Entanglements and Exploits: Sociotechnical Security as an Analytic Framework

Matt Goerzen, *Data & Society Research Institute* Elizabeth Anne Watkins, *Columbia University*
Gabrielle Lim, *Open Technology Fund Research Fellow*

Abstract

The rise of social media platforms has produced novel security threats and vulnerabilities. Malicious actors can now exploit entanglements of once disparate technical and social systems to target exposed communities. These exploits pose a challenge to legacy security frameworks drawn from technical and state-based conceptions of referent objects and bounded information systems. In this paper we propose a new framework of analysis to meet this challenge, Sociotechnical Security (STsec), which acknowledges how the interplay between actors produces emergent threats to participant communities. This exploratory paper offers an overview of sociotechnical systems, explains why these threats and vulnerabilities require us to expand our understanding of security with regards to participatory technology, and how sociotechnical security can be operationalized as a framework for analysis.

1. Introduction

The early 2000s through the Arab Spring in 2011 saw a short period of jubilant expectation around social media platforms and the internet writ large. The current sentiment around the world, however, is one of growing discontent and mistrust (Romm et al., 2019). Technology commentators, policymakers, civil society, and governments from almost every continent and region have expressed concern about the proliferation of unwanted and potentially dangerous content on the internet. Harms that have been linked to social media platforms and other internet-based participatory technologies range from advertising fraud to targeted harassment to terrorist recruitment.

Following the 2019 terrorist attack on a mosque in Christchurch, New Zealand, for example, 8Chan founder Fredrick Brennan expressed regret for creating the platform, where the shooter posted a link to a Facebook livestream of the attack which killed 51 individuals (McMillan, 2019). Speaking to the Wall Street Journal, he said, "It wouldn't surprise me if this happens again." Less than two months later another man who had also used 8chan to post his "manifesto" opened fire in a California synagogue, killing one woman and injuring three others (Owen, 2019).

Unfortunately, platform-linked violence like the incidents in Christchurch and San Diego are not uncommon and are unlikely to subside in the near future. Other forms of problematic and potentially dangerous content also proliferate. False and misleading content distributed on Whatsapp in India led to the murders of over two dozen people in 2018 (Goel et al., 2018). The genocide of the Rohyinga in Myanmar was partially incited by the military through disinformation and false accounts on Facebook (Mozur, 2018). Facebook usage may also correlate with higher rates of anti-refugee attacks in Germany (Müller and Schwarz 2018). Meanwhile, targeted harassment has been linked to self-censorship and a chilling effect on internet-based communications (UN Office of the High Commissioner, 2017). Amnesty International in 2018, for example, declared Twitter trolling of women a human rights abuse (2018), while in India, online harassment is so pervasive a study from the Software Freedom Law Centre found it had effects similar to censorship (2016). More recently, at the G7 meeting of digital ministers in Paris, states and actors from private industry committed to the "Christchurch Call," which aims to eliminate terrorist and violent extremist content online (Christchurch Call, 2019).

In addition, online disinformation is now regularly recognized as a dimension of state warfare in the guise of "hybrid warfare," "information warfare," and "influence operations." Following public attention to disinformation campaigns by foreign agents in the United States, it has been increasingly discussed as a "threat to democracy" (Kavanagh and Rich, 2018; Prier, 2017; Snegovaya, 2015). Although its impacts are debatable and dependent on context, disinformation nonetheless has the potential to exacerbate existing divisions in society and influence public discourse.

As such, there is a growing consensus that the use and prevalence of social media platforms present risks to the well-being and safety of individuals and their communities, warranting some level of intervention. Public discourse regarding these issues has increasingly relied on the language of security. Words like "national security risk" (Morris, 2019), "threat to democracy" (Deb et al., 2017), and "weaponize" are frequently used by academics, politicians, and civil society alike to describe the various

ways threat actors use these platforms (Bosetta, 2018; Howard, 2018; Nadler et al., 2018). In their article on social media, "War Goes Viral," Emerson T. Brooking and P. W. Singer warn that "information could be fashioned into a dangerous weapon" (2016). The security risks in question are often discussed at primarily abstract (the nation state; democracy) or technical levels (the functioning of software and communications infrastructure; the integrity of databases), drawing on established security frameworks such as national security and information security.

But if "security" is to be applied to participatory technologies, and in particular social media platforms, we run into some new challenges. These are complex systems that touch on a range of interconnected social dynamics, profit incentives, psychological variables, technology designs, data storage systems, automated and machine learning algorithms—and all the things that emerge from the messy interactions of those different things. This isn't as simple as just finding bugs in code. Current technical security frameworks, we argue, do not fully address the challenges of participatory information systems, as they tend to relegate human elements as passive recipients of engineering decisions, or individual error as human vulnerability. Meanwhile, national security frameworks often fail to consider the individual as an object of protection, focusing primarily on the security of the state.

We therefore propose a new concept, which we term *sociotechnical security* (STsec). Drawing from sociotechnical studies and security studies, this framework places a group of individuals within a sociotechnical system as the primary object in need of protection. The distinction of this approach is the working assumption that because the social and technical layers are interconnected, so too are the vulnerabilities, exploits, and threats that emerge from such systems. By acknowledging this complex interplay, STsec provides a framework of analysis for identifying how social and technical conditions, which may not be inherently weaknesses or flaws on their own, can be exploited for harm when combined. Finally, an STsec framework provides a way to reconcile the needs of local communities with the global scale of today's platforms.

This paper is divided into the following sections: Section 2 offers an overview of sociotechnical systems and what defines a sociotechnical exploit or threat. Section 3 explains why these threats and their related vulnerabilities require us to expand our understanding of security with regards to participatory technology. Section 4 demonstrates how sociotechnical security as an analytic framework can be operationalized with reference to three examples: targeted harassment, foreign disinformation, and extremist content. Section 5 considers STsec's applications and limitations. Section 6 concludes with pathways for future research.

As this is an exploratory paper, we hope it will inspire others to build upon this framework and apply it to sociotechnical systems beyond social media platforms.

2. What Is Sociotechnical Security?

2.1 Defining Sociotechnical

Taking a sociotechnical lens means viewing both social systems and technological systems as interdependent and co-constituted. While users can be provoked into shifts in relation or organization by changes in material technologies, ultimately what technology means and what it's used for is not prescribed but rather emergent from the interplay between tools and behaviors of users (Orlikowski, 2007; Leonardi, 2012). The term "sociotechnical" can be traced back to the postwar British coal mining industry, where the UK-based Tavistock Institute funded studies of local coal mining concerns, in particular their micro-level work practices, meso-level organizational arrangements, and macro-level social systems (Trist, 1981). This approach bears a resemblance to French sociologist Bruno Latour's Actor-Network Theory, where humans and non-humans alike are considered to be actors, with non-human actors being the objects through which humans think, work, and collaborate, like devices, interfaces, platforms, and instruments (Hutchins 1995; Latour 1998, 1991). Sociotechnical analyses have also been informed by theorists like Donna Haraway, who emphasize that effects in the world are enacted by the complex relations of a dizzying array of entities, and that knowledge is contingent and always locally situated (Haraway, 1991; 1988).

Classical notions of computer security contend with a fairly limited set of attacks an actor could potentially wreak against a computer, boiled down to stealing its data, misusing its credentials, or hijacking its resources (Singer and Friedman, 2014). An altogether different class of malicious action, however, has arisen in recent years, due largely to the increasingly participatory nature of information and communication technologies. In this paper, we focus on the case of social media platforms.

2.2 Sociotechnical Systems & Social Media

Social media platforms like Facebook, Twitter, Twitch, and YouTube, when viewed as sociotechnical systems, present novel challenges to anyone trying to protect entangled communities, technologies, and interests. As humans and non-humans are brought together by these new sociotechnical systems, they relate to each other in new arrangements. Social media brings actors together into novel encounters and proximities: communities who at one time may have been invisible to those that would, if given the

chance, do them harm, are suddenly entangled. They become connected and intertwined by and with powerful messaging, targeting, and automation technologies. Groups are no longer governed by social norms alone, but also by technological norms (Niederer and van Dijck, 2010), bringing about changes in how these groups must think about exposure, protection, and security. These new conceptualizations may be thought of as an extension of the shared imaginaries already at work when groups think of themselves as collectives or communities (Anderson, 2006).

2.3 Vulnerabilities, Exploits, & Threats

All systems possess the capacity to be used, or "exploited," in ways their designers did not foresee. Latour called these unanticipated uses "anti-programs," where users rebel against an object's programmed use (Latour, 1991). In security discourses, these capacities are considered "vulnerabilities." Vulnerabilities that are exploitable in ways that can harm targeted communities therefore constitute "threats." In sociotechnical security, we recognize that vulnerabilities are either primarily technical or social in nature, and when combined by threat actors, constitute uniquely sociotechnical exploits (ST exploits). The question of what constitutes a "threat" is typically subjective, and often contingent on local practices of thinking through potential threats to a system, à la "threat modeling." (This framework resonates with the notion from science and technology studies that knowledge is situated and locally contingent, what's known as "standpoint epistemology.") The harassment of women on digital platforms, for example, has recently become so elevated a threat that it's been listed by the United Nations as a human-rights violation (UN Office of the High Commissioner, 2017). Likewise, false information, which has been rampant since the internet's earliest years, is now considered a viable threat to democracy, and a harm that must be addressed.

In sociotechnical systems, ST exploits are not related to a software bug or the failure of a particular technical component, but lie instead at the intersection of technical and social conditions; a tool that works as intended in a narrow use scenario (say, among a community of users whose goals are aligned with those of the designers) may demonstrate substantial capacity for abuse in another context (say, among an expanded community of users, some of whom may be antagonistic to the designers' use prescriptions).

While we use the term "vulnerability" to describe how latent affordances or conditions in sociotechnical systems can be exploited, this does not imply that they are necessarily flaws or weaknesses that need to be addressed in the same way a software vulnerability might necessitate patching. In many cases, that which manifests as a vulnerability within a sociotechnical system may also serve as a safety mechanism for other users. Anonymous accounts, encrypted messaging, or DDoS protection, for example, can enable the dissemination of problematic information which may result in harm for some communities—even as they protect other users from surveillance, harassment, and censorship. It may well be the case that the vulnerability being subjected to ST exploits offers far greater benefits to the general user base than the harms associated with it. Because of the fluid and sometimes normative way in which threats are conceptualized, we caution against an overly inclusive range of phenomena into this category. This is especially important when impact is not clearly understood. A miscategorization of a threat may result in heavy-handed solutions and regulatory or legislative overreach that may do more harm than the threat being mitigated.

