difficulty of measuring their impact. In the US, they’ve not been well funded. Indeed, one study of media literacy funding put the amount at a mere $12 million between 2006 and 2016 (Fleming 2016).

The 2016 elections and the rise of Facebook and Twitter brought about a renewed interest in propaganda, misinformation, and disinformation and a flurry of commissions (UN 2019), reports (House of Commons 2018), and initiatives (Barr 2017). For the tech giants, supporting media literacy efforts became a natural part of efforts to avoid regulation. Their view was that they were simply the pipes for other people’s content, and it was up to the audience to figure out what was true and what was not (Bell 2016). The companies funded some media literacy groups, including the LAMP project in New York City, which received funding from Google, Twitter and Facebook (Vito 2019). Facebook and a group of funders also gave $14 million to City University of New York and a consortium of other international institutions, including Sciences-Po in Paris, for a News Integrity Initiative (Murgia 2017). Despite these grants, the groups doing media literacy training in the US complain that while interest has soared since 2016, funding has not (author interviews with media literacy proponents 2019).

Myriad groups teach media literacy today. Some also commission research, design curricula, and work with parents, teachers, and legislators trying to pass laws mandating media literacy classes. Librarians are also very involved in teaching media literacy skills (Head et al., 2020, pp. 12-17). However, their efforts tend to be fragmented, and include priorities such as diversity online, representation, identifying disinformation, reasoning, and understanding the role of journalism (Head, Fister, and MacMiIllan 2020).

Some initiatives conflate media literacy with other pressing digital-era matters. Some US groups are promoting legislation to create a legal “right to be forgotten” online for children under the age of 17 (Steinberg 2018). Groups like Common Sense focus on campaigns for digital safety for children and oppose cyber-bullying, though they’ve also surveyed young people about whether they can discern false information online (Robb 2017).
The National Association for Media Literacy Education (NAMLE) was founded by a group of academics in 1997 and tries to raise awareness about the importance of media literacy through conferences attended by parents, teachers, and government officials. Since 2016, public attention and requests for information from government and journalists have skyrocketed, says NAMLE’s executive director, Michelle Lipkin (2019). However, this new interest has not translated into a notable increase in funding for NAMLE. “We’ve yet to see any significant funding that changes the level for the media-literacy playing field generally,” says Lipkin (2019).

Some of the groups now teaching media literacy, such as the Center for News Literacy at the State University of New York, Stony Brook, teach techniques for how to determine whether something is false. They promote techniques that fact-checkers use, encouraging audiences to look “laterally” across the internet in order to see the context of a single piece of information and check it against other sources. Many of today’s news literacy groups were founded by journalists with a passion for educating the public about the importance of journalism and the role of journalists. These include Howard Schneider, the former Newsday editor who went on to direct Stony Brook’s Center for News Literacy, and Alan Miller, a former Los Angeles Times reporter who founded the News Literacy Project. This has led to criticism that some of the current efforts spend too much time glorifying journalism practice and lecturing students about the profession (Hobbs 2010; Maksl et al.). Others take the opposite view: Allison Head and colleagues who have researched college students’ attitudes to the news strongly recommend that college professors discuss the news as part of their teaching and say they have a role to play in helping students understand which sources are reliable (Head et al. 2019). Helping young people engage with the news is part of creating life-long citizen participation and so is essential for society (Malik et al. 2013).

In an attempt to come up with a common framework for measuring media literacy teaching, Rand published a report in the summer of 2019 (Huguet et al. 2019) suggesting a core set of standards and discussing ways to evaluate media literacy programs.

**Selected Media Literacy-Related Competencies Relevant to Mitigating Truth Decay**

Standards
1. Recognize the demand for and be able to search for, access, and retrieve information and media content.\(^a\)

2. Use evidence to investigate questions; devise and implement a plan to fill knowledge gaps.\(^b\)

3. Analyze information from multiple sources and identify complexities, discrepancies, and different perspectives.\(^c\)

4. Evaluate characteristics of information products that indicate the underlying creation processes.\(^d\)

5. Trace and evaluate an argument and specific claims in a text, assessing whether the reasoning is sound and the evidence is relevant and sufficient to support the claims.\(^e\)

6. Students create original works or responsibly repurpose or remix digital resources into new creations.\(^f\)

7. Communicate discoveries in ways that suit the purpose and audience.\(^g\)

\(^a\)UNESCO, 2013.


\(^e\)National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010.

\(^f\)International Society of Technology in Education, undated.

\(^g\)National Council of Teachers of English, “NCTE/IRA Standards for the English Language Arts,” webpage, undated.

All of this may be aimed at the wrong target, however. New scholarship suggests that older audiences are the most susceptible to mis/disinformation (Silverman 2019b). So, while targeting young audiences and helping them think critically is important, it will not address the population that is mainly spreading online mis/disinformation, a
subject that was discussed in the media in the post-2016 period (Newaz 2018; Grinberg et al. 2019).

Impact

Some argue that propaganda and mis/disinformation have little effect, and they criticize news and media literacy efforts for being unable to demonstrate impact. These debates have been around for decades and some of the thinkers of the post-war period answered these questions in ways that are still germane. Media literacy is the quintessential demand-side intervention, placing the onus on audiences to determine what is accurate and what is false and asking audiences to detect propaganda. It’s a long-term fix and expensive to scale and there is no conclusive evidence that it works. Over the years, there have been many debates as to how to measure persuasion effects from new forms of technology (Gerbner & Gross 1976)

In 1962, Ellul pointed out the numerous measurement difficulties: one is that there is inconsistency in what is being taught and which pedagogical outcomes are intended. Attribution of key variables is impossible to determine, and many of the studies rely on survey data and use students as respondents. The groups we interviewed told us they had done no longitudinal studies at all, though some recent research has reported that Stony Brook students who took a news literacy class reported later that they sought out more news sources (Maksl et al. 2017). However, the authors caution that the group was self-selected and took the course because they were interested in the subject.

In their 2019 report, Huguet and colleagues noted that many of the ways of measuring the effects of media literacy programs do not capture their effects fully. Some studies rely on self-reporting while others use multiple-choice assessments and still others rely on performance-based assessments, giving tasks to students who have completed a media literacy training (Huguet et al. 2019; Wineburg et al. 2016). In the early 2000s there was an attempt to use algorithms to assess cognitive decision-making. Students were given tasks to complete based on scenarios and the assessment was done by computer (Katz 2007).

Those who don’t worry about the effects on society of mis/disinformation also note the lack of evidence as to its impact. In the post-2016 period, one of the first papers on the subject of fake news during the US elections was by Allcott and Gentzkow and
concluded that it had no effect (Allcott and Gentzkow 2017). This was quoted widely and cited nearly 2000 times. The research done by these two economists raises some questions:

1) Much of the information that would help us understand the effects of mis/disinformation on voting behavior is held by platforms that won’t share it. For example, if researchers knew exactly where micro-targeting was aimed, they could check it against voting patterns just as researchers have done with Fox News and also to understand the effects of the closure of local newspapers (DellaVigna and Kaplan 2007).

2) Many studies of the effects of mis/disinformation assume that all “fake news” is equal in its impact. But it may be that one kind of rumor or false information is more likely to sway a particular voter than another.

3) It’s not clear if there is a threshold effect or what is the cumulative effects of mis/disinformation. Seeing hundreds of tweets and having them reinforced on Fox may produce a different effect than seeing a single piece of misinformation once. Kathleen Hall Jamieson discusses the reinforcement effect in her study of the 2016 presidential elections (Jamieson 2018). It’s also hard to measure the entire media ecosystem, where incorrect information becomes viral and “ricochets” around the internet (Persily 2017). Benkler and colleagues also found that quality, legacy media outlets inadvertently amplify mis/disinformation sometimes when trying to debunk it. In 2016, it may be that the greatest effect of mis/disinformation may not have been to put forth a particular point of view or piece of disinformation but to sow mistrust and reduce voter turnout (Persily 2017). Given that mis/disinformation is often large scale, or created and spread by state actors, it is hard to see how individual demand-side efforts can solve a structural problem.11

Our point here is that because the impact of persuasion and the effects of propaganda are famously difficult to prove, it is easy to dismiss the critics as lacking evidence. In his 1962 book on propaganda (Ellul 1962, p. 259), Jacques Ellul summarized many of the methodological problems that researchers face when trying to understand media effects and concluded that it’s simply not possible to do surveys or try to determine the individual effects of propaganda. It must be studied in the context of societies that have been exposed to it over many years.

This is similar to the point made by development economists about the limitations of randomized control tests that look only at small development interventions. These

---

11 Allison Head, comment to author, February 2020.
studies of individuals and isolated cases don’t take into account contagion effects, how knowledge is transmitted more generally, or the broader social and political contexts (Deaton and Cartwright 2018). Ellul argued that, rather than focus on individual studies or small groups, we can learn more from looking at entire societies that have been subjected to propaganda (Ellul 1962). From today’s vantage point, Ellul’s point seems correct and so does the IPA’s point about how, in certain context, some individuals may be swayed by multiple exposures to mis/disinformation. The rise of the anti-vaxxer movement and the spate of killings by white supremacists who cite phrases used by “conservative media stars” suggest that consumption of incendiary information spread by the media and/or online has an effect (Peters et al. 2019). The IPA was right in saying that media literacy education cannot be just about looking at facts and images and trying to verify them. The sources of the information, the agenda of those sources, and the broader context of the information also matter. Clearly, teaching people to question what they’re told and to think rationally about the information they’re fed is as essential now as it was in the 1930s. Unfortunately, well-intentioned as these efforts are, it’s unlikely that current efforts at media literacy and education will solve the problems posed by disinformation and propaganda. For starters, they are too small-scale. As well, some have been criticized for being out of date and incompatible with existing curriculum (Lister 2019).

However, how societies decide what is true or how to think about information standards are important subjects. Part of what is powerful about media literacy efforts is not just the skills but the critical thinking they try to impart. This critical thinking is still core to media literacy efforts. This idea of prompting people to think before they forward mis/disinformation is also being tried (Lewandowsky et al. 2017; Pennycook and Rand 2018).

Also important are grassroots efforts to reach classrooms and explain journalism all over the world. The News Literacy Program has revived its Newsroom-to-Classroom initiative so that teachers around the US can bring journalists into their classrooms to talk about how journalists gather information and what makes journalism different from other kinds of information (Miller 2019b). These efforts by the public and by teachers are part of building understanding and letting people know what news and what journalists do, including their practices and ethics. All of these efforts to raise awareness, call out disinformation, and stand by the importance of facts, are critical to the larger mosaic of efforts. We will further explore efforts to raise awareness about journalism and correct and debunk mis/disinformation in the following chapters on fact-checking and community engagement.
Mis- and Disinformation online
Chapter 3

Community engagement efforts

Introduction

Working with audiences and helping them understand the processes behind news-gathering is part of news literacy efforts. So, too, for the community engagement movement. The idea of engaging with audiences has been around for decades—indeed it’s particularly strong in northern Europe and in traditional forms of journalism practice, such as local news-gathering. But what had been viewed as an essential part of the journalism mission and a possible solution to financial woes morphed, in the period after 2016, into a possible answer to the problem of trust.

Causality became confused: The decline of local news and the rise of political polarization were seen as related. Many argued that journalists had abandoned voters in the red states, or the parts of the United Kingdom that voted for Brexit, and so missed the underlying sentiment and anger about globalization and fear of economic dislocation that contributed to the votes for Trump and Brexit (Massing 2016; Massing 2018). This line of thinking led to efforts at many outlets, including The New York Times and The Washington Post, to spend more time on under-covered areas. There was talk of news deserts and spending more money on supporting local news. Among other efforts, the renewed emphasis on community engagement projects led to the expansion of Internews’ US efforts, the creation of new fellowships for reporters...
from underserved areas,\textsuperscript{12} and the work of projects such as that of Andrea Wenzel, who worked with audiences in Kentucky (Wenzel 2018).

The push for local reporting\textsuperscript{13} is similar to what happened after the 1965 Watts riots when newspapers began sending black staffers, including an office messenger, library assistant, and truck driver out to cover poor black communities (Carroll 2017). But it wasn’t until the Kerner commission in 1968 blasted US media for their coverage of the riots and their lack of black reporters (Kerner report 1968) that systematic efforts at hiring journalists of color began with programs like the Fred W. Friendly summer program, which launched the careers of a number of black journalists (Basconi 2008).

Of course, it’s not clear that having a national media outlet cover under-covered areas will boost trust in media or that this trust in media would be generalized to other institutions (Schiffrin et al. 2017; Levi and Stoker 2000). Nor do we have causal data suggesting that people watch Fox or read The Sun or the Daily Mail because they don’t get enough news about their communities. Even so, much of the journalism community and some donors believed that flooding the zone with high-quality journalism would help audiences come back from the brink of Fox News and online disinformation, restore the cohesion of the public, and help solve the problem of political polarization (Bell 2016).

In some ways, the post-2016 hopes were reminiscent of the old debate between media malaise and the virtuous circle (Norris 2000; Hume in Schiffrin et al. 2017; Curran et al. 2018; Strömbäck and Shehata 2010; Newton 1999). Over decades, scholars debated whether television consumption led to cynicism and disengagement or whether the “virtuous circle” described by political scientist Pippa Norris existed. Norris argues that there is a “virtuous circle” in which media exposure can create more trust in society (Norris, 2000). Ultimately, the research came down on the side of the virtuous circle, showing that in many cases exposure to newspapers did lead to political and civic engagement. Many of the people involved in the community engagement movement, of course, were not familiar with the academic debates of the last century. They believed in journalism more as an article of faith, because journalists believe that the solution to many social problems is more journalism (Schiffrin 2016). In short, they hoped that addressing trust in media through engagement efforts might solve for

\textsuperscript{12} The prestigious Nieman program at Harvard announced new fellowships in 2017 https://nieman.harvard.edu/news/2017/12/abrams-nieman-fellowship-for-local-investigative-journalism-established-at-harvard-university/

\textsuperscript{13} Now also being funded by Facebook https://www.facebook.com/journalismproject/coronavirus-update-news-industry-support
political polarization. Restoration of trust could help restore a public conversation, rather than a fragmented and polarized one.

In an interview, one donor summed up this thinking:

After 2016 we, like everyone else, thought trust was the solution and we needed to build trust in the media…. We gave a lot of money to media who were improving capacity to engage with the public. Before 2016, we were sure that engaging better with the public was the only way they would survive the business model [crisis] and then after 2016 we thought more about how the media needs to connect with the public and reach across the divides and connect people. (Ronderos 2019)

The groups that had been doing community engagement work felt that the 2016 elections were confirmation of the importance of their work and a chance to “double down” and expand their efforts (Bourgault 2019). Organizations like Listening Post, founded in New Orleans in 2013, received donor funding after 2016 so they could broaden their community-building activities with readers. Just like the “public journalism” or “civic journalism” efforts of the 1990s (Ferrucci 2017; Wenzel et al. 2019), the intent post-2016 was to bring the voice of community members and audiences into the news-making process.

Adding to the urgency of the 21st century community engagement movement is the financial survival of legacy media. The need to solidify relations with audiences in uncertain times is certainly one aspect of the movement even if it’s often not stated explicitly (Mckenzie et al. 2011; Graves and Glaisyer 2012). The hope that engagement efforts will lead to paid membership or subscription revenue is yet another reason for carrying out such efforts (Hansen and Goligoski 2018). Our interviews highlighted this idea that community engagement will promote both financial health and the health of society.

“Subscriptions are reflective of reminding people of what journalism is and part of living in any given place. You move to LA and get a California driver’s license and also know you should subscribe to the local paper and want to do that. Media has to reflect that they know the community and can give the information that people want” (author interview with Jesse Hardman May 2019). Getting people to pay for news requires creating news that is so relevant to community residents that they are willing to pay for it. Again, there is a similarity to the 1990s emphasis on local “news you can use,” exemplified in the US by the Gannett newspaper chain (McCartney 2007; Prichard 1997).
This chapter will first explore the roots of the community engagement movement within the public journalism movement of the 1990s, and then look at how community engagement was framed as a solution in the post-2016 period. We will present some of our own research on organizations around the world trying to build trust and conclude by looking at whether it’s possible to measure the impact of engagement efforts.

We argue that civic engagement efforts have multiple motivations and may never fully address the question of how to build trust in quality media or help audiences distinguish between quality information online and misinformation. Further, context matters, and so these efforts may be more effective in some places than in others. Trust in media and institutions is relatively high in Sweden and levels of disinformation low, so it stands to reason that engagement efforts are more effective there than in other places. Indeed, recent work by Efrat Nechushtai has uncovered interesting differences between how journalists approach community engagement efforts and how they respond to suggestions/requests from audiences. The German journalists that Nechushtai studied are more likely to view themselves as part of the community and are relatively willing to adapt their coverage to the desires of their audience. By contrast, US journalists viewed themselves as experts and treated engagement more as a chance to explain their work and their role to their audience. (Nehcushtai 2020)

**History of the public journalism movement of the 1990s and similarities to community engagement post-2016**

In this section we discuss some similarities between the public journalism movement of the 1990s with the community engagement movement post-2016 (Nelson 2018; Ferrucci 2017). Contemporary community engagement initiatives have their roots in the 1990s, when “public” or “civic” journalism was born to address worries that changing media-consumption habits, which stemmed in part from new technologies such as cable television, changed the way Americans got their news and inclined them toward an echo chamber (Grabe 2000). Segmentation in consumption patterns meant the public no longer shared common news sources. In addition, people felt increasingly alienated from the political process and more and more detached from democracy. They simply didn’t believe they had the power to change anything, so voting dropped in congressional and presidential elections. And by 1995 (Conte 1996), US citizens expressing “a great deal of confidence” in newspapers stood at only 21 percent, while television news performed only marginally better at 23 percent,
down from 51 and 55 percent respectively in 1988, according to the Yankelovich Monitor (Hume in Schiffrin et al. 2017).

Given the news media’s self-perception of its role as an essential participant in a well-functioning democracy, a body that holds governments to account and exists as a platform for the free flow of ideas and information, journalists worried about this decline in confidence and what it would mean for US democracy (Elvestad, Phillips, and Feuerstein 2018). They wondered how changes in practice would affect trust in media (Grosser, Hase, and Wintterlin 2019). The impact of sensationalism in boosting ratings and the roles of radio commentators, television pundits, and televangelists were likely to affect how audiences understood the news and thought about the problems of the day. The reliance on opinion rather than facts for agenda-setting was worrisome (Carlson 2017; McCartney 1997).

The political context of the 1970s and 1980s was relevant, too. Paid media operatives for Republican presidents Nixon, Reagan, and both Bushes charged the media with liberal bias in an effort to attack press credibility. Rupert Murdoch’s influence was considered to be pernicious. The left, including Noam Chomsky, accused the mainstream press of being pawns of corporate sponsors and expressing the ideology of elites (Herman and Chomsky 1988; Carey in Glasser 1999).

Foundations, scholars and journalists reimagined journalists’ role in public life. Public journalists wanted to re-energize civic participation, public discourse, and the media’s role in democracy by crossing the line from “objective observers” to participants who still hewed to professional journalism standards. Reporters went directly to residents in the communities they covered to ask them which topics they were most concerned about and wanted to see covered by the media. Groups of regional news media formed focus groups and conducted polls to understand local priorities (Rosen 1999). News outlets helped community members write their own opinion pieces (Conte 1996), let readers decide what the local paper covered, and published lists of community concerns. One of the hopes was to change people’s perception of the mainstream media as elites embedded with politicians and government workers (Fawzi 2019).

The public journalism movement was controversial at the time. Among other things, critics worried that forsaking objectivity would damage the media’s credibility and put journalists in the dangerous role of deciding the public agenda rather than reporting on it (Voakes 2004). Much derided by journalists (Voakes 1999), it eventually died (Ferrucci 2017). The rebirth of the movement came partly in response to the financial difficulties faced by US journalists. It was also fueled by the rise of digital technology.
that was, initially, meant to usher in an era of disintermediation. Another change has
been the branding and spread of the “solutions journalism” movement, which focuses
on training journalists and publishing articles that discuss solutions to thorny
problems (Rosenberg and Bornstein 2016; Ferrucci 2017).

Here we confine ourselves to noting the similarities between the public journalism
and community engagement movements. Most notable in this context are the
ambitious goals of both movements, the hope that these efforts would renew civic
e Engagement takes work. A real focus on audience is not a specific action, and it
cannot be achieved by hiring a social media editor. Community engagement in
journalism is about an attitude of embracing the role the audience plays in
consuming and contributing to journalist processes and products. For that
attitude to spread through a news organization requires a clearly stated mission,
strategies that align with that mission, and a plan to assess progress. (Mayer
2011, p. 82)

Many, but not all, of the community engagement activities were a normal part of
journalism practice in the pre-internet era, when building relationships was an
essential part of newsgathering. Journalists showed up in person and developed
sources to gather information (Gans 1979; Donohew 1967). As financial difficulties
worsened and staffing was cut, journalists were no longer able to spend as much time
in communities speaking to audiences and became more cut off from the people they
were, in theory, serving (Starkman 2010; Davies 2009).

Thus, some of the new practitioners, but not all, frame their activities as going back to
the “old way” of doing things. “The engagement space that everyone is now
collaborating in, I think of it more as, “This is how we used to be—which is reporters spending time in the community,” said Jesse Hardman in an interview. “Some people think of it as a separate universe and it has grown since the elections” (Hardman 2019).

However, in interviews with journalists overseas carried out by the author and colleagues, many suggested they were doing something new. Indeed, one difference between the two movements is the technology used, with mobile and digital tools used to communicate with audiences now instead of exclusively in-person interactions (Schiffrin et al. 2017). This has allowed both community engagement practitioners and fact-checkers to hold live events and offer online courses. There has been extensive writing on the different digital tools that promote engagement with audiences. Online commenting was an early form of engagement, and a number of other tools have developed since, including Hearken, which helps newsrooms gather input from their audiences; the BBC’s Your News; and the Knight Foundation Community Information Toolkit (Glaisyer 2019; Nelson 2018). But to practitioners, the process of spending time in the communities is essential.

You can get people and communities to pay for a platform, but you’ll have no idea how long it will last. People can invest in that, for a few hundred or a few thousand dollars a month, but unless you are doing ground-level organizing, spreading out, being on the ground, you are still limiting who your audience will be. Just texting a lot of people won’t build trust. It’s because I showed up at your community meeting, or church, or grocery store. If you aren’t spending time in the community you are limiting the possibilities of what you can do. (Hardman 2019)

**Intellectual support for community engagement efforts**

In the post-2016 period, donor support for community engagement as a solution to the trust problem rose, and even though it was not fully backed up by research showing its effectiveness, by 2018-19 it had garnered intellectual support from heavyweight academics in the communications and journalism fields (Lewis 2019).

Seth Lewis, who had carried out rigorous research on civic journalism, had by 2019 explicitly linked the problem of trust in media with the problem of authoritarianism. He singled out Hearken for praise and called for a new relationship with the public.

Broad distrust of journalism, while neither as recent nor as historically exceptional as people believe, matters at this moment because of what it signals about the institutional weakness of the press, particularly at a time when
journalism’s normative functions for society are needed amid a global uptick in authoritarian tendencies. A radical rethinking of journalism is in order. One place to begin is to consider the potential for a relational style of journalism—a journalism that puts the building and maintaining of relationships with publics it normatively serves at the center of its work....” (Lewis 2019, p. 44)

In a special 2019 issue of the prestigious journal Journalism, several well-known voices argued that rebuilding trust with audiences is essential to the survival and effectiveness of journalism. Noting that because of polarization in the US the public has been replaced by angry “niche-oriented groups susceptible to commercial and political manipulation that encourage people to assume binary interests and pick sides to protect discursive territory,” the University of Wisconsin’s Susan Robinson argued that journalists need to focus on developing audience trust:

How can we convince people it behooves them to be a part of shared public discourses for the sake of collective governing? Journalists who strive for neutrality represent a unique actor in information flows, have the networks to connect mass groups, and yet also hover above niche interests themselves. I propose a networked-based response that re-conceptualizes the journalist/news-citizen/community relationship, moving away from institutional and organizational versions of trust and toward a reconfiguring of the individual journalist–citizen association. (Robinson 2019)

Robinson’s suggested responses include reclaiming the narrative through films such as Spotlight, which came out in 2015 and portrayed the Boston Globe journalists who, under the guidance of editor Marty Baron, tirelessly worked to expose the extent of Catholic priest’ abuse of children. Robinson also proposed building individual relationships: “Remind citizens that journalists are individuals too. Re-envision the brand as pinpricks of relationship-building throughout the media landscape via individual actors who convene disparate groups into deliberation” (2019, p. 58). Robinson also advised journalists to enable citizen journalism and find sources of revenue that are not dependent on advertising.

In the same issue, Katherine Fink of Pace University also urged journalists to spend more time with audiences.

What to do? My suggestion for building trust is modest, low-tech, and unoriginal: journalists should have more conversations with strangers. That is, once a week, every journalist should meet someone new. Go out for coffee, or ice cream, or whatever. It should be face-to-face, because in-person
conversations are better for building trust (Nilsson and Mattes, 2015). Journalists should choose people who have never been sources or are likely to become sources due to their jobs, expertise, or social prominence. Journalists could, for instance, pick random people from their news organization’s list of subscribers. Even better, journalists should seek out news avoiders (Toff and Nielsen, 2018)—their trust in media is obviously so low that they have decided news is not worth their money or time. News avoiders are the ones with whom journalists truly need to ‘engage’. These conversations with strangers should be informal and off-the-record—no notes or recordings allowed. Journalists should aim to talk as little as possible and not try to steer the conversations in self-serving directions. If their conversation partners want to talk about news, fine—but journalists should not see these conversations as means to stories, quotes, or sources. That means resisting the journalistic impulse to interject ‘that would be a great story!’ even when conversation partners say something that truly would be. Journalists should instead focus on active listening (Charon 2001) in order to build trust. (Fink 2019, p. 42)

**Funding for community engagement efforts**

Funding for community engagements has dwarfed that spent on fact-checking (author interviews) though it’s hard to know how much donors gave, often because their own accounting doesn’t have a category for community engagement or because the term is a catch-all that includes other things besides efforts by media outlets. Based on our interviews, we can say that Omidyar committed at least $1.6 million in 2015 alone (Omidyar 2015) and between 2017 and 2018, OSF awarded $7 million to 65 grants related to community engagement with media and media outreach. As mentioned earlier, part of the 2017 awards from the News Integrity Initiative spent $1.8 million on efforts to build trust in news (Owen 2017).

One example of where the funding went can be found in the response of Internews to the 2016 crisis. Internews is a media-development NGO, mostly funded by USAID, which was founded in 1982 by David Hoffman and has done media development work, including community engagement, in more than 70 countries (Hoffman 2013). The 501c-3 non-profit traditionally has worked in low-income and post-conflict situations (Hoffman 2013). After 2016, Internews staffers found that it became easier to raise funds as donors had developed an interest in US news ecosystems. Internews quickly raised nearly $1 million for US engagement programs, including from community foundations, which supported work in Fresno and New Orleans. Internews spent some $350,000 in 2018 on community engagement in the
US and planned to spend between $400,000 and $500,000 in 2019 (author Interview, May 2019).

Interviews with Internews staff support our earlier point that, faced with the 2016 crisis, many players decided to keep doing what they were doing or just do more of it. Jeanne Bourgault, Internews’ president and CEO, read the events of 2016 as confirmation of the work that the organization had been doing globally for decades. A core belief of Internews is that the news media is a vital part of strengthening democracy and that it’s important to focus on the supply side of providing meaningful local news and engaging with audiences to give them information they want and can use. After 2016, Bourgault and colleagues felt that US donors began to understand what has long been recognized in many places where Internews works:

Our feeling generally was to double down on what we have been doing for years. In 2016, all the issues we had been working on for years came home. So now we have an opportunity to talk about solutions in many more places in the US as well as the countries we had been working in. We started going to many more conferences and our talking points were, ‘This is what we have seen overseas in many places. And now it has come home.’ (author interview Bourgault 2019)

Global efforts at building trust through engagement

In the post-2016 period, hopes that community engagement would boost trust in the media could be found internationally as well as in the US. Looking at how engagement efforts are understood and carried out is part of understanding whether they work, so we present here some relevant findings from a 2017 study “Bridging the Gap”, led by this author, of 17 organizations around the world seeking to build media trust. These were mostly niche, non-profits, funded by donors (Schiffrin et. al. 2017) Consistent with much of the research on US community engagement, the efforts of the international groups we studied included spending time with audiences and showing them the inside of newsrooms, going out to meet members of the public, getting audience members to help find and analyze information, training them in reporting skills, and asking them to become members or to contribute to crowd-funding. The organizations we studied were transparent about their finances and gave audiences a voice in editorial decisions. They mostly used mobile to communicate with audiences and half of the organizations we surveyed said they respond to comments online.

14 A complete list of the organizations and the study is available
The organizations that we profiled were mostly, but not all, news organizations that had added a community engagement piece to their journalism work. Journalism was their core activity and their perspective as journalists colored how they viewed their community engagement activities. Of course, small, niche news organizations have always been engaged with their audiences. The difference may be the technology and financing problems that ruptured the traditional patterns of engagement. As we mentioned elsewhere, because of the funding crisis that hit local news, the number of reporters fell dramatically in many places and so journalists had less contact with their audiences. The move online, the rise of aggregation websites and dissemination by Facebook also attenuated the relationship of audiences to journalists. Commenting and other forms of online engagement replaced earlier face-to-face contact.

This trend was observable in the organizations we studied that used digital tools to enable a new kind of relationship. They still spent time with audiences, but the constant online feedback and the speed with which journalists replied marked a difference from the pre-internet days (Schiffrin et al. 2017). Engagement between individual writers and readers took place on social media or in comments under published articles. Several outlets, including Krautreporter (Germany), Južne Vesti (Serbia), and Chequeado (Argentina), described this interaction as important to their newsroom process. They said they try to write prompt and personal responses to readers who relay their thoughts and constructive criticism, and consider replying to be a crucial component of their work. They mostly said they didn’t respond to trolling.

**Political Climate**

Many news organizations we studied described their community engagement in relation to the difficult political climates in which they work. The respondents explicitly said their efforts would rebuild trust and help shore up democracy. They also articulated the view that they were responding to a global turn towards authoritarianism and that re-shaping newsroom practices was a necessary response. Nearly all of the people we interviewed discussed the broader political context and expressed the hope that their efforts at building trust would counter-balance government attacks on media and freedom of expression.

Broadly, the groups we researched believe that producing quality journalism isn’t enough. It must be accompanied by efforts to build trust. This is a dramatic shift from an earlier view that journalists’ chief responsibility is to report the news and implies a major shift in the conception of what journalists’ roles should be. In this sense it’s part of a continuing realignment as journalists and foundations develop a different
conception of journalism (Powers 2018). The organizations’ donors see efforts to build media trust as part of shoring up free societies in places where they are under attack around the world.

The rise of strongmen and authoritarian regimes in many countries where the organizations work forced some of the outlets to reshape their strategies, sometimes by boosting their efforts to engage with their audiences and demonstrate their credibility and trustworthiness. While the circumstances differed from country to country, our research highlighted some points in common: falling trust, government criticism of the media, and the capture of local news outlets by business-people close to their leaders.

As Ilona Moricz, director of the nonprofit Center for Independent Journalism in Hungary, noted, “The general mistrust and uncertainty can be quite hard to bear. I think there is a general mistrust of media and it’s not new.” This sentiment, which she thinks took root in the past decade, has been strengthened by tremendous changes in media ownership. “Twenty years ago, we had predominantly foreign investors,” Moricz says. “Now the majority of ownership is Hungarian and is not media investment but is largely linked to the political elite. Some organizations have started to publish fake news—not just propaganda but deliberate fake news, which is more damaging…. It’s difficult to fight against trends that are centrally supported” (Schiffrin 2017, p. 12).

Room for political debate has also been shrinking in South Africa, where local media has been captured by business elites. The journalists we spoke to cited renewed pressure on journalists after a series of deeply polarizing corruption scandals, including “Guptagate” of 2015, which exposed connections between former President Jacob Zuma and a wealthy Indian family. “It’s scary and frightening for everybody in South Africa because the space has been taken over by a bunch of corrupt people, and they might do anything to stay in power, a bit like America,” one South African journalist told us.

**Traditional journalists’ mind-set colored their community engagement practices**

We observed that the practices of the groups undertaking community engagement efforts, and their measurements of success, were affected by the fact that these were journalists doing the work. For these journalists, publishing credible news on important subjects and holding government to account were key elements of trust building. By doing this, the journalists felt they were modeling good journalistic
standards in societies where these had been eroded and were under attack. While they spoke of the need to publish stories relevant to the daily lives of their readers, they still viewed accountability reporting as essential to their credibility, saying that simply by existing as a watchdog they would make a difference. This, of course, is a classical journalistic view (Coronel 2009). In some cases, they took the view that putting out good content was an essential part of building trust. In cases where self-censorship is common, these outlets encouraged investigative reporting. Two Serbian outlets we interviewed, *Južne Vesti* and *KRIK*, said that they often collaborate on in-depth and data-heavy reporting, especially when investigating government corruption. Covering corruption was seen as a way of restoring trust even if it didn’t have policy impact. Predrag Blagojevič, editor-in-chief of the independent Serbian website *Južne Vesti*, said he did not think that the site’s reporting is having much systemic impact on corruption. Even so, *Južne Vesti* reporters continue to investigate graft, particularly on a local level, and Blagojevič said it was important to continue holding government accountable. Blagojevič believes that reporting the truth, even if it contradicts local media reports, eventually wins over readers. “You have to be persistent in publishing the truth, you can’t lie to people all the time. When they see you publishing the truth, they’ll be able to see the lies in other places. After that it is much easier, people become your correspondent, see you as their protector” (Schiffrin et al. 2017, p. 61).

Nigerian news site *Premium Times* sees its mission of promoting social justice as its own way of fighting the trend of falling media independence. Cofounder Dapo Olorunyomi said trust in Nigerian media has declined because two overlapping crises, of revenue and of ethics, have hampered its effectiveness as a government watchdog. Declining circulation and waning income for many mainstream publications have helped create a media ecosystem dominated by state-sponsored news. As a result, Olorunyomi says, many people have come to believe that traditional media are “in bed with administration and politicians,” leaving readers to suspect that they are not getting all the facts (Schiffrin et al. 2017, p. 13).

Other outlets, too, such as the South African *GroundUp*, said they hoped that publishing factual information would buoy standards across the profession. Nathan Geffen, *GroundUp’s* founder, said the organization is resisting calls to make its reporting more partisan. “We try to be reasonable and we don’t let our politics get in the way of what stories we cover. We are strict about sticking to the ‘rules’ of journalism. We don’t take cheap shots and avoid editorializing. We try to do news reporting whether we are in a sane world or an insane world because that’s the kind of journalism we believe in,” he says (Schiffrin et al. 2017, p. 13).
The interviewees drew a direct connection between attracting a dedicated audience with deep knowledge of their coverage and building their readers’ confidence. Although many were at best skeptical about the general state of public trust in media in their societies, they prided themselves on their content being trusted and often promoted by people who knew it well.

On the political scrutiny and accountability situation – over the last twenty years local journalism in the UK has been marred by monopolisation, closures and asset stripping leading to media deserts spreading across the UK. Many communities have lost a vital source of public accountability and democratic scrutiny.

Trust is bandied around as an important concept. But I think that restoring interest is also important. And that’s our responsibility as reporters and editors, to make local affairs engaging and communicate their relevance and situate them in a wider context. It’s a long process. People have switched off from local news. If you are getting news from a Facebook feed and you are presented with news of a local news or an international disaster you may be more likely to click on the latter (Aviram 2019).

The groups we researched had made a commitment to transparency, including discussing their editorial decisions and story selection. Being journalists, they navigated between two priorities: publishing important stories even if they might draw fewer page views than other more entertaining pieces, and publishing stories that mattered to their readers. They portrayed themselves as doing in-depth accountability reporting, but from their websites it was clear that they tried to post content that would attract audiences.

**Citizen reporting and interacting with audiences**

Other outlets use readers and members of the public as sources in order to build trust. One example: In Dortmund, Germany, *Correctiv* teamed up with a local newspaper and set up a “Crowd Newsroom” to crowdsourcing information on the adverse effects of teacher shortages in the area, such as lesson cancellations in local schools. Thanks to numerous contributions from parents and pupils, it found that children ended up missing far more classes than the ministry of education had acknowledged.

Many of those we interviewed hold events as a way to build a closer connection with readers. In addition to engaging with its audience through social media, for example, *Correctiv* organizes dozens of readings, lectures, and exhibitions across Germany every year, even opening a café in Essen with weekly meetings and lectures. To reach
younger followers, it teamed up with a university in Dortmund for a festival of journalism and new media held in September, 2017, and launched online journalism tutorials and a collaboration with German community colleges to promote citizen journalism.

*Južne Vesti* had plans (eventually unfulfilled) to travel to cities and towns throughout southern Serbia, inviting readers to meet with reporters for coffee. At the time we spoke to editor Lina Ejelat, her outlet 7iber.com was raising funding for a similar project, where its writers will host panels, lectures and talks in local communities to share their stories. In Zimbabwe, the founder of *263Chat* gives lectures and addresses church groups, and in Sweden journalists often interact with community members in formal and informal ways, giving talks and making themselves available for conversations.

A tenet of the community engagement movement is that journalists need to reach across the divide to underserved communities. The international organizations we spoke to emphasized attempts to bring in low-income, working-class groups and, in some cases, political activists.

Ejelat and colleagues tried to mentor numerous citizen reporters. Many of its writers come from activist backgrounds, and while they are deeply connected to their local communities, they often need editing and other help to allow their reporting to resonate. The site sees this mentorship as a crucial part of its mission. “We offer tools to professionalize their writing, make sure they know how to cover a demonstration or how to ask a relevant government body for a comment,” its editor says (Schiffrin et al. 2017, p. 36).

*GroundUp* in Capetown has pursued a similar course. Founder Nathan Geffen says, “We ran workshops with political activists from working-class backgrounds who wanted to become reporters and that was where our initial intake came from. They have less experience and may not have great writing skills, but they bring a certain insight that other news organizations are missing. Now we take a lot of students from journalism school.”

*Bristol Cable* filmed a documentary about inequality in a working-class neighborhood and then offered a series of events around it. “Instead of leaving it there, we worked with our events and community engagement to hold events, screen the film, and [hold] round-table discussions with stakeholders from the community and issues affecting them” (Aviram interview with author, May 2019). Others, like the Center for Independent Journalism in Hungary, offer workshops in data reporting to local
journalists. Chequeado runs online courses on fact-checking techniques, along with classes on more specific topics such as gender and climate change.

Many of these efforts at personal reader engagement are labor-intensive and time-consuming. But the people carrying them out felt that they were at the heart of engaging with their audiences.

Handling corrections

There is overlap between the engagement efforts we studied and the fact-checking movement, observable in the handling of corrections. The groups we spoke with didn’t just believe in being open about their finances but also in the way they handled mistakes and corrections. They discussed the process and said that handling corrections transparently built trust within their audiences. They viewed the internet as a free-for-all where errors, rumors, and disinformation run rampant, and some respondents said they built trust by handling corrections of factual mistakes in an open and professional way. Bristol Cable and Chequeado said further that they apologize for inaccuracies they register, printing lengthy explanations of how factual errors came to be made. Their staffers believe that making public and detailed corrections builds trust. 263Chat founder Nigel Mugamu apologized to readers for a photo that was in poor taste and said that owning up to mistakes improves his readers’ confidence. “I quickly apologize. It’s very important for credibility,” he said.

For instance, on July 18, 2017 Bristol Cable reported a 53 percent increase in the number of council employees employed on the highest salary bands. Two days later they published a correction (The Bristol Cable 2017) which was unusually detailed and clear in that it didn’t just fix the mistake but clearly explained how the mistake was made. The outlet’s founder described that article’s response:

“Our correction got more traffic than the initial story. People really, really appreciated it. We were amazed at how much support we got. People wrote in and said this is why I support the Bristol Cable—because they don’t bury it and are honest about their mistakes” (Schiffrin et al. 2017, p. 43).

Measuring impact

Using journalism standards of impact

The community engagement movement has been criticized for not having rigorous metrics as to what constitutes impact (Ferrucci 2017; Nelson 2018). This was borne out by the groups we interviewed, many of which said they didn’t measure impact, although one group does keep a spreadsheet logging how many people it contacts and
how many come to its events. One major funder of engagement movements said it didn’t have a set of metrics as to what would constitute success. “We were not scientific about it. I did ask grantees how they would know if it was working or not,” said the donor.

The groups we surveyed tended to measure impact in terms familiar to journalists, including having their stories picked up by mainstream media and getting government to act (Hamilton 2016; Protess et al. 1992). They considered changing conditions or policies as a result of their reporting to be a sign of success. They did not have clear criteria for what constituted success in terms of audience. Mostly they were niche organizations with limited audience engagement.

GroundUp’s Nathan Geffen argued that amplification by other outlets was a sign of success: “We don’t publish a story unless it will get republished. We only publish if we think it will be picked up by [a big outlet] like News24, Daily Maverick and TimesLive.”

He also believed—as did others we interviewed—that getting a response from government signaled impact. “We have stories that have impact all the time. In our early days [2012, 2013] we reported on the fact that the biggest township in Capetown didn’t have street lights, and eventually the city listened and switched on the street lights.” Bristol Cable spent three years covering labor exploitation at a local ice-cream parlor. In May, 2019 a “slavery and trafficking risk order” was issued, effectively freeing the workers and putting the ice-cream parlor out of business (Aviram 2019; BBC 2019a).

Indirect impact

Modeling ethical norms

We concluded that the community engagement projects we examined would remain small and would not be able to scale up their efforts in a major way. Although the organizations we surveyed were committed, energetic, and creative, we predicted that they would not become very large or influential. We saw them as more likely to remain niche organizations serving targeted communities, either due to funding constraints or because of outside factors. In Africa, Internet usage is concentrated in the cities. In other cases, it is because the content appeals to a local audience by design. Južne Vesti, for instance, was created to concentrate on southern Serbia and the city of Niš in particular. Editor-in-chief Blagojević said that publishing more international and national news would probably bring in a larger audience, but that the team was not considering this route. These considerations were echoed by Bristol
Cable, which is committed to serving communities in its namesake city, as well as by many US efforts that focus on small areas (Hardman 2019).

Educating the next generation

Many of the organizations we spoke with believe that education is key to having an impact, and so they work to educate the younger generation through mentoring, training, and online journalism courses. For example, Južne Vesti, which had struggled to find reporters in Serbia who could meet its high benchmark for accuracy and rigor, in 2017 hosted an intensive writing workshop for young reporters, with funding from the Serbian Ministry of Information. "We don’t want to be the only ones writing quality, we want healthy competition," Blagojević said. Premium Times, in partnership with the Ford Foundation, conducted an internship program for journalism students at eight of Nigeria’s public universities. The program put 500 university students through a three-day training on investigative reporting, data reporting, and media management, and set them up with mentors.

Because many of the outlets are donor-funded non-profits in difficult markets, they often don’t last more than a few years. As these outlets close or morph into new ones, their founders move around. Added to normal job turnover, this means that journalists working at one outlet bring with them knowledge and expertise when they move to another. In our research we came across several examples of well-known journalists changing organizations and/or launching new ones and, thus, hiring and training the next generation of journalists. For example, several journalists at Nigeria’s Premium Times previously reported for Next234, a now-defunct Nigerian newspaper founded by Dele Olojede, a Pulitzer Prize-winning writer. This kind of cross-pollination between small outlets may have the effect of strengthening independent journalism in their regions even if the individual outlets do not scale (Schiffrin et al. 2017, p. 36).

Research on the effectiveness of engagement is scant

More generally, the question of how to measure the effectiveness of the community engagement movement has bedeviled many of its practitioners (Mayer 2011). Impact is always difficult to measure, and the people doing community engagement work understand this. In her master’s thesis of 2011 Joy Mayer, who is still involved in the field, wrote:

I asked all the journalists I interviewed what they considered to be signs of success in their engagement work. Many said they had no idea. They were proceeding with what felt right, without any accountability for or measures of
success. Some shared benchmarks that are hard for working journalists to track, such as more civil online conversations, more community contributions, more connections with audience and more social media attention. Some are looking at numbers for how many people attend an event, download a resource from the website or comment on a story. Because the engagement culture is so new, several journalists said they felt in the dark as to what success would look like. (Mayer 2011, p. 78)

These difficulties mean that even now there is very little literature available on the effectiveness of the community engagement movement since 2016. Further, it is unclear how the movement might be studied as a whole, since the goals of the different outlets are disparate and there is so much variation across countries. Finding a unifying set of metrics is difficult. Perhaps because of these difficulties, research on the effects of community engagement projects tends to be case-based, focused on a single outlet’s efforts, mostly in the US (Xiao 2019; Schmidt and Lawrence 2019) and not overtly analytical. This was true, too, of the public journalism movement (Nichols et al. 2006, p. 77).

The most rigorous work to date on the community engagement movement has been done by Jacob Nelson and co-authors. His in-depth look at Hearken, a Chicago-based organization that builds community engagement tools and processes, highlights a disconnect between the “idealized” view of the audience and the reality encountered by journalists working to build engagement.

We find a tension between engaged journalism’s conceptualization of ordinary citizens, and the experiences of journalists working to engage them collaboratively in the news. Perhaps the more interesting challenges stem from the tension between engaged journalism’s perception of ordinary citizens, and the reaction of those citizens to actual attempts at engagement. As our findings reveal, the number of citizens willing and able to take advantage of all the engagement opportunities Hearken-equipped newsrooms had to offer was often smaller than the journalists within those newsrooms originally expected. This might mean that “ordinary citizens” by and large are less interested in participating in the news-making process than engaged journalism advocates believe (Schmidt, Nelson, and Lawrence 2019).

The findings are consistent with this author’s findings that engagement efforts tended to fall short of original hopes for engagement with a broad community. Audiences that got involved with media outlets tended to be narrower, more urban, and more prosperous than originally intended by the founders of these engagement efforts.
Part II. Demand-side solutions

(Schiffrin et al. 2017). This was true not just of audience engagement efforts but of some long investigative reporting pieces. This disconnect between the efforts of the journalists and the interests of their audiences could be observed in the disappointing response to many of the donor-funded investigative research projects that failed to garner large audiences in parts of Africa where they appeared (Schiffrin and Powell 2019; author interviews). Whether it’s necessary to have broad reach is unclear. It may be that the traditional approach of agenda-setting for elites is sufficient in some cases, but that is not what the community engagement community believes. Their core belief is the necessity of engaging the public.

What can we learn from the research on public journalism?

When trying to understand how to think about the effectiveness of contemporary efforts to engage with community members and audiences, it may be helpful to look back to the civic/public journalism movements of the 1990s (Rosen 1999; Örnebring 2009; Heflin 2015; Voakes 2004). However, it must be noted that this literature did not study whether public trust in media rose as a result of public journalism efforts.

