Toward an Ethical Framework for Countering Extremist Propaganda Online

Henschke, Adam; Reed, Alastair

DOI
10.1080/1057610X.2020.1866744

Publication date
2021

Document Version
Final published version

Published in
Studies in Conflict and Terrorism

Studies in Conflict & Terrorism

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uter20

Toward an Ethical Framework for Countering Extremist Propaganda Online

Adam Henschke & Alastair Reed

To cite this article: Adam Henschke & Alastair Reed (2021): Toward an Ethical Framework for Countering Extremist Propaganda Online, Studies in Conflict & Terrorism, DOI: 10.1080/1057610X.2020.1866744

To link to this article: https://doi.org/10.1080/1057610X.2020.1866744

STUDIES IN CONFLICT & TERRORISM
https://doi.org/10.1080/1057610X.2020.1866744

Routledge
Taylor & Francis Group

Toward an Ethical Framework for Countering Extremist Propaganda Online

Adam Henschke[a] and Alastair Reed[b,c]

[a]National Security College, The Australian National University, Canberra, Australia; [b]Hilary Rodham School of Law, Swansea University, Swansea, UK; [c]Delft University of Technology, Delft, Netherlands

ABSTRACT

In recent years, extremists have increasingly turned to online spaces to distribute propaganda and as a recruitment tool. While there is a clear need for governments and social media companies to respond to these efforts, such responses also bring with them a set of ethical challenges. This paper provides an ethical analysis of key policy responses to online extremist propaganda. It identifies the ethical challenges faced by policy responses and details the ethical foundations on which such policies can potentially be justified in a modern liberal democracy. We also offer an ethical framework in which policy responses to online extremism in liberal democracies can be grounded, setting clear parameters upon which future policies can be built in a fast-changing online environment.

In 2005, Al Qaida leader Ayman al-Zawahiri told Abu Musab al-Zarqawi, the erstwhile commander of Al Qaida in Iraq, that: "we are in a battle, and that more than half of this battle is taking place in the battlefield of the media. We are in a media battle in a race for the hearts and minds of our Umma."[1] Following the dramatic rise of the so-called Islamic State (IS), and its use of social media to publicize and achieve its goals, policymakers have become increasingly aware of the threat al-Zawahiri alluded to in 2005 – extremists use of online propaganda.[2] Whilst the use of propaganda by extremist groups is nothing new, IS's exploitation of online technology in its media strategy presents policymakers with new challenges. Countering extremist propaganda online has become a priority in counterterrorism and countering violent extremism (CVE) policy.

There has been an explosion of policy responses from governments around the world to tackle the threat posed by extremist propaganda online. In the face of this somewhat rapid roll out of counter-propaganda policies, there has been increasing concern in both the public and political sphere of the ethical implications of such approaches. While upholding individual freedoms is central to modern liberal democracies, democracies place limits on these freedoms in the interest of the wider public.[3] We offer an ethical framework here for these three reasons. First, given the potential negative impacts, rights violations and government overreach that can occur when responding to such propaganda, we need to make sure that such responses have good reasons justifying

CONTACT Adam Henschke adam.henschke@anu.edu.au Crawford School of Public Policy, ANU College of Asia and the Pacific, J. G. Crawford Building, 132 Lennox Crossing, Acton ACT 2601, Australia

them. Second, careless and poorly thought out responses can play a part in ongoing extremist narratives about the lack of values in the given community or nation and can be used to try to justify the extremist's positions. Finally, without an ethically informed response, governments and other relevant decision makers can fall prey to ad hoc decision making which is, in fact, biased, unfair, unjustified or simply inexplicable. In short, ethical reflection makes for better policymaking. The issue we seek to address in this paper is, in countering extremist[4] propaganda online, where do we ethically place limitations on our liberal democratic freedoms? How can we ensure effective policy responses to online extremist propaganda that, at the same time, sit within our values as liberal democracies?[5] We note here the legitimate concerns with focusing simply on *online* extremism.[6] In reality, both online and offline interactions combine to influence the radicalization process. As Brown and Cowls note, "Real-world connections and experiences and peer groups seem to be most important in introducing individuals to extremist ideologies, although the Internet can act as an 'echo chamber' to confirm existing beliefs."[7]