Looking for vulnerabilities across social platforms means not just looking for bugs in code, but also considering the wider social context. Sociotechnical vulnerabilities can range from the macro, such as the participatory nature of contemporary ICTs enabling access to an audience of billions, to the micro, such as the capacity to use a particular feature of a platform's advertising tools to target a particular individual for attack. Well beyond the techno-utopianism which once accompanied the "democratization" of communications (Benkler, 2006), many users today utilize these tools to target vulnerable persons and populations. Harassment efforts can range from individualized targeting to coordinated campaigns. While these issues are not normatively considered a "security" threat by legacy technical computer security frameworks, we argue that ensuring the free and open communications of vulnerable people ought to be considered a key security concern, and that a sociotechnical security lens offers tools to grant entangled actors greater agency.

For example, we can consider the business model of platforms like Facebook and YouTube a multifaceted vulnerability. Digital information distribution models demand that news, for example, be both a market good and a civic good (Ananny and Bighash, 2016): citizens of our representative democracy learn about their world and create mental models of their participation in this world through the news they're exposed to via online platforms (Schudson, 1999), yet this news is distributed through a business model contingent on economies of scale competing in an attention economy (Wu, 2017), serving the needs of advertisers. Any website contingent on advertising cannot be divorced from the need to expose individuals to information designed to change their behavior, and as such, relies on tools built to serve that need. The threat comes about when a malicious actor recognizes an opportunity for an exploit. Disinformation campaigns, for example those waged by the Russian Internet Research Agency in the United States, Bangladesh, and Venezuela (Frenkel et al., 2019), exploit

microtargeting tools (Nadler et al., 2018) to expose specific individuals to disinformation engineered to sow division or promote extremist thinking.

Corporate performance metrics can also be considered vulnerabilities. Stock-based compensation, the expectations of shareholders, and associated key performance indicators (as platforms like Twitter, Facebook, Twitch, and YouTube are all publicly traded or owned by publicly traded firms) create perverse incentives for platform employees (Doctorow, 2018; Stamos, 2019), skewing decisions about how to protect targeted communities. Twitter, for instance, subsisted solely on investment funds for the first twelves years of its operation (Wagner, 2018). These investors' demands for profitability can lead to business practices and design decisions that harbor sociotechnical vulnerabilities, creating novel threat surfaces." The driving need to derive shareholder value from the engagement of users has led to the creation of a system that narrowly treats users as consumers and their interaction as valuable data in service of market transactions (Nadler et al., 2018). In this context, harassment campaigns on a profit-and-loss statement look like "increased engagement," even if that "engagement" is made up of users swarming to the site to attack, and others clamoring to the defense of those embattled. In 2015 Leslie Miley (then an engineer at Twitter) reported to his managers that he'd found a "huge number" of Russia-based spam accounts and bots. His recommendation that these accounts be deleted, however, was reportedly refused by Twitter's growth team, who according to Miley were "more concerned with growth numbers than fake and compromised accounts" (Ingram, 2018).

These designs—of platforms themselves as well as skewed employee incentives—are not strictly necessary for platforms to operate sustainably, but rather, are an outcome of deliberate decisions which could have been made differently. These decisions produce latent affordances which can be exploited by malicious actors, and operationalized to become new types of attack or grant existing types of attack novel scale and reach.

In another example of tools built to serve advertisers being exploited by bad actors, one reactionary Twitter user (who will not be named here) exploited Twitter's "Promoted Tweets" service to reach communities who had blocked his harassing Tweets. The user himself explained that "You can … choose to display ads to followers of specific users, like @Jezebel or @feministing," communities which he targeted for harassment. Although more than one of his targets had previously blocked him, Twitter's promoted tweets feature were able to bypass the site's protections and privacy filters. Writing about the incident for Slate, Jacob Brogan noted, "Given that Twitter gives its customers considerable control over who sees their ads, this loophole could make it easier than ever to employ the site as a platform for abuse" (Brogan, 2015).

There is no lack of evidence of the damage wrought against vulnerable communities by these attacks. Another widely referenced case is #GamerGate, in which journalist and video-game developer Brianna Wu was subjected to months of "brigading"-style harassment and both online and offline threats of violence. Speaking to the *Guardian*, she said "If I'm saying this less on Twitter it's because I feel fanning the flames will endanger my life even more… Every woman I know in the industry is scared" (Stuart, 2014). In a joint statement, the UN experts on Freedom of Expression and Violence Against Women warned that online gender-based harassment can lead to a chilling effect, silencing the voices of those attacked. Special Rapporteur on violence against women, Dubravka Simonovic, emphasized that "Ensuring an internet free from gender-based violence enhances freedom of expression as it allows women to fully participate in all areas of life and is integral to women's empowerment" (UN Office of the High Commissioner, 2017).

The security challenges presented by new arrangements of people and tools aren't limited to social and participatory media. STsec as a framework of analysis has recently been applied to other emergent entanglements of social and technical systems; the swift integration of algorithmic decision-making across systems as divergent as stock trading, garbage collection, and home assistants generates novel vulnerabilities, likewise urgently necessitating a sociotechnical approach to considering and maintaining safety and security (Elish and Watkins, 2019).

3. The Need For An Updated Security Framework

In the context of this paper, we understand security in the broadest sense: as freedom or protection from harm. We understand the need for security as a given—that is to say, that all individuals and communities deserve safety, insofar as their safety does not unduly jeopardize the security of others. However, as we have shown in the previous section, the myriad sociotechnical vulnerabilities and threats that may be used to target individuals and communities have yet to be properly addressed. The following sections provide an overview of current dominant forms of security and why they are ill-equipped to assess potential harms associated with participatory technology.

3.1 The State of Security

Security is both a concept and a practice that can imply very different commitments and processes, depending on how it is formalized. Largely, this relates to how different security regimes posit different "objects" of security. In *Security: A New Framework for Analysis (1998)*, Buzan, Wæver, and de

Wilde develop securitization theory, which describes the process in which a state (the securitizing actor) transforms an issue or event into a security threat, thus allowing for extraordinary measures in order to address the thing that needs to be protected (known as the referent object). While securitization theory was originally applied to nation states, the concept of a designated referent object requiring protection from an identified threat or threats (and the acceptance of such a threat by a defined audience) can be applied to other security frameworks.

For instance, technicalized security regimes such as computer security, information security, and network security emphasize the integrity of property as the referent object. Computer security is chiefly concerned with protecting computers from harm; information security, with protecting information from falling into the hands of an adversarial party; network security with maintaining the prescribed operation of a digital network; and so on. While the security of a human user or participant is typically implied in each of these regimes, the prioritization of non-human elements can often cast the figure of the human as a threat. For instance, the "human factor" is often identified as the weakest component of computer, info, and network security (Sasse et al., 2001)—a factor exploitable by "social engineering" attacks even when a technical system is otherwise impenetrable.[1]

Other prominent security regimes take groups as their object of protection, with "national" and "domestic" security looming large in this category. Here, specialized actors "securitize" (Hansen and Nissenbaum, 2009) different domains against purported threat to the members of the nation or group in question. Typically, this mode of security is extremely top-down, and while it purports to represent the interest of group members, it often operates with a broad brush that can ignore, or even exacerbate, the threats to particular, typically marginalized, members of the political body in question. Even when this is not a major concern, theorists have shown how concepts like "national security" can serve to undermine trust between peers and communities, replacing them instead with trust of top-down authoritative powers—in this way, introducing harms all their own (Schneier and Farrell, 2018; Winner, 2004; Molotch, 2014).

[1] Social engineering, for example, is not about engineering better societies or more altruistic organizations—but instead engineering desired behaviors in social actors by exploiting psychological and social vulnerabilities. Its corrective is usually focused on increasing human compliance with a technocratic "operational security" protocol. See: Hadnagy, 2010 and Thomas, 2002 for more on social engineering in a computer security context.

There are, however, calls towards more human-centric concepts of security. Most notably, the "human security" advocated by certain human rights groups. Here, humans are explicitly the chief object of security—and their freedom from harm is nearly unlimited by a particular domain of threat. The concept has been welcomed by many, but faces issues related to its scope and operationality (Homolar, 2015). "Public security" measures are often concrete by comparison, visibly manifesting as law enforcement, justice systems, emergency response, and other locally embodied interventions. Yet as historical tensions between vulnerable populations and agents of public security attests, these institutions are often unresponsive or even anathema to specific communities' needs and interests.

Each of these conceptualizations has its strengths and weaknesses. Perhaps most importantly, each comes embedded with assumptions about what has value, and what by its absence (or explicit interpretation as a threat) may not.

Moreover, the boundaries are not fixed.

Computer security is increasingly implied by national security, for instance, as critical infrastructure is connected to networks and run by software. Human security too, becomes affected, as migrants are entered into networked biometric databases at borders and other check-points (Latonero and Kift, 2018; Latonero et al., 2019). Indeed as software eats the world, the security of software and the systems that run it seem poised to gobble up other notions of security; conversely, national security happily expands to include emerging domains under its remit.

To be sure, there are shifts towards a more holistic approach. Since 2011, for example, a group of information scientists hailing from Italy and the Netherlands has used the term "sociotechnical security" to argue that system security protocols be grounded in social and organizational needs, rather than limited to technical requirements to protect the computer system itself (Dalpiaz et al., 2011, 2016; Paja et al., 2012). The distinction between that application and our own lies in the boundaries of an organizational network of actors, and how the process of drawing of those boundaries (i.e. deciding who is recognized as an actor and who is not) denotes membership and therefore legibility.

