Automated Influence and the Challenge of Cognitive Security

Sarah Rajtmajer and Daniel Susser*
smr48,dus1042@psu.edu
The Pennsylvania State University
University Park, PA

ABSTRACT

Advances in AI are powering increasingly precise and widespread computational propaganda, posing serious threats to national security. The military and intelligence communities are starting to discuss ways to engage in this space, but the path forward is still unclear. These developments raise pressing ethical questions, about which existing ethics frameworks are silent. Understanding these challenges through the lens of "cognitive security," we argue, offers a promising approach.

CCS CONCEPTS

• **Security and privacy → Human and societal aspects of security and privacy**; • **Social and professional topics**;

KEYWORDS

disinformation, information operations, grey zone conflict, cognitive security

ACM Reference Format:
Sarah Rajtmajer and Daniel Susser. 2020. Automated Influence and the Challenge of Cognitive Security. In *Hot Topics in the Science of Security Symposium (HotSoS '20), April 7–8, 2020, Lawrence, KS, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3384217.3385615

1 INTRODUCTION

As applications of artificial intelligence (AI) proliferate, concerns grow about whether AI-driven systems conform to or disrupt shared values. Nowhere are such concerns more pronounced, especially of late, than in discussions about automation and online influence [e.g., 30, 50, 57].[1] Researchers have long worried that social media platforms, like Facebook and Twitter, which organize and distribute much of the information people regularly access about the news, about one another, and about events near and far, are failing to organize that information in ways that promote important epistemic values, such as accuracy, objectivity, and diversity of perspective.[2] Social media platforms fail to filter out false or sensationalized stories, critics charge [23]. Indeed, because the algorithms selecting and arranging content in social media feeds are optimized for user engagement (i.e., they are designed to deliver information most likely to keep users' attention locked in place, usually on the websites and apps that serve as vehicles for revenue-generating advertisements), gripping, emotionally-charged media are prioritized and promoted [62].

Of course, these concerns are not unique to digital social media; many of the same criticisms were directed at television in the late-20th century, and against radio before that [38]. What specifically worries observers about the emergence of these problems in the digital context is the speed at which media can be amplified and the degree to which information flows are personalized for each individual recipient [34]. More, while production of traditional media, such as television and radio, is concentrated in relatively few hands—and is therefore easier to govern—the production of digital media is diffuse, and thus, from a governance perspective, more complicated to manage. These capabilities, of course, are born of advances in AI. The same massive-scale data collection and inference algorithms behind state of the art recommender systems can equally be leveraged for targeted exploitation of cognitive vulnerabilities. Users of social media platforms generate great troves of personal data through shared content, interactions with others' content, and embedded social networks. They do so consistently and over time, furnishing pattern learning technologies with ample opportunity for sophisticated user profiling.[3]

Beyond concerns about the degradation of socially important epistemic values, like accuracy and objectivity, another set of worries has emerged: namely, can these media, designed to capture and hold attention, be leveraged by foreign adversaries to strategically influence targeted populations? These worries also mirror earlier concerns about radio and television propaganda, but they take on special urgency in digital contexts because of the speed and targeting capabilities described above. In addition, these contexts afford malicious actors relative anonymity, in hand with opportunities for disguise as a peer or trusted source. Concern, both amongst researchers and in the public sphere, has grown especially pronounced since 2016, when state actors (e.g., the Russian Internet Research Agency; IRA) appear to have successfully utilized social media channels to influence voter behavior in elections around the world [29]. According to U.S. Senator Mark Warner, D-Va., "We may have in America the best 20th-century military that money can buy, but we're increasingly in a world where cyber vulnerability, misinformation and disinformation may be the tools of conflict [10]."

*Both authors contributed equally to this research.
[1] See also the work of the Oxford Internet Institute's Computational Propaganda Project: https://comprop.oii.ox.ac.uk/.
[2] The locus classicus of this view is [36]. For an argument against, see [64].
[3] For a recent review of Facebook profiling practices, see https://www.pewresearch.org/internet/2019/01/16/facebook-algorithms-and-personal-data.

HotSoS '20, April 7–8, 2020, Lawrence, KS, USA

ACM ISBN 978-1-4503-7561-0/20/04…$15.00
https://doi.org/10.1145/3384217.3385615

In response to these concerns, researchers and policymakers have devoted significant attention to contemplating how some combination of ethical reflection, market pressure, and legal regulation might be brought to bear on social media companies to change the way they organize and distribute information [e.g., 1, 24, 59]. However, less attention has focused on anticipating, evaluating, and systematizing potential intelligence and national security responses to the threats posed by adversaries leveraging advances in data mining and targeting to shape domestic information landscapes. While it is clear that the security community is deeply concerned about these threats and is investing in research to understand and mitigate them[4], we have not yet seen concerted efforts to articulate ethical and policy parameters for the security community's inevitable interventions in this space, or guidance about the form such interventions should take.[5] This is especially urgent because the ethical frameworks traditionally used to guide military and other national security activities may not be suited to navigating these challenges, at least not in their present form.

To explain why, in the next section we describe how the security community understands looming threats in the information landscape as an evolution of "gray zone" conflict—organized, competitive, non-kinetic interactions between states, and between states and non-state actors. In Section 3, we gather information available in the public domain about how the security community plans to respond to these threats, and we describe potential paths they might take. In Section 4, we examine, briefly, traditional frameworks for assessing the morality of war and war fighting, and describe why those frameworks are ill-suited to navigating the ethical questions these new threats pose. Finally, in Section 5, we argue that the notion of "cognitive security" offer a productive frame for articulating and investigating these questions, and thus suggests a promising way forward.