Voakes’ review of the literature on public journalism notes that “there is no unifying theory or definition, let alone a set of practical guidelines” and that “most of the research has explored whether citizens’ knowledge or attitudes about a public issue or political campaign changed after a public journalism project.” The results are mixed, although one study by Friedland found that public journalism can “intensify connections among networks in the community, sometimes in a way that leads to new connections” (Voakes 2004, p. 32).

Nichols et al. (2006) did a meta-analysis of hundreds of projects and ran regression analyses to understand which story frames, project features, and organizational factors affected which audience responses. The authors began by noting that there was not sufficient research and that the available literature was often polemical or based on case studies. The authors undertook to study the effect of the public journalism movement on three broad subject areas: “1) organization of newsrooms and their effects on individual journalists’ values, norms and behavior; 2) links between public journalism efforts and changes in news content, framing and sourcing; and 3) public journalism’s impact on electoral knowledge and behavior, and on citizen participation in real life.”

The third point is most relevant to our inquiry. Noting that the research on the effect of on civic and political life was “partial and incomplete,” the authors said that “on the whole, studies support a moderate effect of civic electoral coverage in the areas of
voter awareness of issues and traditional forms of political participation…. Although untangling political cause and effect is difficult, there is a case for public journalism efforts and broader public engagement” (Nichols et al. 2006, p. 80).

Similarly, other researchers found correlations between the existence of public journalism projects and increased civic understanding or even participation (Ardèvol-Abreu and Hooker 2017). Trust in citizen journalism also correlated with active political participation (Kaufhold, Valenzuela, and Zúñiga 2010). But there is no relationship between trust and civic engagement (Uslaner 2000).

While the body of literature on public journalism is inconclusive in some ways, better meta-studies based on this literature could still be useful for those seeking to understand better the determinants of media trust.

A literature review of all the research on digitally enabled audience participation concluded with a recommendation for more longitudinal studies as well as more research on non-elite audiences (Engelke 2019). It is to be expected that in coming years more research will surface that tries to determine community engagement journalism’s effectiveness.

**Conclusion**

The idea that fact-checking and accuracy can build trust in media is plausible. If recognizing that something is a lie makes people not trust it, then the opposite—affirming that something is true—should help shore up or restore trust. Moreover, the fact-checking and engagement efforts are laudable and the journalists and media development practitioners who carry them out are working hard to realize the lofty ideal that journalism contributes to democracy, political engagement, and citizen participation and creates a culture of truth-telling and transparency. By improving the supply of quality information, they hope that demand will shift as well. Given the role of habit in news consumption there may be some truth to this (Lee and Carpini 2010). As audiences become accustomed to an idea or develop a habit of reading true information and trusting journalism, they become loyal. Certainly, there is evidence of the virtuous circle that Pippa Norris has written about and the idea that news consumption builds trust (Norris 2000; Mourão et al. 2018; Strömbäck et al. 2016). There are worries that fact-checking can amplify disinformation that doesn’t mean it shouldn’t happen. It just means that the way that fact checks, or corrected information, are presented needs to be done carefully.

Nelson argues that there is a lack of understanding as to why audiences avoid political news, whether due to a lack of trust or lack of interest. Without understanding the
reasons, if journalists just “go with their gut” in trying to address this avoidance the community engagement movements may not be effective (Nelson 2018). Given the research findings that media exposure can indeed promote civic engagement and political knowledge, then it’s essential for the practitioners to act in an informed way. Nelson’s research is central this dissertation’s focus on strategies to combat mis/disinformation online. Even so, we are left with two problems. Is there evidence that engagement with the community will build trust in media or in the political system? If so, then can it realistically serve as a counterweight to mis/disinformation online? Even if community engagement were able to build trust and civic engagement more generally, it’s not clear that this would boost media trust. Even if it did, it’s hard to see how this niche, labor-intensive, movement could expand enough to address the problem of trust in online mis/disinformation. Efforts at building trust, laudable though they are, may be incidental, rather than central to fixing the problem (Schudson 2019; Voakes 2004, p. 29).

After describing efforts in the post-2016 period to build trust in media, one donor concluded:

Media can’t change political realities that have a hundred causes that are not necessarily media. Because of Cambridge Analytica everyone decided that fixing social media could save democracy, but it couldn’t, though it could play a role…The problem was much deeper. People didn’t vote for Trump just because they were manipulated. The disinformation probably helped but it was not the real reason. We were reacting to the fever not to the illness (Ronderos 2019).

Perhaps the answer lies elsewhere: Even if the audience for fact-checking and community engagement is small, and even if an even smaller part of that audience is persuaded by it enough to trust the media and become more engaged and civic-minded, it may still be worth doing. Defending a culture of truth is important, as is signaling that journalism plays a role in keeping public officials and institutions accountable. Showing audiences that journalists are part of their community is surely not a bad thing. Being optimistic about your audience, and hoping they can learn to prefer quality news, and getting them involved in news gathering, is a good thing, too.

Referring to the doctored video of Nancy Pelosi that President Trump circulated in late May, 2019 after it appeared on Fox News, Hardman said, “I don’t know how you compete with this stuff aside from saying, “There is this fountain of information that doesn’t do that and you need it.” These efforts to provide information that responds to the needs of communities and audiences, and give people news that meets their immediate needs, will likely remain small-scale in the absence of large amounts of
funding. But that is all the more reason to scale these efforts. In his 2020 book *Democracy Without Journalism*, Victor Pickard proposes a framework for funding public media in the US and the group Free Press has been active in this field too. The belief that creating a supply of quality information will change consumption and possibly reduce political polarization and demand for politically polarized news has not been fully tested. It’s still worth pursuing as the current financial crisis of news organizations requires bold, large responses. Local news and local relationships are crucial part of the media eco system and will need to be supported as part of the efforts against online mis/disinformation.
Chapter 4

Fact-checking

“At the end of the day we have to put accurate information in the public domain, so when voters go to the ballot box and vote, they are basing their decisions on information that is accurate. That is important for democracy. And we are part of the process. That is what it’s all about—people going to the ballot box and making a decision that is informed by fact. I don’t know if it’s working but if you look around the world, there is a huge demand for factchecking. People are finding it harder than ever to work out what is correct and what is not, and the fact-checking organizations are filling a gap.” —Sushi Das of the ABC/RMIT fact-checking organization.

In the post-2016 period, fact-checking efforts expanded dramatically around the world and became a fix for Facebook, which reached for it as a way to ameliorate the problems it and other platforms had created (Mosseri 2016). On the face of it, the idea of Facebook funding fact-checking seemed absurd to some (Levin 2018a). After disseminating hundreds of millions of falsehoods globally, paying journalists paltry sums to debunk a few thousand of them defied logic, critics argued. Yet the allure of truth and facts persists. Like the transparency movement and other activist movements worldwide, fact-checking is based on the idea that telling the truth and getting facts on the record has power.
In the classic work on the impact of investigative reporting, *Journalism of Outrage*, the authors argued that the impact of investigative journalism can unfold over time and in three phases: individualistic, deliberative, and then substantive (Protess et al. 1992). In other words, investigative reporting’s impact on an individual can be transformed into a broader discussion about how to address the problems revealed by journalists, and then finally into a policy change. The same may be said about fact-checking. Fact-checkers hope that revealing accurate information will change the ideas held by their audience. They hope it will change public conversation (Nyhan and Reifler 2012) or what’s often thought of as agenda-setting or “priming” (Rogerson 2013). More broadly, the foundation of journalists’ work—and that of fact-checkers—is the assumption that if the truth is revealed people will act, governments will respond, civil society will be galvanized, and policies will change (Protess et al. 1992; Sen 1982).

Many of the proposed solutions to the problem of online mis/disinformation involve exposing and describing the problem in the hope that some entity will then take action to solve it. This is true, of course, of many global transparency movements (Florini 2007; Besley and Burgess 2001; Sen 1982) as well as much of the academic literature that is based on the assumption that providing information causes behavioral change. The difficulty here is that the entities that ultimately need to take responsibility are the tech platforms. In the case of Facebook, after a piece of content is debunked, it then needs to be downranked or flagged by Facebook (Mosseri 2016). Yet the fact-checking projects funded by Facebook lack transparency (Levin 2018a). Thus, there is an inherent uncertainty implicit in solutions that rest on providing information in order to provoke action: Whether information will indeed provoke action is unknown.

This is because either the information is unknowable—has someone’s mind really changed?—or because platforms don’t reveal information about actions they’ve taken. Under a 2018 European agreement, the tech companies make monthly reports to the EU Commission about actions they’ve taken to stop the spread of false content ahead of elections, but these lack detail and don’t explain what was done in response to debunking or corrections from fact-checking groups. There is no direct line drawn between flags or verifications provided by fact-checking groups and the response of the platforms. Facebook says that fact-checking causes it to lower the rankings of disinformation shown on the feed (Lyons 2019), but provides no information as to how often or when or how quickly, making it impossible to verify. Some of Facebook’s fact-checking partners have also been criticized for a lack of transparency (Leetaru 2019). The lack of clarity as to how Facebook responds to fact-checking surfaced in a number of our interviews, as people involved in
partnerships and projects with Facebook were unable to explain whether or how Facebook responded to the flags and corrections that their organizations had submitted.

Given this uncertainty, why would an organization proceed with fact-checking? As mentioned earlier, there are three reasons for many of the solutions proposed to address the mis/disinformation problem.

1. Financial incentives. Facebook’s investment in fact-checking has been small (Madrigal 2019), but the company garnered good publicity in the beginning and created the feeling that it was addressing the problem of mis/disinformation online. Facebook’s “war room” featured in multiple articles that portrayed the platform as taking action to fix the problem of online mis/disinformation (Frenkel and Isaac 2018; Scott 2019). The good publicity and the appearance that it was making an effort may also have forestalled regulation, which would be costly and time-consuming for Facebook.

2. Ideology. The default perspective of many in Silicon Valley is libertarianism. While they are mostly socially liberal, they eschew paying corporate taxes, are anti-union, and dismissive of government. Tech titans generally don’t believe in regulation, especially of content. The major tech platforms and companies spent $55 million in 2018 on lobbying in the US. Further, Mark Zuckerberg has said repeatedly that Facebook is just a platform and not a publisher or a media house (Segreti 2016; Thompson and Vogelstein 2019). The platform executives have maintained this position even as they curate information and adjust their news feeds.

3. The expertise of the fact-checking groups. Journalists believe in fact-checking and know how to fact-check. And people do what they know how to do (Nieminen and Rapeli 2018).

In this chapter we will first provide background on the fact-checking movement, then discuss how it was funded in the post-2016 period and highlight some findings from our interviews with international fact-checking groups. We conclude with a summary of some of the research into the effectiveness of fact-checking.

**Background on the fact-checking movement**

In the period after 2016, fact-checkers found themselves again in the limelight. While many like to say that they have a simple, humble, “Just the facts, Ma’am” mission to provide accurate information (Skelton 2019), funding for fact-checking increased
globally after 2016 in part because news organizations hoped that emphasizing accuracy would restore trust in media. In order to regain trust, journalists would need to double down on their role as society’s fact-finding resource and reclaim their social position as providers of accurate, fact-based information. Many fact-checkers avoid using the word “truth,” as it’s considered too loaded, and prefer the term “fact-based” (FactCheck.org; Godler and Reich 2013).

Sushi Das, chief of staff at RMIT ABC Fact Check in Melbourne, noted, “We generally we have a rule here we don’t talk about the truth, we talk about information that is accurate, information that is inaccurate. Misleading information is like a hose. We can’t stop it, but we can’t sit here and not do anything. We need to raise awareness as much as we can” (Das, author interview 2019).

Fact-checking is also viewed by journalists as a means to distinguish themselves from others in the profession (Graves 2018). By highlighting their professionalism and commitment to accuracy, fact-checking journalists signal they believe in providing quality information and are committed to the process of truth-telling.

Core principles of the fact-checking movement include transparency, verification, and dissemination of correct information. They hope that fact-checking serves as an antidote to online disinformation (Wenzel 2019), will help rebuild trust in journalism (Graves 2018, pages 622-3), and reaffirm the value of evidence and scientific method, enlightenment values, and truth in the face of a deluge of mis/disinformation online (Graves 2016, page 74). In some ways the fact-checking movement can be viewed as reinforcing a culture of truth, in which citizens need to be informed so they can make sound decisions. Without reliable information there can be no trust (Arendt 1974; McInerny 1984).

**Fact-checking historically a part of US magazine culture**

Historically, fact-checking was done in-house before the publication of an article and was practiced more at magazines than at newspapers. At newspapers and newswires, the editor or copydesk often pointed out mistakes in copy. The culture of fact-checking as a separate job developed at magazines such as *The New Yorker* and *The Nation*. These magazines were known for the thoroughness of their fact-checking and often employed interns fresh out of college or junior staffers for whom fact-checking was a first step in their career. Beginning as a fact-checker served to inculcate the values of verification and truth-telling early in a career. There is some literature and memoirs about fact-checking, including the memorable Jay McInerney novel *Bright Lights, Big City*, whose hero is a fact-checker (McInerney 1984). The 2018
Broadway play *The Lifespan of a Fact*, based on an earlier book by the same name, portrayed the tensions inherent in the relationship between fact-checker and journalist when a single-minded fact-checker finds numerous errors of fact in a long magazine piece written by a famous, older journalist who embellished the truth to make his article more interesting. The tensions between the two characters drive much of the drama of the play and, ultimately, the editor steps in to mediate the conflict (Kareken et al. 2018).

**Fact-checking organizations today**

Today, some outlets still have in-house fact-checking to vet information before it is published. But the more recent fact-checking movement debunks or verifies information *after* it is published. Some outlets stand alone, such as FactCheck.org, PolitiFact, the UK’s Fullfact.org,\(^\text{15}\) while others form partnerships with media outlets, such as *The Washington Post*’s “Fact Checker,” Decodex in France, and the Australian Broadcasting Corporation’s fact-checking partnership with RMIT University. These fact-checking organizations typically try to expose and correct factually inaccurate statements from public officials, although some, particularly those funded by Facebook, fact-check content found online. Some of the statements to be checked are provided by Facebook to their designated partners and some are chosen by the organizations themselves.

Many of the fact-checkers are former journalists and the movement views itself as part of journalism (Graves 2016). Some organizations pride themselves on close ties to academia (Graves 2016) and many have links to universities, such as Africa Check, which is based at the University of the Witwatersrand in Johannesburg, and the ABC/RMIT fact-checking joint venture, which is based on the RMIT campus. Some groups also do live fact-checking, such as monitoring a political debate or major speech and then trying to produce real-time verification of public statements by politicians (ABC/RMIT and Verificado interviews).

Around 2017-18, donors began to look at the explosion of mis/disinformation about subjects related to public health. The return of polio and measles in many parts of the world focused attention on the role of YouTube and Facebook in spreading misinformation. Accordingly, the TED foundation, Gates, and OSF began funding research on the phenomenon and projects to combat attacks on public health (author correspondence and interviews, 2019). The spread of Covid-19 in 2020 also led to many new efforts to combat health misinformation.

---

\(^\text{15}\) Fullfact.org discloses its funding from Google and Facebook [https://fullfact.org/about/funding/](https://fullfact.org/about/funding/)
Decisions as to how to select which misinformation to fact-check is a key question in the movement and one that is often tendentious. Given the infinite amount of false information online (and in the world), fact-checkers are confined to examining only a miniscule part of it. In his definitive book on the US fact-checking movement, *Deciding What's True: The Rise of Political Fact-Checking in American Journalism*, Lucas Graves explains that, given their limited resources, fact-checking organizations can only verify some information, so they focus on public statements by public officials, often in a prepared speech rather than an off-the-cuff remark, and emphasize subjects that are important to regular citizens. Subjects thought to have broad public interest, such as the state of the economy or declarations by the President about health care policies or education, are more likely to be scrutinized than statements on more obscure topics by less visible officials (Graves 2016).

Graves notes that the growth of standalone fact-checking organizations (as opposed to editing desks within news outlets) can be viewed in the context of ongoing “mediatization,” in which many spheres are influenced by the language and conventions of the media and politicians use media strategies even as they try to “go around the media”. In this case, fact-checking is also a response to the way the internet eroded the gate-keeping role of the journalist, thus confusing audiences which then had to cope with vast amounts of unfiltered information (kindle location 3983). In this confused state, turning to the authority of the fact-checker could be a response. Again, this echoes the argument made by Walter Lippmann and Jacques Ellul that the role of journalists and other experts is to “become a means to sift through and distill the excessive information now available to the masses” (Schiffrin 2018).

Expanding on the general idea that fact-checking is a modern reform movement akin to the US Progressive Era of the early 20th century, in which a subordinate culture is unwilling to adhere to hegemonic power, Amazeen explains why it has expanded in recent years. She gives three reasons: The first is the rise of partisan media coupled with a decline in quality journalism, with he-said/she-said reporting substituting for in-depth coverage. A consequence of this, and Amazeen’s second point, is a feeling of public disempowerment and the belief that it’s become harder for journalists to hold government to account. Finally, new technologies have created so much online information that not only is there confusion as to whether it’s true, but audiences tend to self-select and fall into filter bubbles (Amazeen 2018a, page 9).
The growth of funding for fact-checking

This chapter adds to the existing literature explaining the rise of fact-checking by arguing that examining funding sources is important, particularly since donor agendas can affect the initiatives taken by nonprofit news organizations as well as NGOs more generally (Benson 2018; Scott, Bunce and Wright 2017; Callahan 2017; Schiffrin 2017; Ferrucci and Nelson 2020). Thus, it is impossible to understand the growth of fact-checking initiatives—especially their international growth—without understanding the donors funding them. In the heat of the moment after 2016, donors felt they needed to do something, and in many cases the something included funding fact-checking efforts.

There was a marked rise in overall funding after 2016, from foundation and government sources as well as platforms, especially if you count support for conferences, research into fact-checking, etc. One new source that’s had a big impact is Facebook’s third-party fact-checking program, which pays fact-checkers directly depending on the amount of work they do on the platform, sometimes as much as low tens of thousands of dollars US per month. Meanwhile Google has been supporting fact-checkers and the International Fact-Checking Network through its Digital News Project since about 2015. (Graves 2019)

One of the funders of fact-checking organizations shared his thinking as to how the decision was taken to begin that funding. Tom Glaisyer of the Democracy Fund had been interested in the “digital public square” since 2005. He underwrote some of the research on fact-checking and that research was seminal in the decision by Luminate to fund fact-checking. Of key importance was a 2015 paper by Brendan Nyhan and Jason Reifler, which found that US legislators who were informed by letter before the 2012 elections that they were being scrutinized “were substantially less likely to receive a negative fact-checking rating or to have their accuracy questioned publicly, suggesting that fact-checking can reduce inaccuracy when it poses a salient threat…a promising sign for journalistic monitoring in developing countries” (Nyhan and Reifler 2015). This support for the “deterrent effect” of fact-checking resembles the traditional journalism role of “scarecrow,” i.e. that its very existence can have a deterrent effect (Stencel 2015; Anderson, Bell, and Shirky 2015).

The research was intended to “interrogate” whether fact-checking could “work” and the results suggested three effects:
Part II. Demand-side solutions

- Changing people’s minds. To provide an effective counterweight to misinformation about any given issue, fact-checking must present the relevant facts to the right audiences in a way that encourages them to question misleading claims.
- Changing journalism. A goal of many fact-checkers—and especially of media critics who fact-check news reports—is to encourage journalists not just to report competing claims but to assess them, and to challenge politicians who attempt to mislead the public.
- Changing the conversation. By exposing political deception, fact-checkers and journalists in general may exert pressure on political figures to retreat from misleading claims—and perhaps discourage them from making such claims in future (Nyhan and Reifler 2012).

Funding for this study was provided by the Democracy Fund and encouraged Luminate to proceed with supporting fact-checking (author interview with donor, April 2019). In 2017-2018, Omidyar spent $4.5 million on fact-checking initiatives such as Africa Check, Full Fact, the International Fact-Checking Network at the Poynter Institute, and Chequeado. Some of the grants will end in 2021 (author correspondence). The Open Society Foundation also stepped up funding of fact-checking. Generally, grants from the Open Society Foundation average about $100,000 and projects that mention fact-checking received about $3-4 million each year in 2017, 2018, and 2019 (author correspondence with OSF, 2019). This likely makes OSF the biggest foundation funder of fact-checking.

The role of Facebook

The decision-making process at Facebook is less clear. Their spokespeople don’t reveal their funding decisions and the company’s decisions are communicated through press releases, which they describe as blog posts and are posted on their “Newsroom.” They declined to speak to this author. Critics view Facebook’s efforts on fact-checking as a public relations strategy aimed at inducing other organizations to clean up the mess that Facebook created. Funding fact-checking and efforts to support journalism is an attractive option for social media platforms because it takes the onus off them and places it on journalists and consumers (Bell 2016).

One estimate pegs the amount that Facebook has spent on support for fact-checking organizations in the single-digit millions, to approximately 34 partners worldwide (Madrigal 2019). Organizations contacted individually by the author said they received...
annual payments of $100,000 to $250,000 each to carry out their projects, while Snopes received $406,000 in 2018 before severing ties with Facebook in early 2019 (Green and Mikkelson 2019). Commented one interviewee: “Overall, the money being spent in this space is miniscule” (author correspondence). By mid-2019 Facebook had partnered with 52 fact checking sites from 33 countries, but the company does not release figure as to what it spends and requires it partners to sign Non-Disclosure Agreements (author interviews and correspondence).

Fact-checking around the world

At this writing there are an estimated 227 fact-checking organizations globally (Duke Reporters Lab). Some 65 are members of the International Fact-Checking Network, which has a “code of principles” and holds annual fact-checking summits (Poynter).

What had been thought of as mostly a US phenomenon (Graves 2017) has expanded in the past decade. In 2015, the Poynter Institute for Media Studies founded the International Fact-Checking Network with funds from the Arthur M. Blank Family Foundation, the Duke Reporters’ Lab, the National Endowment for Democracy, the Omidyar Network, and the Park Foundation (Poynter 2019). Funding from Facebook as well as OSF, Luminate, Gates Foundation, and the Google Digital News Initiative helped fact-checking expand globally in the post-2016 period.

Fact-checking organizations around the world share some qualities with their US counterparts, including a commitment to transparency as to where they get their information (though not completely detailing where they get their funding), non-partisanship, and engagement with the public. They also believe that restoring confidence in the reliability of media is essential to its financial survival. They say that fact-checking and truth-telling in politics are essential to stop politicians from lying. They are less convinced that they can solve all the problems of democracy, entirely restore trust in institutions, or stop online mis/disinformation.

In this section of the paper we give some examples of fact-checking sites from around the world, discuss their practices, and assess their work. Interviews with some of the fact-checking organizations illuminate their perception that providing accurate information contributes to democracy. The global fact-checking movement has also inculcated shared values and is another manifestation of the recent trend toward collaboration between organizations that was exemplified by the work of the International Consortium of Investigative Journalists on the Panama Papers (Wardle et al. 2019). Our interviews highlight fact-checkers’ commitment to helping
the public be better informed and uncertainty about the impact of their work, and the complexity of their relationship to the social media platforms. The story of Comprova’s experience ahead of the 2018 elections is telling.

In the fall of 2018, 24 Brazilian news organizations got together to debunk misinformation on social media ahead of the November 2018 elections in a project called Comprova. Funded by Google News Digital Initiative and Facebook, the outlets collaborated on verification/debunking and then disseminated their findings online and through their media outlets as well as through paid online advertisements. In doing so, they hoped to shore up a culture of truth in a society that had been inundated by toxic mis/disinformation and which had had multiple demonstrations in recent years over politics.

Comments by one Brazilian fact-checker support the argument running through this dissertation that in many cases organizations chose solutions to online mis/disinformation based on their feeling that action had to be taken and they would keep doing what they knew how to do: “Since we don’t know the solution [to the problem of online disinformation] we still have that feeling that something has to be done and this is the answer we have now. We know how to do it. It may be silly but if, because of my job, one person didn’t believe in something fake then I did something for the country” (author interview, Débora Ely, May 2019).

Acting was better than not acting and fact-checking might work, or at least wouldn’t cause any harm.

The feeling I had—that everyone shared—was the feeling that misinformation was a huge concern in the elections. In May 2018 we had a lot of protests in Brazil and there was a lot of misinformation about a lot of things. People were going to the supermarkets to buy a lot of food because people were saying the food [supplies] would end because of the trucker strike. So, we had experience with misinformation in Brazil and WhatsApp is way more popular in Brazil than in the US. So, everybody had this feeling that this had already happened and if it had happened before the elections then what would happen during the elections? Everyone had the feeling it would be a massive misinformation campaign and the feeling we all shared was that something had to be done…and the journalists felt we had to do something because we deal with information. We felt something would happen and it was our job to do something about it.
A study of the Comprova effort reported that it was a resounding success, for several reasons described later in this chapter, but noted that the attacks on journalists doing the fact-checking were demoralizing. Indeed, this came across in our interview:

The idea was to create and encourage people to go after trustworthy information, but I don’t think it happened.

In the beginning we could see people were engaged and sharing and writing to us and saying thank you….in the end we were all depressed. The debunking didn’t change anything. What changed was the right-wing got even more mad about the media. Like in the US with Trump. It just gave them more energy to attack the media. We got attacked by right-wing people who said, ‘We are a free country, who is Folha or Veja or Facebook or Google to say what is true or what is not true?’ They called us Communists—and Mark Zuckerberg, too.

I would do it again. Despite everything. Because we don’t know yet the solution [to online mis/disinformation is] and since we don’t know the solution we still have that feeling that something has to be done and this is the answer we have now. We know how to do it. It may be silly but if one person didn’t believe in something fake because of my job then I did something for the country. I am a believer.

The June 2019 study of Comprova released by First Draft, singled out two outcomes: the raising of public awareness about Comprova’s work and the strong ties built between journalists who worked together on the project.

However, the report highlighted a shortcoming: Comprova didn’t have access to WhatsApp, which is widely used in Brazil. Comprova was therefore not privy to what was being published in private groups and was unable to verify or debunk the rumors circulating privately during the 2018 elections on WhatsApp (Wardle et al. 2019). “There is so much out there. It’s a blizzard of information and finding the stuff that is resonating is difficult,” said one fact checker (Skelton interview, June 2019).

This lack of access is consistent with the dilemma fact-checkers generally face in choosing their focus. They worry about which statements to verify and whether they should focus online or on statements by public officials. Some do both. Groups that scrutinize statements by public officials are aware that much mis/disinformation exists online and in conversations held in closed WhatsApp groups (author interviews). This fear of being in the wrong place and of not debunking important pieces of disinformation surfaced in several of the cases we studied.
The fact-checking groups’ choice of partners also affected the impact of their work. Groups funded by Facebook spend more time debunking images and video on social media than on statements by public officials. This is transforming the work of fact-checking organizations and making them more like groups that specialize in verification, such as Bellingcat and Storyful.

“Their work has shifted over to the misleading online content Facebook is paying them to debunk, which is a lot of shitty memes and videos. There is a crisis in the field. We went into this because we care about truth and facts. So the money has shifted the field,” said one interviewee (author interview, anonymous, May 2019).

“Because Comprova wasn't an IFCN signatory, Facebook didn't do anything with our findings. We used claim review so they should have got automatically picked up via Google as a 'fact-check'. Our big frustration is that the work of the coalition did not have greater impact on the platforms. We asked for it during the very first meetings but they kept saying, ‘We can't promise anything,’” said one participant in the Brazilian effort.

It may be that Facebook only responds to its partners’ debunking efforts, because it views this as part of quality control. Or it may be a form of digital capture (Nechushtai 2017; Schiffrin ed. 2020) as Facebook wants to attract fact-checking groups into its fold and one way to do this is holding forth the promise that their work will affect Facebook’s rankings.

The need to be part of the Facebook universe is a reason why one fact-checking group we spoke to is considering joining the Facebook partnership. “We want the tools to search Facebook and the audience and to be in that space...There is so much out there. It’s a blizzard of information and finding the stuff that is resonating is difficult,” one of its members said.

At the same time, he expressed ambivalence about the contradiction inherent in Facebook’s fact-checking efforts. Upon being told about the process for debunking false claims online, the interviewee reported, “Facebook said they would lower the offending posts’ visibility in feeds, that groups sharing it would be warned it was false and they would stop it from being monetized. The fundamental question is, why wouldn’t you take it down?”
Different fact checking organizations that we discuss

**Australia ABC/RMIT** is jointly funded by Australia’s public broadcaster ABC and the RMIT university in Melbourne. It was launched in 2013, but the government did not renew its funding. The organization was revived in 2016 after RMIT gave it a home and research support.

From the mission statement on their website:

“In our impossibly polarised world, facts are now submerged in the mud of fake news, self-serving spin, misinformation and good old-fashioned fearmongering. RMIT ABC Fact Check determines the accuracy of claims by politicians, public figures, advocacy groups and institutions engaged in the public debate.”

Its staff of 10 tests and adjudicates the accuracy of claims and then publishes its findings in multiple formats, including memes and livestreaming of verifications. During the 2019 elections, the ABC/RMIT team worked during the “leaders’ debates” of the two candidates for prime minister and transcribed claims as they came through. “We tried to marry the claims with fact-checks we had already done and on the night we produced a documented story. It got so many views. People were hungry for that live fact-checking. They were keen to learn and they wanted the debate to be fact-checked right there and then,” said Sushi Das, chief of staff.

“At the end of the day we have to put accurate information in the public domain, so when voters go to the ballot box and vote, they are basing their decisions on information that is accurate. That is important for democracy. And we are part of the process. That is what it’s all about—people going to the ballot box and making a decision that is informed by fact. I don’t know if it’s working but if you look around the world, there is a huge demand for fact-checking. People are finding it harder than ever to work out what is correct and what is not, and the fact-checking organizations are filling a gap,” Das said.

“Australia is not as bad as the US and Europe, because we are a small market and the scope to monetize fake news is not there. Some of the worst excesses we’ve seen in Europe and the US have not appeared here. Nothing hair-raising like you get in the US,” said founder Russell Skelton.

**Chequeado (Argentina)**

Launch date: 2010.
Executive director: Laura Zommer, former journalist.
Activities: Funding: Relies on foundation grants to operate and in 2018 accepted $100,000 from Facebook so as to expand the LatAm Chequea network through fellowships and other activities in Latin America.
Activities: Verifies statements made by public officials across the political spectrum. Holds live events, including during the President’s annual address to the parliament. Staffers and volunteers, including academics and journalists, come together to check the president’s factual statements in real time. Chequeado has shared information on how to run live events of this kind with The Guardian and PolitiFact as well as media organizations in Peru (Ojo Público), Brazil (Lupa) and Mexico (Animal Político).

In 2019 Chequeado launched Reverso, along with AFP Factual, First Draft, and Pop-Up Newsroom. Reverso will have more than 95 media and technology companies working together on verifications to be published simultaneously by all partners. All groups will need to sign on to consortium principles and participate in one of the 12 trainings that Reverso will offer in the summer of 2019 in 11 different provinces in Argentina as well as in Buenos Aires. You can find more information in www.reversoar.com. It will start publishing on June 11th.

Youth Outreach: Chequeado staffers collaborate with Model United Nations organizers in Argentina to incorporate fact-checking into conference proceedings and are developing a separate Instagram account for young people, where they hope teenagers will produce posts.

They are working on Medio Flash, a site made by and for teenagers. It is an experiment that aims to bring data-literacy to young people in their own language. It will include workshops in schools in order to make learning easier. Launched their Impact and research work in 2015 in order to study the effectiveness of their efforts.

Do you think you are helping Argentine democracy? (Zommer 2019) I believe that we are. There is always more to do and never seems to be enough, but I believe we are walking in the right direction. This year, the first big step we took was to join in one consortium many media of both sides of the polarization. They might have different perspectives, but we can all agree when information is false. In the same page, our education program works to bring data literacy to more people every year, the innovation program develops new tools to make fact-checking easier not only for us, but for everyone who wants to check.

Africa Check (South Africa), launched 2012.

- The first South African website to focus exclusively on fact–checking.
- A small team dedicated to fact-checking viral claims about South Africa (i.e., its economic and social indicators) as well as claims submitted by readers for verification.
Mis- and Disinformation online

- Funded by foundations, including Open Society, Omidyar, Bill and Melinda Gates, and Shuttleworth.
- Has grown to include regional offices in Lagos, Nigeria; Nairobi, Kenya; and Dakar, Senegal.
- From their website: founded in 2012 with a grant from the International Press Institute, in 2018 had about $1.34 million in funding from Shuttleworth, Bill and Melinda Gates, Omidyar Network.

Verificado: Mexico during the elections

Outlets, including Animal Político, AJ + Español, Pop-Up Newsroom came together during the 2018 elections to collaborate on fact-checking with partners in 28 states and spread information broadly across platforms (Owen).

Opened a WhatsApp line where people could send in questions and then receive an individual answer (Owen 2018).

“A huge success because it brought people together across the political divide,” said Maria Teresa Ronderos, former director of the Open Society Foundation’s Program on Independent Journalism.

The role of Facebook and the tensions in that role

As with several of the initiatives described in this dissertation, getting Facebook and other platforms to act is central to making sure the efforts actually work. Since the platforms are largely unregulated, governments have little leverage. So, it’s not just a question of pressuring Facebook to take action and providing them with information, but also, as with many of the solutions discussed in this dissertation, finding steps the platforms are willing to take. Accordingly, governments, foundations, journalists, and NGOs spend a great deal of time trying to find ways to work with Facebook.

Luminate and Poynter held a meeting before the US elections in 2016 to help Facebook meet some of the fact-checking community. Before that, Luminate had helped set up a partnership between Facebook and the International Fact-Checking Network. “Without that, they [Facebook] would have had no-one to partner with,” said one donor involved in the effort (author interview with participant in 2019, anonymous).
Lack of information

Larger questions about Facebook’s intentions and actions remain unanswered. As a matter of policy, Facebook avoids giving interviews or speaking on the record, so we relied on published accounts from former Facebook staffers, news stories, Facebook’s statements on its website and testimony at government hearings, and our own interviews. Unanswered questions that need further research include why Facebook waited so long to verify/debunk information on its platform. Years before the 2016 elections, Facebook and Twitter could have verified or debunked the top 10 most-viral items. Nor is it clear why Facebook and Google needed to partner with fact-checking organizations at all. Presumably they could have done their own debunking without forming partnerships.

Myriad explanations have been offered as to why Facebook didn’t respond earlier, instead manifesting precautionary behavior. In keeping with our analysis, we would point to Facebook’s ideology: They believed in the service they offered and dismissed the importance of mis/disinformation online (McNamee 2018; Thompson and Vogelstein 2019). As well, Facebook had a financial interest at stake, as admitting they had responsibility could open them up to vulnerability on Section 230 (Thompson and Vogelstein 2019). One fact-checker with close ties to Facebook dismissed the platform’s work, launched in 2016. “It was 100 percent a PR move and they've kept scaling it for more PR around the world, but it is a fundamentally flawed project. They know it internally, but they don't know what an alternative is” (author interview, anonymous).

Whatever Facebook’s true motives, after the Trump election in 2016 it began working with some fact-checking organizations around the world, sending them posts that needed to be checked. Given the volume of information on the platform, only a few items could be selected for further study by fact-checkers. Facebook’s criteria included information that was widely circulated on Facebook itself.

Decisions about what to fact-check

In a 2019 Tow Center report, Michael Ananny described the process:

Concretely, the partnership centers around managing a flow of stories that may be considered false. Here’s how it works: Through a proprietary process that mixes algorithmic and human intervention, Facebook identifies candidate stories; these stories are then served to the five news and fact-checking partners through a partners-only dashboard that ranks stories according to popularity. Partners independently choose stories from the dashboard, do their usual fact-checking
work, and append their fact-checks to the stories’ entries in the dashboards. Facebook uses these fact-checks to adjust whether and how it shows potentially false stories to its users. (Annany 2018)

The choice of subjects to verify has become more complicated by the involvement of Facebook in fact-checking, notes Ananny in a 2018 report for the Tow Center.

Not only has Facebook become a gatekeeper of traffic and a monopolistic source for stories, but it was seen as potentially limiting the complexity and significance of fact-checking that fell outside of Facebook’s prioritizations of misinformation. Indeed, one fact-checking partner wished it could drive fact-checks to Facebook, making Facebook users see what they considered to be were important rebuttals, instead of having to wait for Facebook to drive traffic to the partners.

Beyond this particular partnership, the point stands: if Facebook creates entirely new, immensely powerful, and utterly private fact-checking partnerships with ostensibly public-spirited news organizations, it becomes virtually impossible to know in whose interests and according to which dynamics our public communication systems are operating. (Ananny 2018)

One fact-checking group had been offered $AU 800,000 a year but was reluctant to become a Facebook partner. “We can’t accept Facebook money because how can you take their money and claim to be objective and disinterested and reporting objectively? There is a potential for conflict of interest,” said one interviewee.

This multiplicity of purposes and lack of clarity as to the process and goals for fact-checking, as well as the contradictions inherent in the relationship, were, unsurprisingly, manifested in many fact-checkers’ relationships with Facebook. By 2018, some of these had begun to fray (Levin 2018a). The prestigious outlet Snopes formally broke off the relationship in 2019 (Levin 2019).

**Voluntary Code of Practice**

By 2019, Facebook, Google, and Twitter had signed onto a voluntary partnership with the EU Commission (2019), the Code of Practice on Disinformation, and were submitting monthly reports about their activities. These were sketchy and provided few details. The reports mostly offered information about political advertising and takedowns of false accounts, rather than the impact of debunking.

“We don’t have detailed information about the number of news items that they have demoted. The numbers we have are the numbers written in the
Part II. Demand-side solutions

reports that the platforms provided us during the close monitoring of the implementation of the Code of Practice on disinformation. No more than that. We are always asking for more details and numbers in order to allow an independent assessment of the effectiveness of the actions that are being taken. This is something to develop further.

Further work is needed. We are happy because in terms of transparency of political ads and closure of fake accounts they have done something and we can see what they have done. It remains more difficult for us and for independent assessors to concretely assess the impact of their actions. We ask them for more information on that.” (Alberto Rabbachin, Programme Office, DG CONNECT, Social Media and Media Convergence Unit).

Facebook has said on their blog that they use a mix of human intelligence, third-party fact-checking agencies, and artificial intelligence to help them determine what is false and ensure that false information is repressed (made less visible) in people’s new feeds (Facebook 2018). “In our past experience, once a story is rated as false, we've been able to reduce its distribution by more than 80 percent” (author correspondence, anonymous 2019)

Other accounts by journalists (Silverman 2019a) have come out showing that fake accounts still abound on Facebook, with some reporting that even when they are pointed out, Facebook doesn’t take action (Silverman 2018). It may be that Facebook is using (Lyons 2018) debunking by fact-checkers to lower the rankings on the newsfeed, but it’s not clear precisely or how often (Hautala 2018).

In the absence of detailed information from the companies as to how they process and act on the debunks they receive, we can look at the following examples of Facebook’s behavior in order to learn more about their perspective.

As we were researching and writing this chapter, the doctored video that made Democratic House Speaker Nancy Pelosi appear to be drunk and slurring her words circulated widely online, after being aired on Fox news in the US and tweeted about by President Donald Trump on May 24, 2019 (Wagner 2019). Facebook was urged to remove it but declined, saying that the company “does not require posts to be true” (Kang 2019). In early 2020, they declined to remove a doctored video that made it appear that Pelosi was tearing up Trump’s speech after Trump honored several people at the State of the Union.

In March of 2019, the killing of 50 people at a New Zealand mosque was livestreamed on Facebook, then copied and shared at least 1.5 million times within 24 hours
Mis- and Disinformation online

(O’Sullivan 2019). Although Facebook said it would consider video restrictions (Wong 2019), as of May 2019, videos of the event continued to circulate on both Facebook and Instagram (O’Sullivan 2019).

These two examples suggest that even when mis/disinformation is flagged to Facebook, the platform may not take action. Or, as in the case of the New Zealand shooting, the company may not have the ability to remove the vast quantity of posts that do not meet its standards.

Research on fact-checking

In the introduction and the literature review, we discussed the difficulty of measuring the impact of many of the solutions discussed in this dissertation. That uncertainty has contributed to the fragmentation of the avenues pursued. As well as the methodological problems of attribution, decentralization across countries and the secrecy of the platforms contribute to the difficulty of measuring impact. We also discussed the feeling, conveyed in our interviews, that post-2016 action had to be taken quickly—which meant that ideas were pursued, in many cases, without much study.

Unsurprisingly, then, much of the funding and growth of international fact-checking was carried out with little knowledge of its effects. While support from journalists, fact-checking practitioners, media outlets, and the platforms helped the fact-checking movement expand overseas, it was less clear whether or not it would have an impact or what kind of impact it might have. Much of the expansion was done seemingly without recourse to academic literature. This section discusses some of the research on fact-checking, much of it published after the funding increase for fact-checking began.

A 2018 content analysis of academic research on fact-checking found that research fell into three broad categories: 1) the effects of fact-checking; 2) fact-checking as a profession; and 3) public opinion about fact-checking (Nieminen and Rapeli 2018). The effectiveness of fact-checking as a tool for fighting mis/disinformation online is most relevant to this dissertation so we will focus only on this subject. The authors found 24 studies, 14 of which were surveys and two of which were lab experiments that used US students as participants. All had been released between 2011 and 2018. A content analysis of the studies suggested, “A significant question for fact-checking research concerns whether it can correct false beliefs. Studies offer mixed results...the contrasting nature of results, however, suggests that corrections are ineffective, especially when a person is motivated to believe the untrue statement [...] it seems
possible that fact-checking could have an impact on political knowledge and candidate evaluations” (Nieminin and Rapeli 2018 p. 7).

**Limitations to fact-checking**

Since funders’ post-2016 decision to support fact-checking initiatives, more studies have been done as to their efficacy. There are several limitations to the fact-checking process as a way of fighting mis/disinformation online.

The first is that far more false information is available online than can ever be fact-checked by small teams of fact-checkers around the world. The second is that fact-checks are not necessarily seen by the people exposed to the false information in the first place (Riley 2019; Amazeen 2018a, p. 12; Nyhan and Reifler 2015b). A third is that fact-checking has been found to have very little agenda-setting effect on how digital news covers events (Vargo et al. 2018). Also, because it takes a long time to verify information or prove that it’s incorrect, false information can circulate while the verification process is underway (Graves 2016). During that time, repeated exposure to the false information may persuade audiences that it’s true (Zajonc 1968) and it may enter the lexicon and conventional wisdom even if it’s later disproved (Benkler, Faris and Roberts 2018). It may then continue to circulate.

The far-right’s allegations that Hillary Clinton was corrupt, for instance, ended up migrating to the legacy media and entering the public consciousness. In their 2018 book, *Network Propaganda*, Benkler, Faris, and Roberts detail this and several other examples of how false information that initially surfaced online and on Fox News ended up in *The New York Times* and other legacy newspapers. Another example of mis/disinformation that has entered public consciousness in the US is the false notion that Mexican immigrants to the US are criminals, when in fact they commit fewer crimes as a percentage of the population than US citizens do (Flagg 2019). In this case, accusations by Donald Trump during his candidacy resonated in part because they spread online (Jamieson 2018). In Europe, too, press coverage about migrants, as well as on television, has resulted in worsening attitudes towards migrants even when the coverage is favorable (Benesch et al. 2019).

Social psychology research describes the many elements that go into personal beliefs, including perceptions as to what larger groups think (Allen 1969, 1971). One aspect of group identity is political affiliation (Cohen 2003), and whether people are persuaded by corrections often depends on their political beliefs. Indeed, providing accurate information does not actually make a difference when people are partisan (Bullock 2006). Trump supporters often say they don’t mind whether he has a fact wrong
because he is speaking a larger truth or telling their story. This is borne out by findings that Trump supporters may not withdraw their support for the candidate even when presented with correct information about his false statements. Nyhan and Porter report:

...[O]ur results provide compelling evidence that citizens can accept the conclusions of journalistic fact-checks of misstatements even when they are made by one’s preferred candidate during a presidential election. This information had little effect on people’s attitudes toward the candidate being corrected, however. In other words, Trump supporters took fact-checks literally but not seriously enough to affect how they felt toward their preferred candidate.” (Nyhan and Porter et al. 2019)

Another study on corrections of political information found:

[Partisanship was the largest predictor of the political correction’s effectiveness. When correcting political misinformation, no other factors besides partisanship affected whether the type of correction format increased the likelihood that people understood the candidate’s statement was largely false. Neither interest in politics, dislike for ambiguity, close mindedness, or need for cognition made a difference [...] When a candidate was from the same party as the participant, they were less likely to say that the corrected statement was false. These results reinforce the difficulty of overcoming partisan-driven motivated reasoning. Across all of the models, people find fact-checkers more credible when they correct the opposition and less so when they correct one’s own party.” (Amazeen et al. 2015)

A study of rumor diffusion and correction on Twitter in the 2012 presidential election found that rumor-spreaders formed partisan groups which repeatedly circulated rumors about candidates they opposed even after the rumors had been debunked. Comparable groups spreading corrections did not form. The paper concluded, “Although Twitter has the potential to challenge the flow of misinformation, our analysis showed that the platform mainly served as a conduit for rumors to spread through partisan communities rather than as a self-correcting system involving a larger political discussion community” (Shin, Jian et al. 2017). A study by Facebook researchers found that while corrections can result in deletion of rumors, “Rumor cascades run deeper in the social network than reshare cascades...Furthermore, large cascades are able to accumulate hundreds of Snopes comments while continuing to propagate...We showed how readily and rapidly information can be transmitted via social ties, even when it is of dubious veracity” (Friggeri et al. 2014).
Part II. Demand-side solutions

Some research suggests that those who do trust the media are responsive to changes in journalism practice, such as the publication of corrections and explanations as to where source material comes from (Karlsson, Clerwall, & Nord 2017). Those who do not trust the media may not be swayed by more transparency or by changes in journalism practice (Karlsson, Clerwall, & Nord 2017). Also inconclusive is the recent research on the effectiveness of debunking. It’s not clear that presenting new information to audiences actually helps correct misunderstandings. Contextual cues such as photos may undermine the presentation of factual information “by triggering decision-making heuristics that...prove detrimental to individuals’ ability to process novel political information” (Garrett, Nisbet and Lynch 2013). Depending on the subject, reminding audiences of something—even if it is to say it’s false—doesn’t work. Worse, it may even fix the information in the mind of the audience.

Scholars have noticed in some cases a phenomenon called the “backfire effect” (Schwarz et al. 2007; Nyhan and Reifler 2010; Schwarz et al. 2007; Nyhan and Reifler 2010), although this has become controversial as other studies have found no signs of such an effect (Wood and Porter 2019). Some studies have found that audiences may be open to factual corrections on subjects they feel less strongly about (Nyhan et al. 2019). In their paper on debunking, Chan et al. note that “the debunking effect was weaker when the debunking message simply labeled misinformation as incorrect rather than when it introduced corrective information. Contrary to expectations, however, the debunking effects of more detailed debunking messages did not translate into reduced misinformation persistence as the studies with detailed debunking might also have stronger misinformation persistence” (Chan et al. 2017). Political rumors distributed by email to friends and family are likely to be believed and forwarded when they are consistent with their political predispositions (Garrett 2011). But political fact-checking on Twitter seemed to discourage the sharing of rumors online when the corrections were made by someone familiar to the rumor-spreader (Margolin, Hannak, and Weber 2018).