Even in this socially grounded approach, the context of decision-making is closed. Accountability is accorded exclusively to engineers, whose job it is to model users' potential interactions with the system. While this is a step in the right direction, our STsec approach explicitly redraws the boundaries of an organization to include users of a system, so that "users" aren't merely passive recipients of the engineering imaginary but rather granted agency in "securing" their communities. In our case, the system in

question is expanded, more closely resembling what some have referred to as the "public sphere" (Habermas, 1991) and related notions, such as "counterpublics" (Warner, 2002). It is expressly because sociotechnical arrangements like social media platforms themselves redraw how systems of people and tools are arranged, that an updated STsec approach with an equally re-imagined arrangement of agency is needed. The STsec framework we propose offers an opportunity to proactively imagine not just the potential uses but also harms which may come about from a tool, as well as a framework for users to actively draw upon in service of securing themselves.

3.2 ST Exploits in a Sociotechnical World

Sociotechnical exploits, including harassment, trolling, sockpuppetry, brigading, and algorithmic gaming (Allenby, 2017; Goerzen and Matthews, 2019), defy national security and legacy computer/information/network security approaches. A range of frames are emerging to describe behaviors that take place on a social level but utilize emergent affordances in sociotechnical systems. "Cognitive hacking," "media manipulation," "disinformation," "information warfare," "influence operations," "trolling," and "networked propaganda" are all different frames for behaviors that are horizontal to one another. What unifies them is that these threats activate, and are interpreted by, a mixed company of actors across a number of planes—societal, political, industrial, technical, human, non-human—and also incorporate the relations and associations between these actors (Balzacq and Cavelty, 2016). As security isn't just the notion of being free from danger, but also associated with the presence of an adversary, (Singer and Friedman, 2014: 34) it follows that an expanded toolkit for adversaries necessitates an expanded toolkit for security. These novel kinds of attack necessitate an expansion of security approaches, complimentary to legacy frameworks, which emphasize and integrate both technical and social considerations. Challenges like "soft cyber" attacks (Farrell and Schneier, 2018) cannot be adequately addressed by technical solutions alone (Mahar et al., 2018) and are a poor fit for national security frameworks based on Cold War concepts of nation-state adversaries. In this paper, we propose a framework for security thinking that we hope avoids some of the problems of the regimes as surveyed above, while practically speaking can be understood to modulate and complement them.

Engaging communities to foster greater security practices has become increasingly important across disciplines. In the social computing field and venues like ACM's CHI and CSCW conferences, the social aspects of computing systems are a grounding interest. Psychological, organizational, and behavioral factors are influential elements in the design not only of security protocols, but of entire computer systems. Social norms and cultural concepts such as safety, privacy, and identity, for instance, are key elements of computer practices (Antifakos, 2005; Friedman et al, 2002; Kiene et al., 2018; Lederer et al., 2003; Olson et al., 2005; Palen and Dourish, 2003; Shneiderman, 2000). Human-factors and usable security circles, along parallel lines, have long called for humans to be considered more than the "weakest link" in a computer system (Sasse et al., 2001), and that social values such as peer-to-peer trust, cooperation, and a sense of team be more thoughtfully engaged in the design of computer systems (Das, 2018; McGregor et al., 2017; Sedova, 2018; Watkins et al., 2016, 2017). For example, in one recent study focused on social aspects of computing systems, users were found to be more likely to sign up for tools like two-factor authentication if they were told that their friends were signing up, too (Das, 2018). The distinction between this work and our own lies, in part, in the referent object, or what's being protected: rather than leveraging communities to enhance individuals' practices, we're interested in how communities collectively frame and potentially enhance the security of the group as a whole.

Drawing on the foundational insights of "science and technology studies" and security studies, the STsec framework insists that the technical and the social are inseparably linked—that any framework for assessing the threats to either must consider them together, as a "sociotechnical system." These premises have been more or less explicit in the work of many emerging practitioners who take disinformation as a primary subject of study, such as Data & Society's Media Manipulation Initiative, the Credibility Coalition's Misinfosec working group, First Draft News, Harvard University's Technology and Social Change Project, and the Atlantic Council's DFRLab. In a recent paper, Gioe et al. argue that emergent threats require us to pay particular attention to the social layer of cyberspace, and that "concepts of cybersecurity must be expanded to include societal interaction with cyberspace down to the individual user" (2019). In another vein, political scientist Ron Deibert, taking a human-rights approach, has called for "human-centric cybersecurity," which rather than prioritizing territorial sovereignty, prioritizes the individual and views "networks as part of the essential foundation for the modern exercise of human rights" (2018).

Likewise, we posit that any security framework that takes "protection from harm" seriously must posit the individual and the community in which that individual is situated as the primary object of security. The task of ST security then

is not only to edge out the vulnerabilities in code, but also the vulnerabilities in the social systems that design, inform, and use the code or its applications—with the pointed goal of securing individuals and communities, with regards to their situated evaluation of harms.

4. STsec as a Framework of Analysis

While STsec as an analytics framework shares some similarities with other security concepts, there are differences which require an alternative set of questions to be asked.

First, the primary object in need of protection is the community, which is rarely a cohesive group let alone a unitary actor. This requires acknowledging that there may be tension and antagonism within the various sub-groups that make up the community and user base. Second, in other security frameworks like national security or information security, the objects of protection tend to be static, but within a sociotechnical system the referent object (i.e. the community) is accorded varying degrees of agency, influence, and power. Third, because the individuals and sub-groups that make up the community within the sociotechnical system are able to engage and influence the entire system and one another, the threats are not always external actors and in many cases may be users within the sociotechnical system itself. And fourth, because STsec considers the community (and the individuals that make up the community) the primary object of protection, questions about who or what is responsible and accountable in delivering security must be asked. Unlike in national or computational security, for example, it's not always clear cut who is legally or practically best equipped to deliver protection.

Understandably, because of the complexity of such a system, the strength of sociotechnical security as a concept is in identifying the relationships between the various actors (human and non-human), the positions of power that allow the actors to engage with and influence the system, and most importantly, the relational aspect of vulnerabilities and threats within the system. By clarifying these factors, we may be better able to locate when and where potential threat surfaces may emerge in such a system, and what mitigation strategies are most appropriate.

The following sections will highlight five pertinent high-level questions that must be asked when applying STsec to a given system. We stress that this is not an exhaustive list of questions and that additional lines of inquiry may be fruitful. Ultimately, the goal of conducting an analysis through a sociotechnical lens, however, is to identify the threats, vulnerabilities, risks, and tradeoffs when asking for security with regards to a given group of individuals, which, borrowing from securitization theory, we call the "referent community". In addition to illuminating the most salient and likely threats, these questions will hopefully also provide comparability between various referent communities or case studies.

It's important to note that the following questions do not need to be posed in any particular order. Where an analysis begins depends on the use case and whether it is being conducted after an event or as a proactive exercise with the aim of identifying potential threats and their likelihood of happening. In a proactive analysis, one may start by choosing a referent community as a starting point or a list of known threats. But in analyses where an event has already happened, it may be easier to begin by identifying the ST exploits first, in order to then infer the referent community and the vulnerabilities that were taken advantage of.

For the purposes of our paper, we are using past events with already existing research as examples in answering the following questions. These events include the inauthentic #EndFathersDay campaign to discredit Black feminists and the feminist movement writ large; the dissemination of extremist content following the Christchurch shooting; and "Endless Mayfly," an ostensibly Iranian-linked disinformation campaign spoofing existing media outlets. These cases were chosen to exemplify three broad categories of problematic content: targeted harassment, violent and extremist content, and political disinformation.

For a comparative STsec breakdown of the cases used in this paper, see **Appendix A** (#EndFathersDay), **Appendix B** (Christchurch Attack), and **Appendix C** (Endless Mayfly). In each example, we color code the type of vulnerability being exploited (technical or social), and identify the resulting ST exploit and associated threat.

4.1. What Is The Referent Community Requiring Protection?

Depending on whether this is an ex-ante (i.e., threat modeling) or ex-post (i.e. forensics or incident response) analysis, one may begin by identifying the referent group or one may leave that to the end. In an analysis beginning with a pre-identified community, a tradeoff has to be made when identifying the referent community; a larger, more inclusive community may help identify significant inter-group relations, but incur higher levels of complexity that may render analysis less useful. On the other hand, analysis focused on a smaller, more homogeneous community may miss out on key inter-group relationships that could help identify potential vulnerabilities within the ST system.

In analyses where the threat has already materialized, the referent community is likely to be inferred from the exploits and threats. In the #EndFathersDay case, unknown individuals (later identified as originating on 4chan messageboards) impersonated black feminist activist women to popularize the hashtag #EndFathersDay, arguing that the

holiday of Father's Day ought to be abolished. This inevitably got the attention of conservative media outlets, which used the hashtag to discredit the entire feminist movement. Individuals who exposed the hoax were subsequently harassed, receiving rape and death threats that took a psychological toll (Hampton, 2019). Our referent community in this STsec analysis is therefore the black women who were targeted and harassed as well as the social movement discredited by this campaign.

Another case study examines the dissemination of extremist content in the wake of the 2019 terrorist attack on the Al Noor Mosque in Christchurch, New Zealand, in which 51 individuals were killed and another 50 wounded. The terrorist and his allies additionally produced and disseminating a host of material designed to deliver additional harm in the wake of the attack. From the initial livestreaming of the massacre, to the dissemination of the recording, images and links posted to social media immediately prior to the attack, and an associated manifesto, it is apparent that the attack was designed to garner publicity, amplify a variety of white nationalist references, and otherwise shape discourse (Evans, 2019). The event confounded governance mechanisms around mainstream media, social media, and also a variety of smaller online forums. As such, we determined that a primary targeted community, though somewhat abstract, can be understood in many ways as the "public sphere." Ultimately, because this was a terrorist attack designed to achieve mass visibility, it is reasonable to frame the "public sphere" as the referent community.

It is also worth stressing that there may be competing referent communities. Conservative commentators, for example, view themselves as a group that has been unjustly discriminated against by left-leaning groups in tech (Carl, 2017; Stifling Free Speech: Technological Censorship and the Public Discourse, 2019). Yet another group of users on the same platform, say Muslims living in the U.K., may also see themselves as a target of persecution and discrimination (Tell M.A.M.A., 2016). Because sociotechnical systems are often incredibly complex with a multitude of groups and communities in constant motion and change, competing groups vying for security will likely be commonplace. As policymakers and technology creators, this complexity needs to be acknowledged, including an understanding of which communities are most at risk of violent attack on- and offline.