2 INFORMATION OPERATIONS AS A GRAY ZONE

The post-World War II international legal order was built for a state-centric world, in which organized, conventional militaries vie for control over physical territory. It is premised on distinctions between war and peace, foreign and domestic, public and private, combatant and civilian. Today, these foundational premises are on tenuous ground. The United States military/intelligence apparatus increasingly operates in the context of so-called gray zone aggression—competitive interactions among and within state and non-state actors that fall between the traditional war and peace duality [58]. Gray zone tactics can include cyberattacks, economic coercion and sabotage, sponsorship of armed proxy fighters, military expansionism, and hybrids thereof [4]. These types of interactions have existed throughout military history, as state and non-state actors have leveraged economic pressure, propaganda, and espionage to advance their agendas. The Cold War was a gray zone struggle. But recent technological advances have engendered a significant shift toward gray zone aggression as the norm, rather than the exception [33], particularly with respect to cyber and cyber-enabled threats.

A primary theatre for this type of sub-threshold aggression is the information environment. Information operations (IO) describe the range of offensive and defensive military and government strategies and tactics designed to protect and exploit the information environment. Under the broad IO umbrella, psychological operations (PSYOPs) [45] refer specifically to the planned use of information to influence the emotions, motivations, objective reasoning, and ultimately the behavior of foreign individuals, groups, organizations, and governments [54].

The fundamental mission of PSYOPs has remained the same since World War I, when the US waged its first orchestrated military propaganda campaign. The methods, however, have continuously evolved to reflect changes in social, behavioral and political environments, as well as in response to emerging technologies—from leaflet drops in World War I, to radio broadcasts in WWII, to pirated TV broadcasts in Panama in connection with the 1989 overthrow of Manuel Noriega [56]. The Internet and, in particular, social media has served as a force multiplier for PSYOPs, supporting breadth and depth of reach rapidly and cost-effectively. Content generation, targeting, and dissemination is increasingly automated. Once disseminated, curated content is subsumed by networks of real users, who in turn share it forward through their networks of Facebook friends, Twitter followers, and other digitally-mediated audiences [30, 47]. To understand the scope of these strategies, consider testimony by Facebook's General Counsel to the U.S. Senate about IRA operations in 2015-2017:

> We [Facebook] estimate that roughly 29 million people were served content in their News Feeds directly from the IRA's 80,000 posts over the two years [2015-2017]. Posts from these Pages were also shared, liked, and followed by people on Facebook, and, as a result, three times more people may have been exposed to a story that originated from the Russian operation. Our best estimate is that approximately 126 million people may have been served content from a Page associated with the IRA at some point during the two-year period. [49, p. 5]

Importantly, online activity has also furnished a rich source of data about individuals and populations that can be leveraged by adversaries to customize their messaging. Moreover, that information is updated dynamically as individuals interact with media online, yielding feedback to content creators about what works and what doesn't [34, 63]. An Institute of Land Warfare paper argues:

> The ubiquitous nature of electronic platforms provides a direct link, sans geography or security forces, to influence foreign citizens at a massive scale, with feedback—perhaps even through a user's facial expressions—that provides for the most difficult function of information operations, i.e., measures of effectiveness. With each click on a malign meme, the competitor gains cookies, traffic data and a piece of network map to drive further operations. [53, p. 4-5]

[4] See, e.g., DARPA's Media Forensics and Semantic Forensics programs.
[5] The Pentagon recently adopted "ethics principles for using AI in war," an encouraging sign this may be changing [39].

Thus, while Russia's 2016 social media disruption campaign frequently betrayed itself with messages containing broken English or off-topic cultural references [18], operations have grown increasingly sophisticated, with smoother messaging and content better matched to contemporary American political discourse [42].

3 NATIONAL SECURITY RESPONSES TO AUTOMATED INFLUENCE

In 2018, the US Department of Defense (DoD) released its Cyber Strategy [15], representing a significant shift from the 2015 plan it replaced [14]. The new strategy adopts a more focused and assertive posture, highlighting the "persistence" of Chinese and Russian cyber operations, introducing a new mission to "defend forward"—i.e., to disrupt or halt malicious cyber activity at its source—and concluding with a decisive charge to "prepare for war." The document explicitly speaks to the threat of online manipulation, stating: "Russia has used cyber-enabled information operations to influence our population and challenge our democratic processes" (p. 1). A September 2019 report by the RAND Corporation emphasizes the threat of hostile social manipulation with similar urgency, recommending immediate investment in active defense measures [31].

Likewise, in advance of the 2020 U.S. presidential election, the Department of Homeland Security Cybersecurity & Infrastructure Security Agency has launched the #Protect2020 initiative [7]. Working with national partners to build resilience to foreign election interference, this effort places special emphasis on information activities (e.g., countering disinformation). The Agency proposes that responding to foreign interference requires a "whole of society" approach, and in response it has put forth a number of resources aimed at simplifying and taxonimizing the vast array of media-enabled influence strategies [6, 8].

Ultimately, specific details about how the DoD will effectuate active defense against foreign influence operations are—and will likely remain—out of the public domain. There is sufficient information, however, to sketch out plausible categories of AI-driven response in the information space.[6]

3.1 Response 1: Flag and/or contextualize IO

A first set of plausible responses would develop warning systems to flag suspected computational propaganda, with or without details about the nature of or confidence in the suspected offense. A deployment of this type of approach might take the form of a "confidence score" regarding the source or truthfulness of a piece of content, make transparent the content lifespan (a trail back to the originating account), or might provide alternate or additional information and sources germane to the topic. This course would cohere with recommendations of a 2018 Report of the U.S. Senate Committee on Foreign Relations [9] proposing actions targeted at increasing transparency and building a population more *robust to* disinformation and manipulation.

[6]Of course, political or military action outside the information domain (e.g., access to markets) will likely be considered as well, and there is an emerging literature on so-called cross-domain deterrence, exploring the appropriate use of such approaches in today's complex power landscapes (see [28] for overview). Our focus in this paper is on responses *within* the information environment.

The US has engaged in an effort to counter propaganda (specifically Russian propaganda) with Polygraph.info, a government-funded fact-checking website that seeks to provide truthful, fact-based professional journalism in direct response to Russian disinformation. At this time, the effort is exceedingly small (Polygraph.info has a staff of five people) and compartmentalized, issuing ex post corrections dissociated in time and space from the content they address.