**Indirect impact: conclusion**

It is evident that the number of debunks and verifications provided by the small and underfunded global fact-checking movement is small (Funke 2018). It is also unclear to what extent fact-checking has affected rankings on Facebook and its feed, in the absence of information provided by Facebook as to their impact. Where fact-checking does “work” is when there is a supply-side response: Facebook suppresses or flags information deemed to be false. It’s less certain what effect fact-checking/debunking has on the demand side. In this sense it may be that the main effect of fact-checking
on the platforms is that it serves to pressure them to take action. Further, since Facebook and Google pay for some fact-checking efforts they may have more of a commitment to using the information provided by fact-checkers. However, given that the amounts paid are so small for these giant companies, it’s not clear whether this would have more than a reminder effect. Without detailed information from the platforms, it’s difficult to know what they do with the information provided by fact-checkers. One study has found that engagement with fake news websites and fake news stories has fallen on Facebook since 2016 but risen on Twitter (Allcott, Gentzkow and Yu 2019; Resnick and Ovadya et al. 2018).

In the absence of a counterfactual, we cannot discount the importance of the signaling effect provided by fact-checking. It may help create a culture of truth and demonstrate support for a culture of accurate journalism. It may be that the major impact of fact-checking has not been to build trust or suppress online content that has been debunked and flagged, but to change the profession of journalism. In interviews, supporters of fact-checking say that the international alliances that have formed and the training that has taken place have taught journalists how to be accurate and gotten them to cooperate when they hadn’t previously (author interviews in May 2019 with Ronderos, Wardle, Bourgault).

Like the media-literacy and community-engagement movements, the fact-checking efforts have also trained journalists how to work with young people and brought them into contact with their audience and members of the public (Zommer 2019). The existence and idea of fact-checking have a signaling effect for society and serve as a reminder that facts matter and that truth exists. In some places, fact-checking has spurred politicians to become more accurate in their speeches (Skelton 2019; Nyhan and Reifler 2015a). As noted, in the media literacy chapter, it may also be that “pre-bunking” and “inoculation” are more effective than debunking.

One would hope that the supply of accurate information enabled by fact-checking would build audience trust in the media and institutions. In theory, practitioners believe this could be an effective way to combat mis/disinformation online. The evidence, however, is lacking. Given that the money being spent on fact-checking organizations is limited, and that fact-checking has important effects on the practice of journalism as well as serving a social function and as a morale booster for journalists, it’s likely that there will continue to be limited support for organizations that do fact-checking. Fact-checking will likely remain important for the journalism community and may serve as a deterrent to some politicians and a signal for broader society.
The demand side solutions are based on a hopeful view of audiences and the belief that, when confronted with questionable information, common sense and educated skepticism can prevail. But scaling those fixes is beyond the scope of what journalists and NGOs can do. The speed and volume of disinformation online makes it impossible to fact-check and debunk a significant part of what circulates. Human prejudices, beliefs and lack of quality information also affect what people absorb and trust. Reluctant though we are to take away human choice and censor information, it is the supply side solutions that can scale. Changing an algorithm to reduce circulation of a piece of mis/disinformation or bury it in some way is more efficient than spending hours teaching someone to think carefully before forwarding. Legal remedies, too, can provide incentives for the platforms to suppress more information more quickly or address the problem in other ways. Accordingly, the next part of this dissertation addresses these larger scale interventions. We begin by looking at prospects for start-ups using AI to identify mis/disinformation and we then turn to law and regulation.
Chapter 5

Using AI to filter false content

Introduction

As we saw in the earlier part of this dissertation, governments, foundations, and companies around the world are looking to address the problem of online mis/disinformation. Some emphasize the demand side and believe it’s important to focus on consumer behavior and use media literacy and fact-checking to remind audiences not to believe or forward mis/disinformation. Some try to raise the quality of journalism and support local news in the hope that creating more reliable content will counter-balance the mis/disinformation found online. Later in this dissertation we will discuss legal remedies such as platform liability laws; creation, implementation, and enforcement of online hate speech and defamation laws; and privacy protections. In short, there are myriad solutions aimed at addressing the problem of online mis/disinformation.

Many of these solutions require the platforms to take action. Fact-checkers need to know that once they have flagged content, Facebook will stop showing it to people. The tech giants also need to be coerced into complying with regulation. So, too, for several of the fixes examined in this chapter: Many of the initiatives rely on Facebook, Twitter and Google using their technology or taking action in some way, whether by blocking ‘harmful’ content or downranking false or mis/disinformation, or by keeping the advertising of reputable firms from appearing next to such content. Part of what makes this solution difficult is the companies’ lack of clear definitions as to what is
harmful, the instances when their standards differ from international standards and local laws, and the general lack of oversight or an appeals process.

In some ways the use of artificial intelligence (AI) is about the transmission part of the continuum. Rather than leave it to people to determine what is true and what is false, the use of AI to screen content and block it, or make it less available, is a fix that mostly addresses the transmission of mis/disinformation and, in some cases, hopes that lessening the transmission may reduce the supply. This approach conveniently sidesteps the question of audience trust because the disseminators of false information, the platforms, choose what to highlight to their audiences using both their algorithms and their human judgment. “This is the scariest option,” said one senior Facebook official, reflecting the company’s view that it’s not up to the platforms to curate the information that people see but to let it circulate so that people can choose. This idea that the platforms believe only in free speech and don’t want to get involved in gatekeeping, agenda-setting, or traditional journalistic functions conveniently ignores the fact that the platforms, or their algorithms, make decisions constantly about which information to highlight and circulate (Gillespie 2018).

There is a demand-side element too, as we shall see, because some of the tech fixes—such as those that use ratings and checkmarks to highlight accurate websites—are based on research suggesting that such prompts are an effective “nudge” to audiences (Sunstein 1999; Bryant 2019). There is a hope that the algorithmic, AI fixes will be easy ones (Jahromi 2019). In this scenario, the platforms will clean up the mess that they created and thus obviate the need for government regulators to take action. We’ll examine the incentives at play and whether it’s realistic to expect the platforms to use their power to screen and block mis/disinformation. It’s not likely that at this point the platforms could completely control the display of mis/disinformation even if they wanted to. They have created an unregulated market that is so big it’s now impossible to fully understand and monitor.

While many people would like to see the platforms make the problem go away and stop showing illegal, inflammatory, or inaccurate content, others worry about corporate censorship and the development of a vast surveillance and blocking network with no input from government or from NGOs (Solomon 2019). Groups such as Access Now, the Electronic Frontier Foundation, and the Committee to Protect Journalists are among those taking this view. Many of these groups receive funding from the platforms but that is not why they worry so much about what they call “corporate censorship.” One limitation of this chapter is that while we note this
One option for the platforms is to adopt the technology being developed by small start-ups. In this chapter we will examine companies in the information ecosystem that use Natural Language Processing (NLP) as well as human intelligence to identify and, in some cases, block false or inflammatory content online. Several are bootstrap operations while others live off a range of commercial activities. The firms we interviewed for this chapter have clear incentives: They seek scale and profit from their work. In this sense the objectives of the firms are narrower and the reasons for their pursuit are clearer than the groups trying to change hearts and minds through engagement, education, and persuasion.

We explain how the technology works and the various players in the field and delineate the impediments to the success of this entrepreneurial approach. In order for the firms’ technology to scale, their technology fixes will have to be adopted by Facebook, Google, Twitter, and YouTube. Of course, the platforms already screen, block, and remove fake accounts and content. They develop, copy, or occasionally acquire the AI needed for further screening and it’s not clear whether they will want the kinds of technology developed by the firms we surveyed. In the absence of widespread platform adoption, some firms are hoping that individual users can be persuaded to install browser extensions that help them identify trustworthy sites. User take-up of browser extensions is unlikely to provide the kind of scale that is needed.

We conclude by looking at ratings and standard-setting efforts to see whether these could scale more quickly than AI solutions. We suggest that another advantage to the movement for global standards and rating systems for news organizations is the hope that voluntary efforts will be easier to implement and less restrictive than laws that control online content. Finally, we explore whether controlling digital advertising revenues to certain players can assist in curbing the spread of mis/disinformation.

**How AI can detect false information online**

The use of artificial intelligence to detect false information online has been around since 2003 (Martin 2005). Computer scientists build data sets of incorrect information and then use NLP to see whether features in other pieces of content match those deemed to be false. A number of taxonomies have been published describing the different databases available and the pros and cons of each (Oshikawa, Qian, and
Wang 2018). Features that machine learning can look for and detect, or “extract,” include linguistic features, visual features, or social-context features (Shu et al. 2017; Burgoon et al. 2003; Newman et al. 2003 and Zhou et al. 2004 cited in Gravanis et al. 2019). Appeals to emotion, use of unfamiliar sources, and non-human patterns of disseminating information may all indicate that a piece of content is false. Shu et al. (2017) provide a taxonomy of the different kind of models used to determine veracity as well as what sort of characteristics are extracted. There are features of news content such as the source, headline, body text, and image/video. There are linguistic cues such as inflammatory language and visual images aimed at provoking anger or other emotional responses. Detecting these cues may indicate that a piece of content or website is false. Looking to see whether someone has been using a platform for a long time or has a lot of friends can help determine whether the user is a real person or a bot.

Another way to determine whether a piece of content is false is to examine its characteristics and how it circulates. Twitter’s acquisition of London-based Fabula AI in June 2019 suggested the value it places on this kind of technique (Lomas 2019). Fabula AI uses “geometric deep learning” in order to identify patterns in how information spreads. It does not look at the content of the messages, so the content’s language and encryption status do not affect the ability to detect whether it’s false. AI is being used to do rumor analysis, check rumor veracity, detect and screen fake accounts, and redirect people to reliable sources while shadow-banning objectionable content (Jahromi 2019; Alaphilippe et al. 2019). However, the tech companies don’t reveal what they do because they are afraid of revealing trade secrets. If they made their criteria public, they would become easier to game. Disclosure of algorithms, or their effects, is debated. In France it is required by law (comments to author from Henri Verdier, February 2020)

Three Technical impediments

Experts disagree on how effective artificial intelligence has proven at detecting false content. While some research has found that AI can detect false information with 95 percent accuracy (Gravanis et al. 2019; Shu et al. 2017), others note that human intelligence has to be involved in making the determination. There are several reasons why AI, on its own, can probably not detect all kinds of mis/disinformation online.

First, standards and practices are fragmented, making it difficult to acquire the complete datasets of misinformation needed to teach the machines how to detect it (Oshikawa, Qian, and Wang 2018). Presumably, as those detection efforts grow more
successful, the dataset of false information will shrink, making it more difficult to match it with suspect information.

Second, the creation and detection/prevention of online mis/disinformation is an ongoing arms race. While detection may work in a static world, it is hard to believe that it will work in a dynamic one with shifting techniques. Indeed, much online mis/disinformation is first posted to small fake sites that change constantly. These sites inject their false information into Facebook and Twitter, where it gets circulated further. Tracking the impact of the false information spread through the social platforms is impossible in part because the platforms don’t provide information as to where it circulates (YunChee 2019).

Noting that much of the testing was not done under real-world conditions, but rather with clean datasets that represent only a snapshot in time rather than the dynamic situation referred to above, a 2019 report by the EU concluded that there are major problems with scale and accuracy and that algorithmic mistakes of either “false negatives or false positives can be very costly,” and for this reason human intelligence needs to be deployed (Alaphilippe et al. 2019). Related to the arms-race problem is that the channels by which false information travels also change. The use of WhatsApp in India to incite violence and to spread disinformation during the 2018 presidential elections suggests what can happen when mis/disinformation is driven underground, where it cannot be monitored. Reddit and 4Chan in the US provide a home for extreme, often violent, content to flourish (Law 2019). Suppressing content so that it winds up in dark corners of the internet where it cannot be easily monitored is risky.

Third, many of the sites deliberately emulate the patterns of traditional journalism, raising questions about how to classify something as true or false. Much of the mis/disinformation online contains some true elements, is built upon news stories that covered something true (Shu et al. 2017), or is adapted from satire (Gravanis et al. 2019). How to classify a false news story with reputable sources? What to do with a site like Fox News or RT.com (Russia Today), which spin or slant the facts? What to do with medical disinformation that contains spurious science or false claims put forth by doctors or former doctors like Andrew Wakefield?

Ambiguity presents special difficulties for machine learning that relies on binary distinctions. If NLP can filter out only the false, then most misleading content will still get past the filters. There is no clear bright line between what is mostly accurate and mostly misleading.
Part III. Solutions addressing Supply/transmission

Here is a thought experiment. Early attempts at verification looked at the sources of information. By this standard, the assertion by then-Secretary of State Colin Powell at the United Nations on February 5, 2003 that Iraq possessed weapons of mass destruction would have passed muster (Wilkerson 2018). Indeed, it did pass muster, as many people found it impossible to believe that Colin Powell would lie. In many ways it’s similar to the fact-checking problem. What to do with something that is partly false? As many people listened to Colin Powell’s speech, their judgment and analysis of the incentives at play led them to understand the reasons he might lie, or the likelihood of his having been given false or incomplete intelligence that led him to believe that Iraq had weapons of mass destruction. Could AI have detected what television audiences understood? That’s not clear.

Financial impediments; hard to scale unless adopted by platforms

Unless their technology is adopted by the platforms, it’s difficult to see how any of the start-ups we studied will be able to scale. Yet the chance of adoption by the large platforms seems small. Silicon Valley is characterized by oligopsony and there is little room for innovation (Meacham 2018).

A few large players determine which technology will be used. Indeed, our interviews underscored the financial impediments to the growth and scaling of the entrepreneurial solutions. While some of the firms in this field have been around for years, others are nascent and small. Many got help from friends and family as well as small start-up grants and are now busy trying to develop and launch their tools or scale. Metafact was started with funding from Hanyang University in partnership with Seoul Metropolitan Government, while others, like the Global Disinformation Index, relied on foundations. Others have core business activities and/or ties to intelligence and government, cyber-security, or fraud detection. A few, like Brandwatch, have been around for years and have a core activity of monitoring the web for corporate clients. Some, like Vett News, consider fighting disinformation to be a part of their core activity. Others, such as Truepic, hope for commercial applications for their technology. Truepic verifies photographs taken with its technology and assigns them a unique number stored in the cloud. This will not only help governments, human rights groups, and media organizations that need to verify, say, pictures of atrocities or human rights violations, but could also be extremely useful for insurance companies that are vetting claims and rooting out fraud. Others, such as AdVerif.ai, offer pro bono help to media outlets.

16 Supplied by Chris Stone.
Several hope to be acquired by the big platforms. There are plenty of examples of companies in this sector being acquired: Fabula was acquired by Twitter in spring 2019 (Lomas 2019), and advertising-verification service Moat was acquired by Oracle in 2017 for $850 million (Kafka 2017).

The problem is that Silicon Valley’s oligopsony environment means there are few potential buyers for the numerous firms hoping to be acquired or raise funds. Founders are aware of this disadvantage. In interviews, the start-up founders discussed the difficulty of raising funds and fostering scalable innovation in an environment in which a few large platforms dominate. The platforms’ domination means that not only do they have the most revenue, but they are able to hire programming talent and quickly scale their technology. Where the platforms can’t develop it on their own, they can simply copy or buy it through so-called “aquihires.” This relatively closed environment means it’s harder for the smaller companies to get ahead.

When you start a company, you have a vision for building it and that can get derailed by running out of money and going out of business, or you get gobbled up by a large company that thinks what you are doing is really interesting. Google is very good at building search solutions and using machine learning technology to make content available, and has a lot of talent. Can a start-up do better than the big companies? Usually not. (Ellis 2019)

There may be a way for some of these smaller firms to generate revenue from the technology they have developed, but only a few are likely to scale on their own. Without scale, it’s hard to see how they will have much impact.

“There are probably a handful of scalable [companies] that use machine learning and natural language processing. Extracting information from video and images is difficult and something you can profit from. Companies that figure out how to do that will make some money and the rest of the firms will have a hard time scaling,” says Justin Hendrix (2019) from NY Media Lab.

Impediments: lack of incentive for platforms to act

We saw earlier that there are limits to what technology can do: Even the most effective systems for detecting mis/disinformation are unable to capture everything online. This is related to another impediment: The platforms lack incentives to act.

The ecosystem of small firms exists, in part, because social media companies have no financial incentive to fix the problem themselves and government regulation
addressing the problem has not yet been passed. Or the regulation that does exist, i.e. Section 230 of the Communications Decency Act, mostly protects the tech companies from liability for what users put on their platforms. Many of the companies we spoke to said that Facebook, YouTube, and Twitter have been extremely lax and irresponsible in allowing hate speech and disinformation to contaminate their networks. Many believe that, technically, it’s not a hard problem to fix. But, notes Sagar Kaul (2019), one of the founders of Metafact, a fact-checking platform based in India, “If the platforms try to tackle the problem internally it will be a huge revenue loss for them so they don’t have an incentive to do it.”

Online disinformation is a lot like spam, says leading technology journalist Julia Angwin, and it could probably be solved the way we solved the spam problem: with a combination of legal and technical measures. “First, Congress passed a law (the CAN-SPAM Act) that imposed fines on the senders of spam,” she points out. “Faced with liability, the email industry then set up a ‘black list’ of known spammers that they all agreed to block. Similarly, if there was some legal or financial cost to the platforms, they would likely set up a ‘blacklist’ of disinformation outlets to block. But they currently have no incentive to do so” (2019). Mark Zuckerberg has called for global regulation (Zuckerberg 2019) but observers note that his comments are belied by the amount of time and money that Facebook and other tech giants spend lobbying to avoid regulation (Kang and Vogel 2019).

Angwin has spoken frequently about the fact that profitability comes from engagement and audiences engage with content that generates outrage and surprise. Reducing engagement and eliminating clickbaity mis/disinformation would likely hurt Facebook’s profitability. For this reason there is no incentive for Facebook to take action. Indeed, there is evidence that Zuckerberg has made a decision not to try and reduce polarization on Facebook. (Horwitz & Seetharaman). In a May 2020 article, the authors explain in detail how Facebook began studying ways to reduce polarization on the site.

**Fear of regulation causes paralysis**

There is another problem: Attempts to solve the problem may make the platforms liable for the content they post. We will discuss Section 230 in greater detail later in this dissertation, but a 2019 paper in *Harvard Law Review* explains how in the 1990s the platforms feared that if they blocked certain websites they would open themselves up to the charge that they were making editorial decisions and were therefore publishers, not platforms, and thus liable for the content they displayed and circulated. As a
result, the law that some had hoped would make the platforms block certain sites ended up having the opposite effect (Klonick 2018).

The fear of regulation can force the platforms to move but also cause paralysis. For example, in the US it's much harder to shut down right-wing channels now than it would have been five years ago. Those channels don't just have millions of subscribers now, they possess some political influence and have senators organizing hearings on their behalf. US politicians see sites like InfoWars or NRA TV as useful tools to turn out the base. It may be that what Google fears most is not losing advertising revenue—as these conservative channels don't bring in much ad revenue—but that political oversight could lead to antitrust action.

**Background on ad tech**

In order to fully understand the fixes we present later in this chapter, we first provide some background information on online advertising. The rise of programmatic advertising has meant that, in many cases inadvertently, quality brands find their advertisements appearing on sites promoting mis/disinformation (Braun and Eklund 2019). Assuming that big and respectable brands do not want to run their advertisements next to unsavory content, a number of the start-up founders we spoke to expect that reputable corporate advertisers will push for change and force platforms to crack down on hate speech online as well as mis/disinformation. These firms hope to persuade advertisers to push the platforms to do a better job of screening content and placing advertisements online (United for News, 6).

Currently, most online advertising is sold through vast programmatic buys determined by algorithms. According to United for News, in 2018 67 percent of the $237 billion of online advertising sold was done through programmatic advertising (United for News, 6). This means that companies don’t usually know where their advertising will end up online (Sinclair 2016). Advertising is dropped onto any website at any time and can appear next to unsavory content that promotes violence or hate speech or is completely false. The hope is that advertisers will object and will either pressure the platforms and/or lessen programmatic buying and go back to direct advertising.

In many cases, the platforms themselves use their power to choke off advertising revenue to the creators of false news sites. One famous case is that of the Macedonian teenagers who came to light during the 2016 elections (Silverman and Alexander 2016) or popular YouTube channels that spread mis/disinformation being shut down (BBC 2019b). Brand safety firms and programmatic advertisers have also worked with reputable companies to pressure the platforms to make sure that their advertisements
Part III. Solutions addressing Supply/transmission

don’t inadvertently appear on sites purveying mis/disinformation. These attempts have been complicated by Facebook’s refusal to say where exactly ads appear or to provide information as to the scoring safety risk of those ads. However, groups that are trying to use advertising revenues as a point of leverage say that more can be done. They note that some platforms actively hinder advertising transparency, like Facebook blocking tools that had previously allowed users to see how they were being targeted by advertisers (Merrill 2019).

One downside of this attempt to avoid placing ads near unsavory content is that some reputable advertisers have pulled back from digital advertising or now refuse to advertise on so-called “long tail” sites. This has led to fears that smaller news outlets are now losing what little advertising they had (United for News; Moat 2017). According to interviews, some advertisers now eschew advertising on news platforms altogether, preferring only to advertise on subject-specific sites, like weather or comedy, because these are less risky. For this reason, groups like United for News are trying to re-direct advertising to quality news sites (see below).

The idea of pressuring advertisers to pressure the platforms is an attractive one. Since most of the platforms’ revenue comes from advertising, it seems that the large platforms would be most susceptible to pressure from advertisers. Journalists also hope to promote direct buying of advertising rather than programmatic, so that quality media can benefit from advertising revenue rather than see it go to random websites. It is worth noting that advertisers would themselves be damaged by government bans on micro-targeting or the spread of regulation aimed at protecting privacy, so pushing the platforms to take action would protect both sectors (platforms and brand advertisers) from regulations they don’t want.

**Solutions to the lack of incentives:**

*Naming and shaming*

There are a number of efforts aimed at creating incentives that will push the platforms to deal with mis/disinformation they host. One school of thought holds that increasing reputational risk for the platforms, in some cases by naming and shaming, will prove effective. However, it can be argued that the platforms are beyond shaming because of their ideology and entrenched belief that they are simply providing a public square in which all information can be shared (Levin 2018b; Jarvis 2018; Hiatt 2019). Certainly, Zuckerberg’s October 2019 speech in which he reaffirmed Facebook’s policy of disseminating false political advertising suggests that his views are entrenched.
Mis- and Disinformation online

Using advertiser pressure

Others believe that if reputable companies push the platforms, then the platforms will have a financial incentive to take action. In order to identify cases where advertisements from well-known brands appear next to unsuitable content, so-called “brand safety” firms have developed software to identify such content. Thus, some of the firms that identify mis/disinformation overlap with the attempt to exert closer control over digital advertising to choke off revenue that funds sites linked to mis/disinformation.

Creating a workaround—diverting advertising revenue to support quality outlets.

Finally, there are attempts to divert advertising revenue to support the production of quality news. The United for News Coalition is one of them. The coalition brings advertisers and news outlets together to persuade the advertisers to advertise directly on quality news sites. Understanding they can’t get rid of programmatic advertising, as it’s cheap and convenient for advertisers, they want at least to re-direct some advertising spending to quality news sites (Cobb 2019).

"The initiative reminds advertisers of their financial incentives and tries to divert part of the funding for advertising over to quality content with the philosophy we can’t stop the bad guys but we can try to help the good guys,” says Jennifer Cobb (2019), director of United for News.

The efforts are nascent and it’s hard to know how they will develop, but a handful of large brands are starting to push the platforms for change. In the summer of 2019, a group of large firms launched an initiative—the Global Alliance for Responsible Media—with the goal of protecting consumer privacy and combatting disinformation (Green 2019). It will be interesting to see how far advertisers are willing to go with privacy protectons given that the effectiveness of advertising depends on targeting consumers and that this requires information about them.

Impediment to the ad-based approach

Other groups try to expose the presence of bad actors that are advertising on Facebook 17 in the hope this will not just shame the companies but also generate public outrage. However, there is some question as to whether advertisers will adopt an approach that requires more thought, is more time-consuming, and is therefore more costly.

---

17 See for example: [https://www.techtransparencyproject.org/articles/facebook-allowing-pangolin-sales](https://www.techtransparencyproject.org/articles/facebook-allowing-pangolin-sales)
Publisher John Battelle (2019) says, “I am a huge fan of the philosophy. Direct media buying supports what I think is the most important part of publishing—the direct relationship between publisher, audience and marketer. Programmatic advertising has stripped away the context of an experience with the audience. Advertisements go into unknown places, out of context. Direct advertising will re-establish the connection with the audience.”

However, he points out, “The problem is that programmatic advertising is cheap for advertisers and cheap for publishers to implement, so it’s irresistible—especially for the advertisers, who control the ecosystem."

**Impediments: lack of clarity**

Moreover, it’s not clear to advertisers how effective changing their approach might be, because the platforms give little information about who sees the ads, where they appear, or how much revenue is derived from them. This is frustrating for companies that pay for advertising and don’t know whether the ads are being seen by real people or by bots, in which cases the price should be lower (Braun and Eklund 2019). As well as identifying the sites, groups such as Moat, which does brand and platform analytics, also try to pressure the platforms to release more information as to where ads are placed and how ads are scored internally. However, a 2018 paper (Chiou and Tucker 2018) found that removal of advertising for anti-vaccination sites on Facebook caused a 75 percent drop in the sharing of these sites on Facebook when compared to Twitter in the same time period. In theory, Facebook and Google should not only know their customer but also be able to identify unsuitable ads or ads that are placed next false or unsavory content. Ad buyers should be given the right to update the algorithms in the ad-bidding process every hour if they choose. For example, after a mass shooting both Facebook and the advertiser should be able to immediately ensure that their advert does not appear near false claims about the shooting.

Withholding advertising by reputable brands may not make a dent in the revenue generated by false content online. Most likely, the amount of money involved is too small to have an impact on the larger problem of online mis/disinformation.

However, even if attempts to limit advertising revenue don’t always have a direct impact on the dissemination of mis/disinformation, these efforts could still indirectly affect the problem. First, an alliance of big brands from major multinational companies could signal to the platforms that they need to make more of an effort to block/remove content. Second, if United for News is successful it will have a supply-side benefit of boosting the amount of revenue going to quality outlets and therefore
promote the flow of quality journalism. Flooding the zone with quality reporting is one of the ways that journalists hope to combat the problem of online mis/disinformation. Even YouTube is now funding quality video and placing it higher in search results in the hope that these will crowd out some of the mis/disinformation and violent videos on the site (Jahromi 2019).

Choking off small amounts of advertising could also prove effective in early stages of misinformation campaigns. If Macedonian teenagers realize early on that misinformation about polio vaccines isn’t profitable, they may not bother to pursue disseminating it.

What if it’s not about the money?

On the other hand, as we saw with billionaire Albert Dwoskin (Kirkland 2011) and other funders of the anti-vaxx movement, there are likely to be others willing to step in with more funding for misinformation campaigns (Sun and Brittain 2019). Choking off advertising revenue would not be a disincentive for government actors spreading false information for political or ideological reasons. The same is true for anyone with an agenda who is willing to spend money on generating traffic to sites with mis/disinformation.

Another Solution: Standard-Setting, Labelling Options

Diverting audience attention to quality information is related to the difficult problem of defining what is “good.” Determining “quality” or “good” information cannot be addressed by technology alone, as it requires some value judgement. Even tech solutions need some sort of rubric or guidance. So, in this section we turn to promising efforts that provide whitelists or ratings of different outlets. These solutions also have a demand-side element because once an outlet is rated in some way (similar to the green check that appears in Google searches) it can be used by audiences as well as by firms. For example, Google has created “cards” with reliable information on different people, topics and places so that readers and, during Covid 19, Facebook and Youtube directed audiences to reliable websites.

We explained earlier some of the difficulties faced by data scientists trying to assess the veracity of each piece of content online. Given the impossibility of determining whether something is true or mostly true, some groups have turned instead to evaluating news-gathering standards and processes and making an evaluation as to the credibility of outlets that follow those standards. These are essentially demand-side solutions as, by labelling and/or certifying selected news outlets, the groups hope to help audiences make wiser choices as to what to consume and what to circulate. The
thinking is similar to the idea of food labelling—inform ed consumers will make better decisions. These solutions contain supply-side elements as well, as the hope is that labelling and certifying will boost social pressure on suppliers of online content and bring social norms to bear. The solutions also mesh with the ideology of the problem-solvers, as they come from journalism communities that prefer market-based fixes rather than government legislation. Like the platforms, the journalists who began these initiatives would rather see responsible self-regulation than become vulnerable to government regulation (Lee 2019).

In this section we will discuss two promising initiatives focused on standard-setting and labelling. We will provide background on each and assess their advantages and possible shortcomings.

One is New York-based NewsGuard, which was founded by the never-Trumper Republican and former publisher of the Wall Street Journal Gordon Crovitz, and journalist, media entrepreneur, and founder of American Lawyer Steven Brill. Brill is well-respected as a journalist, has written extensively on US healthcare, and is a left-leaning centrist. The founders hoped that the bipartisan nature of their firm would give their efforts credibility (Crovitz 2019). Initially funded with private investment, the founders hoped that the platforms would pay for the service or acquire and use it. “They created the problem so they should pay to fix it,” Crovitz (2018) wrote.

NewsGuard uses human analysts—many of whom are former journalists—to apply nine basic journalistic criteria to the news and information websites that account for over 90 percent of engagement online. Each site receives a green or red rating and a “Nutrition Label” write-up. NewsGuard launched in the US in 2018 and in 2019 was rolled out in key markets in Europe. Microsoft has licensed NewsGuard globally, including for its Edge mobile browser.

This solution is elegant and already available. It is intentionally market-based and respectful of free speech. It has already had an impact, as news organizations are given the right to reply and some have changed their practices after receiving red ratings from NewsGuard (NewsGuard 2019).

It should be noted that the first-mover advantage of NewsGuard may be tenuous: Other groups could easily set up their own ratings systems. For example, anti-vaxxers already rely on articles and doctors they trust. The anti-vaxxer community may already have its own list of sources its members believe are credible. Restaurant ratings and consumer-product rating services have risen and fallen over the years, often replaced by sites such as Yelp and Amazon, which enable the public to assess goods and
services—although they can be gamed by providers of the products sold online (Picchi 2019).

Moreover, the fact that NewsGuard is a for-profit start-up may affect perceptions of its work (Oremus 2019). Questions have already surfaced as to who will rate the raters (Oremus 2019), with the far right attacking sites such as NewsGuard (Li 2019). Believing in a free marketplace of ideas, Crovitz says he welcomes competition.

It should be noted, however, that other attempts to whitelist sites have run into some difficulties. The Poynter Institute published a list of 515 unreliable news sites on April 30, 2019, got in trouble with conservatives, and ended up taking down the list of sites two days later because it was insecure about its methodology (Allen 2019).

The second effort is the “Journalism Trust Initiative” (JTI), which is being led by Reporters Sans Frontières (RSF) and its partners, Agence France Presse (AFP), the European Broadcasting Union (EBU), and the Global Editors Network (GEN). This group is trying to create a set of international standards as to what constitutes quality news. The outlets that sign up to this set of standards can then be included in whitelists used by advertisers or groups such as NewsGuard, as well as by audience members looking for reliable information. The group hopes to finish its efforts in 2019 and then roll them out in 2020, and is working with the European Standards Authority.

JTI’s view is that watermarking a particular news story and showing its provenance (as some have proposed) is simply too complicated and time-consuming for news outlets that publish multiple stories a day. Instead, JTI focuses on setting standards for the process of news production. Its criteria include information about ownership, editorial policies, whether the outlet is responsive to complaints, and whether it has a mechanism for correcting false information (Steenfadt 2019).

JTI had three committees drafting information on identity to ensure that outlets explain themselves, their news-gathering methods, their editorial processes, and their ethical standards. JTI also drafted a questionnaire for a first self-assessment stage for media outlets, a glossary, a security framework for vulnerable individuals and outlets and a manual for different sets of stakeholders and target groups (i.e., platforms, media development, regulatory authorities and self-regulatory bodies, etc.) that want to subscribe to the standards (Steenfadt 2019).

JTI is not a certification exercise per se, but consists of two different elements. The first is the current phase of developing and publishing the actual criteria, then publishing them within the framework of “official” standard-setting. JTI is
following the guidelines and procedures of ISO, the global standardization body, and 
its European subset, CEN. However, producing a standard means it has to be 
certifiable (Steenfadt 2019).

The second stage of the project will be the certification itself, to be carried out by 
different actors. RSF does not aspire to become the global judge for trustworthy 
journalism. Usually an ecosystem of certification entities evolves around a standard. 
These could be start-ups or existing players, commercial or non-for-profit entities, etc. 
“What we are currently developing alongside the standards document is an entity to 
'certify the certifiers' and thus endorse those who would actually audit media outlets 
around the globe by means of JTI standards conformity assessments” (Steenfadt 
2019).

RSF has also begun a project on mapping media ownership, and in 2019 it had two 
planning meetings with about 120 organizations and people, including French 
regulators, the German government, the European Standards authority, Google, 
Facebook, 60 media outlets and some press councils. US officials from the State 
Department and the Department of Homeland Security have expressed interest in the 
orGANIZATION.

The JTI project faces a few impediments. One is its member organizations’ capacity 
to develop new lines of work when they are already underfunded and face resource 
constraints (Steenfadt 2019). This relates to the question of whether JTI will be able 
to raise funding to support their work. Ideally, the organization would like to have an 
array of donors, so as to maintain its independence and not be beholden to one large 
donor.

Another question is whether news organizations will be willing to subscribe to the 
JTI’s processes. They may feel that the standards and, later, certification are too 
similar to licensing and rating of journalists. Licensing of journalists is common in 
European and other countries but has long been resisted by US journalists, who are 
worried it will impinge on free expression (Strasser 2010)\textsuperscript{18}. The tech giants may go 
along with the journalists’ resistance because they want to forestall regulation.

The certification and standards groups bring human intelligence to the problem of 
identification and verification. Like fact-checkers, they hope that labelling will give the 
platforms an easy set of criteria to follow. They also hope that they will signal false 
content to audiences, which will then reject it.

\textsuperscript{18} A well-respected US group (which was funded by Luminate) is The Trust Project \url{https://thetrustproject.org/}
Conclusion

Fears of algorithmic and platform censorship exist. But in the current system, there is already censorship—without accountability—by the tech giants (Gillespie 2018). Since the platforms are not accountable to the public, it’s not clear where the line of unacceptable content is drawn.

In this vein, the European Union in 2018 persuaded the platforms to sign a voluntary Code of Practice under which they submit monthly reports describing their efforts to combat the spread of misinformation. These include Facebook’s European registry of political advertising, which is discussed in their monthly reports. However, the monthly reports are short on specifics and, as discussed earlier, in many cases do not lay out specific criteria used for downranking or removing information.

The lack of specificity and objective criteria characteristic of the current system of platform censorship is in some ways reminiscent of the censorship that exists in authoritarian regimes. In countries like China and Vietnam there is always a question of what lines can be crossed, and dissidents play a cat-and-mouse game, pushing on the limits to see if the government responds (Shirk 2007; Zhao 2000; Repnikova 2017; Pomerantsev 2015). So, too, as the platforms step up their efforts to limit mis/disinformation, purveyors of mis/disinformation try to game the system and continue to spread their questionable content.

In this world of fragmented solutions to the problem of online mis/disinformation, the small start-ups using AI and NLP are another niche to watch. It’s clear that the entrepreneurs we interviewed see a business opportunity in using people and NLP to identify and possibly remove mis/disinformation online. However, the limitations are all too apparent. The first is that AI is not subtle enough to identify all of the myriad forms of mis/disinformation currently contaminating the information ecosystem. Second, even if it were possible to use the technology at scale, there is little evidence that Facebook, Google, and Twitter would use it—one of many reasons why regulation of the platforms is essential. Third, these tech-based solutions don’t address the larger economic, social, and political reasons that mis/disinformation spread. An entire ecosystem, including programmatic advertising and the anonymity of domain registrants, creates the ideal conditions for disinformation.

“Platforms will not take on active defense of truth-telling institutions [and] even if the platforms wanted to fix the problem, they can only have an important but limited impact in the disinformation landscape. They are a contributor to a massively
organized disinformation system. But the digital ecosystem is broken and the possibilities of gaming the system are endless,” says Alejandro Romero (2019) from Alto Data Analytics.

The groups that are trying to certify news processes and create global standards seem the most promising, but these will need to get news organizations to agree to participate and will need funding to roll out their plans. As well as voluntary industry labelling, efforts to get at the financing of mis/disinformation are also important. Campaigns to dissuade advertisers have had success in Slovakia but another obstacle is when it is governments paying for misleading advertising or advertising on sites that transmit misleading information. (Szakács20)

The AI screening efforts require action on the part of the platforms. They also overlap with the ad-tech business, which is a difficult problem to solve. Part of the reason each side is paralyzed is the unwillingness of any party to move first (Braun and Eklund 2019). For this reason, government regulation may be needed order to force action by the platforms. Effective January 1, 2018, the German government made platforms liable for the posting of illegal speech such as hate speech (BBC 2018). This has caused the removal of some content that should not be taken down, such as a satirical magazine’s Twitter account (Reuters 2018) and a political street artist’s work (Horchert 2018). The implementation of the German NetzDG law and the possibilities for other types of regulation will be the subject of the following chapters.
Mis- and Disinformation online
Chapter 6

Intermediary liability and what the US can learn from Europe

As this dissertation broadens its scope, we turn inevitably to the question of whether and how to regulate the content available on the internet. We believe that regulating the internet is by and large a supply-side solution, like the fixes described in the previous chapter—using AI to screen, flag, and remove content. This is because many of the laws addressing mis/disinformation rest on controlling the supply and/or dissemination of what goes online. For example by making companies liable for harms that result from the information the companies disseminate, they will have an incentive not to disseminate it and that may affect the supply as well. One objection to solutions focusing on the supply of information is that these are a form of censorship carried out by corporations with limited rights of appeal and no clear disclosure. Indeed, our last chapter focused on a range of opaque tactics. Shadow banning, takedowns, and using AI and limited numbers of content moderators to make decisions are implemented by tech companies with little discussion or clarity.

Regulation is different. It is unacceptable for governments to force tech companies to take action without setting clear guidelines for what needs to be done, which legal principles are at play, and what rights to recourse and appeal exist. Governments operate on a set of democratic principles or, at least, a presumption of accountability; that firms don’t. Governments cannot get away with the tactics and lack of transparency that companies rely upon—yet in this case, governments depend on powerful and intransigent companies for cooperation. Thus government regulation will need to involve setting up clear rules, specialized regulators and oversight
principles, and constant supervision of the tech companies. In 2019 some French and UK regulators began to compare regulating the tech companies to banking regulation and supervision. In Europe, a legal structure already exists to undergird more rigorous regulation of tech companies, including the EU’s ability to levy fines, the ability for investigators to bring both civil and criminal charges, and the power given to the federal government as opposed to member countries (Micek 2019).

Are there laws the US can implement, given the First Amendment?

In the US, much of the discussion is not about the principles of regulation, but whether we should regulate at all. Given this controversy, we decided to examine which kinds of laws, if any, might address the problem of online mis/disinformation without harming reasonable standards of freedom of speech. This section is written from the point of view of an academic who supports journalism and human rights principles and is hoping to bridge the gap between those advocates and policymakers who seek more regulation and those in the free-expression community who oppose it. Accordingly, the next three chapters will focus on proposals that may be acceptable to those who seek policy change that does not pose a threat to First Amendment principles. This chapter takes up intermediary liability and proposed modifications of Section 230 in the US. We draw on the scholarship of many in the legal profession as well as interviews done with rights groups, regulators, and members of journalism and freedom of expression groups. The following two chapters (seven and eight) look at regulation of political advertising and the use of defamation laws to counter false information.

Legal remedies are perhaps the most controversial of all, as we will discuss later in greater detail. However, after an extensive review of the literature, we are convinced, along with others, that the laws governing the platforms must be updated (Benkelman 2019) because the internet of today has evolved dramatically from 20 years ago (Citron and Wittes 2018). There may have been a case two decades ago for protecting developing tech companies from liability for most forms of third-party content so that they could grow without having to hire teams of lawyers and moderators, but that is no longer germane today. In the interim, we’ve learned that without liability there may be no way of holding the tech companies to account. After sitting in on yet another panel discussing mis/disinformation in the light of the 2016 elections, one frustrated

---

20 Deep thanks to Peter Micek from Access Now and Ryan Whittington for their careful reading and helpful comments. Some of their points are included in this chapter. They were particularly helpful on the history of Section 230, the current debates about it, and how to understand the German NetzDG law. Peter Micek also emphasized repeatedly the importance of freedom of expression and his worries about attempts to erode the First Amendment.
bystander commented, “Well, if they don’t want government regulation and they
don’t want the companies to be in control then what do they want?”

Given this unpleasant choice we opt for government oversight and involvement. We accept that some kind of regulation is necessary and believe that regulation overseen by the courts and by politically accountable officials is better than the opaque policies currently in place at Facebook, Google, Twitter, and the other platforms. In taking this position, we are in the camp of European governments and societies that are far more likely to trust government than big business (Bayer et al. 2019). This chapter will look at the NetzDG law in Germany, the question of intermediary liability, and some of the laws passed in France since 2016 to see whether they bear lessons for the US.

The almost complete freedom of expression afforded by the First Amendment presents a unique challenge for the US, as the country has no hate-speech laws and very little regulation of social media. To Europeans satisfied with their laws on hate speech and election advertising, the US belief that all ideas—no matter how toxic—contribute to constructive public discourse and healthy democracy is incomprehensible. In today’s information environment—radically transformed by the internet—ideas that counter-speech is a panacea in a free marketplace of ideas no longer hold (Napoli). Historian Richard John argues that the problem is not just the scale and speed of what circulates online, but corporations’ takeover of the public sphere:

It is one thing to encourage a multiplicity of viewpoints, a laudable pluralist impulse that would have earned the approval of James Madison and John Stuart Mill. Yet it is quite another thing to condone the existence of multibillion-dollar digital media platform providers such as Facebook and Google that are legally absolved from any liability for the information that they circulate, deny that they are in fact media outlets by claiming merely to be neutral, algorithm-driven platforms, and have arrogated to themselves the authority to alter at will the media streams of millions of people. A further challenge to cherished assumptions about free expression is posed by the weaponization by business corporations of the free press clause to expand their rights, a legal strategy that has been dubbed ‘First Amendment Lochnerism’—a reference to the early twentieth-century business campaign to mobilize the courts to block labor legislation (John 2019).

Legal fixes very different in US, Europe and rest of the world
The debates about legal solutions are, of course, different in each country, although they share some characteristics. We noted earlier that around the world, governments, businesses, foundations, journalists, and academics are trying a range of piecemeal fixes which will not, on their own, solve the problem of online mis/disinformation. These vary greatly. So, too, for legal regimes. The vast array of standards and regulations answers the question of whether a global internet can exist in the absence of a global standard. There is no longer a universal internet, if there ever was one. Nor is there likely to be a global internet in the future. Silicon Valley’s warnings of a “splinternet” seem increasingly likely to come true, if they haven’t already (Schmidt and Cohen 2014; Ross 2016). Regulations and local practices differ across countries. And there is usually a gap between national laws and Facebook and Twitter’s own guidelines, which are sometimes at odds with national laws and international human rights law.

Just as there are many different regulatory regimes, there are many different lawyers, regulators, and NGOs working to draw up legal principles and examine different aspects of regulating the internet. Antitrust proposals and the idea of “breaking up Facebook” have gotten a lot of attention (Wu 2018; Furman 2019), but these proposals may not address the problem of online mis/disinformation. For example, giving countries “local control” over the platforms could exacerbate online hate speech. Local control of WhatsApp in India could lead to continued incitement against minority communities there (McLaughlin 2018) as well as in other parts of the world. Breaking up the platforms and giving control to local entities could facilitate the circulation of more mis/disinformation in these countries without any recourse by Facebook or other platforms. Nor does it address the problem of politicians who incite violence and use social media to circulate their hate speech. Regulating a multitude of platforms could also be more difficult.

Differences in laws spring from many causes, including differences in history, legal tradition, political traditions, and political feasibility (Funke and Flamini 2019). For instance, the European Commission has focused on taxation, regulation of competition, and protection of privacy. Meanwhile, Germany’s Nazi past has led to their strict laws on hate speech and Holocaust denial; Germany has not only banned hate speech online but has made the social media platforms liable for it. Under German law, opinions are protected but false assertions of facts are not (Buermeyer 2018). The US, for its part, is unlikely to enact online hate speech laws because most

---

21 See for example the quantity of data brokers buying and selling personal information https://www.fastcompany.com/90310803/here-are-the-data-brokers-quietly-buying-and-selling-your-personal-information
US speech is protected by the First Amendment (Benkler, Faris and Roberts 2018) and we have a long tradition of litigation around freedom of speech. Further, many of the internet governance groups in the US, which oppose European-type laws, are funded by Facebook and it may be that the dearth of independently funded voices has affected the quality of public discussion (Glaser 2018). There are also differences within governments. French officials note that the National Assembly’s response to the problem of mis/disinformation was very different from that of the Prime Minister’s office (author interview, October 2019).

Countries also have very different attitudes and laws regarding privacy and intermediary liability. Many countries believe that laws that apply offline need to apply online. Some, such as France and Germany, have updated older laws to reflect the new realities of social media.

The US is one of the few countries that has different laws for platforms than for publishers and broadcasters. In the US, Section 230 of the Communications Decency Act protects intermediaries from civil liability for third-party content that appears on their platforms, as well as the less visible data and traffic going through their systems. Though the law does have statutory exceptions, because it doesn’t apply to matters covered by the Electronic Communications Protection Act, intellectual property claims, and claims stemming from federal criminal law (Kosseff 2019). Other countries offer immunity only when the carrier is unaware of the presence of harmful content online. In Brazil (Zingales 2015) and India (Mittal and Koolwal 2019), platforms are liable for harm if they have been notified that their content is illegal or defamatory and have ignored the warnings.

Notice and takedowns are highly controversial among free expression groups. There are also plenty of legal questions to be answered, such as what kind of notification is required, what knowledge on the part of the companies actually means, and what standards are to be used in deciding what to take down (Hoboken and Keller). We note that there are many cases of problems with the notice and takedown approach used to enforce copyright. However, the notice and takedown approach is used all over the world, and so is likely to remain in use despite the complexities.