4.2 What Are The Features Or Conditions Within The Sociotechnical System That Can Be Exploited To Deliver The Threats?

As mentioned in Section 2, we define a vulnerability associated with an ST exploit or threat as a latent condition or feature that may be exploited for malicious means within a sociotechnical system. Again, this does not imply that these affordances are inherently flaws or weaknesses, only that they may be used in combination as an ST exploit to deliver a threat.

Even the most classic hallmarks of online communications can be exploited by bad actors. Anonymity, for example, long considered a pillar of free expression on the internet, is just one condition that's been exploited to create what are called "sock puppet" or "persona" accounts (such as those used in the #EndFathersDay campaign) where members of a marginalized group or movement are impersonated as a vector of attack, potentially with the side effect of damaging the legitimacy of anonymity itself, and those movements which defend it.

It's important to note that the technical affordance of "anonymity" on Twitter did not alone enable the success of the #EndFathersDay attack; rather, the exploit succeeded by playing to multiple vulnerabilities of both a technical *and* social nature: positioning the accounts seemingly within a social network through the cultivation of adjacencies like mentions and followers, and other uses of "data craft" (Acker, 2018) to manipulate metadata in order to game technical signifiers of legitimacy; the adoption of African-American Vernacular English-like speech, which appeared authentic to some onlookers, to game social signifiers of legitimacy; and the incentivization of media outlets to report on the implied sensational stories in service of their own quest for engagement. Cases like this highlight how platforms struggle to deal with contemporary attacks like harassment campaigns with the policies they have in place, in what can only be described as piecemeal and ad hoc community judgment guidelines (Pater et al., 2016; Marwick and Caplan, 2018).

In the case of "Endless Mayfly," the anonymity provided by Twitter was also exploited, albeit to create deceptive personas that disseminated false information (Lim et al., 2019). As in #EndFathersDay, the dissemination of false information took advantage of a combination of technical conditions (ex. anonymity and low barriers to account creation) and social conditions. The spoofed websites, for example, played on our brain's ability to ignore typos when creating lookalike domains (Smith and Muckli, 2010). Indepnedent.co.uk, for example, will read like the real website's URL, Independent.co.uk, for most people at first glance. When it came to dissemination of the spoofed websites, the operators exploited our inability to determine bot from authentic account (Shao et al., 2018; Stocking and Sumida, 2018; Freitas et al., 2014), and our tendency to only read headlines and share links without clicking (American Press Institute, 2014; Gabielkov et al., 2016).

Finally, a number of ST vulnerabilities are apparent from an examination of content dissemination both during and in the

wake of the 2019 terrorist attack on the Al Noor Mosque in Christchurch, New Zealand. Here, we can understand the limitations of content moderation protocols deployed by Facebook as representing a foundational vulnerability—one composed of multiple smaller vulnerabilities primarily social or technical in nature. For instance, the attacker succeeded in streaming a video of the assault in real time, using Facebook's Live feature. When posting the link to a message board composed of many sympathetic individuals, he took pains to ensure the URL would not render as a live link—writing the domain in the format "http:// www . face book . com / [his page]." In doing so, network analytics deployed by Facebook content moderation teams would be unable to determine the referral source of the incoming traffic—information that may prove relevant when the site in question is known to harbor extremists. Even if an automated system had flagged the video for further scrutiny due to the origination of incoming traffic, moderation practices related to live content possess a further vulnerability of a primarily social nature—in that moderators are limited in their capacity to invest their labor and attention in multiple videos simultaneously (Newton 2019; Klonick 2019). In the case of the Christchurch video, the terrorist spent several minutes engaged in innocuous activity before attacking—a fact that may have led a reviewer to dismiss the live content as innocuous. Even if the moderator had "snoozed" the video in order to check in on it again later, the placement of the attacker's video camera in relation to his weapon could lead a careless viewer—or even a machine learning system—to incorrectly classify the footage as originating in a First Person Shooter (FPS) video game.

When considering the subsequent archiving and dissemination of the video and the shooter's manifesto, a number of additional vulnerabilities come into play: such as the capacity for those intent on spreading the content to host it on archiving websites unwilling to comply with normative, political, and even legal calls to suppress the content; mass attempts to disseminate the archived stream on Facebook and other platforms by progressively iterating the video to evade automated deletion—and at a scale suggestive of a coordinated effort by a well-resourced actor in control of a large number of disposable "throwaway" accounts (in total, over 1.5 million videos were deleted, along with associated accounts [Klonick 2019]); and clear attempts in the terrorist's manifesto to bait journalists into addressing contentious issues and figures that would be likely to further exacerbate political polarization and animosity if amplified in the public domain.

4.3 Who Has Or Can Exploit The Vulnerabilities To Deliver A Threat To The Referent Community?

Identifying threat actors is a difficult task and one that may not be feasible, particularly when a community is threat modeling proactively. In the case of an ex-ante analysis—as in incident response or forensics—this shifts to a question of attack attribution. By analyzing evidence around an in-process or recently concluded attack, an analyst may be more readily able to identify a threat actor. In the #EndFathersDay case, for example, the threat actors were identified as "men's rights activists, pickup artists, and miscellaneous misogynists" (Hampton, 2019) organizing on 4chan and blogs. However, due to the anonymity provided by Twitter, it's impossible to know how much activity was coordinated and how much was due to individuals of their own volition choosing to harass these women. Similarly with "Endless Mayfly," the researchers were only able to attribute with moderate confidence the operations to an Iran-aligned group (Lim et al., 2019).

With Christchurch the threat actors are even more complex, as there were a variety of actors involved: the attacker himself; those disseminating the attacker's manifesto and the video of the massacre; those creating and optimizing content designed to speak to individuals searching for niche terminology used by the attacker; those receiving the disseminated content enthusiastically; and so on. The threat actors here can be recognized as a mix of white nationalists, "edgelords" from various social media platforms, those looking to spread chaos in service of various ends, anti-West groups using the shooting as evidence that Islam is under attack—and even journalists tempted to latch onto salacious references in the attacker's manifesto without recognizing them as bait (Evans 2019). When the referent community is the public sphere, the list of threat actors will typically be as diverse as it is complex.

If a known attacker remains active and the vulnerabilities they exploited remain in play, it may be reasonable to identify them as a threat actor worthy of further attention—or even something like what cybersecurity discourse terms an APT ("advanced persistent threat").[2] However, one must be cautious in this analysis; even in cybersecurity discourse attribution is a process plagued by sociotechnical complexities—as when attackers intentionally stage an attack in a way designed to mislead researchers into mistakenly attributing the attack to another faction (Rid and Buchanan, 2015).

4.4 What Is The Likelihood Of A Threat Manifesting?

If the analysis is being conducted ex-ante, identifying the likelihood of a related threat may be useful. This will require asking whether a set of threats has already materialized and whether it's likely to continue.Because the

[2] Operating in a more sociotechnical vein, Clint Watt of the Alliance for Securing Democracy has proposed the category of APMs, or "Advanced Persistent Manipulators" (Watt, 2019).

number of threats in any given sociotechnical system are likely to be more than one, conducting a risk assessment will allow potential threats to be triaged. Other factors to consider may include the low cost of conducting targeted online harassment, such as the likelihood of impunity and low barriers to account creation, the risk profile of individuals in the referent community, and the perceived threat the referent community poses to a known threat actor.

4.5 Who Is Responsible And Accountable For Providing Security To The Referent Community?

One thing separating STsec from conventional security frameworks is that the organization or individuals responsible for protecting the referent community is not a given, and identifying responsibility will often involve tradeoffs that may not be in the best interest of said community. In national security, for example, it is assumed that national military and intelligence agencies, often in coordination with their international counterparts, will be responsible and accountable for the security of the nation state. In information security, technical workers and their departments will be the primary actors responsible for the protection of the data and information of a given product, service, company, country, or network. Sociotechnical security, however, will depend on the referent community, the socio-political context in which they exist in, their reliance on or need of the technical processes in question, and the potential knock-on effects of addressing the vulnerabilities and threats facing them.

Furthermore, there is often a cost to addressing the vulnerabilities and threats within an ST system. Anonymity and low barriers to account creation may allow for malicious sockpuppets and botnets, but those same characteristics have also allowed freedom of expression to flourish, promoted civic engagement, connected once disparate individuals and groups, and expanded inclusivity. As such, any attempts to address a ST vulnerability or threat must be weighed against what could potentially be lost. Anonymity, interconnectedness, and participatory technology are not in themselves inherent vulnerabilities, but also opportunities for a richer and more engaged society. As such, one can see how difficult the task of assigning responsibility and accountability with regards to securing the referent community can be. In *The Morality of Security (*2019), Rita Floyd develops a theory of "just securitization," which acknowledges that there are always dangers associated with securitization, i.e. framing issues as a national security threat. Floyd cautions: "Raising an issue out of normal politics and into the realm of exceptional politics, where it is addressed by extraordinary measures, may, for example, result in the systematic infringement of key rights, the loss of civil liberties, an increase in police powers, 'othering'/alienation of suspect individuals and groups, the use of lethal force, and because the issue itself is removed from democratic decision-making, a reduction of the democratic process." Because of the potentially harmful effects of securitization, asking for protection necessitates asking ethical and moral questions about how security is delivered.

In the wake of the Christchurch shooting, for example, governments and technology companies have committed to the "Christchurch Call," which is an agreement to eliminate terrorist and violent extremist content online (Christchurch Call, 2019). While an admirable and ambitious commitment, it's unclear exactly who is legally and practically responsible for removing the unwanted content and what mechanisms for accountability will be available. Considering the multitude of motivations, resources, capabilities, limitations, and definitions of what is "terrorist" content, securing the multiple referent communities is an incredibly complex and ongoing task, requiring both offline and online actions to mitigate the spread of violent extremist content.[3]

Frequently, solutions to address problematic content have taken the form of "content moderation" practices. However, these processes themselves, positioned at the intersection of technical design and social nuance, can also harbor exploitable vulnerabilities—as when bands of users join together to flag content as a form of horizontal or "user-generated censorship" to silence legitimate perspectives (Peterson, 2013; Caplan, 2018). Moreover, account banning in relation to Terms of Services can have unforeseen repercussions—particularly when there is a lack of transparency around what those terms actually are—such as fostering perceptions of victimization, legal challenges, or the movement of users onto other platforms where their ideas are not subject to scrutiny and counter arguments by a broader public. A report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression also stressed that while governments and private companies should address hate speech and disinformation, current methods run the risk of undermining free speech on the internet (UN Human Rights Council 2018). David Kaye, the Special Rapporteur, stressed that "Governments should move away from such 'viewpoint regulation' and focus their regulatory efforts, if any, on company incentives to disclose information about the enforcement of their rules" (UN Office of the High Commissioner, 2018).