In the future, one might imagine more comprehensive efforts in collaboration with industry partners to label, fact-check, or contextualize suspected disinformation in real time, in native environments. Facebook has experimented with some of these approaches with (as yet) mixed results. In 2016, they rolled out a "Disputed" feature to help users identify articles that failed to pass a fact-checking standard, but found it failed to curb sharing and may even have had the reverse effect [26]. In 2017, Facebook instead deployed a "Related Articles" feature to provide additional sources for controversial stories, which has been more successful [46]. Notably, these features have thus far focused on news articles—i.e., "fake news." Whether and how they might be applied more broadly to user-generated content raises both technical and ethical questions, discussed below.

3.2 Response 2: Remove or limit the reach of IO

A second class of responses would take a more hands-on, aggressive stance toward online influence efforts, attempting to prevent them from being seen by target audiences in the first place. These responses would align with the "defend forward" posturing described in the most recent DoD Cyber Strategy.

Facebook and Twitter have systematically removed content associated with known disinformation campaigns linked to Iran, Russia, Venezuela, and others [see, e.g., 27]. Twitter has gone so far as to post these censored Tweets publicly to "enable independent academic research and investigation"[17]. The ex post censorship of high-confidence, malicious, foreign accounts has generally not been contested. However, these efforts have come too late to represent meaningful remediation. Removed content has long receded from view; new accounts have come online focused on the topics of the day. Real-time take-down of suspected disinformation would, in theory, operate within moments of content posting. Ideally, it would do so with high precision and recall. In practice, though, things would likely be messier and there would likely be a trade-off between the rapid removal of suspected content and accuracy in classifying that content as disinformation.

In a similar vein, one might imagine more modest approaches to limit the reach of suspected disinformation, short of outright removal. Specific tactics would likely vary across platforms, but generally, would seek to de-emphasize content with lower credibility, stemming amplification efforts. For example, Facebook might choose to de-prioritize suspicious content for appearance in users' News Feeds. If successful, these types of approaches could have significant impact on the speed and reach of influence campaigns, while affording flexibility in cases where detection algorithms are uncertain. At the same time, more aggressive approaches would raise even more pressing free speech concerns.

3.3 Response 3: Defensive Manipulation

In the face of a sufficiently grave threat one might argue for a response in kind—in this case, counter-messaging aimed at the same target population, designed to undo or mitigate the impact of malign IO. This counter-messaging or counter-manipulation could, in theory, leverage similar clickbait, sensationalism, and emotional charge to meet these ends. Attribution of content—that is, the source of counter-messaging—could be more or less transparent.[7] We refer to this family of potential responses, broadly, as *defensive manipulation*.

The Institute of Land Warfare paper referenced earlier describes an assertive vision that appears to fall into this category. Specifically, it proposes that "the U.S. national security enterprise requires AI-driven influence policy [of its own]" [53, p. 10]—one centered on persistent, wholesale content curation aimed at the generation of viral effects:

> ...competitors must be beaten at scale, with truthful content. Given the current state of technology, the DoD has the necessary data to create a better, more truthful firehose. The U.S. military must curate faster than its opponents can lie, by filtering streaming field footage, creating content out of existing mission command feeds and aiding public affairs' functions with chat-bots. The DoD needs radical transparency, though certainly with a selective eye; existing AI capabilities can provide it. [53, p. 9]

If engaged, this type of action could arguably fall within the scope of efforts envisioned by the Smith-Mundt Modernization Act of 2012 [55], which "authorized the domestic dissemination of information and material about the United States intended primarily for foreign audiences, and for other purposes." In practice, the Act served as an amendment to the Smith-Mundt Act of 1948, lifting a longstanding ban on the domestic dissemination of materials produced by the US State Department and the independent Broadcasting Board of Governors, allowing products like Voice of America, Radio Free Europe, and the Middle East Broadcasting networks to reach Americans [40]. However, the Act also contains an express prohibition on any use of appropriated funds to "influence public opinion in the United States." Further, the Act narrowly applied only to the State Department and the BBG, which both operate under existing checks on their authority aimed to prevent propaganda [43].

4 NAVIGATING RESPONSES WITH AI: CRITICAL CHALLENGES

It is worth underlining that Responses 1 and 2, in particular, would rely on robust partnership with industry actors, and likely, buy-in from users themselves. While analyses of these relationships is out of scope for this paper, the technical and ethical challenges we explore have been selected with an eye toward these concerns.

In-practice detection. All three responses outlined above rely on fair, explainable, highly accurate, automated approaches to disinformation detection—an ask that is currently out of reach. "Fake news" detection is an emerging area of active of research[8], but current tools still suffer from relatively low accuracy, work best in circumscribed contexts, and critically, are focused on separating "true" from "false" content, rather than understanding its origin. Not all content disseminated by malign foreign actors is, or needs to be, categorically false. While, not all content generated and spread by legitimate users is true. Indeed, a key strategy of the Internet Research Agency during 2016 elections was to seek out and amplify native content deemed to meet their objectives [21].

The reverse scenario is also cause for concern: in addition to adversaries amplifying user-generated content that serves their ends, they also generate content that users themselves amplify. Such cases raise questions about the timeline for and nature of flagging or removing content. For example, if a user re-shares content that a defensive AI later tags as disinformation, should the suspected malicious content be addressed only at the point of origin, or at the point of sharing as well?

Persistent surveillance. Another set of technical and ethical questions raised by likely future real-time approaches to disinformation detection have to do with persistent, comprehensive surveillance of the social media landscape and massive-scale data mining. Perhaps, aspirationally, precautions could be put into place that enable monitoring individual data without storing it. However, given the complex nature of the task, detection algorithms would likely not operate effectively on individual pieces of posted content alone. Rather, they will need to consider user history, metadata, embedded social networks, and related contexts–information which would need to be in memory. Dual use of such tools, security vulnerabilities, and privacy concerns will all be at stake.