---

22 Twitter has even called itself the free expression wing of the free expression party.

23 The 1998 Digital Millennium Copyright Act (DMCA), which requires platforms to take down infringing material once they have received notice (Bolson 2016). However, the DMCA also has procedures for notifying the poster of the materials so they can file a counternotice and potentially post the material again if the person or business that requested the takedown has not filed a lawsuit (Bolson 2016). In this sense, it provides some redress. We don’t have space to discuss the problems with DMCA but extensive work has been done on this.
Free expression groups deeply skeptical of regulation

Many of the regulations being proposed, as well as those already on the books, are viewed with deep skepticism by lawyers, human rights groups, free expression advocates, and journalists (Funke and Flamini 2019). They note that countries such as Singapore and Malaysia that have passed “fake news” laws have a long history of controlling the media (Berger 2019). They worry that such laws will be used to stifle free speech and online political organizing, and may usher in mass surveillance. Groups like Access Now and the Electronic Frontier Foundation have publicly opposed legal fixes (EFF 2015; Krahulcova and Solomon 2019).

Their fear stems in part from a concern that laws regulating speech will inevitably give rise to full-blown censorship. Brett Solomon, the director of internet governance for the NGO Access Now, says there are three key risks:

1. A mechanism to review content must include methods and tools of a surveillance regime. How can you judge the legality of content without seeing what it is?

2. Removing illegal content calls for a framework, process, and tools—that is, a censorship regime.

3. And in many cases, if certain content is searched for, more acceptable alternative content needs to be served up. That might give rise to a propaganda regime, which swaps out “bad” content for “good.”

According to Solomon, “In the name of protecting democracy, conditions for a surveillance, censorship and propaganda state can arise. This outcome is known as authoritarianism” (Author email, 2018).

However, proponents of legislation argue that surveillance is already taking place—by private companies that are not legally accountable (Syed 2017). Decisions as to what should be blocked or removed are made according to these companies’ “community guidelines” and their somewhat opaque processes. They screen, rank, and block millions of pieces of content without much explanation. In the US, where freedom of expression protections are broad, it may be that Facebook and others are performing a useful service by blocking violent speech, hate speech, or medical misinformation that may be problematic but legal.

We would support the European approach that democratically elected governments need to be involved. The choice is not between no censorship and censorship. The reality is that companies are making decisions on their own. Legislating and regulating the platforms is clearly preferable to this state of affairs, but will require investing
funds in the government agencies that would be responsible for implementing and enforcing regulations. The Internal Revenue Service and Securities and Exchange Commission, among other US agencies, do not have enough resources to carry out their job by investigating complaints and bringing action against violators (Schacht 2018). How to regulate and supervise when government is underfunded and short-staffed is a subject we will discuss later. Ideally, the cost should be borne by the tech companies (Krovitz 2018, author interview) and administrative courts will need to be staffed/educated to make decisions about content.

There is disagreement in the free expression community about the kind of regulations that would be acceptable. Reporters Without Borders (Reporters Sans Frontières), has an executive director, Christophe Deloire (2019), is open to regulation but believes it’s not being done in a consistent and principled way. “I don’t believe that disclosure is enough in this case,” he says. In an attempt to go further than disclosure, RSF’s Information and Democracy project has had 12 countries sign on to a set of principles that can be applied by regulators and by self-regulating platforms (RSF 2018b). The principles, however, have been opposed by human rights organizations. Meanwhile David Kaye and other human rights advocates call for regulation to comply with international agreements on free expression (Kaye 2019a).

Two broad categories of laws regulate online mis/disinformation: direct and indirect. Direct laws are aimed at stopping certain kinds of information, such as hate speech, from appearing online. Indirect laws provide incentives to the platforms to do something—such as the hope that stripping away intermediary liability will induce the platforms to do more to filter and downrank (or suppress) mis/disinformation and hate speech. This chapter will discuss proposals to revise the protections afforded by Section 230 of the Communications Decency Act, which is Title V of the 1996 US Telecommunications Act. Proposed amendments to Section 230 would allow parties to sue intermediaries for harms arising out of information that the platforms helped make available. But the legal changes would not directly mandate actions by the platforms. Nor does the US have a system like those in France and Germany, where penalties and fines can be imposed by the government24. We consider a modification

---

24 Notes Micek: “It would be a sort of proportionate liability, using proximate cause analysis. Information itself doesn’t (usually) harm people. There are knock-on effects. So the intermediary is part of a causal chain. It’s often easier to point to the intermediary as the ‘but for’ cause, but there’s still some jerk who made the video, maybe a different jerk posted the video, and other jerks acted on the video posted, etc. Of course, there are regimes, i.e. for child sexual abuse material, where the existence of the video itself is a harm, but that’s more of an exception. Strict liability often doesn’t work so well on the internet. It’s too easy to hack and spoof, so attribution can become difficult..... Yelp faces this - dentists pissed off that disgruntled customers are writing bad reviews. I would argue that those reviews are neither illegal nor, in the broader sense, harmful, though of course the dentists would
of Section 230 to be long overdue and relatively harmless, as all it would do is open up the possibility of lawsuits against the tech companies and thus make them more risk-averse and less likely to circulate hateful and violent messages. Thus, we consider it to be an indirect law.

Direct laws might include proposals governing disclosure of online political advertising, and regulations on micro-targeting and privacy. Former US Federal Election Commissioner Ann Ravel has proposed that the platforms keep a list of who has paid for political advertising and make it available to the public (Ravel 2019). There are many problems that such disclosure would not address, but it seems a fairly straightforward application of the transparency principles guiding political advertising. However, events are moving quickly and in October 2019, Twitter announced that the company will no longer accept any political advertising at all. The disclosure of political advertising online will be discussed in the next chapter. In our third regulation chapter (Chapter 9 of this dissertation) we will delve into other measures, including counter-speech arguments, defamation lawsuits against the originators of mis/disinformation and other possible supply-side measures.

In writing this dissertation, we learned a disappointing truth: Companies that want to avoid regulation can find multiple ways of doing so. Regulators and researchers have recounted numerous examples in the past year of Twitter and Facebook stymieing government regulators (BBC 2019c; Gesley 2019). This can be done after regulations are passed or during the negotiation process. Just as the US uses lawyers to bludgeon and intimidate during trade negotiations, so do Google and Facebook in discussions with the EU. In interviews, regulators and lawyers said they are tired of going to Brussels because the discussions are counterproductive. The lobbying and resistance of companies like Facebook, Google, and Twitter is one more factor that regulators have to take into account when drafting legislation that can be implemented. 25 What is clear is that new laws alone are not enough. There will have to be new agencies, supervisory bodies, regulators and training for judges and lawyers who understand many of these cases. Disclosure will need to be regulated and companies will have to

---

25 According to the EU Transparency Register, in 2018 Google, Facebook, and Twitter spent a combined €11.5 million lobbying the EU (EU Transparency Register 2019). In the US, the tech giants spent $48 million on lobbying in 2018, up 13 percent from a year earlier. These figures do not include funding to support tech innovation, journalism, or fact-checking, all of which are part of their public relations campaign in Europe and elsewhere.
be required to provide far more information about who is amplified and who is suppressed online.\textsuperscript{26}

**Explaining legal protection from intermediary liability, aka Section 230**

One of the most hotly debated topics globally is the question of how to treat intermediary liability—that is, the legal responsibility for activities performed by platforms’ users. Currently, a number of countries including the US, United Kingdom, and Canada protect the platforms from civil liability for harms arising out of the third-party speech they carry online. Other countries make the tech companies liable if they have been warned about potentially harmful content. The large tech companies have spent a huge amount of time and money trying to protect their immunity. To lose it could mean opening themselves up to lawsuits for damages from individuals or (outside the US) fines from government. We are not in a position to say which proposal for modifying Section 230 would be the most effective in helping solve the problem of online mis/disinformation, but we will try here to summarize recent thinking. More proposals continue to appear.

The debate is often couched as a question of whether the tech companies are platforms or publishers. Those who oppose regulation say that as they are “platforms,” they cannot be held solely responsible for what is put on their sites (Jarvis 2018). Those who are more critical of the tech companies note that these firms curate and choose what to display and how to rank it, and therefore have a publishing function and are, in fact, publishers and/or media companies (LaFrance 2018).

Eager to avoid liability and regulations, the tech companies themselves have claimed different roles at different times, insisting that they are a “platform, not a publisher” (or media company) (Levin 2018b); or are simply the pipe “that carries the information” (Wagner 2016). Either way, they reject responsibility for third-party content that appears on their sites. These tech companies are also conduits and distributors, notes legal scholar Ryan Whittington.\textsuperscript{27}

\textsuperscript{26} Notes Peter Micek: “The ‘Ranking Digital Rights’ results routinely show great deficiencies in corporate disclosure of policies and practices impacting human rights online. The highest-ranked companies generally score around 65/100 on the project’s robust set of indicators. Access Now supports these findings with public letters to the CEOs of ranked companies, requesting response to the results. Following the Snowden revelations, platforms showed enthusiasm to disclose government requests for user data and content removal. Yet -- despite improvements following the 2016 US presidential election and Brexit vote -- reporting remains especially poor on platform enforcement of their community guidelines and terms of service.” (comment to author, December 2019)

\textsuperscript{27} “Facebook is not a publisher but for Section 230. Absent Section 230’s protections, in all likelihood, platforms would have been held to either distributor or conduit liability—not publisher liability” (Whittington comment to author, December 2019)
This discussion has gone on for years. Here we simply note that the language in Section 230 is quite simple:

"No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider." (47 U.S.C. § 230)

Supporters of Section 230 argue that it was essential for innovation and that without it, the internet could not have flourished (Wakabayshi 2019; Khanna 2013). Freeing tech companies from liability for user-generated content boosted the development of “websites like Reddit, Craigslist, Digg, and perhaps all of social media” (Khanna 2013). Indeed, some supporters refer to the regulation as “the 26 words that created the internet” (Kosseff 2019). It also allowed the platforms to moderate content without becoming liable for action or non-action on any particular post.

US courts have largely upheld Section 230 despite multiple lawsuits by plaintiffs who alleged that they were harmed by content posted online. However, one landmark decision limiting the immunity of the tech companies came in a 2008 ruling by the 9th Circuit Court of Appeals, which held that Section 230 could not protect a site called Roommates.com from being prosecuted under US housing discrimination laws (Downes 2008). The court drew a distinction between the parts of the site that were just hosting content versus the parts of the site that were developing content (Downes 2008). Providing dropdown menus on the site, instead of text fields, meant that Roommates.com was acting to guide the content on the site, and so was not entitled to immunity (Downes 2008). The disagreement over the verdict showed

28 From Ryan Whittington: “John Bergmayer at Public Knowledge has some interesting ideas (source) about expanding the development doctrine to cover instances in which platforms pay to create specific content (like a YouTube revenue sharing arrangement), although this would have to contend with the holding in Blumenthal v. Drudge (holding that AOL was not liable for content created by the Drudge Report, even though AOL financially contributed to the promotion of the defamatory content).” (comment to author, December 2019)

29 “Writing for the majority, Chief Judge Kozinski held that the service was not entitled to immunity (the merits are yet to be decided) because the site was a ‘developer’ of the potentially illegal content. In short, Kozinski distinguished free-form text boxes (immunity intact) from drop-down menus that offered only limited choices. The drop-down menus, the court held, are not immune, because they cross the line between hosting and assisting in the development of the content, and Section 230 applies only to the former.” (Downes 2008)

http://cyberlaw.stanford.edu/blog/2008/05/section-230-immunity-and-roommatescom-case
some of the differing views about whether laws that apply offline should apply online or whether cyber exceptionalism should apply. (Goldman 2008).30

On the other hand, other landmark decisions upheld protections for sex trafficking website Backpage.com, which was well-known for its prostitution advertisements. Multiple suits by women and children who had been harmed were unsuccessful, as the courts ruled that Section 230 protected Backpage.com from liability, although the site’s owners later modified some of their practices.31 In response, the Stop Enabling Sex Traffickers Act (SESTA) and Allow States and Victims to Fight Online Sex Trafficking Act (FOSTA) were passed and signed into law in April 201832.

The US debate about modifying Section 230

The current environment favoring regulation of the tech platforms has given rise to debate in the legal community as to whether Congress should pass legislation to modify Section 230. Tech companies argue that Section 230 protects innovation. Others, such as Jeff Kosseff, a cybersecurity lawyer at the US Naval Academy, argue that Section 230 is not perfect but is better than any alternative. Digital rights groups

30 “47 USC 230 was enacted in 1996 during the height of ‘cyberspace exceptionalism,’ the belief that the Internet was unique/special/different and therefore should be regulated differently. 47 USC 230 is a flagship example of such exceptionalism. It creates rules that really differ between the online and offline worlds, such that publishing content online may not create liability where publishing the identical content offline would. The medium matters.

“Despite Congress’ obvious intent, the majority opinion bristles with antipathy towards cyberspace exceptionalism. In numerous places, it expresses skepticism that offline rules should not apply verbatim to online behavior—even though that’s exactly what 230 does in some cases. Kozinski’s not the first judge to rankle at the policy implications of Congress’ cyberspace exceptionalism, but he may have let this hostility override other analytical considerations. “As a result, I think the majority opinion reflects an implicit normative judgment that the Fair Housing Act should trump 47 USC 230, which distorts the statutory analysis. (To be clear, the dissent makes this point too, so I’m not the only one who feels this way). I understand why the majority would go this direction; anti-discrimination goals are normatively very important. At the same time, 230 codifies some very important norms too.” (Goldman 2008) https://blog.ericgoldman.org/archives/2008/04/roommatescom_de_1.htm

31 Originally, the 2014 lawsuit against Backpage.com and several of its parent companies by women alleging they were “repeatedly forced as minors to engage in illegal commercial sex transactions,” was unsuccessful. The first U.S. Circuit Court of Appeals in Boston rejected the women’s claims and in 2017 the Supreme Court refused to hear an appeal (Chung 2017). The CEO of Backpage.com was indicted along with other company officials who had been indicted, under seal, earlier. This occurred before President Trump signed FOSTA-SESTA so the indictment relied on existing federal criminal laws (namely federal money laundering laws and the Travel Act) and Section 230’s federal criminal law exception rather than FOSTA-SESTA. Charges included conspiracy for money laundering and facilitating prostitution. Michael Lacey, James Larkin, and five employees of the site who are charged in the case are scheduled for trial on Jan. 15, 2020, the Associated Press reported (AP 2018). According to news reports, after Backpage.com was shut down, sex workers ended up back on the streets (Steimle 2019) and other sites sprang up to replace Backpage.com, again illustrating the limits on what lawsuits can achieve (Clozel 2019).

in the US largely agree and are opposed to modification of the law (Harmon and Falcon 2019) though some, like Access Now, are open to some amendments (Olukototon, Micel and Mitnick 2019).

Since 1996, however, hundreds of papers have been published calling for changes to the law (Reidenberg et al. 2012, p. 9). Since writing this section, many more proposals have been published and many more guidelines and principles have been suggested (MacCarthy 2020; Hoboken and Keller 2019; Kaye 2019b). Where possible we try to incorporate some of the more recent literature into this dissertation.

One of the most outspoken legal voices calling for Congress to modify Section 230 is Danielle Citron, professor of law at Boston University. Like other legal scholars, she argues that the law has become far too broad, providing protections well beyond the original intent.

The CDA was part of a broad campaign—rather ironically in retrospect—to restrict access to sexually explicit material online. Lawmakers thought they were devising a limited safe harbor from liability for online providers engaged in self-regulation. Because regulators could not keep up with the volume of noxious material online, the participation of private actors was essential. Courts, however, have extended this safe harbor far beyond what the provision's words, context, and purpose support. (Citron and Wittes 2018, p. 455)

To the extent the internet needed a broad liability shield when it was young, it certainly needs it no longer. Innovation on online platforms can at this point coexist with an expectation that platform companies will behave according to some enforceable standard of conduct. (Citron and Wittes 2018, p. 462)

How to modify Section 230 in the US: courts, Congress, quid pro quo

Citron and Wittes argue that one option for a “modest” modification “would be to eliminate the immunity for the worst actors.” As Citron has argued, sites that encourage destructive online abuse or that know they are widely used for that purpose should not enjoy immunity from liability.

---

33 One literature review examined 192 papers discussing Section 230 and found that most were critical of the scope of immunity it grants (Reidenberg et al. 2012, p. 22), with some arguing that Congress created overly broad immunity. The authors noted two periods of literature critical of the law. The first was in 1998-2005, in which scholars cautioned against broad immunity after a 1997 case that was the first to interpret Section 230 as providing immunity to ISPs. Then again between 2007 and 2012 the literature “generally argues that negative consequences have arisen as a result of broad Section 230 immunity” (Reidenberg et al. 2012, p. 9)
Citron and Wittes suggest the following amendment could be added to the law:

Nothing in Section 230 shall be construed to limit or expand the application of civil or criminal liability for any website or other content host that purposefully encourages cyber stalking, nonconsensual pornography, sex trafficking, child sexual exploitation, or that principally hosts such material.

Access Now attorney Peter Micek notes that this law would most likely not apply to Facebook but “rather against smaller, niche platforms (i.e. its competitors). This law could be used to limit Facebook and YouTube’s competition.”

A broader modification to the law might be the addition of the words in italic:

No provider or user of an interactive computer service that takes reasonable steps to prevent or address unlawful uses of its services once warned about such uses shall be treated as the publisher or speaker of any information provided by another information content provider in any action arising out of the publication of content provided by that information content provider.

With this revision, platforms would enjoy immunity from liability if they could show that their response to unlawful uses of their services was reasonable (Citron 2018)

Citron and Wittes note:

Our proposal leaves dramatically more protection in place than is currently accorded the physical operations of newspapers or colleges. The Washington Post, for instance, does not enjoy blanket immunity from having to defend a lawsuit for publishing an article. Color us skeptical that online providers really need dramatically more protection than do newspapers to protect free expression in the digital age—and particularly, that they need that protection for all sorts of actions that have nothing to do with speech. In the world we envision, the CDA's immunity provision would be unavailable to operators only when they cannot make a cogent argument that they are behaving reasonably to stop illegal activity. The consequence of that failure, in our scheme, is not even liability; it is merely the removal of an absolute shield from the possibility of liability (Citron and Wittes 2017, p. 420).

While Citron and Wittes focus on illegal activity, Lee K. Royster (2017) is concerned about harm resulting from false information. He uses the example of the harm that the false “Pizzagate” story did to the restaurant and its owner, and points out,
“Numerous legal scholars, law students, and even federal judges have proposed that the notice-and-takedown scheme of the Digital Millennium Copyright Act (‘DCMA’) be used as a model by which to alter § 230.” He goes on, “It is worth discussing amending § 230 not only because it can provide actual relief to those harmed, but also because reform may negate rash and outlandish actions by those who seek to do others harm based on completely false publications.” However, false information is nonetheless protected by the First Amendment and the reporting system for the DMCA is often abused. There are many examples of how companies use the notice-and-takedown system, aimed at protecting copyright, to hamper their competitors.

Were Section 230 to be further modified, it could happen through the courts, building on the precedent of Roomates.com (which would involve making decisions about what kind of “platform conduct could be and should be treated as content development”) or through congressional action (Hwang 2017, page 13). Some scholars and politicians have suggested that that the large tech companies should have to earn the protection of Section 230.

Republicans in the US have been keen to make sure the big tech companies are not biased against them and have used allegations of bias to influence the platforms (Ghaffary 2019). In June 2019, Republican Senator Josh Hawley (R-Mo) introduced legislation “to remove the immunity big tech companies receive under Section 230 unless they submit to an external audit that proves by clear and convincing evidence that their algorithms and content-removal practices are politically neutral” (Hawley 2019). The outcry from Democrats after this proposal was fierce, as it’s understood to be a way of stopping the tech companies from removing right-wing material (Kelly 2019) (Coasten 2019).

34 (see United States v. Alvarez). “A takedown notice would only be valid where the speech is unlawful, which means it the speech in question would need to fit into the narrow exceptions to the First Amendment (true threats, speech incident to unlawful conduct, or defamation) to warrant takedown.” (Ryan Whittington, comments to author December 2019)


36 Point made by Ryan Whittington.
Similarly, in a paper published in August 2019, Rutgers law professor Ellen Goodman and Ryan Whittington, a law clerk with the Digital Innovation and Democracy Initiative, survey the proposals that currently exist in the world and find that protection from intermediary liability could be given to the platforms in exchange for meeting various conditions. One might be providing more transparency about their curation and moderation procedures. Another would require the platforms to donate to a dedicated fund that would support accountability journalism (Zingales et al, p. 194). The authors also want to require large platforms to develop “detailed, transparent, appealable practices specifically for disrupting coordinated campaigns” that engage in activities that “threaten or intentionally incite physical violence, … that clearly constitute online harassment, or that constitute commercial fraud” (Goodman and Kornbluh 2019). This latter condition is similar to the French and UK governments’ proposals to mandate a “duty of care” (Carnegie Trust 2019).

Unlike Hawley’s legislation, Goodman and Whittington contend their proposal would not interfere with the purported political leanings of the platforms.

“While treating Section 230 protections as a privilege would be a substantial change, such proposals do not discriminate on the basis of viewpoint and require adjudication on an ex post basis,” they write (Goodman and Whittington 2019).

Other ideas….hierarchy of speech

John Bergmayer at Public Knowledge proposes a hierarchy of content so that paid advertising or monetized content could be subject to a different standard than other kinds of content (Bergmayer 2019). This is in keeping with US court rulings which traditionally did not protect commercial speech as much as other kinds of speech, but might not be upheld in court. Senator Mark Warner’s 2018 white paper on disinformation suggests a notice-and-takedown approach so that tech companies are liable for state-law torts for defamatory or illegal deep fake content—but they have to be notified first and the complainant has to get a court judgement, thus protecting due process more than the procedures laid out in the DMCA. This approach of involving

---

37 Peter Micek notes: “In order for these rules to actually ameliorate the situation of at-risk persons, they’ll need to target more powerful entities, the sorts of political parties and governments orchestrating sophisticated campaigns of disinformation. Of course, even with this language, I fear for how TikTok or even Twitter would use this as an excuse to kill hashtags that people use to coordinate protests where someone happens to set a car on fire.” (Micek comments to author, December 2019)

38 Blumenthal v Drudge and AOL found AOL was not liable for defamation despite the fact that it had contributed financially to the Drudge report. http://www.dmlp.org/threats/blumenthal-v-drudge
courts might eliminate some of the abusive reporting found under the DMCA (Whittington, email to author, February 2020). In this way, the Warner idea is similar to French regulations and Germany’s NetzDG, though European governments have authority to levy fines for refusal to take down illegal content. In the US, of course, a modification of Section 230 would simply open the way for court cases for damages from harms (Whittington, comment to author December 2020).

Due process and the right to appeal is a key sticking point for free speech advocates who note that, even with the best of intentions, companies can over-comply. “Companies are very bad at sorting through content, especially when asked to look beyond ‘illegal’ content. Content moderation farms involve underpaid, under-resourced, and over-stressed individuals making split-second decisions over complex and context-specific content. Under pressure from regulators, the safest route is to over-comply and remove anything close to the line,” says Peter Micek.  

Access Now’s position is that the focus should be on privacy and data protection “because the power of much mis/disinformation comes from its ability to target certain people and communities. Take away the data profiling and you take some teeth out of campaigns to skew public opinion” (Micek 2019).

**Approaches to intermediary liability vary around the world**

Outside the US and Europe, countries have adopted a variety of approaches to intermediary liability, in general holding platforms liable for harms if they do not respond to warnings or requests for takedowns (Romm 2019a). In India, the police as well as government and individuals can request takedowns (Bahree 2019). In Brazil, offline laws about hate speech apply online; tech platforms are liable for third-party content if they do not remove it following a judicial order to do so. Individuals may also request takedowns, and if tech companies don’t act, then judges can rule on the takedown notice (de Souza Abreu 2018). In 2019, Australia passed a bill, the Sharing of Abhorrent Violent Material Act, making online companies responsible for violent content posted by third parties (Australian Parliament 2019). The bill stipulates fines of up to AUD 2.1 million for individuals and AUD 10.5 million for companies and 10 percent of annual revenue with company employees facing up to three years in jail.

---

39 Adds Micek, “That’s the root of DMCA abuse, and without adequate notice and appeals systems for affected users—like in the 2017 YouTube Syria takedowns—much legal and important content can be lost. The Syria debacle was at least in part due to new use of machines to identify and act, or suggest action, on content, raising the whole other set of biases and errors of machine-enabled content governance” (Micek, comment to author December, 2019).
Part III. Solutions addressing Supply/transmission

(Cave 2019; Karp 2019). In Canada, internet companies are protected against intermediary liability (Laidlaw and Young 2018) until they are notified of potentially harmful content. Once they are notified “they can be on the hook for defamation claims or even criminal charges” (Kosseff 2019, p. 162).

**European approaches to intermediary liability, notification, and takedown**

As in India, Brazil, and other countries, the EU has pursued a notice and takedown approach. The European E-Commerce Directive of 2000 addresses the various activities of the intermediary service providers “separately” (OECD 2011, p. 15). Describing the evolution of intermediary liability in Europe, Internet law expert Lilian Edwards notes that from the beginning, Internet Service Providers (ISP), and later platforms, made several arguments as to why they should not be held responsible for third-party content (Edwards 2018). Among other points, the ISPs argued essentially that they were just the messenger and that it was physically impossible to check all third-party content; in addition, unlike newspapers, ISPs could not use legal protections such as contracts with freelance journalists. Intermediary liability, they further argued, would hurt the public interest (Edwards 2018). According to Edwards, with these arguments the tech companies “effectively established the ideas of limited liability and ‘notice and take down’ (NTD) as the template for intermediary responsibility, a solution which had remarkable reach for over a decade and remains the pattern of many OECD laws” (Edwards 2018, chapter 9). Remarkably, Article 15 of the E-Commerce Directive actually released the ISPs from responsibility for monitoring third-party content on their sites. Article 14(1)(B) released the ISPs from liability as long as “the provider, upon obtaining such knowledge or awareness, acts

---

40 The Sharing of Abhorrent Violent Material Act creates new offences for content service providers and hosting services that fail to notify the Australian federal police about or fail to expeditiously remove videos depicting “abhorrent violent conduct.” That conduct is defined as videos depicting terrorist acts, murders, attempted murders, torture, rape, or kidnap. According to *The Guardian*: “The bill creates a regime for the eSafety Commissioner to notify social media companies that they are deemed to be aware they are hosting abhorrent violent material, triggering an obligation to take it down” (Karp 2019).

41 In an email sent to the author on February 5, 2020, Edwards clarified what she meant by ISP versus platforms. “ISPs largely did become more generically known as platforms (esp as they moved from just providing access to multiple functions) but ISPs do (some) remain a subset of what I more generally call online intermediaries.”

expeditiously to remove or to disable access to the information” (Article 14, ECD) (Edwards 2018).43

Over the years, some of the original arguments made by the ISPs began to seem less persuasive. In 2000, a French court ruled that Yahoo had to block French audiences from seeing a website that sold Nazi memorabilia. Yahoo had argued that such screening was not possible, but the court gave Yahoo three months to figure out a way to block French users and declared that Yahoo would be fined 100,000 francs for each day it did not comply. In the end, Yahoo found it could block 90 percent of French audiences from visiting the site (The Guardian 2000). Unsurprisingly, in 2001 a US court ruled that the French court order could not apply in the US as it would violate the First Amendment (Kaplan 2001).

As the internet evolved, it also became clear that the ISPs were more than just “innocent messengers” distributing information. In fact, they made money from advertisements placed next to user-generated content (UGC) (Edwards 2018). As worries in Europe about hate speech and toxic speech grew, European attitudes toward regulation broadened. While in 2000, Europe and the US agreed that platforms could have immunity from liability as long as they blocked or removed “illegal or infringing content” (Khan and Murgia 2019), today the EU Commission is weighing a Digital Services Act that could incentivize platforms to remove speech that is unpleasant but not actually illegal—for example, hate speech which is not banned in all EU countries to the degree that it is in Germany, or some forms of incitement to violence (Khan and Murgia 2019). The proposal, floated in July 2019, would set up a centralized European regulator and try to standardize rules about content on the internet.

This is a shift from the period just after 2016, when the EU focus was just on terrorism and illegal speech. Given that it does not have the legal standing to remove

43 Article 14 of the ECD indicates that

"Where an information society service is provided that consists of the storage of information provided by a recipient of the service, Member States shall ensure that the service provider is not liable for the information stored at the request of a recipient of the service, on condition that:

(a) the provider does not have actual knowledge of illegal activity or information and, as regards claims for damages, is not aware of facts or circumstances from which the illegal activity or information is apparent; or

(b) the provider, upon obtaining such knowledge or awareness, acts expeditiously to remove or to disable access to the information."
Part III. Solutions addressing Supply/transmission

speech that is legal, the EU has pursued a strategy post-2016 to pressure the tech companies to change what they permit online, given that the companies have been slow to act on their own (BBC 2016). In 2016, EU officials held repeated meetings with Facebook, Twitter, Google, and Microsoft and got the companies to say they would follow a code of conduct related to illegal hate speech. Under that agreement, the EU, with help from social media firms, began working with NGOs to flag illegal hate speech and alert Facebook, Twitter, and YouTube when they find content they think is illegal. Individual member countries also set up Internet Referral Units (Europol 2016), though these work mostly on terrorist content.

The companies’ performance in fulfilling their responsibility to take down illegal speech under the agreement has improved over time. The first and second monitoring rounds, released in December 2016 (Jourová 2016) and in spring 2017 (Jourová 2017), found the companies were not taking down much of the flagged content. The third round, released in January 2018, found an increase of takedowns of reported content, and noted that the companies were reviewing most notifications within 24 hours. Since then, the response and takedown rates of the tech companies have continued to rise, and the January 2019 report notes that “all IT companies fully meet the target of reviewing the majority of the notifications within 24 hours, reaching an average of 89%,” and that, on average, they’re removing 72 percent of the hate speech reported to them (Jourová 2019).

The EU has continued trying to persuade other platforms to sign on to the code of conduct, and in 2018 and 2019 got Snapchat, Instagram, Google Plus, video-sharing platform Daily Motion, and the French platform jeuxvideo.com to agree.

According to Louisa Klingvall, a policy officer in the Fundamental Rights Policy Unit of the European Commission’s Directorate General of Justice and Consumers:

“We have complete legitimacy to act on stuff that is illegal. We can demand its removal because unlike the case for disinformation, there is a clear legal standard for what constitutes incitement to violence and hatred based on protected characteristics, and that can be restricted in full compliance with the right to freedom of expression.

In 2018 the Commission assessed the need for further regulatory measures to tackle illegal content online. We had several options ranging from no measures


at all, measures to tackle specific types of illegal content such as terrorism, hate speech, or child sexual abuse, or more horizontal measures that would apply to all kinds of illegal content. The outcome of this assessment was to propose legally binding measures to tackle the spread of terrorist content, but not on hate speech, because so far the code of conduct has worked quite well.”

Meanwhile Germany had different ideas, moving ahead with regulation, voting in December 2017 to remove protections for intermediary liability. As one French regulator put it, “The Germans got sick and tired of talking to the platforms and seeing nothing happen. They got pissed off and passed the NetzDG law. They are completely aware that it’s not perfect and they want to update the law” (author interview, October 15, 2019).

Removal of intermediary liability—lessons from Germany

The German NetzDG law, which introduced heavy financial penalties for platforms that do not honor requests to remove illegal content in a timely manner, became effective in January 2018. Germany historically, and for obvious reasons, has strong laws against hate speech and incitement, including particular prohibitions such as a ban on Nazi symbols and Holocaust denial (Buermeyer 2018).

Under the NetzDG law, social media platforms are responsible for enforcing 22 pre-existing statutes in the German criminal code (Tworek 2019). The law has been followed closely by EU officials, who are assessing whether and how aspects of the new law should be applied to other European countries (author interviews). Accordingly, it’s worth examining NetzDG to see how effective the law has been in reducing the amount of mis/disinformation circulating online.

Before it passed, there were many examples of mis/disinformation circulating in Germany, including accusations that Angela Merkel had committed war crimes and photos that had been distorted to increase anti-refugee sentiment (DFRlab 2017). There was also a study finding that right-wing, anti-refugee sentiment on Facebook was associated with violent crimes against refugees (Müller and Schwarz 2018).

However, despite many instances of hate speech circulating online, Facebook rarely cooperated with German law enforcement, and there was a widespread perception

---

46 That study’s authors argued, “We show that right-wing anti-refugee sentiment on Facebook predicts violent crimes against refugees in municipalities with higher social media usage. To establish causality, we exploit exogenous variation in major internet and Facebook outages, which fully undo the correlation between social media and hate crime” (Müller and Schwarz 2018, p. 1).
that “social media were full of illegal content but platforms didn’t care” (Buermeyer 2018).

Germany was the first major power to enact such sweeping restrictions, and after the law passed, Justice Minister Heiko Maas went so far as to say, “With this law, we put an end to the verbal law of the jungle on the internet.” Still, NetzDG was opposed by human rights groups worried about “over-blocking” and “over-censoring,” as well as by the German Journalists Association, which signed on to a statement by academics, lawyers, and civil rights activists opposing the law (Tworek 2019). Article 19 opposed the criminalization of some forms of speech, and the Germany director of Human Rights Watch said the law “turns private companies into overzealous censors to avoid steep fines” (Tworek 2019).

The part of NetzDG that received the most attention was the requirement that platforms with over two million users in Germany remove illegal speech within 24 hours of notification or face fines up to €50 million. NetzDG introduced a system to sanction platform non-compliance and imposed an obligation to remove flagged content over the long term (Buermeyer 2019). The law, which did not create new categories of illegal content (Tworek 2019), included insult, defamation, incitement to crime or hatred, and visual depiction of violence (Theil 2018).

Heidi Tworek, an assistant professor of international history at the University of British Columbia, explains key points of NetzDG:

At the outset, it is important to clarify that NetzDG does not actually create new categories of illegal content. Its purpose is to enforce 22 statutes in the online space that already existed in the German criminal code and to hold large social media platforms responsible for their enforcement. The 22 statutes include categories such as “incitement to hatred,” “dissemination of depictions of violence,” “forming terrorist organizations,” and “the use of symbols of unconstitutional organizations.” NetzDG also applies to other categories, such as “distribution of child pornography,” “insult,” “defamation,” “defamation of religions, religious and ideological associations in a manner that is capable of disturbing the public peace,” “violation of intimate privacy by making photographs,” “threatening the commission of a felony,” and “forgery of data intended to provide proof.”

“NetzDG was really about hate speech that is unlawful and which the platforms didn’t remove. The law tries to force companies to remove speech that is illegal under existing law and which is flagged by users. It is highly controversial because in a way it privatizes law enforcement. It just removes the content without investigating the person who did it,” says Stefan Heumann, co-director of the German think-tank Stiftung Neue Verantwortung. “What has not turned out to be true are the nightmare
scenarios that came from the internet activist communities saying this would destroy free speech on the platform. That was overblown” (author interview 2019).

The law does three things. First, it imposes a new obligation on the platforms for timely removal of illegal speech. It does not prosecute the platforms for being accomplices to illegal speech but simply makes them responsible for removing it in a timely fashion and gives the Federal government the right to prosecute for non-compliance (Buermeyer 2019). The different German states retain the responsibility to prosecute, and law enforcement can still bring charges. Individuals are still allowed to sue over any content that violates the criminal code and platforms have to preserve the content for 10 weeks so that it can be used as evidence (Theil 2018).

Second, the law targets platforms with more than two million users located in Germany, so it applies to Facebook, Twitter, YouTube, Google and several others. There has been debate about whether the rulings would apply to German citizens accessing the internet in other countries (Kinstler 2018). However, this is seen as unlikely, because if it did that would make German law applicable in other nations.47 However, in October 2019 the European Court of Justice ruled that Facebook had to remove globally remarks about a Green Politician in Austria that were defamatory or otherwise illegal (Satariano 2019).

Third: the platforms are meant to provide a mechanism for complaints, and once they receive a complaint they must begin investigating (Tworek 2019). Further, if a platform receives more than 100 complaints each year it must make effective and transparent disclosure, including bi-annual reports (Theil 2018) that describe its content moderation practices (Theil 2018; Tworek 2019). The act details what types of information must be included (Tworek 2019). “Manifestly” illegal content must be removed within 24 hours of the complaint, while “merely” illegal content must be removed within a week of the complaint48 (Theil 2018). Failure to comply can mean fines of up to 50 million euros (Tworek 2019), and exceptions to extend the seven-day period are limited (Theil 2018).

What has happened in Germany since NetzDG was implemented

47 In September 2019, the EU court ruled that the ‘right to be forgotten’ law can only be applied in Europe (Marsh 2019).

48 Heidi Tworek explained the difference in a February 18, 2020 email to the author “Manifestly illegal means content that is prima facie illegal. Some content could be borderline or require more deliberation to determine if it is illegal. Hence the more lenient timeframe.”
Part III. Solutions addressing Supply/transmission

It’s extremely difficult to measure the law’s effect because the disclosure has been both incomplete and inconsistent between the different platforms (Tworek 2019; Gollatz, Riedl and Pohlmann 2018). As a result, there is disagreement as to whether the much-feared “over-blocking” has taken place (Nachten conversation with author, October 4).

Studies of the impact of NetzDG note that the platforms have not provided all the information required. Indeed, Facebook was fined 2 million euros (approx. US$2.3 million) by German authorities in July 2019 for not publicly sharing the full number of hate-speech postings that people reported in the first half of 2018 (Breeden 2019; Escritt 2019). Despite NetzDG’s requirement that platforms set up a complaint mechanism, according to researchers Facebook dealt with this by hiding the complaints button while Google and Twitter made their complaint mechanisms more transparent (author interview 2019). Researchers also disagree as to whether the law resulted in more and/or faster takedowns. The law certainly sped up the removal of content that is illegal in Germany (Heldt 2019). However, “I would be skeptical about any claims that it has had any big impact on hate speech and disinformation,” said Stefan Heumann.

“The social media companies are required to show user requests for each kind of illegal speech,” he continued, “but the government failed to give detailed specifications around implementation and transparency reporting. It’s very important how the companies implement it, and there are big differences between how Google, Facebook, Twitter implemented it. Facebook made it more difficult to file a complaint for NetzDG because people couldn’t find the button. They took down much more content under their [own] Terms of Service, but the NetzDG numbers are comparatively low. Google and Twitter made it easier to file a NetzDG complaint and thus they have reported higher numbers of user requests for takedowns.”

All of this points to the need for more and better reporting from companies both about the volume and nature of the complaints they receive, and about their response. It also means that new legislation should provide more funding for implementation and supervision.

**Impediments in France and Germany—lack of data, slowness, sovereignty**

Governments have been looking to each other in order to share experiences and get ideas, and there is great interest in Germany’s NetzDG law. In both France and Germany, regulators, scholars, and NGOs have pointed to various problems with the law. Some are related to fears of censorship and over-blocking, others relate to the
difficulty of implementing the law. Regardless of the free speech question, problems with implementing the NetzDG law are endemic to the regulation of mis/disinformation. It’s hard to find laws that work, and it is technically difficult to block some kinds of speech and not others.

Another perennial problem is the mismatch between the speed and scale of the social networks and the ability of the government to respond. Virality means that mis/disinformation circulates so quickly that by the time the government can react, it is already too late. Regarding the NetzDG requirement to take down illegal information within 24 hours, German lawyer Ulf Buermeyer notes that 24 hours is a long time on social media and “a week is an eternity” (Buermeyer 2018). By the time illegal content is removed it can already have spread widely. There are added complications about how to take down illegal content that has been modified as it spreads and what to do about sites that resemble but don’t completely replicate the site on which the illegal content originally appeared (author interview 2019). One response could be to shift the burden of proof, so that the tech companies would have to demonstrate why they didn’t remove illegal material from the places to which it has spread. Currently the platforms need to be notified about the content before they can incur liability.

Another obstacle is the reluctance of Facebook and Twitter to follow national laws. Despite German law delineating illegal content, researchers examining the effects of NetzDG found that Facebook tends to follow its own “community guidelines” rather than German law (Heldt 2019; Tworek 2019). These rules simply detail what is and is not allowed on the platform (hate speech, graphic images, selling regulated goods, etc.). In 2018, Facebook allowed the Data Transparency Advisory Group (DTAG) to independently and publicly assess whether the metrics Facebook released in its Community Standards Enforcement Report on content moderation were accurate or meaningful. The DTAG responded by providing fifteen recommendations, including calls for more transparency (Plumb 2019).

The gap between Facebook’s community guidelines and local laws means that applying the community guidelines globally, as Facebook does, could involve contravening some laws in other countries (Heldt 2019). Similarly, making sure that content visible in Germany is not visible in other countries would mean German laws apply in other countries (Heldt 2019). Conversely, NetzDG does not require social media platforms to uphold lawful speech—so, for example, Facebook can remove legal speech such as nudity (Heumann, author email 2019). In many cases Facebook’s standards are tougher than those of national governments (such as in the US), and
Part III. Solutions addressing Supply/transmission

regulations like NetzDG are slowly pushing Facebook and other platforms toward tighter standards everywhere as they try to meet European standards and find it easier to apply those same standards for all. For this reason, it’s important to understand EU laws, as they may become a global standard.

Disclosure is necessary but not sufficient

This tension over corporate censorship and over-blocking highlights again the need for more, and more consistent, information from the tech companies. Many of the laws discussed in this chapter, as well as other fixes discussed in this dissertation, will not work unless the tech giants provide more information to regulators and the public. The call for more information has been taken up by researchers, journalists, academics, and government. There are also calls for more consistency between the different platforms and or standardized reporting as well as more information about takedown decisions (Gollatz, Riedl and Pohlmann 2018).

Disclosure of information helps regulators, but in a world where technology changes constantly, laws requiring particular kinds of information can quickly become outdated (Hwang 2017, p. 27). Some scholars argue that simply revealing data about takedowns is not enough. Kate Klonick, an expert in intermediary liability, believes that Facebook needs to follow due process, ensure representation on its ‘courts’ or ‘panels,’ and guarantee their independence (Klonick and Kadri 2018). A number of scholars and activists have called for disclosure (and other aspects of corporate behavior) to comply with human rights principles. They have suggested that disclosure be standardized across companies and platforms, and discussed the scope

---

49 Groups like https://onlinecensorship.org/ and freedom of expression activist Jillian York, director of the Electronic Frontier Foundation, try to help people whose sites have been taken down without explanation.

50 “Based on statistical reports, indexes and rankings we know that bare numbers have flaws, yet they are being interpreted and weaponised by various stakeholders. The aggregated numbers that Facebook, Twitter and Google have produced and disclosed are difficult to verify because researchers don’t have access to the ‘raw data’. Often times, reports on numbers conceal more than they reveal, as they eliminate important information on context and intent...If we want to better understand how companies make decisions about acceptable and unacceptable speech online, we need a more granular understanding of case-by-case determinations. This would allow to learn more about the scale and scope of policies and practices, and consequently, it would enable us to understand the ways in which the current content moderation systems are affected by NetzDG. Who are the requesters for takedowns, and how strategic are their uses of reporting systems? How do flagging mechanisms affect user behavior? And furthermore, does the NetzDG offer the ground for new reporting mechanisms of content governance online, including a powerful rhetorical legitimation for takedowns? Platforms should team up with social scientists and let academics survey users on questions like perceived chilling effects and self-censorship prior to posting.” (Gollatz, Riedl and Pohlmann 2018)
of what needs to be disclosed as well as what a disclosure regime would look like to regulators, academics, and the public (Kaye 2019b; Micek 2019; MacCarthy 2019).

Other scholars question the ability of judicial systems and regulators to rule on and regulate difficult technical questions (Hwang 2017, p. 29). However, that just means new rules regarding disclosure should include the development of regulatory expertise. Courts and regulators make decisions about complicated industries all the time. Banking and financial markets, pharmaceuticals, and aviation are all regulated and monitored. Saying something is complicated just means it’s even more important to regulate it. It also means that we will need to develop expertise in the legal profession and judiciary on technical matters just as exists now in patent and immigration courts (Micek 2019).

**Platform response to criticism**

The complexity argument is often used by opponents of regulation, who contend that government officials are uninformed dinosaurs incapable of regulating the social media platforms. This tone often seeps into media coverage of congressional hearings, with government officials being treated as lacking the know-how to regulate (Gottlieb 2018; Tibken 2018). Being patronizing is one way the tech companies try to block regulation and undermine their critics. Others include denial, deceit, evasion, lobbying and, finally, taking action. As we described in Chapter One, the firms’ responses have been affected by two preoccupations: economic interest, with the tech firms hoping to avoid costly regulation; and ideology, in that tech founders believe in the Silicon Valley credo of free speech and disruption. This is related to the “exposure effect”: They genuinely believe that their companies are helping the world. They spent years arguing that bringing people together was a good thing, caused the 2011 Arab Spring, and helped spread democratic values. Facebook CEO Mark Zuckerberg, Twitter CEO Jack Dorsey and others have built their empires on the premise that connecting people helps humanity (Grossman 2014).

Zuckerberg has repeatedly emphasized Facebook’s ability to bring people together: “The Facebook community has also shown us that simply through sharing and connecting, the world gets smaller and better,” (Zuckerberg and Sandberg 2012) and in 2017 he went even further and announced Facebook’s goal to “help one billion people join meaningful communities” to “start to strengthen our social fabric and bring the world closer together” (Zuckerberg 2017).

The exposure effect and ideas of escalating commitment apply here. After years of working to connect the world, the internet titans seem to be unable to reverse
themselves. European regulators and NGOs working with Google and Facebook note a difference between lower-level employees in Europe, who are sympathetic to European perspectives on regulation, and their superiors working in Silicon Valley, who are more entrenched in their views and have an inflated sense of their self-worth. It’s become seemingly impossible for them to revise their views.

Apparently, Facebook will do whatever it pleases, following its own guidelines and procedures without necessarily following the laws of the countries in which it works. In many cases this means that Facebook follows policies that are stricter than national laws. One example: Facebook enforces its ban on nudity even in countries, such as France, that allow nude photos under some circumstances (IPNESS 2011). Perhaps the most famous example was the 2016 takedown of the iconic Vietnam War photo of Phan Thị Kim Phúc by photographer Nick Ut, taken as she was fleeing a napalm attack. The photo was originally posted by a Norwegian newspaper editor as part of a series on war photography. After Facebook removed it, an editor from one of Norway’s most well-known newspapers put it up again; his account was suspended. Norwegian Prime Minister Erna Solberg and others protested the takedown; her post was deleted. “First you create rules that don’t distinguish between child pornography and famous war photographs. Then you practice these rules without allowing space for good judgement,” wrote Aftenposten editor-in-chief Espen Egil Hansen in an open letter to Mark Zuckerberg (Ingram 2016).

<table>
<thead>
<tr>
<th>Timeline of Facebook responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>August 2017</td>
</tr>
<tr>
<td>January 2018</td>
</tr>
<tr>
<td>April 2018</td>
</tr>
<tr>
<td>November 2018</td>
</tr>
</tbody>
</table>
January 2019  | Facebook publishes more details about its plan to deal with content removal, calling the system the “oversight board” (Clegg 2019).
March 2019   | Facebook announces a no-tolerance policy for “praise, support and representation of white nationalism and white separatism.”
September 2019| Zuckerberg fleshes out information on the governance system (Ingram 2019). Publishes the charter of the “Oversight Board” which says it will be an independent body with 11-40 members.
October 2019 | In a highly publicized and much-criticized speech at Georgetown University, Zuckerberg says that Facebook will continue to allow false political advertising on the site.