Furthermore, because many of the vulnerabilities are sociopolitical in nature, such as deep-rooted racism and

[3] It's worth noting that computer security discourse often confronts its own entanglements, as in the multi-stakeholder issues confounding efforts to secure vulnerabilities related to IoT devices (Geer, 2014).

misogyny in the case of #EndFathersDay or longstanding geopolitical rivalries in the case of "Endless Mayfly," technological solutions alone are unlikely to address these social cleavages, which will continue to exist so long as the ideology and narratives behind them continue to have pull. Likewise, media literacy, which has been touted as an antidote to false and misleading information by improving critical thinking, does not necessarily remove historical grievances and tensions or negate misogyny and racism; indeed, different epistemic communities can have different conceptions of literacy and source validity (Tripodi 2018).

5. Applications & Limitations

Systematically identifying sociotechnical vulnerabilities and the ways they can be exploited is useful for a number of individuals, namely policymakers, platforms and technology designers, and the communities that are being targeted or affected.

For policymakers, the elucidation of exploitable vulnerabilities will hopefully inform their decisions on what solutions might be feasible and also what their limitations may be. In the "Endless Mayfly" example, each of the exploits take advantage of a mix of social and technical vulnerabilities. A policymaker addressing a similar case may find an STsec analysis informs decisions about whether addressing the social or technical vulnerability will prove more effective, or if both are required. Where a condition, such as anonymity or ease of account creation, is deemed to have high positive value relative to the threat it presents, then finding ways to address the social condition may be more viable.

For social media companies and other technology creators, STsec may be useful in thinking about how their platforms and tools may be used in malicious ways beyond their original prescription. Doing so will potentially preempt abuse, thus avoiding possible costly legal and public relations problems. As technology companies' responsibility for removing unwanted content increases, a systematic way for identifying the ST threats and exploits will hopefully clarify what is technically feasible. STsec analysis could aid communities, activists, policymakers and regulators, and tech companies in incentive alignment. Furthermore, because ST threats are rarely confined to a single platform, inter-company organizations like the Global Internet Forum for Countering Terrorism or the Partnership on AI may find a comparable framework like STsec useful. Such a comparative framework also offers a chance to find patterns in ST exploits and build a database to head off future threats, in a similar fashion to the Common Vulnerabilities and Exposures (CVE) database.

We can already observe a range of activities that can be recognized as sociotechnical security practice, each of which might be usefully informed by a framework such as the one presented here.

The most foundational practices manifest as direct engagement by individuals and communities. Familiarity with known sociotechnical vulnerabilities and threats could enable communities to be proactive about their own security, while giving them agency to call for more accountability and response from the platform or specific actions from the government. We can already see communities rallying to their self-defense against campaigns like #EndFathersDay, often by implementing normative protocols for member identification and initiating their own awareness campaigns targeted at journalists, content moderators, and other informational gatekeepers. STsec may also provide a common framework for sharing and engaging with other similarly targeted communities when collaborating on collective security and threat modeling.

Increasingly, self-organized user communities and a variety of other organizations, such as non-profits and newsrooms, are informed by emerging folk security regimes often referred to as "digital security," "digital safety," and "digital self-defense" practices. These efforts are led by organizations and individual experts who seek to educate, train, or provide tools to others, with the goal of empowering them to be responsible for their own privacy and digital best practices (Wash, 2010; Al-Ameen et al., 2017). Often these programs address populations overlooked in less community-oriented approaches to security, and emphasize threats that are liminal—if not entirely unaddressed—in existent infosec and opsec protocols. Trainings and best practices manuals are offered by organizations like The Committee to Protect Journalists, The Electronic Frontier Foundation, The Engine Room, Equality Labs, the Tactical Technology Collective, and Our Data Bodies, whose Digital Defense Playbook aims to "support organizations and community members involved in intersectional fights for racial justice LGBQ liberation, feminism, immigrant rights, economic justice, and other freedom struggles" (Lewis et al, 2018).

The STsec model may also be useful in framing partnerships between stratified projects and local practitioners designed to promote the health of online communities. An example of this sort of endeavor is the work of CivilServant.io, a data science research group that collaborates with user communities to iterate and evaluate the effects of local moderation practices on community health and engagement (Matias, 2019). These protocols then serve as models for other communities, who might adapt them to their local needs. At a more general level, STsec might be a useful way to think about social science studies aimed at enhancing our

understanding about how to normalize good security practices.

Relatedly, a variety of non-profit consultancies and services have begun to appear, offering tailored services and trainings designed to help newsrooms, non-profits, NGOs, activist groups, and political campaigns mitigate sociotechnical threats, in a manner akin to the security consulting and incident response services initially offered by cybersecurity firms in the 1990s. For-profit firms are also emerging, addressing the effects of disinformation, trolling, and other sociotechnical threats by offering reputation management and pre-crisis PR services to brands and corporations.

While such resources, including trainings and education, are essential, we should also be wary not to normalize a shift in the burden of risk management onto the individual or the marketplace—instead emphasizing the need for a societal, structural approach to securitization around issues like harassment, disinformation, and horizontal censorship which can pre-empt the need for already overtaxed community members to ensure their own safety. Where localized knowledge and threat modeling is essential, we could at least seek to ensure that resources are in place to recognize and reward this labor.

Towards this goal, an STsec analysis might also inform activists, civil rights organizations, policy analysts, and investigative reporting outfits. ProPublica, for example, is already at work identifying ST vulnerabilities and calling for accountability and action from those parties responsible for maintaining and improving the systems in which those vulnerabilities appear. The Citizen Lab at the Munk School of Global Affairs & Public Policy, University of Toronto, likewise, situates their research at the intersection of human rights, global security, and technology, and frequently call on governments and corporations to address issues that affect traditionally overlooked or less powerful groups.

While some communities like Wikipedia already have governance structures that approach a more horizontal model, such arrangements on a larger scale will remain a distant goal given the design and ownership of sociotechnical infrastructure, from ISPs and DDOS-mitigation providers to social media companies. In the interim, however, we should societally be careful not to offload the burden of research, governance, and security onto individuals and communities who are not supported financially or otherwise to engage in this work.

6. Future Research

We also identify a variety of avenues for future research. First, a deeper examination of the relativistic or relational aspect of sociotechnical security—i.e., how vulnerabilities manifest differently for different groups—is needed. Relatedly, there is poor understanding of how cross-platform dynamics can produce unforeseen vulnerabilities of a sociotechnical nature—for instance, how the movement of content from one platform to another might change its nature. For researchers, the use of STsec as an analytical framework could allow comparability across multiple cases or events, potentially leading to a database where known exploits and vulnerabilities may be shared. While myriad case studies have emerged in recent years, there is not yet a standardized methodological way of organizing, hosting, or publishing the data. Furthermore, it's clear from just three cases that the types of exploits and vulnerabilities involved require interdisciplinary research from both the social and computer sciences if we are to produce meaningful analyses.

At a more abstract level, the STsec framework could be directly useful to those engaged in the sort of forensic work discussed in Section 4. Shared knowledge around common ST vulnerabilities and exploits can help ensure researchers don't 'reinvent the wheel' every time they try to make sense of a new incident, giving them clues about what vulnerabilities may be at work and who might be exploiting them. Such a framework can help inform a shared conceptual apparatus for discussing vulnerabilities and tactics, interphasing with similar efforts by some of the practitioners discussed in Section 3.2. In addition to privately maintained inventories of vulnerabilities, exploits, techniques, and tactics, STsec efforts will greatly benefit from public-facing inventories. The Global Media Manipulation Case Book (GMMCB) being developed at Harvard's Technology and Social Change Research Project is poised to provide a carefully curated set of foundational case studies. Meanwhile, Credibility Coalition's Misinfosec Working Group is developing a tactics, techniques, and procedures (TTP)-based framework inspired by computer security practices (Terp, 2019).

Private industry has begun to answer the call for more widely distributed consideration of how software and systems can harbor latent vulnerabilities. One institutionalized practice which has become a norm in computer security research is the "bug bounty program" in which non-employee programmers can be compensated for identifying technical vulnerabilities, or "bugs" (Elazari Bar On, 2018; Ellis et al, 2019). In a similar vein, Ashkan Soltani has called on Silicon Valley to conduct "abusability testing" (Soltani, 2019), in which firms allocate more organizational attention to the potential ways that software products could potentially be exploited by malicious actors. Research is needed to better understand the pros and cons of applying such programs to sociotechnical problems.

Academic researchers' hands are often tied to performing similarly speculative work, as research protocols frown on breaching Terms of Service agreements. This hampers efforts to "reverse engineer" or audit black-boxed recommendation algorithms to better understand problems like polarization, radicalization, and bias. Crucially, reticence on the part of digital infrastructure owners to allow research to proceed unbidden may be a vulnerability. In an instance that resonates with the challenges faced by hacker communities in the 1990s, the CFAA (Computer Fraud and Abuse Act) has even been deployed to frustrate the research (Sandvig v. Barr, 2019). Research should be protected in the form of safe harbor for researchers (Knight Institute, 2018; Korolova, 2018).

Relatedly, it is crucial to interrogate the professionalization of ST security research—and also what types of individuals will be recognized as possessing expertise. Whether it is the work of threat modeling, abusability testing, or more academicized forms of knowledge production, the question of expertise and compensation for labor is particularly important in a discourse where localized knowledge and practical experience is integral.