Contextualization vs Influence. In our discussion of Response 1 (flag/contextualize IO), we pointed to Facebook's "Related Articles" feature as a positive example. However, there is important nuance to consider, around the choice of context to provide, and the specific way in which to provide it—in particular, if the selection of contextualizing content is automated. One might imagine the slippery slope from the first (or second) class of responses to the third, whereby adding context or strategically emphasizing or de-emphasizing particular content effectively becomes an instance of defensive manipulation. To varying degrees, all three responses thus raise related, urgent questions about AI-driven influence policy. How would such an AI be trained? What would it optimize for? Who would decide?

A speculative 2017 essay published by the Atlantic Council brings a worst case scenario to life in an alarming vignette titled the "influence machine," outlining a dystopian future of global information warfare [5]. AI systems generating computational propaganda dominate conversations online, and the information environment devolves into a morass of competing, state-sponsored, manipulative, machine-driven speech. The author cautions:

[7] An example of defensive manipulation might be found in the well-known case of YouTube's "Redirect Method," which sends users searching for extremist/terrorist propaganda to videos that aim to debunk such messages and de-radicalize viewers [22].

[8] See, e.g., https://www.aclweb.org/anthology/C18-1287/ and http://www.fakenewschallenge.org.

> The community of democracies must recognize the serious threats posed by [AI-driven conversation tools], computational propaganda, and weaponized narratives. Democracies must move aggressively to address these threats on multiple fronts, by crafting comprehensive strategies to protect their populations from online propaganda and disinformation, while maintaining the core democratic values of equality and liberty. (p. 2)

With these possible (if hypothetical) futures in mind, we turn now to the values at stake in the decisions the security community—and the public governing it—will soon have to make. How *ought* they respond to threats of online influence? What are the ethical ramifications of each approach? When, if ever, is any approach appropriate?

5 THE ETHICS OF WAR

Traditionally, philosophical reflection on the ethics of military engagement has navigated a course between "realism" on one hand and "pacifism" on the other. Realists hold that war stands completely apart from the rest of human affairs, such that there is no sense in discussing the right and wrong of war; war just *is*. As Michael Walzer describes the realist position: "War is a world apart, where life itself is at stake, where human nature is reduced to its elemental forms, where self-interest and necessity prevail. Here men and women do what they must to save themselves and their communities, and morality and law have no place" [61, p. 3]. At the other end of the spectrum, pacifists argue that all war is unethical, that moral analysis cannot rationalize the human costs of war. As Cheyney Ryan puts it, pacifists "*unconditionally* oppose war" [44, p. 2]. The first extreme excludes ethics from the realm of war; the second excludes war from the realm of ethics. If there is an "ethics of war," then, it lay not at either extreme but rather somewhere along the spectrum.

Just War Theory is the predominant effort to articulate such a middle position. It asks: (1) what are the conditions that justify waging war (*jus ad bellum*), and (2) what are the principles of right conduct in war (*jus in bello*) [25]. Though there is considerable debate amongst scholars about the details, the terms of the debate are more or less settled. Just war is *necessary* (i.e., war must always be the last resort) and *proportional* (the good war achieves must outweigh the harm it causes).[9] Likewise, conduct in war is just so long as it necessary, proportional, and it *discriminates* between civilians and combatants (Ibid.).

Gray zone conflict complicates this picture, as it challenges a number of assumptions about war that are foundational to determining whether any particular war is necessary and proportional, and whether conduct in war is necessary, proportional, and discriminating [12]. First, information operations test the boundaries of the concept of war itself, calling into question the basic distinction between war and peace. The kind of war at the center of Just War Theory is *bloody* war—attacks that involve maiming and killing. IO, by contrast, rarely causes harms that grave, leading some to describe it as a form of "soft war" [19]. As Mariarosaria Taddeo argues, the necessity condition—the condition that war only ever be waged as a last resort—is premised on the enormity of war's harms; absent such harms, it is not clear that the necessity condition holds [51, 52].

Second, IO raises questions about the balancing of good and harm required by the proportionality principle, for the same reason it challenges the necessity principle: the harms imagined by Just War Theory are not (generally) the harms wrought by gray zone or "soft war" tactics [20]. The moral dilemma Just War Theory attempts to resolve arises in the first place because the harms of war are, to use Taddeo's language, *"universal"* harms—i.e., extensive loss of life [52, p. 218]. Because gray zone tactics are non-violent and thus cause far less serious harms, it is possible (assuming war is motivated by a just cause) that on a Just War Theory analysis gray zone conflict is always justified, despite the fact that gray zone or soft war tactics may nevertheless have significant and intuitively unethical consequences (Ibid.). In other words, gray zone conflict threatens to tilt the moral balance of Just War Theory, from an ethical framework that hardly ever sanctions war (i.e., war as a last resort) to one that nearly always sanctions it.

Third, IO complicates the distinction between combatants and civilians/non-combatants. As discussed in Section 2, the "information environment"—the "battlefield" for information operations—permeates civil society. And though propaganda efforts and other PSYOPs and IO operations have always targeted the "hearts and minds" of civilians, today's pervasive information and communications technologies (ICTs) open up vast new possibilities for gray zone efforts. "As the means and opportunities for propaganda continue to expand with technological advances," writes Laurie Blank, "so do the targets and goals of that propaganda" [3, p. 89]. Furthermore, as Taddeo points out, the problem is not simply that it becomes easier (and perhaps more justifiable) to target civilians in soft war contexts; it also becomes easier for civilians to actively participate in these activities. "Civilians may take part in a combat action from the comfort of their homes, while carrying on with their civilian life and hiding their status as informational warriors" [52, p. 218].

Finally, an important implication of the "battlefield" permeating civil society, one which has gone largely unnoticed in the existing literature, is the degree to which U.S. intelligence and national security organizations will have to engage with *us*—with the information U.S. citizens produce, seek out, and consume—in order to mitigate the effects of foreign influence operations. As the discussion in the previous section suggests, it is easy to imagine the security community involving itself—"actively" and "persistently"—in our own domestic information environment. Defensive operations of that kind raise urgent ethical questions about when, how, and to what extent the military ought to engage in influence operations (or counter-influence operations) aimed inward at the populations they aim to defend.