Problems with the responses of the tech companies—threat to national laws

Another worry, among people concerned about corporate power, is that Facebook’s public appeals process will make it harder to use national laws when pursuing complaints against Facebook. After Zuckerberg again announced that Facebook would establish a “Supreme Court,” analysts said such a court could do an end-run around national courts, as these might mandate that plaintiffs must “exhaust” options at Facebook first (Weinzierl 2019). Indeed, Facebook’s announcement in September 2019 (Ingram 2019) that it’s setting up its own appeals process to review takedowns pursuant to its community guidelines was remarkable in omitting any mention of national laws or regulations. Many scholars have pointed out the need to ensure that such panels complement and comply with international human rights laws and norms51 (Kaye 2019b).

Exporting the US model of privatization of justice

Over the years, the US has seen a privatization of justice as companies increasingly require customers and workers to sign away their rights to litigation over matters such

51 Peter Micek adds that “there are also clear, consensus-based international guidelines calling on companies to provide and participate in operational-level grievance mechanisms. The third pillar of the UN Guiding Principles on Business and Human Rights, on the right to effective remedy, places courts and state-based redress systems as the primary provider of access to remedy but also calls on the private sector jointly to participate in and explore non-judicial remedy. The UNGPs, themselves based on UN Special Representative John Ruggie’s ‘Protect, Respect, and Remedy’ Framework, were unanimously endorsed by the Human Rights Council in 2011. With certain safeguards for claimants, who should always retain their right to appeal to the courts while participating in non-state mechanisms, these can bring some relief in ways more efficient and less formal and costly than other mechanisms provide. Our 2013 paper, the Telco Remedy Plan, and David Kaye’s exploration of remedy in the context of the ICT sector’s human rights responsibilities, sketch a basis for such remedial mechanisms.” (December 2019, comment on this chapter)
as workplace disputes or allegations of illegal or deceitful business practices (Silver-Greenberg and Gebeloff 2015). Arbitration has become the norm for credit card companies, banks, and phone companies, as well as nursing homes and car rental companies (Silver-Greenberg and Gebeloff 2015). Investigative journalists have delineated how private firms require customers to agree to mandatory arbitration and how these proceedings are stacked in favor of the companies, as the arbitrators tend to be from the industries they purportedly are judging (Silver-Greenberg and Corkery 2015). A worst-case scenario for this author would be if terms of service are expanded to include binding arbitration in cases where users are dissatisfied with content moderation.\(^52\)

As Facebook proceeds with its own appeals process and its own governance system, national governments will need to ban binding arbitration and ensure that national courts and national laws are ultimately sovereign. Facebook’s plans for its own mechanisms/governance bodies/panels further underscore the difficulties facing governments seeking to regulate large companies.

**Impediment to regulation: the need to appoint one person who is accountable**

Another barrier to legislation is the requirement by some European countries to designate one person who can be held accountable.\(^53\) NetzDG required the tech companies to have such a person in Germany, but France does not yet have such a law for tech platforms, although it does for newspapers. French regulators say that headquartering in Ireland didn’t just help the large tech companies avoid taxes, but also helped them avoid personal accountability and strict data protection enforcement (Neate 2019).

“Every time the large American corporations get sued, they use procedural arguments, such as saying the French court does not have jurisdiction and/or that the claimant did not sue the right company. In most cases, American tech companies do not officially operate their web-hosting services from their local subsidiary, but only their marketing services. This means a French claimant is not allowed to sue Facebook’s or Twitter’s subsidiary in France to have French illegal online content removed. The court summons has to be sent to the subsidiary in Ireland or to the headquarters in the United States, which takes months and deters judicial procedures. Combined with

---

\(^52\) Not having read terms of service on the social media platforms I don’t know if this has already happened.

\(^53\) 29 July 1881 (French Law on Press Freedom): requires that someone in each newspaper, the “directeur de publication,” is legally responsible for its published content.
aggressive tax optimization practices, it gives an impression in Europe that there is impunity for the big networks,” said Paris-based attorney Benoit Huet, who has written extensively on French regulations against the manipulation of information (author interview 2019).

**Impediment: intransigence by the platforms**

The large tech companies do not just employ armies of lobbyists and spend millions to generate good will and good publicity in Europe. They also have found myriad ways to avoid complying with or to fight back against European governments.

Some examples:

- Studies found that Facebook decreased the effectiveness of the NetzDG law by creating two complaint mechanisms and encouraging users to file all complaints, even those about posts in violation of the law, through the standard feedback process. This meant that Facebook only had to include a fraction of its complaints (the ones made through the other complaint mechanism) in their “transparency report” (Gesley 2019).
- Twitter rebelled against French laws by using them against the French government. In April 2019, the company initially refused to promote French public service announcements encouraging citizens to vote because they argued it constituted a political campaign (BBC 2019c). The French government soon convinced the company to reverse the block.
- From *Politico*: "According to the EU, Facebook misled European officials into believing [Facebook and WhatsApp] would not exchange information, thereby allaying concerns at the time of the merger that WhatsApp users could be drawn into Facebook’s web. In 2016, a German court found that the two networks were indeed sharing data, and barred Facebook and WhatsApp from exchanging information about German users. The ban became unenforceable when the GDPR took effect and Ireland became the lead supervisory authority. Now, German authorities say the sharing has resumed and Ireland must crack down. For their part, Irish officials said in a statement they’re satisfied that Facebook and WhatsApp aren’t sharing information for the purposes of ‘friend suggestion or enhanced advertising’” (Vinocur 2019).
- Another method is to turn off information or refuse to supply it. “During the EU elections [Facebook] said ‘we are collaborating’ but then turned off their API so a startup that was trying to track the disinformation was no longer able to do so,” one government official told us (author interview, October 2019).
What the French are doing—hopes for comprehensive regulations

In the wake of 2016, the French government felt it also needed to act, and immediately began to look at election advertising and hate-speech laws. The wake-up call for France, said one official, was not so much the 2016 US elections as the appearance of purportedly hacked Macron emails on May 5, 2017—just before the election news blackout was about to begin. Under Article L49 of the French electoral code, no news that is material to a political campaign can be published or broadcast within 44 hours of a national election campaign (“Article L49” 2011).

The emails released during the hack (some true and others false) made a number of allegations about Macron (Donadio 2017). One purported that he owned hidden bank accounts in the Cayman Islands. Although this rumor was not covered by French media outlets, it spread on Twitter and Facebook (The Observers 2017) and could be seen by French citizens online.

The Macron leaks/hack made the government realize it “had no strong or consistent ways to fight back. People say it was a success in that the leaks didn’t twist the election results. But we were lucky and had a good reaction from the media, which decided not to spread the news. Their argument was that the timing of the leak was too suspicious to be honest or provide information for the vote,” said one official.

Interviews with government officials and regulators in Paris in October 2019 underscored the differing opinions within the government about how to proceed. Some felt it was important to tackle mis/disinformation at the source by confronting the Russian government. Others felt that educating the public not to be deceived by false information would be more effective. Eventually, the general view shifted from focusing on the source of the mis/disinformation toward the idea that people need to be empowered to detect and understand false information (author interview, October 10). In some ways this contradicts Mark Zuckerberg’s contention in his October 2019 Georgetown speech that knowing who is conveying a message is more important than whether the message is true or false (Romm 2019b). On the other hand, the regulators we interviewed seem to agree with Zuckerberg that audiences need to be empowered to make decisions as to what is true and what is false.

The Avia Bill

However, while it may have been the French elections that got Macron worried about mis/disinformation online, the government response was to tackle hate speech rather than mis/disinformation, because regulating hate speech is perceived as being less of a threat to free speech (Piquard 2019). Macron’s advisor drafted a strong hate-speech
law, which was then modified by members of the National Assembly. The so-called “Avia bill”, named after Laetitia Avia, a parliamentarian in Macron’s party En Marche, was approved by the National Assembly, modified by the Senate, and went back to the Assembly in early 2020 for reconsideration. In February 2020, the Senate proposed removing the misdemeanor for companies that failed to respond to complaints (Public Senat 2020). Like the German NetzDG bill, the French bill would fine the platforms if they don’t take down “obviously illegal content within 24 hours.” Anyone, including individuals, legal entities, and administrative authorities, can issue complaints. Unlike the German bill, there is no threshold for how many users the site must have, because the bill encompasses all communication services that offer the public a space to share content (social media companies and collaborative platforms like Wikipedia) and search engines. The fines would be up to 1.25 million euros.

There has been much discussion about the flaws of the Avia bill. Access Now opposes it, and Reporters Without Borders is worried about over-blocking of content; the group has been meeting with the government to discuss the issue. Others say that the judicial review system is far too slow and can’t handle a vast volume of content. The courts are not sufficiently staffed to help affected individuals issue multiple take-down notices or make decisions afterwards about harms caused by the mis/disinformation. Moreover, the European Commission weighed in at the end of November 2019 with its own criticism, contending that at several points the bill was incompatible with European regulation. Among other points, the Commission noted, the bill in practical terms would require the platforms to automate the suppression of content, and so would violate directives on freedom of expression and online commerce (Rees 2019).

Campaign Advertising Laws

In April 2019, the French government also issued Décret 2019-297 on election advertising, which will be discussed more deeply in the following chapter on election advertising. It calls for disclosure of sponsored content over 100 euros and disclosure of the name of the sponsor’s name and amount paid. A 2018 law regarding the manipulation of information also requires the platforms to put in place measures to disclose the funding and source of information in case it’s seen as potentially troubling to public order or affecting the balance of an election (“LOI n° 2018-1202” 2018). This law does not only apply to overt political advertising (banned before elections in any case) but, as one regulator explained: “If you say ‘immigrants from
Africa killed a white woman’ you need to disclose who put the content out and how much was spent on that content.”

When asked whether this law has been effective, one regulator replied that it’s not clear, as its implementation on social media has not been fully monitored. The Parliament and the judiciary are charged with ensuring that the 2018 laws are being followed, but clearly the French government needs to allocate funding for more oversight and regulators to make sure the new laws are implemented.

**Relevant French laws**

29 July 1881 (French Law on Press Freedom): contains a provision that “makes it illegal to disturb public peace through the publication, dissemination, or reproduction of fake news in bad faith” (Boring 2019). This law also requires that someone in each newspaper, the “directeur de publication,” is legally responsible for its published content.

In 2004, Article L49 of the electoral code enforces the election silence law, which states that 44 hours before a French presidential or legislative election, the French media cannot quote any of the candidates or their supporters, or comment on polling until the polls close. On April 14, 2011 article 4 of n°2011-412 extends this rule to the online world.

---

54 According to a French regulator talking about Law 2018-1202, “The law concerns a period of three months before a general election. It has three main consequences, only in the context of a political campaign (three months before the election):

1) If someone identifies ‘fake news’ within this three-month period before a political election that is spread ‘deliberately, artificially or automatically and massively’ on the internet, he can call on a judge. In 48 hours, the judge should decide if it is ‘fake news’ that could affect a ‘general interest debate’ (that is to say, that could affect the election). If it is the case, he can order the ‘fake news’ to be removed. For instance, if fake report such as ‘immigrants from Africa killed a white woman last night in Paris is spread ‘deliberately, artificially and massively,” we can imagine that it will fuel far-right votes. Therefore, you can ask ‘in emergency’ that a judge control the veracity of this information (i), the ‘deliberate, artificial and massive’ spread (ii), and its influence on the political debate (iii). If he considers that it is indeed fake news spread deliberately, artificially and massively that could affect the election, the judge can ask this content to be removed.

2) Platforms have to be transparent about who is paying for sponsored advertisements during a political campaign (during a period within three months before the election). For example, if, on my Facebook wall, an ad which has political content is appearing, the platform is obliged to be transparent about who is paying (and how much) for it.

3) The French Audiovisual Council (TV regulator) can decide, within this period of 3 months before an election, to stop broadcasting foreign television which is suspected of attempting to manipulate the elections, notably by broadcasting fake news that might modify the ‘sincerity’ of the election.” (email to author, February 2020)
Decree n° 2015-125 of February 5, 2015 implements article 6-1 of the 2004 Law on Confidence in the Digital Economy (LCEN) and article 12 of an antiterrorism law passed in November 2014 to allow the government to remove websites containing materials that incite or condone terrorism or display child pornography.

Law n° 2018-1202 of December 22, 2018, “Loi relative à la lutte contre la manipulation de l'information,” popularly known as the “infox” or “fake news” bill. This implements three measures to combat mis/disinformation. First, if an individual identifies a piece of “fake news” within an election season, they can call on a judge to determine whether or not it should be removed within 48 hours. Secondly, the platforms have to be transparent about who is paying for sponsored advertisements. Finally, the French Audiovisual Council can decide to stop broadcasting a foreign television show if it is suspected of attempting manipulation.

Decree n° 2019-297 of April 10, 2019 is issued pursuant to law n° 2018-1202. It sets the threshold at five million unique visitors per month. It provides that the information about the source of advertisements must appear close to the relevant content.

On July 9, 2019, the “Avia bill” is adopted by the National Assembly and, once enforced, would require platforms to remove hate speech flagged by citizens who can directly alert the digital platforms within 24 hours. The platform would have 24 hours to delete a piece of "obviously illegal content" and if not, the platform risks an important sanction (up to 250 000€).

The French approach to regulation is robust and enthusiastic. As regulators delve into the question of online mis/disinformation they feel comfortable proposing regulatory solutions. Even more than regulation, the French want more transparency and for the large tech firms to feel obliged to handle the problems that they have created. Understanding that the judiciary is too slow and the volume of illegal speech too large for government to handle, regulators want algorithmic transparency. In several interviews, regulators compared the social media companies to banks, and argued that government should take a supervisory role. They further argue that the platforms have a “duty to care,” meaning that since the regulators can’t solve every problem, the platforms need to be proactive and solve problems as they arise.

Limitations to the implementation of French regulations

When it comes to mis/disinformation, legal practitioners argue that existing French laws governing freedom of speech are sufficient and just need to be enforced online.
Benoit Huet explained: “When we talk about regulation, we already have tools like defamation laws that are efficient enough to sanction disinformation. They’ve been tried many times in history, and we know they work and are balanced. More regulation on freedom of speech is useless. The main obstacles encountered relate to judicial procedure, which is very difficult to conduct when the perpetrator hides behind the anonymity of the internet, and uses a social network domiciled abroad” (author interview 2019).

According to Huet, the main limitation is simply one of volume. Courts are not equipped to handle the volume of complaints quickly enough. “The answer won’t come from the judicial authority,” he said. “The volume is too big and won’t be quick enough.”

The clout and money of the large tech companies pose another obstacle to implementing regulations.

“Facebook, Google, and Twitter deploy huge means in Europe for lobbying, and created offices in Brussels, France, and Germany. They’ve recruited talented people and have been very good at lobbying. You can feel it strongly in the meetings organized by the European Commission. It’s somehow shocking that these companies are treated like [sovereign] states. When Zuckerberg comes to Paris, he is received by Macron personally as though he was head of state” (Huet interview 2019).

**Will the rest of Europe follow France and Germany?**

French and German regulation is far ahead of the rest of the Europe’s, and it’s unlikely their laws will be adopted by the European Union, European regulators say. In some quarters, France is viewed as wanting to punish US companies (author interview) and Germany’s NetzDG and its strict views on hate speech are viewed as resulting from its Nazi history (author interviews).

“It would take years and years to negotiate a NetzDG for the rest of Europe, because it includes aspects that constitute legal content in other Member States. It is relevant for Germany because of their historical experience. At the same time there is clear need to clarify rules on transparency for online platforms and procedures to remove illegal content,” said one civil servant.

Nor could France’s Avia Bill and 2018 hate speech law be adopted by the EU, in part because the EU can only legislate against illegal hate speech and not false information more generally. However, despite divisions within the EU and the complexity of layering EU regulations on top of local laws, the new EU administration apparently
has an appetite for more regulation of the platforms. Statements made in 2019 by Věra Jourová, vice-president of the European Commission, suggest there will be new proposals for platform regulation. We will discuss in the next chapter changes that might be expected to political advertising regulations, but certainly as regards mis/disinformation there may be a new mandate for regulation given the belief that self-regulation around elections has not worked as well as it should have. The Code of Practice on Disinformation is seen to be flawed and some EU officials say there is a need for far more transparency and information from the large tech companies. In July 2019, the Financial Times reported that the EU Commission was planning “sweeping new rules” that would upgrade the Digital E-Commerce Directive in order to make companies responsible for taking down hate speech and release more information about algorithms and (Khan and Murgi 2019).

In interviews in October 2019, officials said other steps being considered include updating the definition of disinformation to take into consideration that there is often a grey zone between what is true and what is false, as well as the fact that mis/disinformation is often amplified by bots for political purposes. The current definition provided in the December 2018 Action Plan Against Disinformation—“verifiably false or misleading information created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm” (‘Action Plan Against Disinformation’ 2018, 1)—is seen as too limited. Other options include cracking down on bots and fake accounts, limiting micro-targeting, requiring more information transparency (in part to empower users to understand the sources of the information they see online), and boosting the rapid alert and monitoring capacities of the EU (‘Code of Practice on Disinformation one year on’ 2019; Jourová 2019). Decisions about possible revamped rules are likely to be taken after the spring 2020 release of the final report assessing the 2019 EU Parliamentary elections and—if the Commission decides to go down the regulatory road—will need to be voted on by member states.

The next country seen as most likely to regulate is the UK, following the publication of the April 2019 Online Harms White Paper (Department of Digital Culture, Media and Sport 2019). Its proposals are comprehensive and overlap in some respects with ideas proposed by French regulators. In the White Paper, the UK government said its new regulations would encompass search engines, file-hosting sites, and social media platforms, just like the French Avia Bill (Article 19 2019). The White Paper also called for a regulatory agency to oversee compliance. However, given Boris Johnson’s victory in December 2019, it is not clear if, how, or when the UK will regulate. The White Paper itself will almost certainly be updated (Pomerantsev 2019).
Although response from British academics has largely been favorable, there has been criticism by University of Essex law professor Lorna Woods and others, who say that the meaning and responsibilities of platforms’ “duty of care” toward their users need to be clarified (Woods 2019). Damien Tambini, a senior lecturer at the London School of Economics, notes that the proposal further blurs the line between illegal speech and hate speech, and calls for clarification of the differences (Tambini 2019). Lilian Edwards, a leading academic in the field of Internet law, told the Financial Times that the “duty of care” responsibility would likely incentivize platforms to block content that is not illegal, but which could be perceived as harmful.

Many countries already apply national laws governing illegal speech (often including hate speech) to content online. However, there is a grey area between speech that is illegal and speech that does harm. Freedom of expression groups worry that broad mandates from government will expand the definition of what is harmful. Vera Franz tracks the debates closely, as she funds a number of think tanks and advocacy groups as part of her job as deputy director of the information program at the Open Society Foundation. She’s watching to see how far governments will go in mandating filtering by the tech companies. “They are moving towards this for illegal content, but are also considering it for harmful content, which is hugely problematic from a free speech perspective” (email to author, October 2019).

### Relevant European Union regulations

In **1953**, Article 10 of the European Convention on Human Rights stated: “Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers.”

**2000** E-Commerce Directive (ECD). Provides immunity from liability for certain ‘information society service providers’ (ISS providers) that meet the following criteria: qualify as an ISS provider within the terms of the ECD; and one of the following applies: the provider is a ‘mere conduit’ (Art. 12); provides caching services (Art. 13); or provides hosting services (Art. 14 ECD).

In **2005** the European Court of Human Rights argued that article 10 “does not prohibit discussion or dissemination of information received even if it is strongly suspected that this information might not be truthful” (*Salov v. Ukraine*).

October **2018** Facebook, Google, and Twitter sign Code of Practice against disinformation which implements self-regulatory standards to fight disinformation.
January-May 2019 EU requires platforms to submit reports and performs targeted monitoring to ensure the platforms adhere to the Code.

Highlights from the April 2019 UK Government’s White Paper

Department for Digital, Culture, Media & Sport, and the Home Office:

“A new statutory duty of care to make companies take more responsibility for the safety of their users and tackle harm caused by content or activity on their services.

17. Compliance with this duty of care will be overseen and enforced by an independent regulator.

18. All companies in scope of the regulatory framework will need to be able to show that they are fulfilling their duty of care. Relevant terms and conditions will be required to be sufficiently clear and accessible, including to children and other vulnerable users. The regulator will assess how effectively these terms are enforced as part of any regulatory action.

19. The regulator will have a suite of powers to take effective enforcement action against companies that have breached their statutory duty of care. This may include the powers to issue substantial fines and to impose liability on individual members of senior management.

20. Companies must fulfil the new legal duty. The regulator will set out how to do this in codes of practice. If companies want to fulfil this duty in a manner not set out in the codes, they will have to explain and justify to the regulator how their alternative approach will effectively deliver the same or greater level of impact.

29. We propose that the regulatory framework should apply to companies that allow users to share or discover user-generated content or interact with each other online.

30. These services are offered by a very wide range of companies of all sizes, including social media platforms, file hosting sites, public discussion forums, messaging services and search engines.”

Already, free expression advocates worry that NetzDG is inspiring imitators around the world and being copied by Russia, Vietnam, and Venezuela, among others (Mchangama and Fiss 2019). In 2019 Singapore passed a fake news law, while Malaysia repealed its 2018 Anti-Fake News Act. It was deemed too repressive because it made it a crime to produce, publish, or circulate misleading information—with the
government serving as the arbiter (Latiff 2019). However, it must be remembered that these countries controlled free expression and regulated the internet long before Germany passed its NetzDG law.

**One regulatory approach: Take a leaf from banking regulators**

Comparing regulation of tech companies to the regulation of banks comes up from time to time in the literature, and often rests on the idea that the tech companies have both disclosure and fiduciary responsibilities.

> "The UK is a leader in global banking in part thanks to its regulatory environment. [...] Applying similar regulatory principles can improve the economic environment in the UK for digital start-ups and scale-ups while creating more predictability for large incumbent firms." (Furman 2019, p. 6)

"Similar to banks, digital platforms should have information fiduciary obligations [...] Similar to the international banking system, perhaps multinational digital platforms should set up local branches accountable to both domestic and international banking regulations." (Winseeck 2019)

Cédric O, French Minister for Digital Affairs, told Techcrunch, “It’s just like banking regulators. They check that banks have implemented systems that are efficient, and they audit those systems. I think that’s how we should think about it.” (Dillet 2019)

What this might mean in practice is less clear. In the US, the supervisory role for government means that regional Federal Reserve Banks and other regulators monitor the banks. They may have staff working in bank offices, and banks are regularly audited and monitored to ensure compliance.

Evelyn Douek, an affiliate at the Berkman Klein Center For Internet & Society studying international regulation of online speech at Harvard, contrasted the language in the May 2019 interim report by a specially created French mission, “Creating a French framework to make social media platforms more accountable: Acting in France with a European vision,” against that of the German NetzDG law, and argued that the French government was proposing a more mild, preventive approach than the NetzDG law.

In what is likely a reference to heavily criticized German laws, the report notes that the punitive approach incentivizes platforms to over-censor content to avoid liability and therefore “does not seem to be a very satisfactory solution.” It therefore recommends a preventive, “compliance approach” that focuses on creating incentives for platforms to create systems to prevent content
moderation failures. The report draws an analogy to the financial sector: Regulators do not punish financial institutions that have been used for unlawful behavior such as money laundering but, instead, punish those that fail to implement a prescribed prevention measure, whether or not this actually leads to unlawful behavior (Douek 2019).

In an email to this author, Douek explains, “I do think that fines are punishment—my point was more that regulators tend not to focus on punishing for *individual* instances of money laundering unless there is a systemic failure; what regulators seek to do is create rules for and punish non-compliance with *systems* that seek to prevent money laundering. The point the French report was making was that asking companies to absolutely ensure there is no breach of content moderation rules is impossible, but asking them to have reasonable systems for the enforcement of their rules is not.”

My interpretation of supervision is slightly different from Douek’s. Banks around the world have been fined billions of dollars over the past decades for assisting with money laundering and a host of other activities (Flitter 2019). Nor is it accurate to say the financial institutions are “used” for unlawful behavior as, in fact, the banks sell themselves as being able to provide services such as tax avoidance. However, Douek’s point underscores the dilemma also found in the world of internet governance: Activities that are bad for society are not always illegal (racist speech, trolling of women, tax avoidance).

In its report, the French mission wrote:

> [A]pplied to social networking services, this type of intervention involves identifying the few generic obligations likely to create the right incentives, particularly by increasing the effectiveness of political dialogue with Member States’ political institutions and their civil societies. This implies imposing strong obligations on the transparency of key systems unobservable from the outside, i.e. the moderation system (and procedures for developing and updating the terms of use that underlies it), as well as the use of algorithms for targeting and personalising the content presented (Interim mission report 2019, p. 16).

These comparisons between the tech and financial sectors do not mention the robust supervisory role of bank regulators. But Douek is right that the French report and others are calling for the tech companies to put in place processes to enforce regulations.
As for what these processes could look like, we suggest that supervision is a form of public auditing as opposed to private auditing. Ideally, audits verify three things: company processes; company compliance with the processes; and, through spot checks, whether the compliance processes are working. If a spot check of 10 instances of supposed compliance finds that three were not actually in compliance, then that would suggest the system isn’t working.

How would this work with regard to the platforms? There are many areas that governments could supervise: Are the companies complying with laws regarding takedowns, disclosure of paid advertising, or the presence of political advertising? What processes are in place to ensure compliance? How often does the process in place not work?

Imagine a “know your customer” regulation. Is a single department in charge? What documentation is required from the customer? Can it simply be a referral? In the case of banking, a banker may say, “They were referred by a Swiss bank,” but this would be unacceptable documentation. If the process requires a customer to submit a set of pro forma documents and some bank officers routinely say that the pro forma documents satisfy the requirements, then these decisions need to be audited. First, to check the decision, so that it can be reversed if needed. And second, the system itself needs to be tested. This is part of why regulators need to know details about how many complaints Facebook and Twitter get and how they respond when they get them. In essence, the tires need to be kicked on the system itself. Supervision is essential but it will require far more detailed information from the tech companies, which is why transparency of algorithms, at least to regulators, is so important.

**Conclusion: modifying section 230 may incentivize the tech companies to remove more harmful speech**

Much of the anxiety about mis/disinformation online focuses on the consequences for democracy. We will tackle the question of elections and political advertising in greater detail in the next chapter. In this chapter we reviewed governments’ efforts to hold the tech companies responsible for the false information that circulates on social media. Globally there seems to be increased appetite for more regulation; increasing the liability of the tech companies is viewed as one way to get the companies

---

55 Thanks to my economist husband for sharing his ideas in the next four paragraphs.
themselves to reduce the amount of false information flowing online. In the US, discussion about Section 230 is based on the assumption that the threat of lawsuits will force the tech companies to take action on content removal in a way they hadn’t previously. But we’d suggest that given the companies’ propensity for ignoring or breaking laws, it seems sensible to look at European regulatory regimes which ensure that any penalties or fines include a clause that mandates higher fines for repeat offenses. For example, the FTC in 2019 levied a US$5 billion fine on Facebook after the company repeatedly violated a 2011 consent decree on how it would handle users’ data (FTC 2019; Lapowsky and Kelly 2019).

Lessons for the US?

In some ways, the EU and France are learning from Germany’s experience. The lesson for the US is that the tech companies need to be forced to share data and comply with regulation. Laws alone are not sufficient. They need to be backed up by enforcement, and supervision will be critical. NetzDG was intended to provide a strong incentive to the platforms to take down illegal content, but regulators still complain they don’t have enough information to determine the processes used by the platforms or the outcomes of their efforts. As Claire Wardle notes, the platforms should not “mark their own homework” (Wardle 2019) but if they don’t provide basic information then the government can’t mark it either.

Recognition of the need for stronger laws and implementation is just beginning. What at first was thought to be adequate, simply passing regulation, is now understood to be necessary but not sufficient. Laws without supervision leave intact too many incentives and opportunities for non-compliance, especially given the information complexities and asymmetries. The analogy to bank regulation is sound. Passing banking regulation has not been enough; auditing and supervision of financial institutions is essential.

Limitations: What if the problem is actually Fox News?

Some scholars have noted that the mis/disinformation being spread online and on the outer fringes of the internet is picked up and amplified by more centrist media outlets in the US. Limiting online intermediary liability in the US would presumably incentivize the tech companies to prevent such information from spreading, but might not address how mis/disinformation enters mainstream discourse once it has appeared online and is then pushed on to Fox, repeated in books, and amplified by legacy media outlets.
In their book *Network Propaganda*, Benkler, Faris, and Roberts (2018) show that many of the political conspiracy theories online originate from and are spread by the right. There is no equivalent level of disinformation and propaganda in the Democrats’ sphere of news and information. However, while false stories may begin in *Daily Caller*, *Zero Hedge*, *Breitbart*, and on Fox, they are then amplified by a mainstream media ill-equipped to defend itself from falsehood. As journalists try to get “scoops” and “balance” the information they receive—practices intrinsic to professionalized legacy media—reputable outlets inadvertently spread the disinformation peddled by the right. Benkler, Faris and Roberts found that the media’s need to provide balance meant that, during the 2016 elections, it gave air time to many of the egregious falsehoods disseminated on Breitbart and the *Daily Caller*, *Zero Hedge*, and *Infowars*. This problem may not be solved by modifying Section 230. For even if the US summoned the political will to pass appropriate legislation regarding big tech and the social media platforms, we would still be left with dishonest Republican operatives and their capture, and creation, of right-wing media outlets. As Guy Berger from UNESCO put it, “The problem in the US is a combination of bad attitude and false facts, and addressing intermediary liability won’t solve this” (author interview 2019).
Chapter 7

Regulation of online election advertising\textsuperscript{56} — getting beyond transparency

In looking for solutions to the problem of online mis/disinformation, we have to confront the problems posed by online political advertising. As in other areas, self-regulation by the tech giants has not worked well. Examples of targeted (and non-targeted) misinformation during elections abound around the world, but it was concern about Russian mis/disinformation targeted at US voters in 2016 as well as at voters in Europe that seems to have galvanized regulators.

At this writing, the EU is studying regulations for online political advertising to be announced in 2020, following the planned release of its assessment of the 2019 European parliamentary elections. Reaching supranational agreements will be difficult, if not impossible, as different countries have very different kinds of election advertising laws. It is also clear that in the US, government regulation will be slow to come, and that the tech giants are acting more quickly than government.

Regulating online political advertising is complicated by several factors. One is the need to determine what can be regulated and which regulations can be enforced, since many kinds of speech relate to politics. Paid political advertisements are subject to regulation in some countries, but these comprise only a small part of the universe of politically related speech.

\textsuperscript{56} I am grateful to Lili Levi, from University of Miami, who sent pages and pages of helpful comments and new ideas that helped shape this chapter.
In the US, we have two further complicating factors: our First Amendment protections have become extraordinarily broad and now include the rights of companies to speak, limits on what entities can be asked to disclose, protections for false speech, and, in many instances, protections for violent speech. Further, the US has a dysfunctional campaign finance system that has allowed money to erode its democracy, and a Federal Elections Commission that has been more or less paralyzed by partisan division and is unlikely to take action.

In this chapter, we first discuss the research showing that high-quality information is important for democracies to function. Then we look wistfully at the laws being discussed in Europe versus those in the US, and try to come up with ideas that could realistically be carried out in the US. Most realistic may be the strengthening of disclosure requirements and registries of groups and individuals that have paid for political advertising. The US may be able to invoke consumer protection laws or laws that protect consumers from false advertising, or might strengthen regulations around the use of data as California did in 2019. Some lawyers believe that broadcast regulations could be applied to online advertising too. We will also touch on some of the measures put in place by the tech companies. We argue that disclosure alone is not enough, and we survey the shortcomings of the ad archives set up by the tech giants. We propose some principles for regulation of online political speech as well as a voluntary public service program that would improve political information online during election periods.

**Background: the role of information in elections and governance**

There is a venerable pedigree to the idea that quality information provided by disinterested parties—often the media, academics, scientists, government agencies, and other researchers—is essential for healthy societies. For example, the scholarly literature reinforces a view that journalists have long held: By playing a watchdog role (Coronel 2009) and helping frame the agenda for public discussion (Norris 2009), journalists provide information that helps societies understand key policy questions and make sound decisions.

Because people get information from many different sources and filter that information through their own pre-conceived ideas, the political impact of media persuasion is highly contested. It's exceedingly difficult—if not impossible—for researchers to parse the precise effect of online political advertising, talk radio, Fox news, amplification of messages by politicians, or Russian propaganda (Schudson 2019). Some contend it is the reinforcement of the messages that makes a difference, arguing that television and legacy media coverage, as well as the debates between
Part III. Solutions addressing Supply/transmission

candidates, reinforced the lies spread by Russian disinformation campaigns in 2016 (Jamieson 2018). Others maintain that exposure to “fake news” makes no difference. Yet research by political scientists and economists has added to our knowledge of the effect of information on our societies. This recent body of work has repeatedly shown that not only does the media help educate the public and increase political participation, but it produces higher rates of government responsiveness and lower rates of government corruption (Besley and Burgess 2001; Weingart 2009; Adsera, Boix and Payne 2003). The press can highlight problems in governance and expose corruption.

Changes in voting patterns suggest that voters respond to new information. Studies in Brazil show that politicians who are revealed as corrupt are less likely to be reelected (Ferraz and Finan 2011). This effect is particularly pronounced in areas with more radio coverage (Ferraz and Finan 2008). In their 2008 paper, Ferraz and Finan note, “Our paper lends strong support to the value of information and the importance of local media in promoting political accountability. Thus, our findings are consistent with an emerging empirical literature that examines the role of information flows in shaping electoral accountability and public policy” (Ferraz and Finan 2008, p. 706).

There has also been a spate of research examining what happens when newspapers close. One well-publicized paper published in 2009 found that after one local paper closed in Ohio, voter turnout fell and the reelection of incumbents increased (Schulhofer-Wohl and Garrido 2009). This result is consistent with research done by Gentzkow, Shapiro, and Sinkinson (2011) who found that voter turnout rises when a first newspaper is introduced into a market. They argue that closings of monopoly papers will lead to declines in voter turnout. A more recent study found that closure of local newspapers led not just to lower voter turnout, but also to fewer challengers in mayoral elections (Rubado and Jennings 2019).

In the US, new research on the “Fox News effect” has found that the channel’s arrival in a region leads to electoral gains for the Republican Party (DellaVigna and Kaplan 2007; Martin and Yurukoglu 2017). The presence of Fox News also correlates to cuts in government spending. A 2019 paper, “Partisan Media and Fiscal Policy Choices: Evidence from U.S. Cable News Channels” (Ash and Galletta), describes the effects that media bias can have on local fiscal policies using the example of Fox News. It finds that while exposure to less explicitly biased news channels like MSNBC results

---

57 Chapters 2 and 9 of Cyber-War: How Russian Hackers and Trolls Helped Elect a President explain how traditional media-effects theories such as “framing and priming” apply to what happened in 2016.
in voting patterns that expand local government revenue, exposure to Fox News leads to a decrease in local government financing (Ash and Galletta 2019).

In other words, research suggests that media exposure affects political preferences and voting. Certainly, the presence of quality information is essential to democratic processes. The absence of quality information can simultaneously distort voting patterns, election outcomes, and government policies. However, quality information alone is not enough; such information needs to be easily accessible.

One of the drawbacks of the internet is that there is now so much information that it’s hard for people to distinguish quality information from what is not (Grosser, Hase and Wintterlin 2019; Tambini 2017; Xia et al. 2017). The proliferation of low-quality content and targeted mis/disinformation online has become an impediment to the circulation of quality information—during election periods as well as at other times.

Indeed, sociologist Zeynep Tufekci has warned, “It is increasingly clear that more speech can in fact threaten democracy. The glut of information we now face, made possible by digital tools and social media platforms, can bury what is true, greatly elevate and amplify misinformation and distract from what is important” (Tufekci 2018b).

Economists have been surprised to learn that more competition has not in fact benefited the quality of news. Rather, there has been a race to the bottom in some markets as media outlets scramble to compete for a shrinking pie of revenues. Not only has quality in many cases worsened—in part because of diminishing resources—but echo chambers have been aggravated as the media market has segmented (Mancini 2012). Further, public discourse and politics have been complicated by the rise of the notion of “alternative facts” (Blake 2018) and difficulty in reaching agreement on government policies when policy makers hold incompatible ideas about underlying data (climate change, for instance, or whether migrants commit more crimes than the general population).

All of this has enormous implications for the conduct of free and fair elections (Tambini 2017). Voters targeted by large volumes of personalized advertising or mis/disinformation before elections may become confused and/or may not turn out to vote (Allcot and Gentzkow 2017; DiFranzo and Gloria-Garcia 2017). Other researchers have documented paid advertising aimed at voter suppression in the 2016 election and the 2018 midterms. Tactics included ads giving false information about voting times and location of polling places, as well as “calls for boycott and voter intimidation or threats” (Kim 2018).
On the other hand, experiments by Facebook in 2010 and 2012 suggest that messages encouraging people to go to the polls do boost voter turnout (Bond et al. 2012; Rosen 2012). Using social pressure, such as threatening to publish the names of abstainers, is even more effective (Haenschen 2016). Social media seems able to affect voter turnout both positively and negatively. However, no one has been able to prove definitively what the effect on recent elections has been. Nor have the tech companies released detailed data as to who saw what messages in the 2016 elections, so it’s hard to pinpoint their precise impact.

**Nearly all speech is political**

Further complicating the discussion about regulating political speech online is the fact that many kinds of speech are political or related to politics in some way. Regulating online political advertising does not address organic speech—that is, unpaid messages disseminated by supporters of one political party or candidate or cause. Nor would regulation of advertising address violent or defamatory speech that can create an atmosphere of fear and intimidation, with political and social consequences. US law does not regulate false political speech, such as doctored videos, that are spread deliberately. All of these categories (online political advertising, issue messaging, violent and inciteful speech, organic political speech) may well affect elections and society more generally. The expansive interpretations of the First Amendment make it difficult to regulate most kinds of content, so we will narrow our focus to online political advertising and messaging and why we think it deserves scrutiny.

**What is different about online political speech?**

Political advertising is directly aimed at swaying elections. It is designed to be more effective at doing so than more generalized political speech. Some studies suggest that audiences are more receptive to persuasion during election periods (Huber and Arceneaux 2007) and more likely to believe information associated with a particular candidate (Newton 1999). As discussed in chapter 2, there is research suggesting that the reputation of the messenger is what helps make falsehood credible. Thus, messages from candidates or campaigns may be more persuasive than other kinds of messaging.

There is also research suggesting that false political ads may be reinforced by the presence of other false information. Political advertising on its own may have an effect, but it may be even more effective when reinforced by false statements from

---

58 Thanks again to Lili Levi for helping me think through these two paragraphs.
political campaigns (Jamieson 2018). Or false political messages may have more resonance when they are amplified by reputable media outlets (Faris, Benkler, and Roberts 2018).

Much of the discussion about potential harm from online advertising and the need for regulation is based on the assumption that such advertising can sway elections. If it didn’t, then why would people pay for such advertising? While media effects are notoriously hard to determine and the effectiveness of political advertising is often unclear (Allcott and Gentzkow 2017; Brockman and Green 2014; Wong 2018; Newton 2019), the fears of harms are so great that governments are no longer willing to hold off.

As with many other forms of misinformation, there is a blurring of lines between what is completely false, what is misleading, and what is taken out of context. The taxonomies created by Clyde Miller, Claire Wardle, and others are relevant here. Further muddying the waters is that politicians like to spin and make themselves look good. Cherry-picking of data, exaggerating one’s achievements, and inventing falsehoods about one’s opponent all predate the internet. In normal circumstances this could be dismissed as politics as usual, but falsehoods during an election may be more salient.

Yet, while election advertising and messaging may be influential, the US does not have many laws governing it online. Giving out illegal information about voting times and places is illegal, but no federal regulation requires truth in political advertising. State laws trying to require truth in advertising have been struck down.59

What may be different about online political speech than offline?

Given the rapid growth of online political advertising and messaging we are convinced this hands-off attitude is no longer appropriate. There are several characteristics of online political advertising that make it worthy of attention: its volume, its virality, and the fact that it’s highly targeted. As well, online ads are more

59 Ohio law makes it a crime to “[p]ost, publish, circulate, distribute, or otherwise disseminate a false statement concerning a candidate, either knowing the same to be false or with reckless disregard of whether it was false or not, if the statement is designed to promote the election, nomination, or defeat of the candidate” (Tau 2014), and gives the state election commission the right to enforce the law, but this was not upheld by the courts. Montana has also been unable to uphold laws against lying in political campaigns (Vasilogambros 2019).
partisan and ideological than ads on television. They are less informative, less issue-based, and more focused on mobilizing the base (Fowler et al. 2018, p. 33).

**Volume**

Although campaigns still spend most of their money on television advertising, spending on online advertising is increasing rapidly.

In the US alone, political advertising is expected to be around $10 billion in 2020, with $2.8 billion spent on online political advertising and Facebook expected to sell 59 percent of all the digital political advertising (Bruell 2019). Another estimate puts Facebook’s online political advertising revenue at $420 million, which is some 0.5 percent of its projected 2020 revenues of $84 billion (Ivanova 2019).

Because digital ads are so much cheaper than broadcast, the number of advertisements is staggering. On an average day in 2016, Trump’s digital campaign ran 40,000 to 50,000 variations of their ads on Facebook, and at key moments that number was much higher; on the night of the third presidential debate in October, the campaign ran 175,000 different ads on Facebook (Lapowsky 2016).

Although the number is mindboggling and vastly exceeds both print and broadcast, it is unclear how much each voter was exposed to this advertising—again, because data has not been released by the tech companies.

**Virality**

Not only is the sheer quantity of online advertising different from print and television, but the speed at which information circulates on social media is much faster, with content that promotes anger or outrage becoming more viral (Shu et al. 2017; Bakir and McStay 2017). Facebook has said it will not fact-check political advertising, but even where corrections and fact-checks exist, they will likely arrive too late to make a difference (Zajonc 1968; Benkler, Faris, and Roberts 2018). As we discussed in the chapter on fact-checking, the people who see incorrect information are not necessarily the people who see the corrections that may follow. In the case of election advertising or messaging, by the time voters see corrections they may already have made a decision on whom to vote for, or even have cast their ballot. The frustration felt by fact-checkers at Brazil’s Comprova stemmed in part from the fact that their corrections weren’t seen by most voters. Since we began writing this chapter, the US
primaries of 2020 presented a number of examples of false or misleading political messaging online that Facebook refused to take down.\textsuperscript{60}

**Microtargeting**

Part of the power of online political advertising may lie in the fact that it’s so highly targeted. An advertisement or message that is modified or targeted may be more persuasive than one that is not (Konitzer \textit{et al.} 2019). It’s telling that Google said in November 2019 (Culliford 2019) that it would no longer sell political advertising based on location or voting behavior such as party registration or turnout.\textsuperscript{61} Facebook, however, not only micro-targets its advertisements, but has staff working closely with—indeed, embedded in—campaigns in order to help shape their advertising (Kreiss and McGregor 2017) Banning microtargeting is one solution that has been proposed particularly by freedom of expression groups that are reluctant to regulate content. However, microtargeting and using audience information are core to Facebook’s revenue model, so it’s hard to believe the company will stop even if it no longer officially charges political parties for the service (Ghosh 2019).

As well as political advertising tied to a campaign, there are of course targeted messages that, while not overtly connected to a campaign, may help to sway voters. During the 2018 mid-term elections, Facebook ads targeted at swing-state voters linked philanthropist and financier George Soros with former President Obama and Hillary Clinton. These advertisements, paid for by right-wing groups, also accused Soros of being a Nazi collaborator and of paying for migrants to come to the US and Europe (Mergler and McLaughlin 2019). Whether or not such advertisements dampen turnout or help elect Republicans has not been documented by researchers, but it’s hard to see why such targeted and false messages—whether paid or unpaid—should be allowed to circulate widely without being regulated in some way.

Many governments agree, and so strengthened protection of personal data is starting to be used to combat targeted political misinformation online. In April 2019, the European Commission introduced new sanctions, amounting to 5 percent of the annual budget of any European political party or foundation that violates data

\textsuperscript{60} After several days, Facebook finally labeled a deceptive video about Democratic front-runner Joseph Biden as partly false: https://www.nytimes.com/2020/03/09/technology/manipulated-biden-video-trump.html. Twitter labeled the video as misleading after “it had been viewed more than five million times and retweeted more than 20,000 times.” Twitter also labeled the video as “manipulated” For another example from October 2019 of Facebook running a false ad about Biden, see here: https://www.nytimes.com/2019/10/08/technology/facebook-trump-biden-ad.html

\textsuperscript{61} Google still said that false videos were not a violation of its policies: https://www.reuters.com/article/us-alphabet-google-advertising/google-bars-elections-ads-from-using-political-leaning-voter-data-idUSKBN1XU2WT
Part III. Solutions addressing supply/transmission

protection rules in their campaigning (European Commission 2019). The General
Data Protection Regulation, implemented in the EU in 2018, made it more difficult
for campaigns to perform micro-targeting (McGee 2019). More complaints to
national GDPR regulators about discriminatory targeting and other GDPR violations
may also help address the problem of microtargeting of false political advertising
(Morrow interview 2019; Jus-Burke 2019). Overall however, the GDPR includes
carve-outs for political advertising and these were upheld recently by a Spanish court
(Morrow, author email, 2019).

In the US, expanding privacy regulations are seen by some as more realistic than
overhauling political advertising and campaign finance laws\(^\text{62}\). A May 2020 report
argued that a federal privacy law “can protect internet users from the discrimination
inherent to automated content optimization and limit the viral spread of harmful
messages.” (Maréchal, Mackinnon & Dheere). They cite an apt quote from Alex
Campbell on the Just Security website “Absent the detailed data on users’ political
beliefs, age, location and gender that currently guide ads and suggested content,
disinformation has a higher chance of being lost in the noise.” \(^\text{63}\)

Where the world is actually moving: funding disclosure and ad registries

Given the volume, virality, and microtargeting of online political advertising, interest
in its regulation has increased. The US has moved slowly compared to the European
Union, which asked the tech companies to set up a digital ads registry in 2019. Many
European countries, as well as the UK, already have robust election-advertising and
equal-time laws in place that they can build on. European countries maintain a host of
election laws not seen in the US, especially since the US revoked the Fairness
Doctrine in 1987 and, in 2000, eliminated two corollary rules which gave people who
had been attacked free time to reply\(^\text{64}\) (Pickard 2018; Leweke 2001).

For example, the UK has strict campaign spending limits that apply both to parties
and to candidates (Anderson 2019). The shorter campaign periods in parliamentary
democracies also mean less spending is needed for elections (author interview James
Deane). Some countries have “election silences,” a period before an election when
campaign advertising is illegal, as in France’s ban on advertising 32 hours ahead of the
election (Caruso-Cabrera 2017). However, this rule is hard to enforce online (Musiał-
Karg 2019).