Other foundational questions relate to issues around the disclosure of ST vulnerabilities and exploits, particularly when public discussion of these issues could be seen to perpetuate or amplify threats (Phillips, 2018). Some insight may be on offer from computer security research, where public disclosure of vulnerabilities was (and remains) a hotly contested subject (Arora et al, 2006; Schneier, 2007). Computer security discourse, in general, harbors a rich array of concepts and protocols that might be adapted to the challenges of ST security. Moreover, it will be valuable to understand where conflicts of a sociotechnical nature have already emerged in computer security discourse—particularly around issues like the sale of vulnerabilities to states, the sale of hacking tools to human rights abusing governments, and the ethics of providing security services to companies and organizations that might threaten the safety of community members.

We can also see value in expanding the remit of what we see as informative research; artists like Constant Dullaart, who created and managed a large botnet to write poetry on Instagram pages, demonstrate the ease of deploying influence campaigns, while also revealing the vulnerabilities and exploits required for their operation. Work by explorative artists can be seen as furnishing valuable knowledge regarding the processes of manipulation; by recognizing these practices as a form of exploratory security research designed to discover and elucidate ST exploits and vulnerabilities, and perhaps even formalizing it, those seeking to improve systems or defend communities might gain insight into foundational threats.

Central to a sociotechnical lens is the idea that the social and the technical are irreducible, that those affected by problems are experts in their own situation, and that their perspectives need to be included in any truly comprehensive research agenda. The "threat modeling" process that forms the backbone of any practical security protocol tacitly recognizes this, in that those threatened are best able to understand what constitutes a risk. Thus, any research that proceeds on this subject matter would benefit from centering the perspective of those affected, recognizing their local expertise—and then syndicating and compiling the insights gained from multiple perspectives to understand how specific vulnerabilities and harms might stratify particular contexts. Security should not be a process that is solely bottom-up or top-down. Ideally it would be horizontal, with local communities sharing best practices with one another, empowered to implement any requisite technical changes accordingly. Security is best understood as a moving target (Bilar et al., 2013). This work is continuous.

7. Acknowledgements

The authors would like to thank the Data & Society Research Institute for supporting this work, and in particular danah boyd, Madeleine C. Elish, Charley Johnson, Robyn Caplan, and all the members of the Media Manipulation Initiative for their invaluable input and collaboration. The authors would also like to thank the Columbia Tow Center, Emily Bell, Susan McGregor, Kelly Caine, Gabriella Coleman, and Franzi Roesner for their support of the research leading to this work.

8. References

Anderson, B. (2006). *Imagined communities: Reflections on the origin and spread of nationalism.* Verso books.

Al-Ameen, M.N., Watkins, E.A., Lowens, B., Roesner, F., McGregor, S.E. and Caine, K. (2017). Evaluating Online Security Training for Journalists Through the Lens of Learning Science. USENIX ASE Lightning Talk.

Allenby, B. R. (2017). The Age of Weaponized Narrative or, Where Have You Gone, Walter Cronkite? *Issues in Science and Technology,* 33(4), 65-70.

Amnesty International. (2018). Troll Patrol Findings.

Antifakos, S., Kern, N., Schiele, B., & Schwaninger, A. (2005, September). Towards improving trust in context-aware systems by displaying system confidence. In Proceedings of the 7th international conference on Human

computer interaction with mobile devices & services (pp. 9-14). ACM.

Ananny, M., & Bighash, L. (2016). Why drop a paywall? Mapping industry accounts of online news decommodification. *International Journal of Communication*, 10, 22.

Arora, A., Nandkumar, A., Telang, R. (2006). Does information security attack frequency increase with vulnerability disclosure? An empirical analysis. *Information Systems Frontiers, 5*(8), 350-362.

Balzacq, T., & Cavelty, M. D. (2016). A theory of actor-network for cyber-security. *European Journal of International Security*, *1*(2), 176-198.

Benkler, Y. (2006). *The wealth of networks: How social production transforms markets and freedom.* Yale University Press.

Bilar, D., Cybenko, G., & Murphy, J. (2013). Adversarial dynamics: the conficker case study. In *Moving Target Defense II* (pp. 41-71). Springer, New York, NY.

Brogan, J. (2015, May 6). Famous Troll Targets Activists With White-Supremacist Promoted Tweets. *Slate*.

Brooking, E.T., and Singer, P. W. (2016, November). War Goes Viral: How Social Media is Being Weaponized Across The World. *The Atlantic*.

Bossetta, M. (2018). The Weaponization Of Social Media: Spear Phishing And Cyberattacks On Democracy. *Journal of International Affairs,* 71(1.5), Special Issue: Contentious Narratives: Digital Technology And The Attack On Liberal Democratic Norms: 97-106.

Buzan, B., Waever, O., and de Wilde, J. (1998). *Security: A New Framework For Analysis*. Boulder, CO: Lynne Rienner.

Caplan, R. (2018). Content or Context Moderation? Data & Society Research Institute.

Carl, J. (2017, August 15). How to Break Silicon Valley's Anti-Free-Speech Monopoly. *National Review.*

Checkoway, S., McCoy, D., Kantor, B., Anderson, D., Shacham, H., Savage, S., Koscher, K., Czeskis, A., Roesner, F. & Kohno, T. (2011, August). Comprehensive experimental analyses of automotive attack surfaces. In USENIX Security Symposium (Vol. 4, pp. 447-462).

Christchurch Call. (2019). https://www.christchurchcall.com.

Dalpiaz, F., Paja, E., & Giorgini, P. (2011, September). Security requirements engineering via commitments. In 2011 1st Workshop on Sociotechnical Aspects in Security and Trust (STAST) (pp. 1-8). IEEE.

Dalpiaz, F., Paja, E., & Giorgini, P. (2016). *Security requirements engineering: designing secure sociotechnical systems*. MIT Press.

Das, S. (2018). *Social Cybersecurity: Reshaping Security through an Empirical Understanding of Human Social Behavior*. USENIX Enigma 2018.

Deb, A., Donohue, S., and Glaisyer, T. (2017, October). *Is Social Media a Threat to Democracy?* The Omidyar Group.

Deibert, R. J. (2018). Toward a Human-Centric Approach to Cybersecurity. *Ethics & International Affairs*, 32(4), 411-424.

Doctorow, C. (2018, May 7). The Engagement-Maximization Presidency. *Locus Mag*.

Elazari Bar On, A. (2018). Private Ordering Shaping Cybersecurity Policy: The Case of Bug Bounties.

Elish, M.C. and Watkins, E.A.. (2019). When Humans Attack: Rethinking Safety, Security, and AI. *Data & Society Points*.

Ellis, R., Huang, K., Siegel, M., Moussouris, K., & Houghton, J. (2017). *Fixing a hole: The labor market for bugs. New Solutions for Cybersecurity,* 122-147.

Epstein, R. and Robertson, R.E. (2015). Search engine manipulation effect (SEME). *Proceedings of the National Academy of Sciences Aug 2015*, 112 (33): E4512-E4521.

Evans, R. (2019, March 15). Shitposting, inspirational terrorism, and the Christchurch mosque massacre. *Bellingcat.*

Farrell, H.J., and Schneier, B. (2018, October). Common-Knowledge Attacks on Democracy. Berkman Klein Center Research Publication No. 2018-7.

Floyd, R. (2019). *The Morality of Security: A Theory of Just Securitization.* New York, NY: Cambridge University Press.

Friedman, B., Hurley, D., Howe, D. C., Nissenbaum, H., & Felten, E. (2002, April). Users' conceptions of risks and harms on the web: a comparative study. In *CHI'02 extended*

abstracts on Human factors in computing systems (pp. 614-615). ACM.

Freitas, C.A., Benevenuto, F., Ghosh, S., and Veloso, A. (2014). Reverse Engineering Socialbot Infiltration Strategies in Twitter.

Frenkel, S., Conger, K., & Roose, K. (2019, January 31). Russia's Playbook for Social Media Disinformation Has Gone Global. *The New York Times*.

Gabielkov, M., Ramachandran, A., Chaintreau, A., Legout, A. (2016, June). Social Clicks: What and Who Gets Read on Twitter? *ACM SIGMETRICS / IFIP Performance 2016*.

Geer, D. (2014, May 7). Security of Things. Keynote, Security of Things workshop.

Gioe, D. V., Goodman, M. S., & Wanless, A. (2019). Rebalancing cybersecurity imperatives: patching the social layer. *Journal of Cyber Policy*, 1-21.

Goel, V., Raj, S., & Ravichandran, P. (2018, July 18). How WhatsApp Leads Mobs to Murder in India. *The New York Times*.

Goerzen, M. and Matthews, J. (2019). Black Hat Trolling, White Hat Trolling, and Hacking the Attention Landscape. FATES.

Golebiewski, M. and boyd, d. (2018). Data Voids: Where Missing Data Can Easily Be Exploited. Data & Society Primer.

Groll, E. (2019, May 14). How Pro-Iran Hackers Spoofed FP and the News Media. *Foreign Policy*.

Habermas, J. (1991). *The structural transformation of the public sphere: An inquiry into a category of bourgeois society.* MIT Press.

Hadnagy, C. (2010, December 21). *Social engineering: the art of human hacking*. Wiley.

Hampton, R. (2019, April 23). The Black Feminists Who Saw the Alt-Right Threat Coming. *Slate*.

Hansen, L., & Nissenbaum, H. (2009). Digital disaster, cyber security, and the Copenhagen School. *International studies quarterly*, 53(4), 1155-1175.

Haraway, D. (1988). Situated Knowledges. *Feminist Studies, 14(3),* 575-599.

---. (2013). *Simians, Cyborgs, and Women: The Reinvention of Nature.* Routledge.

Homolar, A. (2015). Human security benchmarks: Governing human wellbeing at a distance. *Review of International Studies*, *41*(5), 843-863.

Howard, P. N. (2018, October 18). How Political Campaigns Weaponize Social Media Bots. *IEEE SPectrum*.

Hutchins, E. (1995). *Cognition in the wild*. Cambridge, Mass: MIT Press.

Ingram, M. (2018, January 31). Twitter has been ignoring its fake account problem for years. *Columbia Journalism Review*.

Kavanagh, J. and Rich, M.D. (2018). *Truth Decay: A Threat to Policymaking and Democracy*. Santa Monica, CA: RAND Corporation.