Just War Theory, the ethical framework normally used to guide and evaluate efforts by the military and national security apparatus, appears, then, to offer little guidance here. In what remains, we suggest that examining these questions through the lens of "cognitive security" may provide a (partial) path forward.

[9] Just War theorists often stipulate other *ad bellum* principles, such as just wars being motivated by just causes and having reasonable prospects for success. As Lazar argues, however, necessity and proportionality are the only strictly necessary conditions for just war (even at the same time as they are not—even jointly—sufficient) [25].

6 TOWARD AN ETHICS OF COGNITIVE SECURITY

There is much work to be done adapting ethical frameworks, such as Just War Theory, to the normative challenges of gray zone conflict, and we can only tackle a small part of that task here. Importantly, we put aside questions about how to evaluate the ethical stakes of offensive IO—i.e., if and when it is justified to deploy influence operations against foreign populations. With respect to *defensive* operations, however—e.g., which, if any, of the three responses to foreign influence campaigns, outlined above, the military and national security apparatus ought to adopt in cases where IO is targeted at US citizens—progress might be made by casting the question in terms of *cognitive security.*

To our knowledge, the term cognitive security was introduced by Rand Waltzman—an analyst at RAND Corporation—first in a 2015 report[10] and then in testimony before the United States Senate Armed Services Committee's Subcommittee on Cybersecurity in 2017 [60].[11] In his testimony, Waltzman makes an urgent case that our reliance on internet platforms for information and communication is making us increasingly vulnerable to outside influence [60]. The internet and social media, he argues, have created "a qualitatively new landscape of influence operations, persuasion, and, more generally, mass manipulation" [60]. As our interactions with and reliance on these technologies evolve, "old models are becoming irrelevant faster than we can develop new ones. The result is uncertainty that leaves us exposed to dangerous influences without proper defenses" [60]. Collectively, Waltzman refers to these vulnerabilities as challenges to cognitive security.

While Waltzman's rhetoric is often dramatic (e.g., "Today, thanks to the Internet and social media, the manipulation of our perception of the world is taking place on previously unimaginable scales of time, space and intentionality" [60]), his assessment parallels concerns emerging from many quarters (discussed in Sections 1 and 2, above). In order to respond to these challenges, Waltzman envisions a "whole of nation approach": government, academia, and industry engaged in a continual arms race to influence and protect from influence large groups of users online. Driving these efforts, he advocates "a new field of cognitive security," related to but distinct from existing cybersecurity programs. Notably, he suggests that:

> Although COGSEC [cognitive security] emerges from social engineering and discussions of social deception in the computer security space, it differs in a number of important respects. First, whereas the focus in computer security is on the influence of a few individuals, COGSEC focuses on the exploitation of cognitive biases in large public groups. Second, while computer security focuses on deception as a means of compromising computer systems, COGSEC focuses on social influence as an end unto itself. Finally, COGSEC emphasizes formality and quantitative measurement, as distinct from the more qualitative discussions of social engineering in computer security.

[10]Report cited in Waltzman (2017).

[11]It is worth noting that the term "cognitive security" is sometimes used in a very different way—namely, as a marketing term for cybersecurity products driven by artificial intelligence. See, for example, https://www.ibm.com/security/artificial-intelligence.

The specific details of Waltzman's proposal are less important for our purposes than the framing. In an important 2005 essay, philosopher Helen Nissenbaum examines differences between what she terms, on one hand, "technical computer security," and on the other hand, then-emerging discussions about "cybersecurity" [35].[12] For Nissenbaum, the differences between these two terms have less to do with the techniques they involve (they largely overlap) than with the worries that motivate them. Understanding those differences is key to determining what strategies (technical and otherwise) are justified in their name.

Security is safety from harm, Nissenbaum argues, and the harms imagined by proponents of technical computer security differ from the harms imagined by proponents of cybersecurity. Technical computer security is concerned with protecting information systems and their users from a particular set of harms, usually described in terms of "availability, integrity, and confidentiality" [35]. Which is to say, the techniques of computer security are designed to ensure that information is available when users want it, that the information is uncorrupted, and that it is only available to those authorized to access it. By contrast, cybersecurity is broader in scope. While securing information systems is part of cybersecurity's mandate, the harms cybersecurity advocates imagine are graver: they are not merely threats to individuals, but rather to society at large. For example, cybersecurity efforts aim to defend against:

> "Threats posed by the use of networked computers as a medium or staging ground for antisocial, disruptive, or dangerous organizations and communications; [...] Threats of attack on critical societal infrastructure, including utilities, banking, government administration, education, healthcare, manufacturing, and communications media; [...and] Threats to the networked information system itself ranging from disablement of various kinds and degrees to—in the worst case—complete debility." [35]

The "moral force" behind arguments in favor of taking drastic action to advance cybersecurity derive from the severity of the threats it is assumed to protect us from. And Nissenbaum argues that the threats around which cybersecurity is oriented are not simply larger threats than those technical computer security focuses on (in the sense that the former have to do with threats to society, while the latter has to do with threats to individuals); cybersecurity threats are cast in *existential* terms. Following the terrorist attacks of September 11th 2001 in the United States, the need for cybersecurity was understood as essential to protecting the very existence of the American Republic [35]. Ethically, far more drastic—and perhaps invasive—actions are warranted to protect society from an existential threat than are warranted to protect individuals from less dire ones. For example, as is now well-known, following 9/11 the national security apparatus put into place a dragnet surveillance program, which tracked nearly all phone calls and text messages sent and received through major US carriers. It is arguable whether or not such a program was justified (we do not think it was), but whatever one's position, Nissenbaum's argument highlights the

[12]Nissenbaum renders the term "cyber-security," with a hyphen. We adopt today's more common, non-hyphenated formulation.

fact that *it is only possible to justify* such a program in the name of securing Americans against a truly existential threat.