---

\(^{62}\) A 2020 Gallup poll found that 72% of Americans oppose microtargeting of political ads
https://news.gallup.com/opinion/gallup/286490/oppose-micro-targeting-online-political-ads.aspx


\(^{64}\) These laws were known as the "personal attack rule" and the "political editorial" rule.
A number of countries place ceilings on what can be spent, and ban foreign funding of campaigns: 13 EU members, for instance, completely ban the acceptance of foreign donations (Berzina 2018). The International Institute for Democracy and Electoral Assistance has found that of 32 OECD countries, all but 12 limit politicians’ spending on political campaigns. Even in those 12, spending ends up being minimal compared to the US, for various reasons. Government limits on election advertising and/or government support for advertising and equal-time rules during campaigns can help minimize the need for spending (Waldman 2014).

There are also rules dealing with public-service messaging. Messaging ostensibly about a government program can easily be construed as promotion for the incumbent (Young 2004). For this reason, many governments limit the kinds of messaging allowed from government agencies during election periods (Young 2004; Young 2005; Young 2007). Regulators in Australia, Canada, and the UK, among others, have strict rules about government spending on public service announcements, advertising such as tenders or job openings, and messages that provide information about government activities. They attempt to walk a line between informing voters about what government has done and unfairly advantaging the incumbent.

French election-spending laws ban foreign funding, place ceilings on spending, require disclosure of spending, and prohibit paid advertising for three months before an election. In France, online political campaigns must disclose their donors and how much they receive from them (Tidman 2019). This disclosure rule applies to all information that concerns the public debate. “It doesn’t matter if it’s a click farm or an advertisement: The platform has to specify who put up the content. During election season the amount of money spent has to be revealed, too. If, for example, you put a statement online that says something like, ‘An immigrant from Africa killed a white woman last night,’ the platform has to disclose who put the content up and how much money they spent” (author interview with French regulator, October 2019). If the statements seem false, then it’s possible to go to a judge who decides whether they should be removed. French laws against false information during elections date to 1881 and are the most robust in Europe. They were augmented with the 2018 bill against the “manipulation of information” (Fiorentino 2019). It’s unlikely they will be replicated in other countries, but there is increasing interest outside France in regulation.

The landmark April 2019 Online Harms report from the UK government’s Digital, Culture, Media and Sport Committee described the electoral law as ‘not fit for purpose’ and called for more regulation, asserting that the platforms have a “Duty of
Part III. Solutions addressing Supply/transmission

Care” to their users (House of Commons 2019). UK researchers expect new regulations to be announced, though are uncertain what they will look like. Given Boris Johnson’s electoral victory in December 2019, it’s unlikely that honesty and transparency of election campaigns will be a priority. However, the government has signaled a rejection of US-style First Amendment absolutism (Tambini 2019).

Despite having detailed election advertising laws in place, European countries have seen plenty of examples of online mis/disinformation. Russian interference was a major worry in a number of elections, including the EU parliamentary elections (Apuzzo and Satariano 2019), the 2017 German parliamentary elections, during the Catalan referendum of 2017 (Alandete 2017), and the Brexit referendums (Kirkpatrick 2017). But, some of the most notorious examples of mis/disinformation came from supporters of the Tory party, the most famous being the advertisement on the side of a bus implying that if the UK left the European Union, it would free up money to help National Health Service: “We send the EU £350 million a week. Let’s fund the NHS instead.” There was also ongoing anti-immigrant and anti-Islamic messaging, as well as Euro-skeptic messaging online (Marchal et al. 2019).

New regulation expected from the EU Commission in 2020

The pressure put on the tech companies after these elections may have contributed to their steps to control mis/disinformation ahead of the 2019 EU parliamentary elections. Interviews in Brussels in November 2019 found that regulators felt the elections had gone well, but that more could be done to safeguard against online interference (Mogherini et al. 2019; Bentzen 2019). New regulations are expected after the EU publishes its own assessment (originally scheduled for early 2020) of how the 2019 elections went and whether the disclosure policies were effective. Already, officials have said they plan to implement more precise transparency requirements for political parties and to place further sanctions on the infringement of Europe’s data protection laws (author interview 2019). In November 2019, then-incoming EU Vice President Vera Jourová said the EU would try to implement regulation that would steer clear of any form of censorship and instead require all online political campaigns to be transparent about whose interests they represent (Schulze 2019).

Indeed, the new Political Guidelines issued by the president of the European Commission in September 2019 promised a European Democracy Action Plan, saying:

Our democratic systems and institutions have come increasingly under attack in recent years from those who wish to divide and destabilise our Union….We
should develop a joint approach and common standards to tackle issues such as disinformation and online hate messages. I will put forward a European Democracy Action Plan. It will address the threats of external intervention in our European elections. It will include legislative proposals to ensure greater transparency on paid political advertising and clearer rules on the financing of European political parties. (von der Leyen 2019, 20-21)

Other countries, including Ireland, are also considering a raft of transparency, labelling, and reporting requirements (Quain and Cheallacháin 2019). On November 5, 2019, the Irish Cabinet noted that self-regulation has been insufficient and approved a proposal to regulate transparency of political advertising during election periods (McGee 2019). It argued that paid online political advertising will require labelling, and identified transparency as crucial in combatting disinformation and ensuring quality information for voters. Further, the government said that the burden of proof for identifying political advertising will be on the seller, and that the proposal is “an interim measure until the establishment of a Statutory Electoral Commission which will oversee a wider reform of the electoral processes” (gov.ie 2019).

In addition, a meeting in Ireland of the parliamentarians from 10 countries who serve on the International Grand Committee (IGC) on Disinformation and ‘Fake News’ also concluded on November 7, 2019 that self-regulation is insufficient and that tech firms cannot act on their own to deal with “harmful content, hate speech and electoral interference.” The group recommended a moratorium on micro-targeted online political advertising. Signatories to the principles are Ireland, Australia, Finland, Estonia, Singapore, the UK and the US. The Committee said, “There is need for full transparency regarding the source, targeting methodology and levels of funding for all online political advertising but such controls should not be interpreted as a blanket ban on advertising relating to the political sphere” (Houses of the Oireachtas 2019). However, the group did not say how such a moratorium will be implemented or what the next steps are.

**US regulatory efforts on disclosure of political advertising**

Because of its very different system, the US is far less likely to adopt new regulation on advertising or other contested subjects anytime soon. We have long and expensive elections campaigns, a polarized and paralyzed Federal Election Commission, and court decisions that have dramatically expanded First Amendment protections for speech. This includes limits on what entities can be told to say, and may be used to argue against disclosure of political advertising. The US also has a president with
right-wing allies who have weaponized political speech on- and off-line and are therefore unlikely to regulate. While US federal regulations have stalled, states have moved ahead with both truth-in-political-advertising laws and disclosure laws, many of which have been struck down by the courts.

Although they have not passed, proposals have also been introduced in the Senate that include the labelling of online political advertising, maintaining a registry of buyers of political advertisements, and stronger enforcement of existing rules barring foreign spending on elections. A bipartisan bill, “The Honest Ads Act,” was introduced in the US Senate in October 2017 and reintroduced by Senator Amy Klobuchar in May, 2019 (Blake 2019). 7

Another proposed federal law was introduced in November 2019: The “Thune/Warner Filter Bubble Transparency Act” would bar large social companies from using algorithms to collect data or create targeted advertisements or messages to users without the user’s consent. Users would need to be informed when an algorithm is collecting data on them and would have the choice to switch to an “input-transparent algorithm” instead (Corrigan 2019). Political advertising could still be presented by the algorithms, but it would not be targeted.

Although transparency laws seem relatively straightforward, the tech companies have resisted them. Google lobbied against the Honest Ads Act in 2017 (Haenschen and Wolf 2019) and delayed implementing even weaker versions of disclosures. Facebook’s attempts to undermine regulation have been well documented by The New York Times and others. One example: after the Honest Ads bill was proposed, Facebook put pressure on the bill’s co-sponsor, Amy Klobuchar, and enlisted New York Senator Chuck Schumer to strong-arm sponsor Mark Warner (Frenkel et al. 2018). Opposition to disclosure regulations has not just come from the tech giants. Republicans of all stripes have come out against them, in one case four former FEC commissioners arguing that the volume of content makes it impossible to police (Smith, Goodman, and Toner 2019). Even the Washington Post, which went to court arguing that the regulations would be too onerous, has opposed them. It must be noted that in the US, laws that govern offline political advertising do not apply online. Indeed, political advertising in the US is protected speech and standards of truth in advertising currently don’t apply to political advertising online. 8

---

65 Joan Donovan has described this weaponization. More information available at: https://www.cigionline.org/data-governance-digital-age
The Honest Ads Act, sponsored in October 2017 by Senators Mark Warner and Amy Klobuchar, as well 26 other senators, including the deceased John McCain, would mandate that social media companies disclose the identities of funders of political and other ads and their target audiences. The bill would:

- Amend the definition of electioneering communication (set by the Bipartisan Campaign Reform Act of 2002) to include paid internet and digital ads.
- Require digital platforms to maintain a public file of all political communications purchased at more than $500. It would include:
  - A digital copy of the ad and a description of the audience it targets.
  - The number of views generated, the dates and times of publication, the rates charged, and the contact information of the buyer.
- Require online platforms to make efforts to ensure that foreign actors don’t purchase political ads in order to influence the American electorate.

White paper by Sen. Mark Warner, August 2018:

- US government must prepare for more disinformation campaigns at election times.
- It’s imperative to label bots, determine origin of posts and accounts.
- Platforms should be liable for state-law torts (defamation, false light, public disclosure of private facts, and failure to remove fake audio/video content)
- Data Transparency bill is needed.
- Data protection legislation: no personal data should be processed unless in keeping with the law. First-party consent should be required, not third-party.

Given the lack of federal regulation, states have passed two kinds of laws. Some prohibit false statements in campaign advertising and the others require disclosure of who placed the ads. By 2014, the most recent tally available (Listes and Underhill 2014), 27 states had passed laws prohibiting certain kinds of false statements in online campaign advertising. Four of those state laws have been overturned by US federal courts (Vasilogambros 2019). These laws prohibit false information about “claims of incumbency, statements about endorsements, providing voter information about where and when elections are held, and veteran status” (Listes and Underhill 2014).

While courts have not upheld truth-in-advertising laws, Ravel and Wood (2018) note that the courts have consistently upheld regulations requiring the disclosure of political advertising, so using consumer protection laws may be a way to require more disclosure of political advertising. A number of states have passed their own disclosure requirements (Glazer and Haggin 2019). These include reporting
requirements and visible disclaimers saying who paid for an advertisement (Born 2019). As of December 31st, 2018 at least 4 states have enacted legislation that directly requires disclosure in political advertising and 35 “have statutes that address disclosures for online political ads in some form” (Kansas Legislative Research Department 2018, p. 13).

New York, Maryland, California, and Washington have passed their own funding disclosure laws (New York Governor’s Office 2018; Sanders 2018; Douglas and Nichols 2019 DISCLOSE). New York’s Democracy Protection Act requires disclosure of all online ads, advertiser verification, and registration with the NY Board of Elections, as well as an online archive. Seattle’s 1977 law required organizations that accept money for political advertising to provide information about the money’s source (Sanders 2018). The laws caused some tech companies to stop accepting political advertising in those states. As of July 2018, Google no longer accepts state and local political advertising in Maryland (Wiggins and Romm 2018).

State laws, too, have been challenged in court. Maryland’s 2018 Online Electioneering Transparency and Accountability Act was struck down by the courts in early 2019 for impinging on the First Amendment rights of the eight press plaintiffs (including The Washington Post and The Baltimore Sun; Wiley Rein 2019). The Fourth Circuit Court of Appeals affirmed the decision in December 2019.66 Opponents of new regulations argue that they are not necessary and will be onerous for media outlets (Glazer and Haggin 2019) and will likely cause the tech companies to stop accepting political advertising. As always, it’s not clear whether forcing disclosures violates First Amendment protections, say some scholars. The ACLU and Planned Parenthood won a case in 2014 against a law in North Carolina requiring doctors to show women seeking abortions an ultrasound of their fetus and deliver messages about the ultrasound.67 Citizens United, however, mandates disclosure of the sources of political funding.68

66 The decision can be found here: https://scholar.google.com/scholar_case?case=15107509448251661044&q=washington+post+v+mcmanus&hl=en&as_sdt=40006&as_vis=1

67 See Stuart vs. Camnitz: https://globalfreedomofexpression.columbia.edu/cases/stuart-v-camnitz/

68 Funding for political advertising is "political speech" and receives the highest protections, as per the Supreme Court in Citizens United, but sources of funding for political advertising should be disclosed. The 5-4 decision was approximately 180 pages, 90 of which consisted of a dissent. In affirming the BCRA’s requirement for corporations to disclose their spending in advertisements, the Court supported its holding in opining that, “[t]he First Amendment protects political speech; and disclosure permits citizens and shareholders to react to the speech of corporate entities in a proper way. This transparency enables the electorate to make informed decisions and give
Limitations to the tech company ad library

While regulation in the US has been fragmented and incomplete, the corporate response has had its own limitations. Since 2018, Facebook, Google, and Twitter have launched archives of political ads in the US, rolling them out in more countries in 2019 (Singer 2018; Lomas 2019c).69 Since May 2018, Facebook and Twitter have verified identities and locations of political ad buyers and disclosed this information to their audiences. They also have rules about banning foreign ads. Facebook ads targeting people in the US about “social issues” can only be bought by officially recognized US citizens (Daskal 2019). Still, inconsistency remains between the platforms (Ravel 2019), and studies of the ad archives created by the tech companies show that vast amounts of information are missing (Edelson and Day 2019). The archives (now known as ad libraries) are criticized for being inconsistent and providing inadequate information about content or who was targeted (Edelson and Dey et al. 2019; Leerssen et al. 2019).

A recent paper summed up many of the limitations:

Firstly, we discuss scoping: ad archives have faced difficulty in defining and identifying, at scale, what constitutes a ‘political advertisement’.
Secondly, verifying: ad archives have proven vulnerable to inauthentic behaviour, particularly from ad buyers seeking to hide their true identity or the origin of their funding. Thirdly, targeting data: ad archives do not document in meaningful detail how ads are targeted or distributed (Leerssen et al. 2019, p. 2)

A number of researchers and legal scholars have made comprehensive suggestions as to how the archives can be strengthened and how regulations can assist.

Regarding the scope of the registry, Wood and Ravel suggest strengthening the “capture [of] any political advertising, including disinformation, that is promoted via paid distribution on social media, as well as all other online political advertising. Second, existing loopholes in transparency regulations related to online advertising should be closed. Congress has a role here as it has prevented regulatory agencies from acting to require disclosure from so-called dark money groups” (Wood and

---

69In 2019 the name was changed to the Ad Library.

---

proper weight to different speakers and messages.” The Court also found that the disclaimers aid voters by ensuring that voters are fully aware and dissolve confusion about who is speaking the political message.
Ravel 2018, p. 1228). They also suggest a number of government nudges, such as a
default setting that would not expose audience members to political advertising at all
unless they opt in (Wood and Ravel 2018, p. 1269). The Treasury Financial Crimes
Unit, the FCC, and the FBI’s Cyber Crimes Division would all be able to enforce the
new regulations (Wood and Ravel 2018).

In terms of disclosure, Wood and Ravel note that since Facebook already screens for
guns, drugs, and weapons advertising (Wood and Ravel 2018, p. 1262), it makes sense
for them to further eliminate advertiser anonymity (Wood and Ravel 2018, pp. 1267-
1268). They suggest that platforms that accept paid political advertising institute a
“know your customer” rule and keep information about their customers on file with
the government—or release it if requested.

Rutgers law professor Ellen Goodman concurs: "Large platforms should be required
to implement Know Your Customer procedures, similar to those implemented by
banks, to ensure that advertisers are in fact giving the company accurate information,
and the database should name funders of dark money groups rather than their opaque
corporate names" (Kornbluh and Goodman 2019). One limitation: just as with
offshore bank accounts, it may be possible to register which US entity is paying for
online political advertising, but it’s impossible to know whether that US entity is
getting its funds from overseas. Another unresolved question is whether a know your
customer rule would affect issue advertising or just political campaign advertising
(Glazer and Haggin 2019).

A third set of recommendations argue that the tech companies should provide far
more information as to who has been targeted by online political ads (Wood and
Ravel 2018; Edelson and Dey et al. 2019; Leerssen and Ausloos 2019). Although the
tech companies have made some changes since Wood and Ravel made their original
2018 proposal, they note, “None has provided audience identifiers to enable counter
speech. Nor have they joined together or formed a co-regulatory arrangement with
the government. Some are attempting to ‘nudge’ users, but none has provided an opt-
in or opt-out for narrowly targeted political content. As it stands, without co-
regulation or comprehensive industry self-regulation, any positive reforms they make
may be changed at any time, with no accountability” (Wood and Ravel 2018).

As always, compliance by the platforms is essential (Hern and McIntyre 2019). As
Paige Morrow and others have noted, the lines between organic and paid speech are
blurred, and restricting one could lead to more of the other. Morrow further writes,
“Regulating online political advertising could lead to increased use of astro-turfing as
campaigners resort to activity that is not categorised as advertising [in order] to increase their reach” (Morrow 2019).

**Principles of regulation: precautionary and do no harm**

Given so much uncertainty as to the influence campaign messaging has on voters, we suggest some principles and desired outcomes need to be applied to the regulation of online political advertising. The first is that regulation should be precautionary. Given the possibility of harm from false or misleading political advertising, regulations are appropriate. If scientific evidence emerges later that such regulation is not needed, it can be revoked.

At the same time, regulating when one doesn’t have all the evidence in hand means a different kind of regulation. So a basic principle has to be that regulation does no harm; any regulation should add to the greater good if it’s effective but cause no harm if it turns out to be unnecessary. Penalties and mandates, for instance, might be varied to address different types of content. In his analysis of the UK government’s “Online Harms White Paper,” Tambini (2019) points out that the government makes recommendations about content that is illegal and content that is not. He suggests that regulators impose fines for actions that are illegal and require disclosure for content that may be harmful but is not illegal. This distinction could be helpful in the US context, where there is a very broad view of what is considered legal—and thus protected—speech.

Ideally, regulation in the US should aim to achieve four goals:

*Preserve the best of journalism and the best of social media.* At its best, journalism excels at gathering quality information from audiences and then quickly disseminating information on important subjects. The beauty of social media is its ability to quickly disseminate information to many people. But if the information is of low quality, then the false information becomes a public harm. In that case, social media’s reach not only has no value but in fact becomes a negative value. In an ideal world, such contamination would be barred or taxed, but First Amendment principles prohibit this. Therefore, something needs to be done to guarantee the quality of information that is distributed widely.

*Provide access to broadcast.* There should not be a market for access to the marketplace of ideas. The right to convey factual and relevant information for elections should be open to all political parties and candidates, not just to those who can pay for it. This is why many countries have laws giving equal time to different political parties or providing free advertising on public broadcasters. In the US, it is most likely that such
provisions would be weighted to the two mainstream parties. The space given to lesser parties or candidates can be determined by a basket of factors that could include their polling numbers, votes garnered in previous elections, number of donors, and other criteria, such as those used to determine who is included in debates before presidential primaries.

Distribute the same information to recipients. Another important principle is that audiences should receive the same information. One of the media’s key roles has been to provide quality information about contemporary issues so that the public and policy makers can make decisions about what matters. The role of journalism in agenda-setting helps create a marketplace of ideas so that government and the public can discuss and pursue policies that are necessary and effective (McCombs 1997). For this to work there has to be commonly shared information. The reason micro-targeting can be dangerous is not just that it invades the user’s privacy, but also that the information provided is shaped for the intended target. Politicians may tailor their messages to their audiences, but imagine a message delivered to just one Facebook user that is completely different from the message delivered to a user next door with different political beliefs. Or a deep-fake video shown to one person or community and not another. In his book, *This is Not Propaganda*, journalist Peter Pomerantsev describes how in the 2016 Trump campaign, the Five Star movement in Italy, and the Vote Leave groups in the UK microtargeted potential voters with completely different ideas and desires (Pomerantsev 2019). Five Star published posts on Facebook devoted to entirely different subjects, one complaining about potholes and the other about immigration. Vote Leave targeted people who thought the EU was cruel to animals and then broke that down into sub-groups, “sending more graphic ads with photos of mutilated animals and gentler ads with pictures of cuddly sheep to others” (Pomerantsev 2019).

Find a way to make high-quality information salient. Because of the crowding-out problem— that is, the excess of low-quality information online— there has to be a way to highlight or show high-quality information to voters. In order for this to happen, there need to be flows of solid information, as well as some kind of labeling or checkmark system. Google News indicates types of articles, and labels whether they are on a topic that has been fact-checked (Gingras 2016). In December 2019 Facebook announced it would consider labeling political ads that have not been fact-checked (Gilbert 2019) but so far, labeling efforts are basically invisible to the average internet user. While the tech companies have said repeatedly that they are platforms not publishers, it’s clear that they make editorial judgements constantly. Rather than simply circulating/highlighting paid false content, they need to find a way to bury it.
That burial should also be done in a transparent and predictable fashion and, ideally, would discourage people from trying to circulate false information to begin with.

The above principles should make clear that disclosing sources of political advertising, keeping a list of who has paid for ads, and strictly enforcing bans on foreign funding of elections are the absolute minimum of what is needed to protect the quality of election advertising. However, in the US two other factors further complicate regulation. One is the First Amendment, which is understood to allow far more speech of all kinds than in other parts of the world. As Zuckerberg intimated in his October 2019 speech at Georgetown University, false political advertising is protected speech in the US. So are hate speech and other kinds of speech that would be barred in Europe and elsewhere. As we discussed in the last chapter, the power of the First Amendment makes regulation more difficult and often means that regulators fall back on disclosure instead of regulation. 70

A second problem in the US is the dismal state of campaign finance laws generally (Flavin 2015; Haenschen and Wolf 2019). Intense lobbying by corporations, the US Supreme Court decisions on corporate speech (Citizens United v. FEC), and reluctance among elected officials and politicians to restrict campaign financing have all contributed to an environment in which money has a powerful influence on elections (Thurber 2018). Further, scholarly research suggests that Google and Facebook make arbitrary and opaque decisions as to what information they allow online, and lack a transparent appeals process regarding campaign and election-related news (Kreiss and McGregor 2019).

Why activists do not want to see a ban on political advertising online

Given how hard political advertising is to manage, some have proposed banning micro-targeting or political ads all together (though this does not address the problem of ‘organic’ speech). Indeed, faced with a morass of false political advertising and a barrage of public criticism, Twitter announced in October 2019 that it would ban all paid political ads, while Google said it would ban microtargeting of political advertising. Still, banning or restricting online elections advertising is controversial. Many activists on the left as well as the right believe that online political messaging
Part III. Solutions addressing Supply/transmission

and advertising are essential for fundraising as well as movement-building. For the left, bans on such advertisements may be particularly painful, as they would favor incumbents, particularly since in the US the right has an entire media infrastructure at its disposal, including Fox television (McGowan 2019). Tara McGowan, of digital political strategy group Acronym, goes so far as to say that a ban on political advertising would help President Trump win the 2020 presidential election. Moreover, she argues, “any efforts to apply pressure on Facebook to do this would severely harm Democrats’ chances of winning next November” (McGowan 2019).

More disclosure ideas

Disclosure is the minimum. Law professor Lili Levi suggests some basic ideas as a starter. The social media platforms should:

- reject false political advertising that does not contain a picture or voice of the candidate;
- adopt disclosure-based rules that would provide context for ads not attributable to the campaigns;
- adopt disclosure-based rules that indicate that candidate ads have not been checked for accuracy (or have been deemed suspect by the platform’s fact-checkers);
- maintain better and more complete archives of all political advertising—there has already been research (cited above) on the shortcomings of the archives as they exist.

Our proposal: a Voluntary Fairness Doctrine

Regulation requiring broad disclosure of the funding sources and registration entities that pay for advertising and messaging online is not enough. We also propose a voluntary code of conduct for the provision of election news during US election campaigns. Election periods are a time when the tech companies need to promote quality information and, in doing so, they would help strengthen US democracy. Since they seem to believe that companies are more capable than governments—and that social media is a force for good—this is a chance for them to test their beliefs.

We suggest that the tech companies look at and apply broadcasters’ principles on providing equal access and equal information (Zuckerman 2019; Pickard 2018). Under a “Voluntary Fairness Doctrine,” the tech companies could take on responsibility for providing equal information to voters at no cost. For example, they could provide equal amounts of campaign advertising for free to Democrats and Republicans and target all US citizens who use their services in the months before the 2020 election. Each item would be labelled as to where it comes from. Ideally, each item would be
fact-checked before appearing. But if not, each piece of information should be labelled with a disclaimer that the information within has not been factually verified; a list of verified websites would be provided to encourage audiences to use their critical thinking skills and visit reputable websites to check facts that might be fabricated or presented out of context (Berger, author interview 2019).

Alternatively, if the companies don’t want to be accused of selling time/space/prominence to the highest bidder, they could simply accept payment with a cap on the amount spent per sponsoring entity or cause. Both approaches are complicated, but not impossible. The question is whether this might discriminate against those who can’t even pay the entry-level costs despite a fixed ceiling.

**Providing Free Advertising to Registered Political Parties**

A different approach would be to provide free ads equally to registered parties only, since providing free ad slots to other entities promoting various causes might be too demanding (Berger 2019). Or the tech companies could abolish political advertising altogether and disseminate equal amounts of news about each campaign’s position on key issues to US citizens for six months before the election. Given Facebook’s push toward supporting news, perhaps information supplied by campaigns could be included in that initiative, though it would need to be clearly labelled. If the social media platforms don’t want to simply disseminate campaign information, they could curate posts on a particular topic and then explain where the candidates stand on each issue, providing links for more information from trusted news sources. This would involve curation and making editorial decisions, but the companies already do that with a mix of algorithms and moderators. Publishing such explainers would be a public service and would level the playing field both in terms of the dissemination of such information (supply) and the access of audiences to this information (demand). The platforms already evaluate and reject some political advertising and make editorial decisions, so our proposal would not be much of a reach (Mac and Hirji 2019).

The public-service duties would simply be giving the tech companies a new function, which is to consciously add relevant information about public affairs to the online conversation rather than simply ranking or suppressing it. This use of counter speech is not dissimilar to what YouTube does with its Creators for Change program, which finds and fosters influencers who can post quality videos in order to counter false and inflammatory videos (Jahromi 2019).

We will discuss counter speech proposals in the following chapter. Here we note that critics of platform power may feel that the job of curating political speech should not be given to the large tech companies. However, the tech companies are already circulating political speech globally (including false information) and are profiting from it. Why not give them a public-interest mandate as well? This was how US
spectrum allocation was determined in the past and is part of the mandate of the major European public service broadcasters as well (Pickard 2018). We believe our proposal is not just constructive but also realistic, because it accepts the key role the tech companies play but builds on it by imposing a public-service requirement rather than simply punishing them. Ensuring equal access to quality information is perfectly in keeping with First Amendment principles.

In the fall of 2019, the tech companies released their own policies as to how they will deal with political advertising (see appendix C). However, these proposals fall far short of what is needed to help fair and democratic elections function. None of the proposals made in the autumn of 2019 provide equal time, none provide ranking as to salience, and none have said they will provide equal information. None have fully been able to avoid foreign funding or properly implement comprehensive know your customer policies. Only Google has said it will avoid micro-targeting of political advertising (Feiner 2019). Encrypted, closed groups and messenger apps where political messaging and advertisements spread should also be regulated. This is likely to be extremely contentious in democratic societies and will be opposed by groups devoted to freedom of expression. The battle to change American voting laws and to allow for more disclosure and transparency has been going on for decades (Allen and Stevens 2019). Setbacks like the Supreme Court justice in the Citizens United Case in 2000 aggravated the situation by allowing floods of corporate money into US elections (Mayer 2016). Drastic changes will need to be made and disclosure and labelling is only a part of a larger solution (Cagé 2020).

Limiting the spread and impact of political advertising online necessarily involves massive reform to a political system that is largely driven by money. For this reason many of the proposed solutions are work-arounds that try to tackle a part of the problem by avoiding the agencies or laws that are part of the problem. Since Congress and the Federal Election Commission are paralyzed then we turn to the states. When the courts overturn the states efforts on first amendment grounds then reformers try to use consumer protection laws. In this chapter, I’ve proposed voluntary efforts that the tech companies could do but those would almost certainly strengthen their power. There simply don’t seem to be any good options that can actually be implemented. Calls for limiting micro-targeting have public support in the US but it’s not clear whether laws like the California law can be extended and strengthened. As it stands, it is far too weak as it is an “opt-out” law that gives consumers the right to opt out of having their data sold. Far better, and stronger, would be to have an opt-in law so that data is only collected if consumers opt-in to have it collected. Future laws would also need to explicitly discuss political advertising. In any case, strengthening privacy laws
are certainly another effort that should be tried to limit the spread, or at least the effectiveness, of online political advertising.
Chapter 8

The failure of the “Marketplace of Ideas” and the insufficiency of “counter-speech”\textsuperscript{71}

In the last chapter we looked at online political advertising and how the policies of Facebook and other tech companies have distorted the information ecosystem by allowing entities with the most money or the best PR operation to dominate the discussion (Roose 2020). We also proposed a counter-speech solution: that Facebook and others provide political news to audiences at no cost in a way that resembles the traditional obligations of public service broadcasters.

We based this proposal on the traditional argument that the most effective way to counter false speech is to provide counter-speech: that is, to increase the supply of quality speech in the hope that reasoned debate will prevail and good ideas will drive out bad ones. Accordingly, we will use this chapter to consider some of the traditional views of counter-speech and the marketplace of ideas, and to see how these principles play out in a world where speech is increasingly online and there are few barriers to entry. We agree with legal scholars who argue that counter-speech is no longer a sufficient countermeasure to mis/disinformation and so will explore a related topic: legal suits aimed at protecting journalists who are maligned online.

\textsuperscript{71} Thanks to Phil Napoli for his comments on this chapter.
This chapter will summarize the counter-speech arguments, touch on initiatives intended to bolster counter-speech, and then discuss examples of journalists who have resorted to legal remedies in cases where they have been so harassed or defamed that they have been unable to fulfill their role as providers of information and ideas.

Although we put forward a counter-speech recommendation in the previous chapter, we are not starry-eyed about its effectiveness. It’s one thing to argue that counter-speech is the ethical path to take. It surely is. It’s another to argue that it’s adequate for addressing online mis/disinformation. It may be the correct moral choice, but is often not as effective as we would like. Like fact-checking and efforts to build trust in media, counter-speech efforts are worthwhile, but should be viewed realistically.

The arguments for the “marketplace of ideas” and counter-speech arguments

The idea that there is a marketplace of ideas has become conventional wisdom and is frequently cited by legal and media scholars (Nunziato 2019; Miller 2017; Hopkins 1996). Adherents of this view often argue that the best way to fight bad ideas is to promote good ones so that a discerning public can choose the best.

Journalists understand the concept to mean that all ideas should be put on the table of public discourse and the best ideas will win out. Intrinsic to the very notion of a “marketplace of ideas” is equating ideas to goods. Economists who want to limit government regulation on principle argue that a market is a market, there is no inherent difference between ideas and goods, and therefore markets for ideas and markets for goods should be treated in the same way (Coase 1974). Others, such as legal scholar Cass Sunstein, note that all kinds of communications are regulated and point out that arguments insisting they shouldn’t be regulated are often based on the economic interests of the groups arguing against regulation (Sunstein 2019).

For space reasons, we are not going to rehash decades of debate about the marketplace of ideas. We’ll simply note that regulation is essential for markets to flourish. There are myriad examples, from health and safety regulations, to speech-related regulations such as fraud and truth-in-advertising laws that protect consumers.

---

72 Or as Joseph E. Stiglitz put it: “That idea was disposed of by research by Arrow and Stiglitz in the 1960’s and 1970’s, showing that ideas and goods are fundamentally different, and that there was a presumption that the marketplace for ideas would not work well. Their research helped explain why we don’t simply allow a free market for ideas in commerce: cigarette companies cannot simply say cigarettes are safe for your health, or pharmaceutical companies that opioids are non-addictive. After all, so free marketeers might argue, free competition of ideas would mean that those who have better and alternative views concerning cigarettes or opioids would eventually win out” (author interview 2020).
against deception (Stiglitz 2020), to tort laws that ensure accountability if someone is injured by a product.

We talked earlier about the importance of ideology and how it has affected the choice of fixes to online mis/disinformation since 2016. The widespread belief in the marketplace of ideas has contributed to the view that more and better-quality speech is the answer to false speech found online. This route has been especially appealing to tech companies interested in avoiding regulation and to groups that believe in freedom of expression and worry that regulation would ultimately impinge on freedom of speech (Waldman 2017).

In his now infamous Georgetown University speech of October 2019, Zuckerberg said the company’s policies were “inspired” by American speech traditions, although he understands the First Amendment protection of speech “doesn’t apply to companies”73 and Facebook has the right to remove legal speech. “We think people should be able to see for themselves what politicians are saying. And if content is newsworthy, we also won’t take it down even if it would otherwise conflict with many of our standards,” Zuckerberg said at the time (Romm 2019b).

In January 2020, Facebook affirmed that it won’t fact-check or vet political advertising or remove false content if it’s newsworthy. His staff echo the point. “Our approach is grounded in Facebook’s fundamental belief in free expression,” said Katie Harbath, the company’s policy director for global elections (Matney 2019).

In doing so, Facebook seems to espouse a flawed idea that has become conventional wisdom: that the marketplace of ideas is competitive, and just as “leaving it to the market” is the best way to ensure the economy serves all, the best way to fight bad ideas is to counter them with good ones, letting a discerning public choose the best.

One theme running throughout this dissertation is that many of the proposed fixes have been undertaken because they are relatively easy to do. Although there are plenty of good reasons to support counter-speech arguments, their appeal is bolstered by the fact that it’s easier to add more speech than to try to pass regulations over opposition. In our interviews, one French regulator commented that countries like Hungary and Poland prefer “counter-speech” arguments and are all for supporting the media

73 His exact words were, “At the same time, I know that free expression has never been absolute. Some people argue internet platforms should allow all expression protected by the First Amendment, even though the First Amendment explicitly doesn’t apply to companies. I’m proud that our values at Facebook are inspired by the American tradition, which is more supportive of free expression than anywhere else.”
because their media are already captured and pro-government (author interview, October 2019). Another noted that given the lobbying might of the tech giants, it is hard in any case to get regulations approved (author interview 2019).

**Examples of counter speech — another supply-side solution**

As we’ve suggested, regulation fits into our supply-side model because much of it is aimed (directly or indirectly) at affecting the supply and transmission of mis/disinformation. In Chapter 3, we discussed some journalists’ efforts to build bridges with audiences and provide quality news in the hope that this will increase demand for high-quality information. Governments, companies, and individuals are also making efforts to supply accurate information or counter-narratives to offset mis/disinformation online. Indeed, the EU’s 2016 code of conduct for tech companies “includes commitments not only to assessing and removing hate speech that is found to be illegal, it also recognizes the value of independent counter-speech against hateful rhetoric and prejudice” (Klingvall interview 2019).

One example of counter-speech is the RespectWords initiative, an EU-funded local effort to combat hate speech by bolstering ethical codes for the media (OECD 2017). This was launched in 2017 and included independent media stations from seven countries: Spain, Austria, Germany, Greece, Hungary, Italy, and Slovenia. It produced radio broadcasts aimed at raising awareness on migration, minorities, and human rights in local contexts. The initiative included drawing up guidelines for reporting on migrants and minorities, which were presented in the European Parliament in October 2017. These included choosing language carefully, bearing in mind connotations behind words, challenging stereotypes and generalizations, acknowledging the complexity of stories, stressing the importance of context, and avoiding sensationalism and hate speech (Media against Hate 2017). The EU’s Stratcom also supported measures to counteract ISIL recruitment propaganda and Russian disinformation (Fotyga 2016), and to describe the EU’s achievements in countries where the EU was accused of being corrupt (Jozwiak 2015), however the project was opposed by the Dutch government (Rettman 2018).

Other strategies include the use of hashtags to overwhelm a particular message or person, or even the use of photos or images to overwhelm speech. After brutal images of the 2019 murder in Utica, NY of teenager Bianca Devins were posted on Instagram and shared to sites like 4chan and Twitter, other users tried to protect Bianca’s image by flooding the same hashtags with artwork and pleasant images instead (Leigh-Cooper 2019). This is how counter-speech can be used as a force for good, to drown out and block offensive speech and videos.
Impediments to counter-speech efforts: drowning out, fact-checking, market distortion

These examples are encouraging, but when thinking about the problem of mis/disinformation online we are persuaded by the point made by Nabiha Syed that “too often we analyze the problem of fake news by focusing on individual instances, not systematic features of the information economy. We compound the problem by telling ourselves idealistic, unrealistic stories about how truth emerges from online discussion. This theoretical incoherence tracks traditional First Amendment theories, but leaves both users and social media platforms ill-equipped to deal with rapidly evolving problems like fake news” (Syed 2017, p. 338).

And despite these inspiring stories, some scholars argue that counter-speech arguments no longer work in this day and age because there are flaws in the many “fundamental assumptions that underlie [the counter-speech] perspective” (Napoli 2018, p. 61). These flaws include the ideas that “individuals are capable of discerning between true and false information”, “[individuals] place greater value on true news and information than they do on false information”, and that “a sufficient number of those exposed to false information also will be exposed to the countervailing true information” (ibid.). Another point, made by Napoli, is that it is much cheaper to create and spread mis/disinformation than it is to create quality journalism or investigative reporting74.

As we’ve discussed throughout this dissertation, fact-checks are usually not seen by the people who saw the original mis/disinformation, or by many people at all relative to the number of people on the internet (Alto Data Analytics 2019). And there are a number of studies suggesting that when people do see corrections, they don’t necessarily change the views they formed initially (Nyhan et al. 2019).

Problems arising from internet culture

Drowning-out and threats of violence have to be understood as part of the larger online eco-system. Efforts to combat mis/disinformation naturally run into resistance, not just from tech companies that fear regulation but from foreign governments or ideologues determined to spread false information. Free expression groups worry about censorship and over-blocking, and US courts have taken a broad view of what is permissible under the First Amendment (Nossel 2017). But what happens when the

74 See for example Democracy’s Detectives by James T. Hamilton which includes a cost-benefit analysis of investigative journalism and notes that investigative reporting is time-intensive and expensive.
right of speakers to say whatever they want online, even if it's untrue or violent, collides with the personal safety or well-being of the person targeted?

**Crowding out of speech and incitement to violence**

Violent and hateful speech online is a particular form of crowding-out that can have consequences not just for politics but also for society. In Chapter 2, we wrote about the rumor of a conspiracy in a Baltimore restaurant that Clyde Miller and colleagues dissected and corrected in one of the weekly bulletins published by the Institute of Propaganda Analysis in the 1930s. Fast forward to 2016, when a barrage of online disinformation about Hillary Clinton running a pedophile ring in the basement of a pizza parlor drove Edgar Maddison Welch to show up at the pizza parlor with a military-style assault rifle in order to rescue the children he thought were being held in a sex-slave ring (Robb 2017).

In the US, with its prevalence of military-grade weaponry in the hands of citizens, mis/disinformation, violent speech, and incitement may be particularly dangerous. It’s one thing to call for violence against your political opponent in a country where people don’t have guns. It’s quite another to call for such attacks in a country with an estimated 120.5 guns for every 100 residents (Ingraham 2019). Indeed, the last couple of years have seen a dramatic increase in mass shootings carried out by young men, many of whom spent hours online talking to fellow white supremacists and being exposed to false information about immigration and other social issues (Cai and London 2019). And yet US laws on incitement are extremely narrow, protecting most kinds of speech and requiring a high bar before such speech can be banned. Clearly, this kind of speech has political consequences. Equally clearly, in the US it is difficult to regulate and much of the decision-making about taking down has been left to the tech companies.

As mentioned in the previous chapter, US courts have in some cases ruled that in order to meet the bar of illegality, threats have to meet a “reasonable listener” standard and the victim has to know of the threat (Rothman 2001). But these standards make no sense in the online world. Online threats are dangerous because it makes no difference whether a reasonable listener would understand the threat to be a threat. What matters is whether one angry, irrational, or unstable person hears the message and decides to act on it. The notion that the Pizzagate conspiracy was just hyperbole and not meant to be understood by a reasonable listener as a call to action simply doesn’t matter when in fact there was at least one armed man who decided that

---

75 Frequently cited is the use of radio to incite killing in Rwanda. Among other places, Philip Gourevitch discusses the topic in his classic book *We Wish to Inform You that Tomorrow We Will Be Killed with Our Families* (1998).
it was a call to action. Similarly, a victim would not need to know of an online threat for it to be dangerous. In fact, not knowing of the threat could make it more dangerous, because then the intended victim cannot take precautions.

Or consider the online marketplace for hiring assassins (Merchant 2020) or the website with names of doctors who carry out abortions (Vitiello 2000). Whether or not the intended victim knows of the threat doesn’t make a difference, as it doesn’t affect how dangerous the threat is or whether someone exposed to it will become violent. We mention these examples because an atmosphere of menace and threat undermines the argument that counter-speech is an effective response to online mis/disinformation — at least, on its own. An atmosphere of intimidation can also discourage those who might engage in necessary counter-speech.

We noted in previous chapters that exposing online mis/disinformation is one way of trying to combat it. It’s also a preferred method for those who want to avoid regulation of speech. Their hope is that by exposing the groups that put falsehoods online, the public will come not to believe them. Journalists point to the untruths told in the Brexit campaign and the lies that Donald Trump spreads in his speeches and through Twitter. Fact-checkers in Australia told us that their fact-checking causes some politicians to be more careful about their public statements. Peter Cunliffe-Jones and his colleagues work with African health officials to stamp out misinformation about vaccines. Many of these people believe that exposing online falsehoods and the people behind them is the best way to preserve a culture of truth, educate the public to understand what constitutes the truth, and preserve free expression. But despite these efforts, Brexit still happened, Trump lies daily, and the spread of misinformation in early 2020 about the coronavirus shows that simply exposing untruths is not sufficient.

**The Chilling Effect of cyber harassment**

<table>
<thead>
<tr>
<th>Forms of cyber harassment from OSCE/IPI and TrollBusters reports:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>-Cyberstalking:</strong> Sending numerous unsolicited messages to another person that cause the other person distress, anxiety or other forms of harm (McCully 2019).</td>
</tr>
<tr>
<td><strong>-Sending intimidating, threatening or offensive messages:</strong> Sending grossly offensive, aggressive and threatening messages that put an individual in fear of their personal safety or security (McCully 2019).</td>
</tr>
<tr>
<td><strong>-Online impersonation and trolling:</strong> Publishing harmful messages and communications about a person sent to a third party rather than directly to the victim that can subsequently come to the knowledge of the victim and cause</td>
</tr>
</tbody>
</table>
anxiety, stress or fear. Revenge porn, fake profiles, and photoshopped images are examples of this type (McCully 2019).

- **Online harassment campaigns (including “pile on” harassment):** A number of individuals sustaining a campaign of harassment, whether it be coordinated or arise spontaneously (McCully 2019).

- **Doxing:** Researching and broadcasting private or personally identifiable information about an individual (McCully 2019).

- **Online attacks:** Being hacked or having data stolen or compromised (Ferrier 2019).

Under international law, journalists (and everybody else) have the right to free expression. In reality, online harassment of journalists, particularly of women, has had a chilling effect on speech and the ability of journalists to express themselves. A survey of 597 women journalists around the world by the International Institute for Women’s Studies and TrollBusters found that 37 percent had changed their coverage to avoid certain stories as a result of online harassment, 63 percent said they’d experienced harassment online, and 58 percent said they had been harassed in person (Ferrier 2018).

Journalists in the US and elsewhere are regularly attacked and threatened online, especially those who cover the far right, expose communities of trolls, or write about disinformation (Wilson 2018; Sabbagh and Murphy 2019). In these cases, there is often little protection and what does exist tends to be inconsistent. A Committee to Protect Journalists survey of women journalists being harassed online found that in many cases newsrooms did little to help them (Westcott 2019).


Under international human rights law, the right of journalists to express themselves freely is recognized by Article 19(2) of the International Covenant on Civil and Political Rights (ICCPR), which provides that “[e]veryone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice.” This provision protects all forms and means of expression, including “electronic and internet-based modes of expression.” (McCully 2019, 29) https://www.osce.org/representative-on-freedom-of-media/413552?download=true
Supply-side solution: suing online harassers

Accordingly, where the online marketplace of ideas is no longer fully functional and harassment and threats prevent journalists from doing their work, some are resorting to lawsuits. In these cases, journalists, politicians, and others who have been abused online are not suing the tech companies for harms caused by third-party content, but instead are identifying and then suing the individuals who have put damaging information online.

The laws they rely on vary according to the local legal framework. Cyber-harassment violates Article 19 of the UN treaty that protects the right to gather and impart information (McCully 2019). In Germany, individuals are liable for putting out illegal or harmful information online and can be sued. Compared to the US, there are far broader protections in Europe for individuals’ rights to privacy, dignity, and reputation.

Under international law any interference with the right to freedom of expression needs to meet three criteria: (author interview with Jonathan McNUlly, January 2020)

--- Is there a legal basis for the interference? That may be pursuant to a country’s defamation laws, civil code, penal code, and judicial decision;

--- States need to have a legitimate reason to interfere with the rights of the speaker, For example: the interference has to be for the purpose of respecting the rights or reputations of others or for protecting national security, public order, public health or morals. For example, defamation laws are for the purpose of protecting the right to reputation of the person being harassed, as well as their right to privacy/dignity.

--- Is the interference necessary/proportionate? Here the question is whether the least restrictive legal measure was imposed for the purpose sought, e.g. criminal defamation would be more likely to be disproportionate than a civil remedy.

Where online harassment laws don’t exist, some journalists are even using defamation laws to go after their harassers. What’s remarkable is that in the past, journalists have opposed such laws because they are so often used to stifle free expression.\footnote{After I published an article on this subject in Columbia Journalism Review in February 2020 both Bruce Shapiro and Claudia Dreifus wrote to tell me about another example of a journalist suing for defamation, John Henry Faulk. Writes Bruce Shapiro “John Henry Faulk was [a] Texan leftie-populist radio personality - pal of Woody Guthrie,}
“Defamation would be more suitable as a response when you don’t have other options and other kinds of tort or legal remedy (such as harassment or cyber-stalking) available. Defamation is so emotive in media circles, and that is because it has been systematically used to suppress legitimate journalistic expression,” said Jonathan McCully, a legal advisor to the Berlin-based Digital Freedom Fund, an organization that funds strategic litigation on digital rights.

However, despite the past stigma, some journalists are using defamation laws and in doing so they may change norms. In Finland, Peru, and South Africa journalists use libel and defamation laws to pursue their online critics and trolls. This section will discuss these cases and the prospects for US journalists to follow suit.