Klonick, K. (April 2019). Inside the Team at Facebook That Dealt with the Christchurch Shooting. *The Verge.*

Knight Institute. Knight Institute Calls on Facebook to Lift Restrictions on Digital Journalism and Research. (2018, August 7). https://knightcolumbia.org/

Korolova, A. (2018, December 18). Facebook's Illusion of Control over Location-Related Ad Targeting. Medium.

Latonero, M., & Kift, P. (2018). On Digital Passages and Borders: Refugees and the New Infrastructure for Movement and Control. *Social Media + Society*.

Latonero, M., Hiatt, K. Napolitano, A., Clericetti, Giulia, Penagos, M. (2019). *Digital Identity in the Migration & Refugee Context*. Data & Society Research Institute.

Latour, B. 1988. *The Pasteurization of France*. Cambridge, Mass.: Harvard University Press.

---. 1991. Technology is Society Made Durable. Pp. 103-131 in *A Sociology of Monsters: Essays on Power, Technology, and Domination*, edited by John Law. London, Boston: Routledge & Kegan Paul.

Lederer, S., Mankoff, J., & Dey, A. K. (2003, April). Who wants to know what when? privacy preference determinants in ubiquitous computing. In *CHI'03 extended abstracts on Human factors in computing systems* (pp. 724-725). ACM.

Leonardi, P.M., Nardi, B.A. and Kallinikos, J. eds., (2012). Materiality, Sociomateriality, and sociotechnical Systems: What Do These Terms Mean? How Are They Different? Do

We Need Them? by Paul Leonardi, in *Materiality and organizing: Social interaction in a technological world*. Oxford University Press. 25-48.

Lewis, T., Gangadharan, S. P., Saba, M., Petty, T. (2018). *Digital defense playbook: Community power tools for reclaiming data.* Detroit: Our Data Bodies.

Lim, G., Maynier, E., Scott-Railton, J., Fittarelli, A., Moran, N., and Deibert R. (2019). *Burned After Reading Endless Mayfly's Ephemeral Disinformation Campaign*. Citizen Lab.

Madrigal, A. C. (2019, May 15). The Telltale Signs of a Fake 'Atlantic' Article. *The Atlantic*.

Mahar, K., Zhang, A. X., & Karger, D. (2018, April). Squadbox: A tool to combat email harassment using friendsourced moderation. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (p. 586). ACM.

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

Matias, N. (2019). Reddit mods: let's test ideas for preventing harassment and fact-checking the news. Civilservant.io.

McGregor, S. E., Watkins, E. A., Al-Ameen, M. N., Caine, K., & Roesner, F. (2017). When the weakest link is strong: Secure collaboration in the case of the Panama Papers. In 26th {USENIX} Security Symposium 17) (pp. 505-522).

McMillan, R. (2019, March 21). After New Zealand Shooting, Founder of 8chan Expresses Regrets. *The Wall Street Journal.*

Melton, M. (2018, November 27). Climate Change and National Security, Part I: What is the Threat, When's It Coming, and How Bad Will It Be? Lawfare Blog.

Molotch, H. (2014). *Against Security: How We Go Wrong at Airports, Subways, and Other Sites of Ambiguous Danger-Updated Edition*. Princeton University Press.

Morris, M. (2019, March 27). U.S. Government Declares Grindr a National Security Risk. *Fortune*.

Motti, V. G., & Caine, K. (2015, January). Users' privacy concerns about wearables. In *International Conference on Financial Cryptography and Data Security* (pp. 231-244). Springer, Berlin, Heidelberg.

Mozur, P. (2018, October 15). A Genocide Incited on Facebook, With Posts From Myanmar's Military. *The New York Times.*

Müller, K., & Schwarz, C. (2017). Fanning the Flames of Hate: Social Media and Hate Crime. *SSRN Electronic Journal*.

Nadler, A., Crain, M., & Donovan, J. (2018). *Weaponizing the digital influence machine: The political perils of Online Ad Tech*. Data & Society Research Institute.

Newton, C. (2019, February 25). *The Trauma Floor.* The Verge.

Niederer, S. and Van Dijck, J., 2010. Wisdom of the crowd or technicity of content? Wikipedia as a sociotechnical system. *New Media & Society*, 12(8), pp.1368-1387.

Olson, J. S., Grudin, J., & Horvitz, E. (2005, April). A study of preferences for sharing and privacy. In *CHI'05 extended abstracts on Human factors in computing systems* (pp. 1985-1988). ACM.

Owen, T. (2019, April 29). What we know about the 19-year-old charged with murder in California synagogue shooting. *Vice News.*

Paja, E., Dalpiaz, F., Poggianella, M., Roberti, P., & Giorgini, P. (2012, September). STS-Tool: sociotechnical security requirements through social commitments. In *2012 20th IEEE International Requirements Engineering Conference* (RE) (pp. 331-332). IEEE.

Palen, L. and Dourish, P. (2003). April. Unpacking privacy for a networked world. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 129-136). ACM.

Pater, J. A., Kim, M. K., Mynatt, E. D., & Fiesler, C. (2016, November). Characterizations of online harassment: Comparing policies across social media platforms. In Proceedings of the 19th International Conference on Supporting Group Work (pp. 369-374). ACM.

Peterson, C. (2013). User-generated censorship: manipulating the maps of social media. MIT Thesis.

Phillips, W. (2018). *The Oxygen of Amplification.* Data & Society Research Institute.

Prier, J. (2017). Commanding the Trend: Social Media as Information Warfare. *Strategic Studies Quarterly,* 11(4), 55-85.

Orlikowski, W.J. (2007). Sociomaterial practices: Exploring technology at work. *Organization studies,* 28(9), pp.1435-1448.

Rid, T. & Buchanan, B. (2015). Attributing cyber attacks. *Strategic Studies (1-2)*38, 4-37.

Romm, T., Dwoskin, E., & Timberg, C. (2019, April 22). Sri Lanka's social media shutdown illustrates global discontent with Silicon Valley. *The Washington Post*.

Roesner, F., Kohno, T., & Molnar, D. (2014). Security and privacy for augmented reality systems. *Communications of the ACM*, 57(4), 88-96.

Sandvig v. Barr - Challenge to CFAA Prohibition on Uncovering Racial Discrimination Online. (2019, May 22). ACLU.

Sasse, M. A., Brostoff, S., & Weirich, D. (2001). Transforming the 'weakest link'—a human/computer interaction approach to usable and effective security. *BT technology journal*, 19(3), 122-131.

Shneiderman, B., 2000. Designing trust into online experiences. Communications of the ACM, 43(12), pp. 57-59.

Schneier, B. (2007). Full Disclosure of Security Vulnerabilities a 'Damned Good Idea.' CSO Online.

Schudson, M. (1999). *The good citizen: A history of American civic life*. Harvard University Press.

Sedova, M., (2018). *Surfing the motivation wave to create security behavior change*. USENIX Enigma 2018.

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K. C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature communications*, 9(1), 4787.

Singer, P. W., & Friedman, A. 2014. *Cybersecurity: What everyone needs to know*. OUP USA.

Smith, F.W., Muckli, L. (2010, November). Nonstimulated early visual areas carry information about surrounding context. *Proceedings of the National Academy of Sciences Nov 2010*, 107 (46): 20099-20103.

Snegovaya, M. (2015). *Putin's Information Warfare In Ukraine: Soviet Origins Of Russia's Hybrid Warfare.* Institute for the Study of War.

Software Freedom Law Center. (2016). *Online Harassment: A Form of Censorship*.

Soltani, A. (2019). Abusability Testing: Considering the Ways Your Technology Might Be Used for Harm. USENIX Enigma 2019.

Stamos, A. (2019, January 19). @alexstamos. "Stock-based comp in tech creates harmful incentives, exacerbates gender and racial comp disparities (which are underreported), and every successful company has several super-rich early employees who believe they earned it but are considered lottery winners by everybody else." Twitter post. https://twitter.com/alexstamos/status/1086806903886209024

Stifling Free Speech: Technological Censorship and the Public Discourse. Hearings before the Subcommittee on the Constitution, Senate. (2019).

Stocking, G. and Sumida, N. (2018). Social Media Bots Draw Public's Attention and Concern. Pew Research Center.

Stuart, Keith. (2014, October 17). Brianna Wu and the human cost of Gamergate: 'every woman I know in the industry is scared.' *The Guardian*.

Tell M.A.M.A. (2016). *Annual Report: A Constructed Threat: Identity, Intolerance and the Impact of Anti-Muslim Hatred*. London: Faith Matters.

Terp, S-J. (2019). Misinformation has stages. Medium: Misinfocon.

Thomas, D. (2002). *Hacker Culture.* University of Minnesota Press.

Tripodi, F. (2018). *Searching for Alternative Facts.* Data & Society Research Institute.

Trist, E. (1981). The evolution of sociotechnical systems. *Occasional paper*, *2*.

UN Human Rights Council. (2018, April 6). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression.

UN Office of the High Commissioner. (2017, March 8). UN experts urge States and companies to address online gender-based abuse but warn against censorship.

UN Office of the High Commissioner. (2018, 19 June). Landmark report by UN expert urges governments and internet firms to ensure freedom of expression online.

Wagner, Kurt. (2018, Feb 8). Twitter made a profit by cutting costs, not by growing its business. *Vox*.

Warner, M. (2005) *Publics and Counterpublics.* Zone Books.

Watkins, E. A., Al-Ameen, M. N., Roesner, F., Caine, K., & McGregor, S. (2017). Creative and Set in Their Ways: Challenges of Security Sensemaking in Newsrooms. 7th USENIX Workshop on Free and Open Communications on the Internet (FOCI 17).

Watkins, E. A., Roesner, F., McGregor, S., Lowens, B., Caine, K., & Al-Ameen, M. N. (2016, October). Sensemaking and Storytelling: Network Security Strategies for Collaborative Groups. In *2016 International Conference on Collaboration Technologies and Systems (CTS)* (pp. 622-623). IEEE.

Watt, C. (2019). Advanced Persistent Manipulators, Part One: The Threat to the Social Media Industry. Alliance for Securing Democracy.

Winner, L. (2004). Trust and terror: the vulnerability of complex socio-technical systems. *Science as Culture,* 13(2). 155-172.