With this as background, we can ask the following: what are the harms cognitive security aims to protect us from, how dire are they (are they existential?), and consequently, what kinds of responses are warranted? Such questions are, at least in part, empirical, and we are not prepared to answer them here. We suggest, however, that the security community must become prepared to answer them, as the kinds of issues discussed throughout this paper become part of its purview. If some—like Rand Waltzman—are to be believed, the threats to cognitive security posed by digital influence campaigns are, indeed, existential. That is why, to their minds, actions such as those outlined in "Response 3," above, may be warranted. In our view, the case is less clear. What is clear is that the bar for such justifications is very high.

Furthermore, the issue is not only the *severity* of the harms cognitive security promises to protect us from; questions remain about the *nature* of these harms too. As we've seen, the harms of war that ethicists have traditionally contemplated differ in kind from those at stake here. Information operations do not threaten grievous injury and death, at least not directly. Rather, they threaten informational harms—harms scholars working in other domains have analyzed carefully. Meeting the ethical challenges of combating automated influence therefore requires integrating insights from privacy and information ethics, AI ethics, and the ethics of influence into existing cyberwar ethics frameworks. While pursuing that project in earnest is outside the scope of this paper, we hope our discussion will draw attention from scholars working in these spaces to the places where they overlap.

6.1 Privacy and Information Ethics

Although much remains unknown about how the security community is preparing to counter emerging threats in the information environment, one thing is clear: responding to the threat of automated influence requires collecting, storing, and processing vast amounts of data about civilians, including U.S. citizens. The Snowden revelations demonstrated that, historically, the intelligence and national security communities have subordinated privacy concerns to issues of national security, justifying widespread, indiscriminate surveillance in the name of national defense. If military conflict continues to evolve in the direction of increasing engagement in the information space, questions will be raised anew about how to balance these values.

As these issues develop, lessons from the vast literatures on information privacy and information ethics will be invaluable. For example, privacy scholars have long argued—contrary to prevailing opinion—that privacy and security are not necessarily competing values, and that treating them as such often produces a false binary for policymakers [48]. When considering the informational harms *beyond* privacy harms that IO and cyber conflict threaten, the information ethics literature can provide useful insights as well. Taddeo has begun the work of extending Just War Theory frameworks to more robustly contemplate informational harms [52]. Discussions about war ethics and discussions about privacy and information ethics would be much enriched by further efforts in this vein.

6.2 Ethics of Algorithms/AI

The same is true with respect to the emerging literature on the ethics of algorithms/AI ethics. Many of the ethical issues raised by the incorporation of AI-driven technologies into military conflict mirror issues raised by the use of AI in civil society contexts, and the rapidly developing discussions around AI ethics and policy in the latter context can offer guidance for the former. For example, the question of how to attribute causal and moral responsibility for actions carried out by machines in military conflict—a longstanding subject of debate in cyberwar ethics [e.g., 11, 13]—could benefit from related discussions in the ethics of algorithms and AI [for an overview, see 32]. A related, but distinct issue, which has been the subject of significant technical, policy, and ethics research around algorithms and AI is the degree to which the operations of automated systems can be made intelligible to human users [e.g., 37]. If AI-driven systems are used to counter IO in the ways described above—by flagging or contextualizing suspected disinformation, removing or suppressing it, etc.—it will be relevant to discussions about the ethics of such strategies whether or not the users generating and consuming affected content can be made to understand how it is being influenced. There are no doubt many more examples of overlapping issues—e.g., discussions about algorithmic bias [e.g., 2, 16] and trust in automated systems [e.g., 41]—which would mutually benefit from dialogue between AI ethics scholars and military ethics scholars.

6.3 Ethics of Influence

Finally, the issues raised by the preceding discussion point to the special relevance of the ethics of influence to emerging debates in the ethics of military and gray zone conflict. As discussed, AI ethics and policy scholars have begun to address concerns about the use of algorithms to curate, personalize, and target information at individuals—to influence, even manipulate them—and have raised important ethics and policy questions about these practices. If war ethics scholars aim to offer meaningful guidance to the security community about how to respond ethically to the emerging information threats described above, this nascent literature is an indispensable starting point for understanding them. At the same time, while AI ethics and policy scholars often point to uses of these technologies by adversaries, they have not addressed head-on the complex ethical issues destined to emerge from attempts by the intelligence and security communities to mitigate them. Thus if scholars working in these fields hope to fully grapple with the challenges of automated influence, they have as much to gain as they have to offer by engaging with the ethics of war.

7 CONCLUSION

Few issues in recent years have captured public attention like the threat of foreign influence campaigns. While technology ethics, law, and policy scholars have explored these problems, the conditions that enable them, and strategies for utilizing norms, laws, code, and market forces to solve them, too few scholars have trained their sights on the inevitable entrance of the intelligence and national security apparatus into the fray, and on its ethical consequences. Our aim in this paper has been to motivate such efforts by describing what is known, publicly, about how the security community is

thinking about and planning to engage, to discuss why existing military ethics frameworks are unsuited to the task of evaluating such engagement, and to suggest a path forward by casting these questions in terms of cognitive security. The security community must now decide how to incorporate the challenges of cognitive security into its work.

REFERENCES

[1] Mike Ananny. 2016. Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness. *Science, Technology, Human Values* 41, 1 (Jan 2016), 93–117. https://doi.org/10.1177/0162243915606523

[2] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104 (2016), 671–732. https://doi.org/10.2139/ssrn.2477899

[3] Laurie R. Blank. 2017. *Media Warfare, Propaganda, and the Law of War.* Cambridge University Press, 88–103. https://doi.org/10.1017/9781316450802.009

[4] Hal Brands. 2016. *Paradoxes of the Gray Zone.* Technical Report. Foreign Policy Research Institute.