Perhaps the most famous case is that of Finnish journalist Jessikka Aro, who became known for her investigative reporting exposing Russian troll farms. She was harassed extensively for years as a result and in 2016 worked with Finnish prosecutors to bring suits against four of her harassers, three of whom were journalists from a Finnish right-wing, pro-Russian alternative media site. In 2018 the Helsinki district courts ruled in Aro’s favor (McCully 2019). Ilja Janitskin was convicted on 16 counts and sentenced to 22 months in prison on criminal charges related to his website, while Johan Backman was given a one-year suspended jail sentence for aggravated defamation and stalking. Asta Tuominen, the third defendant, was convicted of aggravated defamation. Ultimately, Aro was awarded legal costs and $155,000 in damages (BBC 2018b).

The Finnish journalists’ union commented that the court’s 176-page verdict demonstrated its familiarity with tactics used by online harassers, including their use of tagging and targeting (inciting others to attack) and deceptive image manipulation (Pettersson 2018). However, Kimmo Nuotio, a law professor from the University of Helsinki, criticized the prosecutor’s approach by arguing that it did not sufficiently take into account the distinct nature of online harassment, which aims to harm the individual’s work more than harming the individual. Thus, such harassment is a public matter. (McCully 2019, p. 18).  

Alan Lomax and the rest of the pop-front left. He had a national radio program, ...on CBS Around 1955 Faulk was listed in Red Channels as a Communist, and lost his radio show. He sued Red Channels for libel and defamation, and after a long fight won - it was a crucial victory and pretty much ended the radio blacklist.” The New York Times obituary is here: https://www.nytimes.com/1990/04/10/obituaries/john-henry-faulk-76-dies-humorist-who-challenged-blacklist.html

78 Many of these new forms of harassment have been addressed by recent French laws (Quintin 2019).
In South Africa, defamation suits have long been used against the press, including during the apartheid period. More recently, renowned cartoonist Zapiro, the Avusa media house — publisher of the South African *Sunday Times* — and the paper’s former editor were sued in 2010 for 5 million rand (730,000 USD) by President Jacob Zuma for the 2008 publication of a political cartoon by Zapiro depicting Zuma preparing to rape a woman labelled as “the Justice System” (Zapiro 2008; Mundy 2010).

Now, legal changes have apparently made it easier for individuals — including journalists — to sue government officials for defamation on social media. In 2019 South Africa’s former Minister of Tourism, Derek Hanekom, sued former President Zuma for 500,000 rand (nearly 36,000 USD) after Zuma tweeted in July 2019 that Hanekom was an “enemy agent.” In September the court ruled in favor of Hanekom, saying the tweet was “untrue, defamatory and unlawful” and ordering Zuma to apologize to Hanekom, remove the tweet within 24 hours, and pay his legal costs and an unspecified amount of damages (Mtshali 2019; Nicolson 2019). Zuma was denied leave to appeal (Broughton 2019; Nicolson 2019) but at this writing it’s not clear if he has complied with the court ruling.

In 2018, two prominent anti-apartheid journalists, Anton Harber and economics editor Thandeka Gqubule, also filed civil defamation charges, after Winnie Madikizela-Mandela of the Economic Freedom Fighters political party gave an interview alleging that both had worked for the South African secret police during the apartheid era. The allegation was picked up and repeated by the EFF party, run by Julius Malema (Dawood 2019). Gqubule and Harber are well known in the profession and have garnered dozens of awards. Harber founded one of the leading anti-apartheid newspapers in South Africa (*The Mail & Guardian*) and Gqubule worked there as a reporter. Both went on to other jobs in journalism and Gqubule is now economics editor of the government South African Broadcast Corporation (SABC). The interview first ran on the *Huffington Post* in 2017 and the site took the interview down and apologized. But both Harber and Gqubule said in court filings that they became the subject of hundreds of vicious tweets, including ones calling for their death (Amato 2018). They sent multiple cease-and-desist letters to the EFF but these were ignored. Finally, worried for their reputations, they decided to file for civil defamation charges and asked for a total of R1 million in damages, which they would donate to charity. The Gauteng High court ruled in their favor on January 24, 2020 saying that the EFF had not provided any evidence of the allegation (Gerber 2020).
“As a journalist I have a deep aversion to defamation litigation,” Harber said in an email (author correspondence, January 8th, 2020). “We want the freedom to say strong things, even push the bounds of acceptable speech, and not have defamation law used to silence us — so we have to give others the same respect. We only embarked on this civil action because we could see no other way to stop things being said that were hurtful and harmful, even dangerous.”

Harber and Gqubule were not the first to go to court over government misbehavior online. The Court in 2019 ruled that Malema and the EFF violated the Electoral Code of Conduct in harassing journalist Karima Brown, but it didn’t make him apologize or pay the damages she asked for (CPJ 2019).

Legal changes have apparently made it easier to sue. In South Africa previously, lawyers advised plaintiffs to file complaints with the Press Council ombudsmen rather than start a costly and drawn-out lawsuit. However, the laws have evolved so that plaintiffs can now receive an apology rather than just money (or a ruling that something is untrue), and this may pave the way for other actions (Rabkin 2019). Plaintiffs can also now make an “application” rather than a “trial action,” which also speeds things up (Rabkin 2019).

**Peru: Gustavo Gorriti also suing for defamation**

Peruvian investigative journalist Gustavo Gorriti has several defamation cases underway and says that just starting the legal process caused four of his defamers to issue retractions.

“I don’t think a lawsuit should be the answer of first resort, but when you have a determined criminal organization of trolls and a mob attacking you, then of course you should be able to ask for legal protection. Journalists should take a much more determined approach to fighting back,” Gorriti said [telephone interview, January 2020]

Gorriti says that the harassment started after he reported that President Alan Garcia had received payments from Brazilian construction firm Odebrecht. The payments were exposed as part of the cross-border investigations into the Lava Jato scandal (Gallas 2019). Garcia’s supporters, and those of former president Alberto Fujimori, began to attack Gorriti, accusing him of bringing the “Jewish mafia to Peru” and of being in league with George Soros. The attacks worsened after Garcia’s death in April 2019, as prominent politicians correlated Garcia’s suicide with Gorriti’s reporting (CPJ 2019). First, he was attacked online with memes, tweets, and inappropriate reports and take-down requests targeting his response video on Facebook (Nalvarte 2019a).
Then, in June 2019, a small crowd began picketing his office, some with signs linking Gorriti to the Communist Party and Soros (Nalvarte 2019b). Gorriti began by exposing the trolls and blogging about them (IDL-Reporteros 2019), documenting their links to Fujimori (Arce and Wade 2008), who had been jailed for embezzlement and ordering killings while he was in office. But Gorriti then decided to sue for defamation and is asking for damages and jailtime. He doesn’t expect to get any funds but says that’s not the point of the lawsuit.

**Journalists often reluctant to sue**

Not many journalists are willing to resort to civil damages suits. They are expensive, time consuming and can result in more harassment. Michelle Ferrier, the founder of Troll Busters, notes, “There have been very few cases internationally. I advocate it as a tactic, knowing full well and with caution that this route is fraught with retaliation and not-so-good outcomes” (author interview 2020).

At Rappler, the feisty Philippines news site that has extensively covered the killings and human rights violation of President Rodrigo Duterte's drug war, the editors refuse to go after their harassers with defamation suits because they oppose criminal libel laws (email from Rappler managing editor Glenda M. Gloria, January 15, 2020). That is the position of many free-expression advocates, too. In his report for the OSCE, Jonathan McCully notes that a number of European and international human rights bodies — including regional courts, the European Court of Human Rights, and international government agencies — are opposed to criminal sanctions for defamation. “This criticism highlights that criminal penalties for harm to reputation have a greater potential to exert a chilling effect on free expression and notes that criminal defamation laws at a global level are frequently abused to silence opponents and critics” (McCully 2019, p. 20).

McCully warns that using such laws may inadvertently expand how such laws can be used and ultimately backfire against journalists. “Journalists are not only public watchdogs, they are bloodhounds. That is their role in a democratic society. We expect them to be persistent as part of their newsgathering role, to keep asking questions and to keep poking into places where the light does not reach. If journalists rely on harassment laws to go against trolls, they should also consider how these laws may also be used or weaponised against them” (McCully, telephone interview, January 2020).

Moreover, McCully believes that such lawsuits by journalists interfere with the rights of others. “Anytime you take a defamation or harassment case against someone
because of something they have said or done online, including trolling, then you are interfering with that other person’s right to express themselves freely online” he says (author interview 2020). “Their speech may be harmful or offensive, but the starting point is that it is protected. When a court upholds your case against this other person, the State is then also interfering with the right to freedom of expression.”

So, if journalists shouldn’t use defamation laws to defend themselves, what should they do? McCully and other free-expression advocates call for criminal defamation laws to be replaced by civil defamation laws. “Criminal laws should only be used in worst-case scenarios such as a threat to life, threat of violence, or egregious hate speech. In these cases, criminal penalties may be more likely to meet the proportionality test,” says McCully. “It is generally recognized that defamation should be decriminalized, and international law is going in that direction. For example, fines may be a more proportionate response than putting someone in prison for defaming another person. However, civil penalties can be disproportionate also. It is really for states to work out how to regulate this harmful speech in a softer way. It starts with educating people about harms and providing alternatives to laws that don’t result in the speaker having to pay huge amounts of money to defend themselves.”

McCully concluded his report for the OSCE/IPI with a number of recommendations, several of which have already been instituted in France (McCully 2019). As well as suggesting that criminal defamation laws be replaced by civil defamation laws, they call for recognizing online harassment as a threat to journalism and setting up “swift, low-cost, low-burden remedies” (McCully 2019, p. 39) for journalists who have been harassed.

French laws a model for other countries...

In France, such laws are in place and being used by journalists to protect themselves. In 2014, French law defined online harassment (e-Enfance n.d.) for the first time. Until then, online cases had been processed following offline harassment laws. A 2018 law “against sexual and sexist violence” closed most loopholes surrounding online harassment by broadening its definition (gouvernement.fr 2018). French law protects people, including journalists, from doxing, defamation, insult, and harassment online and offline (Boring 2019). However, cyber-harassment still occurs, as evidenced by a report commissioned by a French NGO (RespectZone 2018), which reports that French journalists in 2017 received 364,000 abusive messages online from 43,300 different users.
When this harassment is brought to light, people are sometimes held accountable. In February 2019 it was revealed that 30 prominent members of the French media were involved in a social media cyber-harassment campaign against women and they all lost their jobs (Timsit 2019). The attention this case drew helped lead to the successful passage of the 2019 online hate speech law, “the Avia bill,” in France’s lower house of parliament. This law requires social media platforms to take down hate speech within 24 hours of being notified, and may require the appointment of a special court to deal with online speech matters (Rosemain and Pineau 2019).

The case of Julie Hainaut in Lyon illustrates a successful use of the law against cyber-harassment. Hainaut, a journalist, was doxed and harassed online after writing an article about a colonial-themed bar in Lyon, “The First Plantation,” which opened in 2017 (Boring 2019). After she made five complaints in two years, the government finally prosecuted the case. One of her harassers was found guilty on the same day the Avia bill was being debated and was sentenced to six months in jail and fines totaling 7000 euros (Boche 2019; Council of Europe 2019).

"These laws on defamation and harassment work pretty well, as they preserve free speech and ensure journalists can fight back against trolls and harassers,” said Paul Coppin, legal director for Paris-based Reporters without Borders. “However, we still need sensitivity training for police and magistrates, as many seem not to fully grasp the impact of online harassment on victims and are not prepared to process complaints as quickly and seriously as they should be." (email to author, January 2020)

The new French laws taking into account the reality of cyber-harassment make it easier for journalists (and others) to sue in Europe. More generally, Europe has a broader view of what kinds of speech can be punished and a more robust defense of the rights of privacy and the rights of the individual than the US. In Europe, attacks on reputation and credibility can be litigated, so journalists have more protection than in the US. Scott Griffen, deputy director of the International Press Institute, notes, “In continental Europe, from a fundamental perspective, there is a clear balance between the right to reputation and personal dignity on the one hand and the right to free expression on the other. While freedom of speech in the U.S. enjoys priority over other considerations, constitutional law in Europe generally does not foresee a hierarchy of rights in this sense. So, in Europe it’s easier to adopt stronger protections against online harassment” (author interview, January 2020). Even so, it’s not clear whether this approach can scale given the vast amount of harassment online.

The US Example: less use of defamation laws but some recent cases
In the US, defamation laws have been used, but not very often, against people putting out false information (Seidenberg 2017). According to Sarah Guinee, a research fellow at the Knight First Amendment Institute at Columbia University, attacks on journalists are often aimed at eroding their credibility and reputation—essential for gainful employment—so plaintiffs have grounds to sue for financial damages. However, as we will explain below, potential plaintiffs often hesitate to use defamation suits against their harassers. The reasons have to do with US law and the personal and financial cost of protracted lawsuits.

**Broad protection afforded by the First Amendment**

As discussed in the previous chapter, US courts take a very broad view of the First Amendment, so the bar for a lawsuit is high. For example, hyperbole, opinions, satire, and imaginative invective are all protected under the First Amendment (Seidenberg 2017). “Because our definition of ‘fake news’ is knowingly limited to intentional or knowingly false statements, it is reasonable to conclude that such statements would satisfy the intent requirement for defamation claims. However, courts have generally afforded ample ‘breathing space’ to defamation claims involving satire or parody”79 (Klein and Wueller 2017). Speech directed at, or about, public citizens is also protected (Ruane 2014). Much less protection is afforded activists and journalists than given to people spreading false claims (Dorf and Tarrow 2017). Citing a 2012 Supreme Court case, US v. Alvarez, Dorf and Tarrow (2017, p. 20) note that “falsity alone may not suffice to bring the speech outside the First Amendment.”

“The US is a massive outlier when it comes to defamation and free expression laws. This is, on the whole, positive. In the US it’s quite difficult for someone to win a defamation case as a plaintiff, especially when the plaintiff is a public figure. This allows for much more robust coverage of politicians and others in public life,” says Scott Griffen (author interview 2020).

By contrast, some scholars argue that there is room for the use of defamation lawsuits as a remedy (Nunziato 2019). For example, “malice” does not always have to be shown in lawsuits by private individuals (Seidenberg 2017). Putting aside the legal arguments (and this is not a legal dissertation) there are many other reasons why people do not sue for defamation. Plaintiffs may be put off by the fact that the

79 As PhilNapoli notes: on Facebook where accounts label themselves as satire/parody as cover to disseminate all kinds of disinformation; and more often the end user has no idea that the content emanated from a satire/parody source? (comment to author in December 2019) .
Part III. Solutions addressing Supply/transmission

lawsuits are expensive, can drag on for years, and end up calling attention to the false information (Seidenberg 2017; NPR 2016).

**Lawsuits expensive and can produce a backlash**

There are plenty of reasons why journalists might not want to sue. Sarah Guinee notes that most of the journalists harassed online are women and are often reluctant to go to court. Lawsuits can be costly, time consuming, and emotionally draining. They can drag on for years. Much of the harassment is aimed at eroding the credibility of the journalists; defamation suits can open up their personal lives to scrutiny.

“To sue someone on defamation charges, you have to show that something is false,” says Guinee. “If there was a coordinated campaign saying that the journalist slept with someone who leaked information to her, are you really going to litigate your sexual history? That is not really worth it” (author interview, January 2020).

She has found in her research that “journalists were discouraged from pursuing legal remedies. They are told either to do journalism investigations about their harassers or ignore the harassment.” She points out, though, that investigating and exposing harassers is “kind of win for the harassers, because the journalists are putting their energy into that instead of the reporting that incensed the harassers in the first place.” Being told to investigative attackers is especially impractical for freelancers, as they may not be able to place the pieces they write, Guinee notes.

With US law giving far less protection to journalists than to publishers and all kinds of speech broadly protected by the First Amendment, “I’d be very surprised if journalists bringing defamation lawsuits against harassers would be a successful tool in the US generally. It’s just very difficult,” says Scott Griffen (author interview 2020).

**Some US lawsuits: against Alex Jones and white supremacist outlet**

Nonetheless, despite the obstacles, successful suits have been launched against the white supremacist newspaper, *The Daily Stormer*. And families whose children were killed in the 2012 Sandy Hook massacre sued Alex Jones, the right-wing Infowars publisher, for defamation over his repeated claim that the event didn’t take place and was fabricated by groups seeking repeal of the Second Amendment (Nunziato 2019). Jones argued that his speech is protected by the First Amendment and has sought dismissal of the cases. In December 2019, a judge ordered Jones to pay legal costs and denied his request for dismissal (Vigdor 2019). Harvey Weinstein and other alleged #MeToo perpetrators are both facing defamation lawsuits and using them against accusers. Both sides are using defamation as a way of establishing the facts of allegations that are often too old to be prosecuted. (Jacobs 2020).
Changes to Section 230 would allow more lawsuits in the US

If Section 230 were modified, that could open the way for more lawsuits against the tech giants, as they could be found liable for harms caused by third-party content. This legal change would provide an incentive for journalists who want to sue (Royster 2017). “It may be necessary to hold third parties (websites that disseminate this misinformation) liable by allowing recovery for individuals in the instances in which the fake news stories are also defamatory. Courts should use the ‘notice and takedown’ provision of the Digital Millennium Copyright Act of 1998 (DMCA) as a model for imposing liability on websites that aid in the spread of fake news that contains defamatory content” (Royster 2017).

Others have argued in favor of a product liability argument, contending that using consumer protections laws may pass muster with courts that take a broad view of the First Amendment and generally don’t like restrictions on corporate speech (Nicas 2019).

Conclusion

Counter-speech arguments are another necessary measure in the ongoing attempt to counter mis/disinformation. We’ve argued, however, that the marketplace of ideas doesn’t function fully (if it ever did) and therefore counter-speech arguments on their own will not suffice. The counter-speech argument assumes that people will see efforts to counter misinformation and that they will change their ideas based on this new information. This may happen in some cases, but probably not enough to make counter-speech effective on its own.

We also argued that in order for the market for information to work, three conditions need to be met:

- The market needs to be transparent. However, given lack of transparency into who sent and who received which messages in the first place, it is difficult to target counter-speech efforts and to make sure that accurate information reaches the audiences exposed to mis/disinformation.
- The marketplace of ideas needs to be open to everyone. Instead, preference is given to those who pay.
- Finally, it cannot be a market where intimidation rules. Yet many parts of the online universe are full of violent speech and threats, and those who provide accurate information or expose the troll farms are often unable or unwilling to express themselves.
Part III. Solutions addressing Supply/transmission

Under these circumstances, counter-speech on its own is only partly effective. So we conclude that the marketplace of ideas needs to be supported by regulation. In chapters six and seven, we explored different legal frameworks that could help underpin the flow of information and ideas. These laws could help the flow of information but run the risk of curtailing free expression. The ability to get the right balance will depend on the robustness of our democracy overall, and the independence of the judiciary and the bureaucracy that helps oversee such regulations. Other institutions that can help check and balance the regulators, such as the media and academia, will also need to play a part.

Many societies have taken the view that special attention needs to be paid to hate speech and speech that can incite violence or civil disorder. Some countries have regulated those forms of speech through libel laws, bans on hate speech or, like Germany, holding tech companies responsible for third-party content. It’s clear that, because of Section 230, the US is an outlier compared to other parts of the world. We described how in countries where journalists have no other options, journalists are trying the supply-side measure of repressing false speech by using defamation lawsuits against their harassers/trolls online. Scholars such as Dawn Nunziato believe that such suits are a way to address the malfunctioning marketplace of ideas (Nunziato 2019, p. 1560). However, groups that work with journalists say lawsuits are often not a realistic option for many journalists as the expense, emotional toll, and risk of a backlash are too great (author interview with Sarah Guinee, 2020; email from Michelle Ferrier). They also point to the opportunity costs as journalists spend time exposing and reporting on their harassers.

We are left feeling again that the answer lies in changing laws to make it easier to sue cyber-harassers for damages. Other countries may want to emulate the French laws that follow the OSCE recommendations. However, as always, it’s not just about adding new laws, but about enforcing existing ones and making sure there are enough ways to implement the new regulations.
At this writing, the Covid-19 pandemic is also being called an “infodemic” as false information is circulating widely, causing a great deal of uncertainty and mistrust. To someone who has spent the last few years trying to understand the problem of mis/disinformation, the cast of characters is familiar: Some of the falsehoods come from targeted campaigns by the Russians and Chinese (Barnes, Rosenberg, and Wong 2020), others are spread by Alex Jones and those who seek to profit from selling cures (Bellingcat, 2020). Politicians such as Donald Trump and Jair Bolsonaro are attacking science, journalism and anyone who disagrees with them and passing on substantial amounts of incorrect information in their press conferences and on Twitter. Far-right groups in the US—including three brothers who are gun rights activists—are taking out Facebook ads calling on people to “open the states” rallies. These extremists are being helped by a close associate of conservative US Secretary of Education Betsy DeVos (Stanley-Becker & Romm 2020). In short, the Covid-19 infodemic is the usual toxic mix of greed and political opportunism, enabled by tech giants such as Facebook and YouTube and the companies that profit from encrypted closed sites where such falsehoods and incendiary speech travels. There is, as always, also the role of the audience. Many people who forward incorrect medical information do so out of confusion or a desire to help (Larson 2018; Wasserman et al. 2019). In a situation where the science is still not settled and doctors and scientists are uncertain, it is understandable that people who don’t have scientific training become confused about what precautions to take. But it’s worse when the confusion of a pandemic is aggravated by cynical (and often hypocritical) people who should know better, such as Fox News denying the seriousness of the epidemic even while canceling Rupert Murdoch’s 89th birthday party on March 11th and stocking Fox’s offices with hand sanitizer (Smith 2020a; Smith 2020b). New research (Simonov et al. 2020; Bursztyn et
al. 2020) has correlated Fox viewership with higher death rates and lower rates of social distancing.

This dissertation is predicated on the idea that facts exist, that scientific methods help societies arrive at truth and that evidence and research are essential in all areas of life. This dissertation takes as a first principle that mis/disinformation online (and off) matters to society. distrust and fear make it impossible for governments to function. As we’ve seen around the world, the consequences are not just political polarization, but even violence.

Shrugging off online mis/disinformation and saying that lies have always existed and are therefore not something to worry about (Allen and Stevens 2019) is an inadequate and irresponsible response. Clearly, I am not alone in thinking this. The regulators who have begun to take steps to address the problem also worry about the effect that the barrage of misinformation online is having on democracy.

There are, of course, a number of studies that suggest that media persuasion effects are not as great as people imagine (Newton 2019; Allcott and Gentzkow 2017). There are notable scholars who have spent their careers researching media persuasion and media effects and parsing whether political advertising or news articles, online messaging, or false tweets change minds. Social psychologists have spent decades looking at confirmation bias, motivated reasoning, heuristics, how beliefs are formed, and how they can be changed.

As discussed in Chapter One, there is debate about the importance that source or message credibility plays in belief formation or change. More recently studies suggest that trust in media/government/institutions is affected by political beliefs, particularly in the US. So too, Newton describes that in the “Standard Model” factors such as age, demographics, income, education matter too (Newton 2019). Audience characteristics play a role. There will always be some people more susceptible than others to false information. An April 2020 study of how audiences access and rank news about Covid-19 found that while trust in scientists and experts was generally high across countries, people with low levels of education answered more questions about Covid-19 incorrectly (Nielsen and Fletcher et al. 2020).

One reason that it’s hard to measure effects from online mis/disinformation in a systematic way, and draw decisive conclusions that can applied to other cases, is that the defining characteristic of much of the mis/disinformation being circulated online is experimentation. Different falsehoods are put on social media or niche conspiracy sites or espoused by Alex Jones on Infowars, and some spread and are taken up broadly while others are not. It’s impossible to know in advance what will work.
Consider the 2004 “Swift Boat” ad attacking John Kerry’s war record. Paid for in part by a Texas billionaire named Harold Simmons, the ads were placed on television in swing states. In a tight election, the intent was to persuade even a few voters (Solman 2004; Morain 2008). The same was true in the US presidential election in 2016. For this reason, it is impossible to say that mis/disinformation has no effect. There are things we don’t know—how much one piece of false information matters or whether it’s an entire ecology. Whether one lie that catches on makes a difference. Contrived experiments can’t test this. As political scientist Alex Hertel-Fernandes put it, “Many of these studies focus on the marginal effect of, say, one additional ad and not the combined effects of a campaign's worth of advertising—which is likely to matter more when you start thinking about aggregate effects. Another important caveat is that these studies tend to focus on voting behavior and not, say, awareness of issues in the news and so can't really speak to the effects media coverage has on agenda-setting” (email to author, April 2020).

Apart from the general difficulties measuring media effects there is the problem, we keep returning to, of the tech companies refusing to share information as to who has been targeted, how and with what messages. This information is not available in the Facebook Ad library. Frankly the lack of information is a stalling tactic and an attempt to protect tech company revenues. This attempt to obfuscate creates uncertainty. But uncertainty as to the precise effects cannot become an excuse for inaction. There are many fixes being tried that will address part of the problems with little or no downside risk and that will be good for society. The earlier chapters of this dissertation cover some of the local and small-scale efforts but, while writing, it became clear to me that there need more comprehensive policies as well as regulation.

As a senior lecturer in a policy school, I wanted to finish this dissertation with an understanding of policies that could be implemented to help combat the spread of mis/disinformation. In this, I feel I succeeded: We conclude with a list of solutions that should be low-hanging fruit and there are many more that can be implemented, too.

**Qualifications/Limitations**

Trying to keep to a list of things that would make sound policies, I focused on problems that something can be done about. In the course of my research, I came across a number of methods being used to tackle the supply of mis/disinformation online that are simply too big for the scope of this dissertation. For example, some have argued that the problem of mis/disinformation coming from Russia is essentially a political problem and has a political solution, but this is not a dissertation about
geopolitics. Privacy legislation may help address the problem of online mis/disinformation particularly if it prohibits micro-targeting of political messaging, but an entire dissertation could be written on the GDPR, so I didn’t attempt to tackle that question in detail. I didn’t discuss debates about encryption and surveillance that are germane. Doubtless, countries with less economic inequality and better education policies than the US have more trust in government, institutions, and media. But I am not an economist, so I didn’t write about economic inequality and its role in creating mistrust.

I had also limitations in the access to sources. No one from the tech companies spoke to me except for one highly placed source who insisted on background conversations only. Later, I was able to speak to one person at Facebook who works on political advertising. In the beginning of my research, I contacted friends of friends, people I’d met, and the company spokespeople. After a while I gave up trying to email contacts at Facebook and Google. Friends would tell me about their wonderful friends at Facebook and swear they’d be willing to speak, but they never were. They’d pass me on to colleagues who’d promise to “find someone” but never did.

Another limitation is the geographic focus of this dissertation—as much as I tried to keep an eye on global trends and bring in lessons learned from Europe, this is not a global dissertation but is focused mostly on the US and Europe. The regulators and government officials in France and Brussels were extraordinarily welcoming. They spent hours with me, explaining their laws and points of view and policies. I met some of the civil servants in Brussels repeatedly over several years. These in-person interviews were essential, because the English-language press coverage of EU and European actions has not been particularly comprehensive and the EU websites are often hard for an outsider to parse. Through the time spent with me, these civil servants and regulators demonstrated a commitment to transparency and accountability, and eagerness to help a PhD student that was completely different from the indifference, platitudes, spin, and evasion of the tech companies I approached. The academic work done by colleagues who have closely examined new laws in Europe and the UK and added their own critiques has been invaluable, and I relied on that too.

Despite my attempts to address the problem of online mis/disinformation in platforms generally, in the end this dissertation mostly turned out to be about Facebook, because it is the biggest platform and a reference point for everyone looking at the circulation of online mis/disinformation. Conversations about “the platforms” often ended up being conversations about Facebook. In part that’s
because many of the regulators, like France’s ARCEP, worked on the principle that it was important to deal with the big companies. And, as little as we know about Facebook and Youtube, we know less about WhatsApp, TikTok, and information flows on encrypted platforms. Given the increased intransigence of Facebook, including Zuckerberg’s May 2020 statements on Fox television against Twitter’s fact-checking of Donald Trump’s falsehoods, I think focusing on Facebook (and it’s subsidiaries) turned out to be germane. Also, recent research affirms the reach of Facebook (along with television) as a major source of news making it all the more important to focus on the most powerful forms of media (Kennedy & Prat 2020).

Things changed constantly as I was writing, and I lived in fear of getting out of date. Teaching at a policy school, my goal was to inform policy makers and provide something useful. Serving on a foundation board I wanted to make sure I learned something that could be the basis of advocacy or grant making. I cancelled my Facebook account in 2018 but I rejoined in 2019 because I wanted to understand how things were changing. As I wrote, I was encouraged by the words of Michael Schudson, who said that even explaining the different solutions and providing a taxonomy would be a worthwhile endeavor.

Researching and writing this dissertation has made me feel different things at different times. In the beginning I just wanted to understand the different solutions to mis/disinformation that are being proposed. Then I got frustrated and depressed because it seemed so many of the fixes wouldn’t really work. Or, as with disclosure of the financing of election advertising, they are part of battles that have been fought for decades and never seem to be resolved.

But now, having come to the end of the dissertation, it’s clear there are many regulations and policies that need to be implemented and there are good examples, especially from France, that should be emulated widely. Nearly every fix I addressed in this dissertation has a role to play in the mosaic of solutions. Many will require government intervention and regulation. They will also require the tech giants such as Google, Twitter, and Facebook to spend heavily. It’s not acceptable for them to say the scale is too big for them to address. They earned massive profits, so they need to plough a lot of it back into solving the problems they created.

Frankly the current system of journalists and academics and civil society groups spending vast amounts of time and effort to identify abuses and flag them for the tech companies to then possibly act on, is absurd. More generally, Joan Donovan has described efforts to research, monitor and debunk misinformation as “cleaning up the mess” that the tech companies created and Princeton politics professor Andy
Guess has talked about the opportunity costs involved. Press coverage of abuses is important but the hope that this will generate enough public outrage to then impel the companies and policy makers to act is a rather round-about way of getting change.

There are many analogies being made to the production and dissemination of false information online. At panels, speakers compare it to environmental pollution, contamination, toxic waste, tainted meat and all kinds of other undesirable things. Sometimes the comparisons are a bit forced but the larger principle holds true: tackling a systemic process on a case-by-case basis is inefficient and not how governments and regulators deal with other other forms of oversight. For this reason, writing the dissertation, made me realize the importance of processes, procedures, auditing and regulations. It is simply too convenient for the tech companies to create a huge problem for society and then only spend a fraction of the time and finding needed to dxi it.

When thinking about solutions there are many in the US who want to avoid a battle over the first amendment and thus seek regulations that do not address the content of the speech. Some support monitoring or even regulation of behavior online for example—more identification of bots or banning microtargeting of information.

They may worry that any regulatory proposal to address online mis/disinformation represents an attack on free speech or the “thin end of the wedge” of censorship. In fact, the right to free speech is not the same as the right to virality (as Jack Dorsey has pointed out) and societies have always made judgements about which speech should be protected. Societies have made distinctions between commercial speech and the rights of newspapers, and have considered the rights of the listener. Now that we are in a new information era, it is time to rebalance.

These are difficult decisions that become even more difficult when they are politicized or tainted by special interests such as the tech companies and their lobbyists. Hiding behind the First Amendment in the US has become an excuse to try to shut down critics, hobble political opponents, and protect profit. But groups with special interests should not be allowed to run, or ruin, a discussion that must be had by citizens and government.

It is heartening that some democratic governments, particularly in Australia, Canada, France, Germany, Ireland, and at the EU in Brussels, are thinking deeply about balance and harms and have made a conscientious decision to find the right balance. Clearly, deep distortions in the media ecology have been aggravated by the freedom from liability given to the tech companies, and by a social-media business model that
derives profit from outrage and anger. For these extremely profitable monopolies to say they do not have the resources to fact-check or moderate content beggars belief; it suggests that without modification of Section 230 in the US, the companies will not be incentivized to act.

But modifying intermediary liability is only one part of the solution. The problem of online mis/disinformation includes a range of harms caused by a range of bad actors and will require many solutions. No single fix discussed in this dissertation will solve the problem. Some will address content, others will address the business model of the tech companies. Still others require changes in audience behavior. There is a role for everyone in this fight: government, companies, audiences. Because I wanted to finish this dissertation with ideas that could be implemented, I focused my conclusions on policy proposals.

This is a brief checklist of suggestions without a lot of detail because history, context, and arguments for and against these ideas can be found throughout the dissertation.

Addressing the Demand Side:

- **Media and news literacy efforts**—Should be scaled up, systematized funded properly and offered in every school. Assessment criteria and benchmarks of skills to be offered at every age should be implemented. This will require federal and state laws, funding, and training of teachers.
- **Fact-checking**—Helps establish a culture of truth and can force public officials to be more careful about what they say. The tech companies pay for it (even for groups that are not fact-checking claims made on social media) but this should be done through an intermediary organization so the tech companies don’t have influence in what gets checked. Further, the tech companies should have a systematized way of using the information given to them and letting people know what their systems are and how false content is treated after it has been debunked (labeled, downranked etc).
- **Community engagement efforts**—More funding is needed to help journalists connect with their audiences. Such efforts can give underserved communities voice and access to needed information. There are many good ideas for how these local initiatives can be established: as well as philanthropy, Free Press in the US has highlighted local efforts such as cooperatives and public radio and broadcasting. There are hundreds of hyper local sites around the US. Without funding, many have failed—it’s almost impossible to make much money from
hyper local outlets—but many endure thanks to volunteer efforts and revenue from other sources\textsuperscript{80}.

Addressing supply and transmission:

- **Content Moderation**—Essential, and more is needed. The tech companies need to be transparent about hiring, hire more staff, be clearer about how decisions are made, and respond quickly to queries. The Facebook oversight board will barely address the volume of disputes that arise and will arise.

- **Using algorithms to rank information**—Essential but should be transparent so that the public and government know the criteria being used. Should be provided to regulators. In general, there needs to be far more algorithmic transparency. Again, French law could be a model and there will need to be a regulator set up to audit companies and their algorithms\textsuperscript{81}.

- **Addressing financial incentives**—Limiting micro-targeting, making sure that purveyors of mis/disinformation do not make money from people clicking on their sites or buying their products. This will likely need to be regulated. Strict rules about what kinds of ads can be sold as well as removal of incentives to make money off clickbait, outrage, anger. The May 2020 revelations about Facebook’s shutting down internal efforts to make the experience less polarizing suggest that regulation will be essential. (Horwitz & Seetharaman).

- **If politically feasible, the introduction of a US version of GDPR or the extension of the California privacy law passed in 2019. However, rather than having consumer opt-out of having their information collected it should probably be made opt-in. An end to micro-targeting of political advertising would help solve the transmission/virality problem. Prominent voices in the US have called for limiting micro targeting of political advertising on Facebook (Maréchal, Mackinnon & Dheere 2020).**

- **Intermediary liability**—Modification of Section 230 would help incentivize the tech companies to take down information that can cause harm.

Election reform—Less money spent on advertising, limit to the length of time when advertisements can appear, disclosure of all funders/amounts spent, sign-off by the candidates showing they’ve approved the ad. Provision of warning labels or a ban on

\textsuperscript{80} This can include events, subscriptions, selling services, finding partners. For more information please see: https://www.cima.ned.org/wp-content/uploads/2019/03/FightingForSurvival_March_2019.pdf

\textsuperscript{81} For more information see https://www.isdglobal.org/isd-publications/algorithm-inspection-and-regulatory-access/
manifestly false claims. The goal should be to have online a system much more like that in Europe, with low advertising spends and requirements for balanced reporting. This would require a dramatic shift in the US, and since passing such laws is unlikely the tech companies should immediately start implementing best practices from France and northern Europe as well as adapting laws on disclosure and truth-in-advertising that some US states have already adopted and providing feeds of trustworthy election-related news from respected sources. Supply Side—protecting or enhancing information provision

- France has now put in place laws against cyber-harassment; these laws help protect journalists covering controversial topics. They could be replicated in other countries, following the recommendations of the OSCE and the IPI.
- Philanthropists and foreign aid agencies are continuing to fund quality journalism in countries around the world. These are not large-scale but still add greatly to the news ecologies of the countries where these outlets are present.
- In the last couple of years, Craig Aaron, Tim Karr and Victor Pickard at Free Press, and James Deane from BBC Media Action have all come up with broad, bold proposals to support public-interest media. Drew Sullivan and Ellen Hume have drawn up a proposal for a special fund to support investigative journalism. These should be funded by taxes and by a pot of funds created by the tech companies. They need to be taxed and then a portion of their taxes can be allocated to news provision. This would be done through supervisory organizations and the governance details are already being discussed.
- Given the collapse of advertising revenue during the Covid 19 pandemic, bold, large-scale initiatives for supporting media need to be implemented quickly. Public broadcasting and legacy outlets are also a key part of the media ecosystem and they need to be supported too (Kennedy & Prat). Foundations and philanthropists in particular should think big and not only support small innovations and niche outlets. In the short term, philanthropists should fund advocacy and efforts to improve media policies and government spending. Supporting the expansion of public broadcasting and radio as well as local sites should be the goal. Taxes from the tech companies should be used to pay for it so as to prevent the companies from choosing where their money is allocated.
- Australia and France in 2020 have used competition law to require Google and Facebook to pay for the news they circulate, succeeding where the European Union’s attempts to use copyright law have failed. This revenue will help support the supply of quality news and so these laws should be emulated widely (Jones 2020; Chavern 2020).
As can be seen from this list, there are many policies that can and should be implemented not just to fight online mis/disinformation but to secure a healthy media eco-system into the future. The time to begin is now.
Conclusiones

En el momento de redactar las páginas finales de esta tesis, la pandemia del Covid-19 ha sido definida también como “infodemia”, ya que la información falsa circula ampliamente en la sociedad. Para alguien que ha pasado los últimos años tratando de entender el problema de la desinformación, el elenco de personajes es familiar: Algunas de las falsedades provienen de campañas dirigidas por los rusos y los chinos (Barnes, Rosenberg y Wong 2020), otras son difundidas por Alex Jones y por aquellos que buscan obtener beneficios de la venta de curas milagrosas (Bellingcat, 2020). Políticos como Donald Trump y Jair Bolsonaro atacan a la ciencia y dicen lo que quieren en conferencias de prensa y en Twitter. Grupos de extrema derecha en los Estados Unidos -incluyendo tres hermanos que son activistas de los derechos de las armas- están sacando anuncios en Facebook llamando a la gente a mítines para promover la “apertura de los estados”. Estos extremistas están siendo ayudados por una estrecha colaboradora de la conservadora Secretaria de Educación de los EE.UU. Betsy DeVos (Stanley-Becker & Romm 2020). En resumen, la infodemia del Covid-19 es la habitual mezcla tóxica de avaricia y oportunismo político, posibilitada por gigantes de la tecnología como Facebook y YouTube, y las empresas que se benefician de los sitios cerrados encriptados donde viajan tales falsedades y discursos incendiarios. También hay que tener en cuenta, como siempre, el papel de la audiencia. Muchas personas que envían información médica incorrecta lo hacen por confusión o por el deseo de ayudar (Larson 2018; Wasserman y Morales 2019). En una situación en la que la ciencia aún no está asentada y los médicos y científicos no están seguros, es comprensible que las personas que no tienen formación científica se
confunden. Pero es peor cuando la confusión de una pandemia se ve agravada por personas cínicas (y a menudo hipócritas) que deberían saber más, como sucede cuando Fox News niega la gravedad de la epidemia al tiempo que cancela la fiesta del 89º cumpleaños de Rupert Murdoch el 11 de marzo, y llena sus propias oficinas con desinfectante para manos (Smith 2020a; Smith 2020b).

Esta tesis se basa en la idea de que los hechos existen, que los métodos científicos ayudan a las sociedades a llegar a la verdad y que las pruebas y la investigación son esenciales en todos los ámbitos de la vida. Esta disertación toma como primer principio que la desinformación en línea (y fuera de ella) es importante para la sociedad. Demasiada desconfianza y miedo hacen imposible que los gobiernos funcionen. Como hemos visto en todo el mundo, las consecuencias no son sólo la polarización política, sino también la violencia.

Encogerse de hombros ante la desinformación en línea y decir que las mentiras siempre han existido y que, por lo tanto, no son algo de lo que preocuparse (Allen y Stevens 2019) es una respuesta inadecuada e irresponsable. Claramente, no soy la única persona que piensa esto. Los reguladores que han comenzado a tomar medidas para abordar el problema también se preocupan por el efecto que el aluvión de desinformación en línea puede tener en la democracia.

Hay, por supuesto, una serie de estudios que sugieren que los efectos de la persuasión de los medios de comunicación no son tan grandes como la gente se imagina (Newton 2019; Allcott y Gentzkow 2017). Hay notables estudiosos que han pasado sus carreras investigando la persuasión y los efectos de los medios de comunicación, y analizando si la publicidad política o los artículos de noticias, los mensajes en línea o los falsos tweets hacen cambiar las opiniones. Los psicólogos sociales han pasado décadas observando el sesgo de confirmación, el razonamiento motivado, la heurística, cómo se forman las creencias y cómo pueden ser cambiadas.

Como se discutió en el Capítulo uno, hay un debate sobre la importancia que la credibilidad de la fuente o del mensaje juega en la formación o en el cambio de las creencias. Estudios más recientes sugieren que la confianza en los medios de comunicación/gobierno/instituciones se ve afectada por las creencias políticas, particularmente en los Estados Unidos. Así también, Newton describe que en el “Modelo Estándar” factores como la edad, la demografía, los ingresos, la educación también importan (Newton 2019). Las características de la audiencia juegan un papel importante. Siempre habrá algunas personas más susceptibles que otras ante la información falsa. En un estudio realizado en abril de 2020 sobre la forma en que el público accede y clasifica las noticias sobre el Covid-19 se comprobó que, si bien la
confianza en los científicos y los expertos era en general elevada en todos los países, las personas con bajos niveles de educación respondían incorrectamente a más preguntas sobre el Covid-19 (Nielsen y Fletcher y otros, 2020).

Una razón por la que es difícil medir los efectos de la desinformación en línea es que la característica definitoria de gran parte de la desinformación que circula es la experimentación. Diferentes falsedades son puestas en los medios sociales o en los sitios de conspiración de nicho, o son propugnadas por Alex Jones en Infowars, y algunas se difunden y son adoptadas ampliamente mientras que otras no lo son. Es imposible saber de antemano lo que funcionará. Considérese el anuncio de 2004 de “Swift Boat” que ataca el historial de guerra de John Kerry. Pagado en parte por un multimillonario de Texas llamado Harold Simmons, los anuncios fueron puestos en la televisión en los estados indecisos. En una elección reñida, la intención era persuadir incluso a unos pocos votantes (Solman 2004; Morain 2008). Lo mismo ocurrió en las elecciones presidenciales de los Estados Unidos en 2016. Por esta razón, es imposible decir que la desinformación no tiene ningún efecto. Hay cosas que no sabemos - si una pieza de información falsa importa o si es parte de toda una ecología. O si una mentira que se pone de moda marca la diferencia. Los experimentos artificiales no pueden probar esto. Como dijo el politólogo Alex Hertel-Fernandes, “muchos de estos estudios se centran en el efecto marginal de, digamos, un anuncio adicional y no en los efectos combinados de la publicidad de una campaña, lo cual es probable que importe más cuando se empieza a pensar en los efectos agregados. Otra advertencia importante es que estos estudios tienden a centrarse en el comportamiento de los votantes y no, por ejemplo, en la conciencia sobre los temas de las noticias, por lo que no pueden hablar realmente de los efectos que la cobertura de los medios de comunicación tiene en el establecimiento de la agenda” (correo electrónico al autor, abril de 2020).

La incertidumbre en cuanto a los efectos precisos no puede convertirse en una excusa para la inacción. Se están intentando muchas soluciones que abordarán parte de los problemas con poco o ningún riesgo de retroceso, y que serán buenas para la sociedad.

Como profesor veterana de una escuela de política, quería terminar esta investigación con una comprensión de las políticas que podrían aplicarse para ayudar a combatir la propagación de la desinformación. Y que creo que esto se ha conseguido al concluir con una lista de soluciones que deberían ser fácilmente operativas, y hay muchas más que pueden ser implementadas, también.
Cualificaciones/Limitaciones

Tratando de mantener una lista de cosas que pueden contribuir a pensar en una política de actuación sólida, naturalmente dejé de lado otras cuestiones relacionadas con temas más generales. En el curso de mi investigación, me encontré con una serie de métodos y actuaciones que se utilizan para hacer frente a la oferta de desinformación en línea y que son simplemente demasiado grandes para el alcance de esta disertación. Por ejemplo, algunos han argumentado que el problema de la desinformación procedente de Rusia es esencialmente un problema político y tiene una solución política, pero esta no es una disertación sobre geopolítica. Asimismo, la legislación sobre privacidad puede ayudar a abordar el problema de la desinformación/mala información en línea, en particular si prohíbe el microtargeting de los mensajes políticos, pero se podría escribir una investigación completa sólo sobre la RPI, por lo que no intenté abordar esa cuestión. No he discutido tampoco los debates sobre encriptación y vigilancia que son relevantes. Sin duda, los países con menos desigualdad económica y mejores políticas educativas que los EE.UU. tienen más confianza en el gobierno, las instituciones y los medios de comunicación. Pero no soy economista, así que tampoco he escrito sobre la desigualdad económica y su papel en la creación de desconfianza.

Por otra parte, ha habido para la realización de este trabajo algunas limitaciones en el uso de fuentes. Nadie de las empresas tecnológicas me habló, excepto una fuente de alto nivel que insistió en mantener conversaciones sobre el contexto de los temas solamente. Más tarde, pude hablar con una persona en Facebook que trabaja en publicidad política. Al principio de mi investigación, me puse en contacto con amigos de amigos, gente que había conocido, y con los portavoces de la compañía. Después de un tiempo dejé de tratar de enviar correos electrónicos a los contactos en Facebook y Google. Los amigos me hablaban de sus maravillosos amigos de Facebook que juraban que estarían dispuestos a hablar, pero nunca lo estuvieron. Me pasaron a colegas que prometían “encontrar a alguien” pero nunca lo hicieron.

Otra limitación es el enfoque geográfico de esta disertación. Por más que traté de mantener en mi horizonte las tendencias globales y traer las lecciones aprendidas de Europa, esta no es una investigación global, sino que se centra principalmente en los EE.UU. y Europa. Los reguladores y funcionarios gubernamentales de Francia y Bruselas fueron extraordinariamente acogedores. Pasaron horas conmigo, explicando sus leyes y puntos de vista y políticas. Me reuní con algunos de los funcionarios en Bruselas repetidamente durante los últimos años. Estas entrevistas en persona fueron esenciales, porque la cobertura de la prensa en inglés sobre las acciones de la UE y
Europa no ha sido particularmente exhaustiva y los sitios web de la UE son a menudo difíciles de analizar para un extraño. Durante el tiempo que pasaron conmigo, estos funcionarios y reguladores demostraron su compromiso con la transparencia y la responsabilidad, y su afán por ayudar a un estudiante de doctorado contrastaba completamente con la indiferencia y la evasión de las empresas tecnológicas a las que me dirigí. El trabajo académico realizado por colegas que han examinado de cerca las nuevas leyes en Europa y el Reino Unido, y que han añadido sus propias críticas, ha sido inestimable, y también me ha apoyado mucho en eso.