Wu, T. (2017). *The attention merchants: The epic scramble to get inside our heads*. Vintage.

Appendix A: Sociotechnical Breakdown of #EndFathersDay[4]

Background:	In 2014, unknown individuals (later identified as originating on 4chan messageboards) impersonated black feminist activist women to popularize the hashtag #EndFathersDay, arguing that the holiday of Father's Day ought to be abolished. This got the attention of conservative media outlets, which used the hashtag to discredit the entire feminist movement. Individuals who exposed the hoax were subsequently harassed, receiving rape and death threats.[5]		
Referent Community:	African-American feminist activists and the legitimacy of their movement as perceived by outsiders and media commentors.		
Threat (Unwanted Event)	**ST Exploit (vehicle)**	**Vulnerabilities (feature or condition)** Blue = technical vulnerability Red = social vulnerability	**Threat Actor/Operator**
Damaged public perception/legitimacy of black feminism movement	Personas posing as activists sharing false and inflammatory content by adopting shibboleths of a particular group by: 1) using photos of women lifted from the internet and; 2) adopting shibboleths of a particular group ("African-american vernacular" sentence structure and radical feminist vocabularly).	- Ease of account creation - Anonymity of Twitter accounts - Ease of account creation - Anonymity of Twitter accounts - Vernacular is perceived as a signal of belonging to a particular group	4chan trolls/bad actors
Sowed dissent within the movement by using Twitter's amplification algorithms to capitalize on "existing rifts in the online feminist movement related to race and class"[6]	Manipulating data to game the right followers and hashtags to appear part of a movement and spread their message along particular channels of activists.	- Position within a Twitter social network, i.e. adjacencies like mentions and followers, are part of a suite of data points algorithms use to amplify messages to a specific group - Humans perceive these data points and algorithmic sharing as signals that a person as being part of a group and facilitates a	Platform's sharing algorithms

[4] These charts are by necessity reductive, and can vary in their level of analytic granularity according to the conditions of the case—and the familiarity / expertise of an analyst in a given subject. For use in threat modeling, it follows that some areas will be more or less relevant than others and will necessitate greater or lesser detail. Furthermore, it may be prudent in some cases to simplify for the sake of brevity.

[5] Hampton, 2019

[6] ibid

		message being spread throughout that group[7]	
Disinformation initially spread on Twitter was amplified throughout the media ecosystem	Tweets were crafted to draw attention via outrageous vitriol	- Media business models dependent on sensationalist headlines for clicks/engagement[8] - Platform distribution of Tweets and re-Tweets lowered cost of amplifying messages for media to notice and pick up	Media/advertising companies
"	Tweets were crafted as political bait for right-wing media personalities by opposing a holiday associated with patriarchal family values	Right-wing media pundits looking for strawmen cases to discredit left-wing feminism - Platform distribution of Tweets and re-Tweets lowered cost of amplifying messages for media to notice and pick up	Media pundits
Activiststs who fought to 'out' the fake hashtag were subjected to harassment, including rape and death threats and calls to their places of employment	Bad actors called for activists' personal identifying information to be posted online, i.e. "doxxing"	- Failure for platforms to moderate doxxing efforts - Platform distribution of information makes it easy to share identifying information	Platform companies

[7] Acker, 2018
[8] Doctorow, 2018; Nadler et al., 2018; Stamos, 2019

Appendix B: Sociotechnical Breakdown of Christchurch Attack

Background:	On March 15, 2019 a terrorist attack at a New Zealand mosque was livestreamed in its entirety on Facebook. Content moderators removed the original video and the attacker's page in the wake of the attack. However, copies of the video, images of the shooter's weapon scrawled with white nationalist references, and an associated manifesto were subsequently spread widely on social media platforms including Facebook—often through the use of strategic iteration to avoid automated detection systems—despite attempts by authorities, platforms, and communities to limit the material's reach.		
Referent Community:	Muslim users potentially traumatized by the content's spread; Those subject to copycat attacks, or attacks by individuals bolstered by disseminated content; Facebook Content Moderation teams; Communities at risk for radicalization; Journalists; The public sphere writ large		
Threat (Unwanted Event)	**ST Exploit (vehicle)**	**Vulnerabilities** Blue = technical vulnerability Red = social vulnerability	**Threat Actor/Operator**
Facebook livestream is not identified by content moderation system as worthy of scrutiny due to its connection to a known problematic source (in this case, 8chan's /pol/ board)	Shared link to Facebook subjected to "linkbreaking" (i.e. https:// www . face book . com/...)	- Hyperlinks pasted into URL bar cannot be tracked by destination website (in this case Facebook) to originating source (in this case 8chan). - Inability to identify co-ordinated incoming traffic	Christchurch attacker
Video streamed uninterrupted, and saved by watchers for further dissemination	- Innocuous activity in the opening sequence of a video. - Configuration of video to appear (potentially to both casual human observers and algorithms) as simulated (in this case, as a first person shooter game)	- Content moderators do not have the capacity to watch an entire video, they need to "snooze" and check in periodically - Difficulty to predict what will happen next - Difficulty to correctly recognize what is happening (in this case, believing the video to be footage from a First Person Shooter video game) - Those watching the video might not flag it to content moderators - Algorithmic content moderation processes cannot accurately assess content or even nominate it for flagging, unless it conforms to an already established hash or fingerprint	Christchurch attacker

Secondary Dissemination of Video	Uploads of video versions iterated to circumvent automated detection, using multiple accounts. In this case, Facebook noted 1.5 million accounts attempted to upload the video)	- Hosting and archiving services beyond the reach of regulators and unwilling to conform to normative or legal efforts to remove content - The ability to download a livestreamed video - Inability for automated recognition systems to identify slightly modified variations on fingerprinted content - Capacity to create "throwaway" accounts, due to unenforced or lax ToS policies (in this case, Facebook's Real Names policy)	Unsure. Potentially involved coordinated, state-backed activity.
The broad amplification of white nationalist keywords and phrases increasing the ease of engaging with the ideas	- Shooter wrote legible, searchable terms on his weapons - Threat actors can use SEO (search engine optimization) techniques to fill data voids, effectively capturing attention	- Ease of disseminating images using social media platforms and other participatory media - Trusting results associated with a search engine regardless of quality[9] - The lack of rigorous and critical content on every searchable subject - Search engine's necessity to return content to a searcher even if it is low quality (Note, these latter two vulnerabilities combine to produce a class of sociotechnical vulnerability known as "data voids"[10])	- Christchurch attacker - Broader white nationalist network
Heightened polarization; Resentment of those who sympathize with referenced figures and deem them to be unfairly attacked; Sustained attention on a given issue or figure	- Namechecking contentious / storied figures in manifesto (i.e. Candace Owens and PewDiePie)	- Engagement metrics as value - Human appetite as controversy - "If it Bleeds it Leads" (i.e., editorial preference for spectacular content) - Technical optimization for engagement (i.e. through trending and recommendation systems)	- Christchurch attacker - Designers of engagement systems - Journalists

[9] Epstein and Robertson, 2015
[10] Golebiewski and boyd, 2018

Appendix C: Sociotechnical Breakdown of Endless Mayfly[11]

Background:	Endless Mayfly is a seemingly Iran-aligned network of personas and social media accounts that spreads falsehoods by spoofing news outlets like the *Atlantic*, *Haaretz*, or *Foreign Policy* and disseminating them over Twitter and third-party sites via personas crafted to look like journalists, activists, and students. The content itself amplified narratives critical of Saudi Arabia, the United States, and Israel, and targeted real journalists and media outlets.		
Referent Community:	Journalists. Although it's reasonable to assume that the operators behind the disinformation campaign were targeting journalists in order to reach a larger public audience, this analysis will use the targeted journalists as the referent community.		
Threat (Unwanted Event)	**Exploits (Vehicle)**	**Vulnerability (Feature or condition)** Blue = technical vulnerability Red = social vulnerability	**Threat Actor/Operator**
False information that presents as factual information	Spoofed websites made to look like legitimate news outlets	- Websites spoofed were primarily well-established and credible news outlets, likely playing on people's trust in these media brands - Target websites' code is public, and was likely scraped and then replicated on a lookalike domain - URL redirects would send visitors to the lookalike root domain to the real website being impersonated	Endless Mayfly
	Typosquatting (intentional registration of domains that take advantage of typographical variants of an already existing domain)	- Human cognition (ex. mistaking indepnedent.co.uk for the actual news outlet, independent.co.uk)[12] - Name space (not all TLDs were registered by the websites being spoofed) - Anonymity of domain registration	Endless Mayfly
Inauthentic amplification of false information on Twitter	Automated Twitter accounts (bots)	- Humans are not great at distinguishing bots from authentic accounts[13] - Ease of account creation - Anonymity of Twitter accounts	Endless Mayfly
	Personas posing as journalists, activists, or students sharing links and screenshots of the spoofed websites.	- Human tendency to only read headlines and share links they haven't read[14] - Anonymity of Twitter accounts	Endless Mayfly

[11] Groll, 2019; Lim et al., 2019; Madrigal, 2019
[12] Smith and Muckli, 2010
[13] Freitas et al., 2014; Shao et al., 2018; Stocking and Sumida, 2018
[14] American Press Institute, 2014; Gabielkov et al., 2016

False information amplified on legitimate websites.	Personas posing as journalists, activists, or students writing for unaffiliated third-party websites.	- Weak editorial standards (ex. ModernGhana.com or TheSeoulTimes.com published inauthentic content containing false information from the personas) - Ease of account creation and publication (ex. Buzzfeed Community or Medium accounts were created in the personas' names to publish misleading and false information)	Endless Mayfly
Dissemination of false information to real journalists and activists.	Personas posing as journalists, activists, or students contacting real journalists and activists.	- Personas posed as individuals with commonalities with their targets (ex. posing as a journalist or student interested in Middle East politics) - Anonymity of Twitter accounts - Direct messaging (personas reached out to targets via Twitter direct messaging) - Hidden nature of Twitter mentions (beginning a Tweet with an account's handle hides that tweet from the default profile view, allowing the persona to contact as many people on Twitter while keeping their spam-like behavior somewhat hidden)	Endless Mayfly