[5] Matt Chessen. 2017. *The MADCOM Future.* Technical Report. The Atlantic Council. http://www.atlanticcouncil.org/images/publications/The_MADCOM_Future_RW_0926.pdf

[6] CISA. 2018. *Foreign Influence Taxonomy.* Technical Report. US Cybersecurity and Infrastructure Security Agency. https://www.dhs.gov/sites/default/files/publications/19_0717_cisa_foreign-influence-taxonomy.pdf

[7] CISA. 2019. *The #Protect2020 Initiative.* Technical Report. US Cybersecurity and Infrastructure Security Agency. https://www.dhs.gov/cisa/protect2020

[8] CISA. 2019. *The War on Pineapple.* Technical Report. US Cybersecurity and Infrastructure Security Agency. https://www.dhs.gov/sites/default/files/publications/19_0717_cisa_the-war-on-pineapple-understanding-foreign-interference-in-5-steps_0.pdf

[9] CoFR. 2018. *Putin's Asymmetric Assault on Democracy in Russia and Europe: Implications for U.S. National Security.* Technical Report. US Senate Committee on Foreign Relations. https://www.foreign.senate.gov/imo/media/doc/SPrt_115-21.pdf

[10] Jack Corrigan. 2017. Social media is 'First Tool' of 21st-Century Warfare, Lawmaker Says. *Nextgov.com* (Sep 2017). https://www.nextgov.com/cio-briefing/2017/09/social-media-first-tool-21st-century-warfare-lawmaker-says/141379/

[11] David Danks and Joseph H. Danks. 2017. *Beyond Machines: Humans in Cyber-operations, Espionage, and Conflict.* Oxford University Press, 177–198. https://doi.org/10.1093/acprof:oso/9780190221072.003.0010

[12] Randall R. Dipert. 2010. The Ethics of Cyberwarfare. *Journal of Military Ethics* 9, 4 (Dec. 2010), 384–410. https://doi.org/10.1080/15027570.2010.536404

[13] Randall R. Dipert. 2016. *Distinctive Ethical Issues of Cyberwarfare.* Oxford University Press, 56–72. https://doi.org/10.1093/acprof:oso/9780190221072.003.0004

[14] DoD. 2015. *Cyber Strategy.* Technical Report. US DoD. http://www.defense.gov/home/features/2015/0415_cyber-strategy/Final_2015_DoD_CYBER_STRATEGY_for_web.pdf

[15] DoD. 2018. *Cyber Strategy.* Technical Report. US DoD. https://media.defense.gov/2018/Sep/18/2002041658/-1/-1/1/CYBER_STRATEGY_SUMMARY_FINAL.PDF

[16] Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. *ACM Transactions on Information Systems* 14, 3 (Jul 1996), 330–347. https://doi.org/10.1145/230538.230561

[17] Vijay Gadde and Yoel Roth. 2018. Enabling further research of information operations on Twitter. https://blog.twitter.com/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter.html

[18] Adam Goldman. 2018. *Justice Dept. Accuses Russians of Interfering in Midterm Elections.* https://www.nytimes.com/2018/10/19/us/politics/russia-interference-midterm-elections.html?module=inline

[19] Michael Gross and Tamar Meisels. 2017. Introduction. In *Soft War: The Ethics of Unarmed Conflict.* Cambridge University Press.

[20] Michael L. Gross. 2017. *Proportionate Self-Defense in Unarmed Conflict.* Cambridge University Press, 217–232. https://doi.org/10.1017/9781316450802.017

[21] Matthew Hindman and Vlad Barash. 2018. *Disinformation, "Fake News" and Influence Campaigns on Twitter.* 62 pages. https://kf-site-production.s3.amazonaws.com/media_elements/files/000/000/238/original/KF-DisinformationReport-final2.pdf

[22] Peter Holley. 2017. YouTube is tricking people who search for ISIS videos. https://www.washingtonpost.com/news/innovations/wp/2017/07/24/youtube-is-tricking-people-who-search-for-isis-videos/

[23] Charlotte Jee. 2019. Twitter and Facebook won't remove false Trump campaign ads about Biden. *MIT Technology Review* (Oct 2019). https://www.technologyreview.com/f/614549/twitter-and-facebook-wont-remove-false-trump-campaign-ads-about-biden/

[24] Kate Klonick. 2018. The New Governors: The People, Rules, and Processes Governing Online Speech. *Harvard Law Review* 131 (2018), 73.

[25] Seth Lazar. 2017. War. *Stanford Encyclopedia of Philosophy* (2017). https://plato.stanford.edu/archives/spr2017/entries/war/

[26] Kalev Leetaru. 2017. The backfire effect and why Facebook's 'Fake News' warning gets it all wrong. https://www.forbes.com/sites/kalevleetaru/2017/03/23/the-backfire-effect-and-why-facebooks-fake-news-warning-gets-it-all-wrong/

[27] Cristiano Lima. 2019. Facebook, Twitter take down disinformation campaigns linked to Iran, Russia, Venezuela. https://politi.co/2G1gNQG

[28] Jon R. Lindsay and Erik Gertzke (Eds.). 2019. *Cross-Domain Deterrence.* Oxford University press.

[29] Ryan Lucas. 2018. New Reports Detail Expansive Russia Disinformation Scheme Targeting U.S. (Dec 2018). https://www.npr.org/2018/12/17/677390345/new-reports-detail-expansive-russia-disinformation-scheme-targeting-u-s

[30] Alice Marwick and Rebecca Lewis. 2017. *Media Manipulation and Disinformation Online.* 106 pages.

[31] M. Mazarr, A. Casey, A. Demus, S. Harold, L. Matthews, N. Beauchamp-Mustafaga, and J. Sladden. 2019. Hostile Social Manipulation: Present Realities and Emerging Trends. (2019). https://www.rand.org/pubs/research_reports/RR2713.html

[32] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The Ethics of Algorithms. *Big Data Society* 3, 2 (Dec 2016).