A pesar de mis intentos de abordar el problema de la desinformación en línea en las plataformas tecnológicas en general, al final esta tesis se centra sobre todo en Facebook, porque es la mayor plataforma y un punto de referencia para todos los que miran la circulación de la desinformación en línea. Las conversaciones sobre “las plataformas” a menudo terminaban siendo conversaciones sobre Facebook. En parte eso se debe a que muchos de los reguladores, como la ARCEP de Francia, trabajaron bajo el principio de que era importante tratar con las grandes empresas. Y, por poco que sepamos sobre Facebook y Youtube, sabemos menos sobre WhatsApp, TikTok y los flujos de información en plataformas cifradas.

Las cosas cambiaban constantemente mientras escribía, y he vivido con el temor de quedar desactualizado. Enseñando en una escuela de políticas, mi objetivo era informar a los responsables políticos y proporcionar algo útil. Como miembro del consejo de una fundación, quería asegurarme de que aprendía algo que pudiera ser la base para la defensa de proyectos o la concesión de subvenciones. Cancelé mi cuenta de Facebook en 2018, pero me reincorporé en 2019 porque quería entender cómo estaban cambiando las cosas. Mientras escribía, me animaron las palabras de Michael Schudson, quien me comentó que ya sólo exponer las diferentes soluciones al tema de la desinformación y proporcionar una taxonomía merecía el esfuerzo que estaba haciendo.

Investigar y escribir esta tesis me ha hecho sentir cosas diferentes en momentos diferentes. Al principio sólo quería entender las diferentes soluciones a la desinformación que se están proponiendo. Luego me sentí frustrada y deprimida porque parecía que muchas de las soluciones no funcionarían. O, como en el caso de la divulgación de la financiación de la publicidad electoral, los temas eran parte de las batallas que se han librado durante décadas y que nunca parecen resolverse.

Pero ahora, habiendo llegado al final del trayecto, está claro que hay muchos reglamentos y políticas que deben aplicarse y hay buenos ejemplos, especialmente de Francia, que deberían ser emulados ampliamente. Casi todos los intentos de arreglo del
problema que he abordado en esta disertación tienen un papel que desempeñar en el mosaico de soluciones. Muchas de ellas requerirán la intervención del gobierno y la reglamentación. También requerirán que los gigantes de la tecnología como Google, Twitter y Facebook gasten muchos más recursos en afrontar este problema. No es aceptable que digan que la escala es demasiado grande para ser abordada. Han obtenido enormes beneficios, y deben invertir parte de ellos en resolver los problemas que crearon.

Algunos se preocupan porque cualquier propuesta regulatoria para abordar la desinformación en línea representa un ataque a la libertad de expresión o el “filo de la cuña” de la censura. De hecho, el derecho a la libertad de expresión no es lo mismo que el derecho a la viralidad (como ha señalado Jack Dorsey) y las sociedades siempre han juzgado de una u otra forma qué discurso debe ser protegido. Las sociedades han hecho distinciones entre la expresión comercial y los derechos de los periódicos, y han considerado los derechos del oyente. Ahora que estamos en una nueva era de la información, es hora de reequilibrar.

Se trata de decisiones difíciles que se hacen aún más difíciles cuando están politizadas o manchadas por intereses especiales, como los de las empresas tecnológicas y sus grupos de presión. Esconderse detrás de la Primera Enmienda en los EE.UU. se ha convertido en una excusa para tratar de cerrar las críticas, de hacer cojear a los oponentes políticos y de proteger los beneficios. Pero no se debe permitir que los grupos de intereses especiales dirijan o arruinen una discusión que debe ser sostenida por los ciudadanos y el gobierno.

Es alentador que algunos gobiernos democráticos, en particular en Alemania, Australia, Canadá, Francia, Irlanda y la UE en Bruselas, estén pensando en serio en el equilibrio y los daños, y hayan tomado decisiones trascendentales para encontrar el equilibrio adecuado. Es evidente que las profundas distorsiones de la ecología de los medios de comunicación se han visto agravadas por la exención de responsabilidad otorgada a las empresas de tecnología y por un modelo de negocio de medios de comunicación social que obtiene beneficios de la indignación y la ira. Que estos monopolios extremadamente rentables digan que no tienen los recursos para comprobar los hechos o moderar contenidos indeseables sugiere que sin la modificación de la Sección 230 en los EE.UU., las empresas no se verán incentivadas a actuar.

Pero modificar la responsabilidad de los intermediarios es sólo una parte de la solución. El problema de la desinformación en línea incluye una serie de daños causados por una serie de malos actores y requerirá muchas soluciones. Ninguna de
las soluciones discutidas en esta tesis resolverá el problema. Algunos abordarán el contenido, otros abordarán el modelo de negocio de las empresas tecnológicas, y habrá quienes se centren en los cambios en el comportamiento de la audiencia. Hay un papel para todos en esta lucha: gobierno, compañías, audiencias. Con la idea de acabar estas conclusiones con propuestas que pudieran ser implementadas, a continuación se detallan algunas de ellas.

Esta es una breve lista de sugerencias sin muchos detalles porque la historia, el contexto y los argumentos a favor y en contra de ellas se han abordado ya a lo largo de la disertación.

**Abordar el lado de la demanda:**

- **Esfuerzos de alfabetización mediática y noticiosa:** Deberían ampliarse y financiarse adecuadamente y ofrecerse en todas las escuelas. Deberían aplicarse los criterios de evaluación y los puntos de referencia de las aptitudes que se han de ofrecer a todas las edades. Esto requerirá leyes federales, financiación y capacitación de los maestros.

- **Comprobación de hechos:** Ayuda a establecer una cultura de la verdad y puede obligar a los funcionarios públicos a ser más cuidadosos con lo que dicen. No sólo las compañías tecnológicas deberían pagar por ello (incluso a aquellos que están verificando los hechos alegados fuera de las redes sociales) sino que las compañías tecnológicas deberían tener una forma sistematizada y transparente de utilizar la información que se les da.

- **Esfuerzos de compromiso con la comunidad:** Se necesitan más fondos para ayudar a los periodistas a conectarse con su público. Eso puede dar a las comunidades menos atendidas voz y acceso a la información que sienten que necesitan.

**Abordar el suministro y la transmisión:**

- **Moderación del contenido:** Esencial, y se necesita más. Las empresas de tecnología deben ser transparentes en la contratación, contratar más personal, ser más claras en la toma de decisiones y responder rápidamente a las consultas.

- **Algoritmos para clasificar la información:** Son igualmente esenciales, pero deben ser transparentes para que el público y el gobierno conozcan los criterios que se utilizan. Deberían facilitarse a los organismos reguladores.

- **Abordar los incentivos financieros:** Limitar el microtargeting, asegurándose de que los proveedores de información errónea/desinformación no ganen dinero con las personas que hacen clic en sus sitios o compran sus productos. Es probable que esto
deba ser regulado. Son necesarias normas estrictas sobre los tipos de anuncios que pueden venderse, eliminación de los incentivos para ganar dinero con el clickbait, con la indignación y la ira.

- Responsabilidad de los intermediarios: La modificación del artículo 230 ayudaría a incentivar a las empresas tecnológicas para que retiren la información que pueda causar daños.

- Reforma electoral: Menos dinero gastado en publicidad, limitación del tiempo en que pueden aparecer los anuncios, divulgación de todos los financiadores/importes gastados, firma de los candidatos demostrando que han aprobado el anuncio. Provisión de etiquetas de advertencia o prohibición de afirmaciones manifiestamente falsas. El objetivo debería ser tener en línea un sistema mucho más parecido al de Europa, con bajos gastos de publicidad y requisitos de presentación de informes equilibrados. Esto requeriría un cambio drástico en los Estados Unidos, pero como las empresas de tecnología no están reguladas en este momento, deberían empezar inmediatamente a aplicar las mejores prácticas de Francia y el norte de Europa, así como a adaptar las leyes sobre la divulgación y la verdad en la publicidad que algunos estados de los Estados Unidos ya han adoptado.

Lado de la oferta - proteger o mejorar el suministro de información

- Francia ha puesto en marcha leyes contra el ciberacoso: estas leyes ayudan a proteger a los periodistas que cubren temas controvertidos. Podrían ser replicadas en otros países, siguiendo las recomendaciones de la OSCE y el IPI.

- Los filántropos y las agencias de ayuda extranjera siguen financiando el periodismo de calidad en países de todo el mundo. No lo hacen a gran escala, pero aún así contribuyen en gran medida a la ecología de las noticias de los países en los que estos medios están presentes.

- En los últimos años, Tim Karr de Free Press, Victor Pickard, y James Deane de BBC Media Action han presentado propuestas amplias y audaces para apoyar a los medios de interés público. Drew Sullivan y Ellen Hume han elaborado una propuesta para un fondo especial para apoyar el periodismo de investigación. Estos deberían ser financiados por impuestos y por una reserva de fondos creada por las empresas de tecnología.

- Australia y Francia en 2020 han utilizado la ley de competencia para exigir a Google y Facebook que paguen por las noticias que difunden, y han tenido éxito donde los intentos de la Unión Europea de utilizar la ley de derechos de autor han fracasado.
Estos ingresos ayudarán a apoyar el suministro de noticias de calidad, por lo que estas leyes deberían ser emuladas ampliamente (Jones 2020; Chavern 2020).

Como se puede ver en esta lista, hay muchas políticas que pueden y deben aplicarse no sólo para luchar contra la desinformación en línea, sino también para asegurar un ecosistema de medios de comunicación saludable en el futuro. El momento de ponerse en marcha es ahora.
BIBLIOGRAPHY


Abiteboul, Serge; Frédéric Potier; Côme Berbain; Jean-Baptiste Gourdin; Jacques Martinon; Gilles Schwoerer; Aude Signourel; Sacha Desmaris; Pierre Dubreuil and Benoît Loutrel. “Creating a French framework to make social media platforms more accountable: Acting in France with a European vision.” Interim mission report, submitted to the French Secretary of State for Digital Affairs, May 2019.


Anderson, Christopher; Emily Bell and Clay Shirky. ”Post-industrial journalism: adapting to the present.” *Geopolitics, History, and International Relations* 7.2 (2015).


Barnidge, Matthew. “Exposure to political disagreement in social media versus face-to-face and anonymous online settings.” *Political Communication* 34.2 (2017): 302-321.


*BBC*. “Twitter blocks French government with its own fake news law.” 3 April 2019c.


Bell, Emily. “Facebook is eating the world.” *Columbia Journalism Review*, 7 March 2016.


Berkowitz, Bonnie; Adrian Blanco; Brittany Renee Mayes; Klara Auerbach and Danielle Rindler. “More and deadlier: Mass shooting trends in America.”
Mis- and Disinformation online

*Washington Post, 5 August 2019. Retrieved 1/1/20:
https://www.washingtonpost.com/nation/2019/08/05/more-deadlier-mass-shooting-trends-america/?arc404=true*


Chen, Brian. “‘Fingerprinting’ to track us online is on the rise. Here’s what to do.” The New York Times, 3 July 2019.


Citron, Danielle. “Section 230’s Challenge to Civil Rights and Civil Liberties.” Knight First Amendment Institute, 6 April 2018.


Cooper, Kelly-Leigh. “Bianca Devins: The teenager whose murder was exploited for clicks.” BBC, 21 July 2019.


Council of Europe. “Police Failure to Respond to the Serious Threats a Journalist Received Online.” 19 December 2019. Retrieved 1/31/20: https://www.coe.int/en/web/media-freedom/detail-


Curran, James; Stuart Soroka; Sharon Coen; Toril Aalberg; Kaori Hayashi; Zira Hichy; Shanto Iyengar; Paul Jones; Gianpietro Mazzoleni; Stylianos Papathanassopoulos; June Woong Rhee, Hernando Rojas, David Rowe and Rod Tiffen. “Reconsidering 'virtuous circle' and 'media malaise' theories of the media: An 11-nation study.” *Journalism* (2014): 1-19.

Daskal, Jennifer. “Facebook’s ban on foreign political ads means the site is segregating speech.” *The Washington Post*, 16 December 2019.


Dreyfuss, Emily and Issie Lapowsky. “Facebook is changing newsfeed (again) to stop fake news.” Wired, 10 April 2019.


Engelke, Katherine. “Audience Participation in Professional Journalism; a Systematic Literature Review.” Draft paper for the 2019 ICA.


Feiner, Lauren. “Google changed its targeting policies to shine a light on political ads, but campaigns are now eyeing more opaque platforms.” CNBC News, 8 December 2019.


Ferrucci, Patrick and Jacob Nelson. “‘When money is offered, we listen’: foundation funding and nonprofit journalism.” Columbia Journalism Review, 10 January 2020.

Ferrucci, Patrick; Chris Vargo and Toby Hopp. “Civic engagement, social capital, and ideological extremity: Exploring online political engagement and political expression on Facebook.” New Media & Society, 2019.


Flitter, Emily. “Standard Chartered Fined $1.1 Billion for Violating Sanctions and Anti-Money Laundering Laws.”


Funke, Daniel and Alexios Mantzarlis. “We asked 19 factcheckers what they think of their relationship with Facebook and here is what they told us.” *Poynter*, 14 December 2018.


Funke, Daniel. “In the past year, Facebook has quadrupled its fact-checking partners.” *Poynter*, 29 April 2019.


Gerber, Jan. “‘Disingenuous' EFF had no evidence Gqubule and Harber were Stratcom agents, court finds.” *News24*, January 2020.


Gottlieb, Jed. “If the US Congress can’t understand Facebook, they definitely can’t fix it.” Quartz, 13 April 2018.


Harrison, Sara. “Twitter’s disinformation data dumps are helpful- to a point.” Wired, 7 July 2019.


Horwitz, Jeff and Seetharaman, Deepa “Facebook Shut Efforts to Become Less Polarizing” Wall Street Journal, May 27, 2020 A1

https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf


Hudson Jr., David. “Free speech or censorship? Social media litigation is a hot legal battleground.” ABA Journal, 1 April 2019. Retrieved 1/10/20:


Ingraham, Christopher. “There are more guns than people in the United States, according to a new study of global firearm ownership.” Washington Post, 19 June 2018.

Ingram, Matthew. “The Facebook Supreme Court will see you now.” Columbia Journalism Review, 19 September 2019.


Jang, S. Mo; Ruofan Xia; Jo-Yun Queenie Li; Jijun Tang; Tieming Geng; Hwalbin Kim and Chin-Tser Huang. “A computational approach for examining the


https://abc14news.com/2020/04/19/australia-forces-google-facebook-to-pay-for-news-content-report/


https://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=71674


Kafka, Peter. “Oracle will pay more than $850 million for Moat.” *Vox*, 20 April 2017.


Khan, Mehreen and Madhumita Murgia. “EU draws up sweeping rules to curb illegal online content.” Financial Times, 23 July 2019.

Khanna, Derek. “The Law that Gave Us the Modern Internet—and the Campaign to Kill it.” The Atlantic, 12 September 2013.


Knight Foundation. “Knight prototype fund awards $1 million to 20 projects to improve the flow of accurate information.” Press release, 22 June 2017.


Kozlowska, Hanna. “Facebook will have a Supreme Court–like body within a year.” *Quartz,* 15 November 2018.


Leetaru, Kalev. “We need to know more about Facebook and the fact checkers.” *Real Clear Politics*, 13 January 2019.


Levin, Sam. “‘They don't care': Facebook factchecking in disarray as journalists push to cut ties.” *The Guardian*, 13 December 2018a.


Loi n° 2018-1202 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information. Retrieved 1/10/20:
https://www.legifrance.gouv.fr/eli/loi/2018/12/22/MICX1808389L/jo/texte

Loi n°2011-412 du 14 avril 2011 - article 4. Retrieved 1/10/20:
https://www.legifrance.gouv.fr/affichCodeArticle.do?cidTexte=LEGITEXT000006070239&idArticle=LEGIARTI0000006353094&dateTexte=&categorieLie n=cid


Madrigal, Alexis. “This is how much fact-checking is worth to Facebook.” The Atlantic, 1 February 2019.

Maksl, Adam; Stephanie Craft; Seth Ashley and Dean Miller. “The Usefulness of a News Media Literacy Measure in Evaluating a News Literacy Curriculum, Journalism & Mass Communications Educator, 72.2 (2017): 228-241.


Massing, Michael “How not to cover America.” *The American Prospect*, 10 April 2018.


McCarthy, Suzanne. “Regulating political advertising in the UK— truth or consequences?” London School of Economics and Political Science, Risk and Regulation 33 (Summer 2017): 26-29.


McGee, Harry. “Politicians from across the world meet in Dublin to quiz tech giants on ‘fake news’.” The Irish Times, 6 November 2019.


Miller, Alan. “Stop the misinformation virus: Don’t be a carrier.” *Medium*, 22 April 2019a.

Miller, Alan. “NLP’s Newsroom to Classroom program brings journalist visits back to schools.” *News Literacy Project*, 2 July 2019b.


Miller, Clyde. “How to detect and analyze propaganda... An address delivered at Town hall.” New York, Town Hall, Inc., 1939.


Moyer, Melinda. “Eye contact quells online hostility.” *Scientific American*, 1 September 2012.


Mullin, Benjamin “Knight Foundation awards $1 million to projects aimed at fighting misinformation.” *Poynter*, June 22 2017.


Nalvarte, Paola. “Defense of the journalistic investigation can be more complex than the investigation itself, says journalist Gustavo Gorriti.” Knight Center for Journalism in the Americas, 1 April 2019b. Retrieved 1/31/20: https://knightcenter.utexas.edu/blog/00-20706-defense-journalistic-investigation-can-be-more-complex-investigation-itself-says-journ


Nicolson, Greg. “Court dismisses JZ defamation appeal bid as legal woes mount.”
*Daily Maverick* 7 November 2019. Retrieved 12/20/19:


https://reutersinstitute.politics.ox.ac.uk/sites/default/files/inline-files/DNR_2019_FINAL.pdf


NPR Staff. “What Legal Recourse Do Victims of Fake News Stories Have?” NPR, 7 December 2016. Retrieved 1/10/20:


Ohlheiser, Abby. “This is how Facebook’s fake-news writers make money.” The Washington Post, 18 November 2016.


O’Sullivan, Donie. “Seven weeks later, videos of New Zealand attack still circulating on Facebook and Instagram.” CNN, 2 May 2019.


Ravetti, Chiara; Mu Quan and Yana Jin. (2017) “A dragon eating its own tail: Information about air pollution, uncertainty and household adaptation in Beijing.” Unpublished manuscript.


Resnick, Ovadya; Aviv Ovadya and Garlin Gilchrist. “Iffy quotient: A platform health metric for misinformation.” White Paper published by the Center for Social Media Responsibility at University of Michigan’s School of Information, 10 October 2018.


RSF. “RSF and its partners unveil the Journalism Trust Initiative to combat Disinformation.” 3 April 2018a.

RSF. “Democratic leaders give historic commitment based on Declaration on Information and Democracy.” 11 November 2018b.


*Salov v. Ukraine* App No 65518/01 (European Court of Human Rights, 6 September 2005).


Schmidt, Thomas; Jacob Nelson and Regina Lawrence. “Engaging the Everyday Citizen: Rhetoric and Role Performance in Participatory Journalism.”


Segreti, Giulia. “Facebook CEO says group will not become a media company.” *Reuters*, 29 August 2016.


Serrano, Carlos. “To slow the spread of false stories on WhatsApp, this Colombian news site is enlisting its own readers.” *Nieman Lab*, 20 March 2017.


Silverman, Craig. “Facebook removed 2 billion fake accounts but the problem is getting worse.” *BuzzFeed News*, 24 May 2019a.

Silverman, Craig. “Old, online, and fed on lies: How an aging population will reshape the internet.” *BuzzFeed News*, 3 April 2019b.

Silverman, Craig. “Meet the people who spend their free time trying to fix Facebook.” *BuzzFeed News*, 15 August 2018.

Silverman, Craig. “Facebook says its fact checking program helps reduce the spread of a fake story by 80%.” *BuzzFeed News*, 11 October 2017.


Simons, Margaret; Rodney Tiffen; Doug Hendrie; Andrea Carson; Helen Sullivan; Denis Muller and Brian McNair. “Understanding the civic impact of journalism: A realistic evaluation perspective.” *Journalism Studies* 18.11 (2017): 1400-1414.


*The Observers*. “How we debunked rumours that Macron has an offshore account.” *France 24*, 5 May 2017.


Thomasson, Emma. “German hate speech law tested as Twitter blocks satire account.” *Reuters*, 3 January 2018.

Thompson, Nicholas and Fred Vogelstein “Inside the two years that shook Facebook - And the world.” *Wired*, 12 February, 2019.


Tibken, Shara. “Questions to Mark Zuckerberg show many senators don't get Facebook.” *CNET*, 11 April 2018.


Van der Linden, Sander; Anthony Leiserowitz; Seth Rosenthal and Edward Maibach. “Inoculating the public against misinformation about climate change.” *Global Challenges* 1, 1600008, 2017.


Westcott, Lucy. “Why newsrooms need a solution to end online harassment of reporters.” *CPJ*, 4 September 2019.


Winseck, Dwayne. “Ready, Fire, Aim: Why Digital Platforms are Not Media Companies and What to Do About Them.” Presented at the International Association of Media and Communication Researchers, Madrid, Spain, July 8-12, 2019.


Yun Chee, Foo. “Google, Facebook, Twitter rapped for not doing enough in EU fake news fight.” Reuters, 17 May 2019.


Zhang, Shan-gen; Liqun Cao; Yuning Wu and Feng Li. "Demystifying confidence in different levels of the police: evidence from Shanghai, China." Policing and Society (2018): 1-14.


Zuckerberg, Mark. “A Blueprint for Content Governance and Enforcement.” Facebook, 15 November 2018.

https://about.fb.com/news/2012/05/organ-donation-friends-saving-lives/

INTERVIEWS LIST

Angwin Julia. Personal interview. 7 February 2019.
Aviram, Alon. Personal interview. 29 May 2019.
Blagojevic, Predrag. Email interview. Summer 2019
Bourgault, Jeanne. Phone interview. 30 May 2019.
Buermeyer, Ulf. Email interview. October 2019.
Cobb, Jennifer. Personal interview. 8 July 2019.
Coppin, Paul. Phone interview. 15 January 2020.
Crovitz, Gordon. Personal interview. 16 April 2019.
Darracq, Jean-Baptiste. Personal interview. 23 October 2019.
Das, Sushi. Phone interview. 27 May 2019.
Deloire, Christophe. Personal interview. 23 November 2019.
Ellis, Joe. Personal interview. 23 January 2019.
Ely, Débora. Phone interview. 28 May 2019.
Franz, Vera. Email interview. October 2019.
Glaisyer, Tom. Phone and email interviews. April 2019.
Hendrix, Justin. Personal interview. 5 March 2019.
Hertel-Fernandes, Alex. Email interview. April 2020.
Heumann, Stefan. Phone interview. October 2019.
Huet, Benoit. Personal interview. 10 October 2019.
Lalwani, Nishant. Phone interview. 21 December 2018.
Lipkin, Michelle. Phone interview. 25 July 2019.
Mantzaris, Alexios. Personal interview, August 2019
Micek, Peter. Email comments, December 2019.
Reichman, Amos. Personal interview. 23 October 2019.
Ronderos, Maria Teresa. Phone interview. 24 May 2019.
Rouse, Marjorie. Phone interview. 30 May 2019.
Skelton, Russell. Personal interview. 31 May 2019.
Solomon, Brett. Personal interview. 4 September 2018.
We also carried out interviews with three or four regulators in Brussels and Paris, many of whom are not authorized to be quoted publicly. A complete list is available on request.
APPENDIX A. List of companies interviews

Appendix A to Chapter 5: List of companies interviewed

Brandwatch—monitors the web and sells the information to companies that care about their brand. Merged with Crimson Hexagon in 2018 and will continue to be called Brandwatch.

What is it? “We work with the language of attribution rather than defining what is true or false. We are not an encyclopedia or research institute. We can’t tell you what the GDP of Peru is, but we can give you tools to determine where a story came from and where it’s likely to go. We give people a more structured way to understand sources and make intelligent decisions. That’s how it should be. The world shouldn’t abandon a posture of neutrality towards information. I am not comfortable with saying what is true and what is not,” said Paul Siegel.

“We tell clients, ‘The underbelly of the internet is coming after your brand. If you don’t buy our software you won’t be able to deal with it.’ It’s a question of long-tail risks, which means that if you are in the normal range of conspiratorial language, it probably won’t affect you but there is always a risk that one of these stories will explode and become a national thing. It’s like throwing the dice. The chances of each roll are low, but it can happen. You never know if you will have a PR explosion.”

How funded: Privately held, VC-funded as well as private clients.

Staff size: Over 550

Launch date: 2006
Future plans: Sees a market for selling services to government, government contractors, and enterprises that are targeted by disinformation.

-------

Factmata: “I think legislation will drive the platforms to do things and is needed urgently. It’s a shame that it has come to regulation and that they haven’t taken it on properly.” -Dhruv Ghulati

What is it? Software company that provides AI tools to detect specific types of disinformation. For businesses, they monitor their brand online to eliminate undesirable information about them.

How funded? Has raised $1.8m from seed funding.

Staff size: 10

Launch date: 2017

Future plans: They are developing browser extensions and an app for individuals.

-------

Global Disinformation Index: “Right now we have the most brand-unsafe environment in the history of advertising. It’s the Wild West. Platforms have no incentive to actually secure themselves. We’re trying to catalyze grassroots support and get the advertising buyers to have a say.” Danny Rogers

What is it? A UK non-profit trying to make an “AI-powered classifier which can identify junk domains automatically” and would then work with programmatic ad networks so that they have a “dynamic blacklist of sites, thereby choking off funding for disinformation networks.”

“The goal is to have a couple of products. One is a self-updating blacklist of junky open web sites that are worth blocking in the ad exchange. This can be used by the ad-tech community to allow them to block ad buy on junky sites. No one company wants to take a stand or say ‘this is good or bad’…We want the Global Disinformation Index to be the ones to take on the risk. We have no skin in the game and can provide transparent, neutral ratings that platforms and the brand safety community can use,” Rogers said.

The second product is a manual assessment of higher-production-quality sites that can be used by the tech industry, governments, NGOs, and others to glean judgments about the journalistic integrity of the more sophisticated disinformation threat actors.
**Funding:** $50,000 grant from Knight Foundation. Now has $1 million in seed money. Other funders include USAID and Luminate.

**Staff size:** Three co-founders.

**Launch date:** 2018

**Future plans:** Is a member of the Reporters Without Borders “technical advisory committee” and is working with the European Standards Organization to develop consensus-based standards for media outlets that opt in. Eventually, this could lead to certification of outlets that meet certain standards of transparency and other criteria.

----

**NewsGuard:**

**What is it?** Ratings system that assigns red or green ratings and explanatory “nutrition label” write-ups to thousands of news sites around the world, based on nine basic criteria of journalistic practice relating to credibility and transparency. Is available as a news-literacy tool through Microsoft and through its browser extension for Chrome, Safari, Edge, and Firefox, which more than 200 public libraries in the U.S. have downloaded onto the computers used by their patrons. It launched in the US in 2018 and in the UK, Germany, France, and Italy in 2019. It also has a brand-safety product called BrandGuard that provides marketers with a whitelist of generally reliable news websites safe for programmatic advertising and a blacklist of sites that misinform.

Founded by journalists Steven Brill (founder of *American Lawyer* and CourtTV) and Gordon Crovitz (former publisher of *The Wall Street Journal* and an early investor in *Business Insider*).

**How funded:** Raised $6 million with the Publicis Groupe advertising agency as the lead investor. Charges the digital platforms to grant their users access to the ratings and nutrition labels instead of charging the actual publications being rated.

**Staff size:** Total of 50, including 40 journalists covering the US, UK, Germany, France, and Italy.

**Launch date:** Launched in the US in 2018 and in the UK, Germany, France, and Italy in 2019.

------
Vett News: “People would like a tool that when they are reading an article would allow them to understand if it’s fake or not, good or not, opinion or not, validation or confirmation” -Paul Glader

Paul Glader got interested after he wrote a very widely circulated piece on forbes.com about what to look for.

What is it? Like NewsGuard, they currently provide a Chrome extension that rates news sites based on their reliability; green for trustworthy and red for unreliable.

Funding: Boot strap, not yet ready to raise funding.

Staff size: About five.

Launch date: 2017

Future plans: Get a browser into schools and libraries and the ad-tech market and be bought by Facebook or Twitter.

-------

Vidrovr: “A lot of the video people watch [online] is not used or found from a search perspective. [Our company wants] to infuse intent from the domain in the region.” -Joe Ellis

Instead of passively receiving information, Ellis hopes that Vidrovr can help audiences as well as video producers become more active about what they see.

“You are at the behest of what the algorithm chooses to show you. Radical transparency will build trust so audiences can know who you are and what you have done. We think that will help combat misinformation and help companies monetize their video, allowing them to be transparent and provide intent into user search,” says Ellis, who took leave from his Columbia PhD program to run Vidrovr.

What is it? Software that helps companies and news outlets index, annotate, and search their videos.

Funding: Got tech start start-up money aimed at student projects and then VC money, raised $1.25 million.

Clients: Noting that they can’t disclose most of their customers, Ellis says Vidrovr "work[s] with some of the largest broadcasters in the US."

Staff size: 7
Launch date: 2016

Future plans: Switch video viewing from social sites to an OTT (personal), more mobile-based platform.

Global Intellectual Property Enforcement Center (GIPEC): Eric Feinberg was in advertising technology and became angry after he began finding unsavory content online, including ISIS posts calling for attacks on US troops. Feinberg says he took out a patent on what he calls “anticipatory intelligence.”

How it works: His software looks for words like “caliphate,” beginning with hashtags, and then trails it through the web. “My systems dig through all accounts using it.” Now Feinberg has a faux account and so the algorithms push pro-Jihadi content to him as well as to ISIS sympathizers.

Feinberg notes that he is “not going after the top, it’s the peer-to-peer, the sympathizers…You’ve got ISIS, radical jihadists getting radicalized in Indonesia, Bangladesh, Philippines….Facebook’s algorithm has connected the world for radical jihad.”

What is it? Monitors where clients’ ads are going online and generally helps its clients protect their brands online. Combines their research and reports with entities such as the Digital Citizens Alliance. (Digital Citizens Alliance is a nonprofit 501(c)(6) organization. Its IRS form doesn’t indicate its donors, simply stating that says DCA’s revenue comes from “program services”).

Funding: Bootstrap, looking for capital.

Staff size: Unknown

Launch date: 2015

Future plans: Hopes to be licensed, funded and to work with social media companies to reduce extremist content on their platforms.

-------

Metafact:

“Metafact is creating disinformation defense solutions for newsrooms, brands and organizations. By leveraging next-gen technology like advanced AI, we help newsrooms to understand if a certain discourse around a particular topic is genuine or
is a targeted campaign trail orchestrated to change public opinion or inflict financial damage by analyzing pattern and bucketing data sets. Detecting bots that spread false claims so profiling them is of paramount importance. Profiling human-run bot-like accounts is tougher, yet achievable with a claims-first approach. By being able to detect a claim as soon as it’s uploaded online, our tool is able to track the interaction of bot accounts and influencers with such content way before any other tool is able to detect it as a threat. By using our claim-first approach we can proactively detect, monitor, and defend brands from disinformation attacks before they gain momentum and inflict financial losses.” -Sagar Kaul

**What is it?** A company that builds AI-based disinformation detection and defense solutions.

**Funding:** Bootstrap, friends and family. Obtained a grant for $20,000 from Hanyang university in South Korea and completed an accelerator program in Ireland through the National Digital Research Centre (NDRC) and sponsored by government. Metafact was the first start-up selected from outside of Ireland and got 30,000 euros and gave away some equity. Has been working on IBM Watson platform since it was selected for IBM Global Entrepreneur program that provides IBM Cloud credits.

**Staff size:** 7

**Launch date:** Spring 2017

**Future plans:** Launching the tool for businesses and corporations and developing an AI-enabled media literacy app for kids: “Media literacy has to play an important part. We’re developing a mobile app that limits what kids can see on their phones. AI needs to be in the app and understand the needs of the kids and the app can give recommendations for reading material.” -Sagar Kaul (This idea was put on hold while developing the main tool.)

---

**AdVerif.ai:** “Terrorism and violence are traditionally [what companies are trying to eliminate]. Our focus is more fake news [that is] more challenging for technology to detect.” -Or Levi

**What is it?** The company uses NLP to create a blacklist for publishers and advertisers to protect their brand from being associated with fake content.

**Funding:** Bootstrapped, raising funding.

**Staff size:** Three

------

Truepic: “We want to make sure anyone in the world with a smartphone has the ability to capture an image and prove its contents are real.” - Mounir Ibrahim.

What is it? Image verification technology. Truepic has several products. Users can download the free app and whenever they take a picture the system will log the time, date, location, and pixilation and assign it an encrypted code that will be stored in the cloud. Truepic also has developed a remote inspections platform—known as Truepic Vision—for enterprise clients in insurance, banking, and lending.

Funding: Not profitable but generating revenue, raised $8m in 2018

Staff size: 30

Launch date: 2014

Future plans: Have commercial applications

------

Alto Data Analytics: “Alto was not created to research disinformation. We just found it.” - Alejandro Romero

What is it? Alto provides actionable insights based on public data. “We help our clients understand the world faster, better.” The company provides AI and machine-learning-based software that harvests, indexes, analyzes, and visualizes public data. Alto also offers services from a team of experts.

Funding: Private, no external funding / VC funding.

Staff size: More than 100 people.

Launch date: End of 2012

------

Semantic Visions: “We don’t focus on analysis of online social networks, but we focus on online news. In our experience the disinformation and propaganda start on news sites and blogs.” - Frantisek Vrabel

What is it? A large, speedy database and a web-mining system used for risk assessment and monitoring. Has roots in the defense industry and uses open-source intelligence.
Funding: Work for corporate clients (real-time risk detection solution integrated into SAP Ariba business commerce platform) pays for the work on disinformation. No government funding or funding from the platforms. Recently won a $250,000 grant from the US Department of State to help fund the development of cutting-edge new technology to combat disinformation online. The grant is going to come from the US Department of State’s Global Engagement Center.

Staff size: 25

Launch date: 2005

Acknowledgements: research for an earlier draft of chapter five was made possible with support from the Nasdaq Educational Foundation. Portions of this chapter appeared online in 2019 as a SIPA white paper and as a policy brief published by the German Marshall Fund in Washington DC.
APPENDIX B.

Information about outlets interviewed for *Bridging the Gap* (as of 2017):

**Bristol Cable:** A British Citizens’ media co-operative (with 1,850 members and counting) founded in 2014 that publishes one story per day. Funded by grants, co-op member contributions, crowdfunding and ads, and they employ 5 full time reporters and 4 part time.

**GroundUp:** A South African nonprofit independent news agency, launched in 2012 that outputs about five stories daily, is funded by foundation grants, and has seven full-time reporters and 15-20 freelancers.

**263chat:** A Zimbabwean media producer housed mostly on social media, launched in 2012 that creates short videos and social media posts. They employ 11 people full time and receive funding from ads from local businesses, video production, and online campaigns.

**Južne Vesti:** A news platform focused on southern Serbia, founded in 2009 that publish about 10-20 daily stories online. They receive funding from foundations, government grants, and advertising and employ 20 people, 8 of them being full time reporters.

**Krautreporter:** A German independent news co-op founded in 2014 that publishes one long story per day, employs 14 people, and receives funding by crowdfunding ($1.2M in starter capital) and member subscriptions.
Correctiv: A German nonprofit investigative newsroom founded in 2014 with a staff of about 50 (including data journalists, software developers, freelancers) that publishes detailed investigative stories and receives funding from grants, reader donations, book publishing, and crowdfunding.

Premium Times: A Nigerian news website with a nonprofit investigative arm, founded in 2011 that has a staff of about 50 and receives funding from foundations, advertising, marketing intelligence, and data sales.

KRIK: A Serbian on-profit investigative network founded in 2015 focused on corruption investigations and compiling a database of Serbian politicians’ assets. They have 11 staff members, two of whom are volunteers and their funding comes from 90% donors and 10% from crowdfunding.
APPENDIX C. Tech companies’ political advertising policies

Facebook

Updates to Ads About Social Issues, Elections or Politics in the US

August 28, 2019

By Katie Harbath, Public Policy Director, Global Elections, and Sarah Schiff, Product Manager

Update on October 16, 2019 at 9:00 AM PT:

Based on feedback from advertisers, we are working to make the new US “Paid for by” ad disclaimer creation process more efficient. To help ensure all advertisers have time to go through the authorization process, we are extending our deadline to become compliant to November 7, 2019.

Originally published on August 28, 2019 at 3:00 AM PT:

People should know who is trying to influence their vote and advertisers shouldn’t be able to cover up who is paying for ads. That’s why over the past few years, we’ve made important changes to help ensure more transparency and authenticity in ads about social issues, elections or politics.

Today, we’re sharing additional steps we’re taking to protect elections and prepare for the US 2020 election. Those steps include strengthening the authorization process for US advertisers, showing people more information about each advertiser and updating our list of social issues in the US to better reflect the public discourse on and off Facebook.
New Disclaimer Requirements

In 2018, we started requiring advertisers to get authorized before running ads about social issues, elections or politics. We also save those ads in an Ad Library so they’re publicly available for seven years.

The authorization process already requires advertisers in the US to provide identification to confirm who they are and where they are located. Advertisers must also place a “Paid for by” disclaimer on their ads to communicate who is responsible for them. Despite these requirements, there are a number of cases where advertisers have attempted to put misleading “Paid for by” disclaimers on their ads. That’s why, starting mid-September, advertisers will need to provide more information about their organization before we review and approve their disclaimer. If they do not provide this information by mid-October, we will pause their ads. While the authorization process won’t be perfect, it will help us confirm the legitimacy of an organization and provide people with more details about who’s behind the ads they are seeing.

Advertisers will have five options for providing more information, three of which demonstrate they are registered with the US government. If they choose one of the three government resource options, they will be allowed to use their registered organization name in disclaimers and the “i” icon that appears in the upper right-hand corner of their ads will read “Confirmed Organization.”

In addition to providing their US street address, phone number, business email and a business website matching the email, they must provide one of the following:

Tax-registered organization identification number (i.e. EIN)

A government website domain that matches an email ending in .gov or .mil

Federal Election Commission (FEC) identification number

We also want to ensure advertisers who may not have those credentials, such as smaller businesses or local politicians, are able to run ads about social issues, elections or politics. Advertisers can also choose one of the following two options:

Submit an organization name by providing a verifiable phone number, business email, mail-deliverable address and a business website with a domain that matches the email.

Provide no organizational information and rely solely on the Page Admin’s legal name on their personal identification document. For this option, the advertiser will not be able to use a registered organization name in disclaimers.
For advertisers that choose one of these two options, the “i” icon will read “About this ad” instead of “Confirmed Organization.”

The “i” icons help people on Facebook and Instagram better understand who’s trying to influence them and why. Now, with one tap, people will not only see information about the ad, but they’ll be able to see the information Facebook confirmed, such as whether an advertiser used an EIN or FEC identification number. This will allow people to confidently gauge the legitimacy of an organization and quickly raise questions or concerns if they find anything out of the ordinary.

**Updating Social Issues in the US**

We are also refreshing the list of social issues in the US to a list of 10 categories, rather than 20 distinct subject areas. As we noted when we first announced this policy, the list is meant to be fluid to reflect the public discourse around social issues on and off Facebook that seek to influence public opinion through advocacy, debate or discussion. We also left it intentionally broad as we worked to refine it over time. Today’s update will bring the US list in line with issue lists in countries who have recently held elections. The shift from 20 subjects to 10 categories does not mean that our authorization process will be less restrictive. We’ll continue to capture a range of topics encompassed by the 10 referenced categories. For instance, in the Civil and Social Rights category, we will continue to proactively detect and review ads on topics like freedom of religion, LGBTQ rights and women’s rights.

You can see more examples in our Help Center.

We have also improved our enforcement based on feedback. For example, in the case of ads that discuss, debate or advocate for environmental issues, ads that merely encourage people to recycle or highlight sustainable products won’t require these additional steps in order to run. If an ad goes further, however, and advocates for or against things like legislation or a ballot initiative, the authorization requirement will continue to apply. As noted, the categories are evolving, so even while we narrow the policy in some areas, we may expand it in others.

These efforts to refine the policy will also apply outside of the US in places that have applicable categories.

**Looking Forward**

Over the coming months, we’ll share more information on our efforts to make elections safer and provide greater transparency on the ads and content people see on Facebook. These updates will include:
More enhancements to our Ad Library, such as making it easier to track and compare spending of US presidential candidates

Expanding our policy to prohibit ads that expressly discourage people in the US from voting, a commitment we made as part of a recent civil rights audit (found on page 20 here)

Requiring all Pages for national candidates or elected officials to go through Page Publishing Authorization, which requires that Page administrators turn on two-factor authentication and verify their primary country location so that we can confirm these Pages are using real accounts and are located in the US

Exposing more information about a Page, such as the business or organization behind it

We know we can’t tackle these challenges alone. That’s why we’re calling for sensible regulation and working directly with governments, watchdogs and regulators.

While our efforts to protect elections are ongoing and won’t be perfect, they will make it harder for advertisers to obscure who is behind ads and will provide greater transparency for people. We’ll continue to share updates as we take steps to protect people ahead of the 2020 US election and beyond.

**Twitter**

From Jack Dorsey’s Twitter account, Oct 30, 2019:

We’ve made the decision to stop all political advertising on Twitter globally. We believe political message reach should be earned, not bought. Why? A few reasons…

A political message earns reach when people decide to follow an account or retweet. Paying for reach removes that decision, forcing highly optimized and targeted political messages on people. We believe this decision should not be compromised by money.

While internet advertising is incredibly powerful and very effective for commercial advertisers, that power brings significant risks to politics, where it can be used to influence votes to affect the lives of millions.

Internet political ads present entirely new challenges to civic discourse: machine learning-based optimization of messaging and micro-targeting, unchecked misleading information, and deep fakes. All at increasing velocity, sophistication, and overwhelming scale.
These challenges will affect ALL internet communication, not just political ads. Best to focus our efforts on the root problems, without the additional burden and complexity taking money brings. Trying to fix both means fixing neither well, and harms our credibility.

For instance, it’s not credible for us to say: “We’re working hard to stop people from gaming our systems to spread misleading info, but if someone pays us to target and force people to see their political ad…well…they can say whatever they want!”

We considered stopping only candidate ads, but issue ads present a way to circumvent. Additionally, it isn’t fair for everyone but candidates to buy ads for issues they want to push. So we’re stopping these too.

We’re well aware we’re a small part of a much larger political advertising ecosystem. Some might argue our actions today could favor incumbents. But we have witnessed many social movements reach massive scale without any political advertising. I trust this will only grow.

In addition, we need more forward-looking political ad regulation (very difficult to do). Ad transparency requirements are progress, but not enough. The internet provides entirely new capabilities, and regulators need to think past the present day to ensure a level playing field.

We’ll share the final policy by 11/15, including a few exceptions (ads in support of voter registration will still be allowed, for instance). We’ll start enforcing our new policy on 11/22 to provide current advertisers a notice period before this change goes into effect.

A final note. This isn’t about free expression. This is about paying for reach. And paying to increase the reach of political speech has significant ramifications that today’s democratic infrastructure may not be prepared to handle. It’s worth stepping back in order to address.

**Google**

An update on our political ads policy

Scott Spencer

VP, Product Management, Google Ads

Published Nov 20, 2019
We’re proud that people around the world use Google to find relevant information about elections and that candidates use Google and search ads to raise small-dollar donations that help fund their campaigns. We’re also committed to a wide range of efforts to help protect campaigns, surface authoritative election news, and protect elections from foreign interference.

But given recent concerns and debates about political advertising, and the importance of shared trust in the democratic process, we want to improve voters' confidence in the political ads they may see on our ad platforms. So we’re making a few changes to how we handle political ads on our platforms globally. Regardless of the cost or impact to spending on our platforms, we believe these changes will help promote confidence in digital political advertising and trust in electoral processes worldwide.

**Our ads platforms today**

Google’s ad platforms are distinctive in a number of important ways:

The main formats we offer political advertisers are search ads (which appear on Google in response to a search for a particular topic or candidate), YouTube ads (which appear on YouTube videos and generate revenue for those creators), and display ads (which appear on websites and generate revenue for our publishing partners).

We provide a publicly accessible, searchable, and downloadable transparency report of election ad content and spending on our platforms, going beyond what’s offered by most other advertising media.

We’ve never allowed granular microtargeting of political ads on our platforms. In many countries, the targeting of political advertising is regulated and we comply with those laws. In the U.S., we have offered basic political targeting capabilities to verified advertisers, such as serving ads based on public voter records and general political affiliations (left-leaning, right-leaning, and independent).

**Taking a new approach to targeting election ads**

While we’ve never offered granular microtargeting of election ads, we believe there’s more we can do to further promote increased visibility of election ads. That’s why we’re limiting election ads audience targeting to the following general categories: age, gender, and general location (postal code level). Political advertisers can, of course, continue to do contextual targeting, such as serving ads to people reading or watching a story about, say, the economy. This will align our approach to election ads with long-established practices in media such as TV, radio, and print, and result in election
ads being more widely seen and available for public discussion. (Of course, some media, like direct mail, continues to be targeted more granularly.) It will take some time to implement these changes, and we will begin enforcing the new approach in the U.K. within a week (ahead of the General Election), in the EU by the end of the year, and in the rest of the world starting on January 6, 2020.

Clarifying our ads policies

Whether you’re running for office or selling office furniture, we apply the same ads policies to everyone; there are no carve-outs. It’s against our policies for any advertiser to make a false claim—whether it's a claim about the price of a chair or a claim that you can vote by text message, that election day is postponed, or that a candidate has died. To make this more explicit, we’re clarifying our ads policies and adding examples to show how our policies prohibit things like “deep fakes” (doctored and manipulated media), misleading claims about the census process, and ads or destinations making demonstrably false claims that could significantly undermine participation or trust in an electoral or democratic process. Of course, we recognize that robust political dialogue is an important part of democracy, and no one can sensibly adjudicate every political claim, counterclaim, and insinuation. So we expect that the number of political ads on which we take action will be very limited—but we will continue to do so for clear violations.

Providing increased transparency

We want the ads we serve to be transparent and widely available so that many voices can debate issues openly. We already offer election advertising transparency in India, in the EU, and for federal U.S. election ads. We provide both in-ad disclosures and a transparency report that shows the actual content of the ads themselves, who paid for them, how much they spent, how many people saw them, and how they were targeted. Starting on December 3, 2019, we’re expanding the coverage of our election advertising transparency to include U.S. state-level candidates and officeholders, ballot measures, and ads that mention federal or state political parties, so that all of those ads will now be searchable and viewable as well.

We’re also looking at ways to bring additional transparency to the ads we serve and we’ll have additional details to share in the coming months. We look forward to continuing our work in this important area.