[33] MWI. 2018. *Rule of Law in the Gray Zone.* Technical Report. Modern War Institute. https://mwi.usma.edu/rule-law-gray-zone/

[34] Anthony Nadler, Matthew Crain, and Joan Donovan. 2018. *Weaponizing the Digital Influence Machine: The Political Perils of Online Ad Tech.* 47 pages. https://datasociety.net/wp-content/uploads/2018/10/DS_Digital_Influence_Machine.pdf

[35] Helen Nissenbaum. 2005. Where Computer Security Meets National Security. *Ethics and Information Technology* 7, 2 (Jun 2005), 61–73. https://doi.org/10.1007/s10676-005-4582-3

[36] Eli Pariser. 2011. *The Filter Bubble: What the Internet Is Hiding from You.* Penguin Press. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1118322

[37] Frank Pasquale. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information.* Harvard University Press.

[38] Neil Postman. 2006. *Amusing Ourselves to Death: Public Discourse in the Age of Show Business* (20th anniversary ed ed.). Penguin Books.

[39] Associated Press. 2020. U.S. military adopts new ethics principles for using AI in war. https://www.latimes.com/world-nation/story/2020-02-24/pentagon-adopts-new-ethical-principles-for-using-ai-in-war

[40] Elspeth Reeve. 2013. Americans finally have access to American propaganda. https://www.theatlantic.com/politics/archive/2013/07/americans-finally-have-access-american-propaganda/313305/

[41] Heather M. Roff and David Danks. 2018. "Trust but Verify": The Difficulty of Trusting Autonomous Weapons Systems. *Journal of Military Ethics* 17, 1 (Jan 2018), 2–20. https://doi.org/10.1080/15027570.2018.1481907

[42] Kevin Roose. 2018. *Facebook Grapples With a Maturing Adversary in Election Meddling.* https://www.nytimes.com/2018/08/01/technology/facebook-trolls-midterm-elections.html?module=inline

[43] Gabe Rottman. 2012. New government "propaganda" bill a positive step for First Amendment. https://www.aclu.org/blog/free-speech/new-government-propaganda-bill-positive-step-first-amendment

[44] Cheyney Ryan. 2016. Pacifism. In *The Oxford Handbook of Ethics of War*, Seth Lazar and Helen Frowe (Eds.). Oxford University Press, Oxford. https://doi.org/10.1093/oxfordhb/9780199943418.013.21

[45] Peter Schoomaker. 2005. *Field Manual No. 3-05.30: Psychological Operations.* Technical Report. US Army. https://fas.org/irp/doddir/army/fm3-05-30.pdf

[46] Deepa Seetharaman. 2017. Facebook drowns out faek news with more information. https://www.wsj.com/articles/facebook-drowns-out-fake-news-with-more-information-1501754403

[47] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature Communications* 9, 1 (Nov 2018). https://doi.org/10.1038/s41467-018-06930-7

[48] Daniel J. Solove. 2011. *Nothing to Hide: The False Tradeoff Between Privacy and Security.* Yale University Press.

[49] Colin Stretch. 2017. Social Media Influence in the 2016 U.S. Election. Hearing before the United States Senate, Committee on the Judiciary, Subcommittee on Crime and Terrorism.

[50] Daniel Susser, Beate Roessler, and Helen Nissenbaum. 2018. Online Manipulation: Hidden Influences in a Digital World. *SSRN* (2018). https://papers.ssrn.com/abstract=3306006

[51] Mariarosaria Taddeo. 2012. Information Warfare: A Philosophical Perspective. *Philosophy & Technology* 25, 1 (March 2012), 105–120. https://doi.org/10.1007/s13347-011-0040-9

[52] Mariarosaria Taddeo. 2016. Just Information Warfare. *Topoi* 35, 1 (April 2016), 213–224. https://doi.org/10.1007/s11245-014-9245-8

[53] Christopher Telley. 2018. *The Influence Machine: Automated Information Operations as a Strategic Defeat Mechanism.* Technical Report 121. The Institute of Land Warfare, Arlington, VA. 23 pages.

[54] Catherine Theohary. 2018. Information Warfare: Issues for Congress. https://fas.org/sgp/crs/natsec/R45142.pdf
[55] Mac Thornberry. 2012. The Smith-Mundt Modernization Act of 2012 (H.R. 5736). https://www.govinfo.gov/content/pkg/BILLS-112hr5736ih/pdf/BILLS-112hr5736ih.pdf
[56] Jared Tracy. 2018. 100 years of subterfuge: the history of Army psychological operations. https://www.army.mil/article/199431/100_years_of_subterfuge_the_history_of_army_psychological_operations
[57] Zeynep Tufekci. 2014. Engineering the Public: Big Data, Surveillance and Computational Politics. *First Monday* 19, 7 (Jul 2014). http://firstmonday.org/ojs/index.php/fm/article/view/4901
[58] USSOCOM. 2018. *Defining Gray Zone Challenges.* Technical Report. US Special Operations Command.
[59] Siva Vaidhyanathan. 2018. *Antisocial Media: How Facebook Disconnects Us and Undermines Democracy.* Oxford University Press.
[60] Rand Waltzman. April 2017. The Weaponization of Information: The Need for Cognitive Security.
[61] Michael Walzer. 2006. *Just and Unjust Wars: A Moral Argument with Historical Illustrations* (4th ed ed.). Basic Books, New York. OCLC: ocm71165547.
[62] Tim Wu. 2016. *The Attention Merchants: The Epic Scramble to Get Inside Our Heads* (first edition ed.). Alfred A. Knopf.
[63] Karen Yeung. 2017. Hypernudge: Big Data as a Mode of Regulation by Design. *Information, Communication Society* 20, 1 (Jan 2017), 118–136. https://doi.org/10.1080/1369118X.2016.1186713
[64] Frederik J. Zuiderveen Borgesius, Damian Trilling, Judith Möller, Balázs Bodó, Claes H. De Vreese, and Natali Helberger. 2016. Should We Worry About Filter Bubbles? *Internet Policy Review* (2016). https://doi.org/10.14763/2016.1.401