This paper provides an in-depth ethical analysis of key policy responses to online extremist propaganda. This analysis will identify the ethical challenges faced by policy responses and detail the ethical foundations on which such policies can be justified (if they can be) in a modern liberal democracy. [8] This paper sets out an ethical framework in which policy responses to online extremism in liberal democracies can be grounded, setting clear parameters upon which future policies can be built in a fast-changing online environment.

Section 1: Existing Responses to Extremist Propaganda Online

Overview of Responses to Online Extremist Propaganda

Five key responses have emerged amidst growing concerns over the role of online extremist propaganda: i) Disruption, ii) Redirection, iii) Counter-Messaging, iv) One-to-One Dialogue, and v) Education and Media Literacy. While these responses comprise key lines of effort to address online extremist propaganda, they are not without ethical implications and considerations.

Disruption

The objective of "disruption" as a response to extremist propaganda is to reduce the supply of such propaganda. In essence, disruptive approaches seek to tackle propaganda at its source by reducing or eliminating its availability and, therefore, its potential impact in the process. There are two main approaches to disruption: the removal of content, and the hiding/filtering of content to limit/prevent access.

The most prominent example of content disruption is the work of Europol's Internet Referral Unit (IRU). The IRU seeks to reduce "the level and impact of terrorist and violent extremist propaganda on the internet… [identifying] and refer[ring] relevant online content towards concerned internet service providers and support[ing] member states with operational and strategic analysis."[9] The IRU monitors and identifies extremist content online, and shares this information with partners organizations such as

internet providers and social media companies, for them to remove the content from their own platforms themselves. The IRU does not take any enforcement action themselves, they identify and bring this to the attention of service providers, highlighting how this content violates the service providers own terms of service. This is on top of what the service providers identify and take down themselves.[10]

Another way of reducing the supply of extremist content online, rather than removing the content, is preventing access to it in the first place online. Peter Neumann highlights two ways in which states would be capable of this. The first is nationwide filtering of internet traffic, possible as most internet traffic flows through a limited number of Internet Service Providers (ISPs). In the case of case in China and Saudi Arabia, all internet users are only able to access the internet via government controlled ISPs, which filter according to government policy.[11] The second, is "by manipulating search results or deleting recommended links or suggestions for websites and videos that are known to promote terrorism or hate speech."[12]

Furthermore, there are subtler disruptive techniques that do not remove online content, but make it more difficult to find. In response to criticism of the presence of videos on Google/YouTube that were deemed offensive, but did not necessarily violate their policies, the company introduced new steps that they explained would ensure the videos "appear behind an interstitial warning and they will not be monetized, recommended or eligible for comments or user endorsements. That means these videos will have less engagement and be harder to find."[13]

Redirection

Many actors consider the removal of all extremist material from the internet as too blunt an instrument for both practical and ethical reasons. The "redirect method" piloted by Jigsaw takes a different approach.[14] Rather than block or filter content, the redirect method targets internet users that search for violent extremist content with advertising that promotes counter-narrative content. It does this by using the same tools used for online advertising, and redirecting "them towards curated YouTube videos debunking ISIS recruiting themes."[15] This content is then connected to individuals that search for certain key words through text adverts, image adverts, or skippable video adverts, rather than changing their actual search results.

Counter-Messaging

Counter-messaging is aimed at reducing the demand side. The idea behind counter-messaging is to create and disseminate messages that counter the impact of extremist propaganda. Such communication campaigns are divided into three categories:

1. Government strategic communications focusing on raising awareness of what the government is doing and correcting misinformation.
2. Alternative narratives, to undercut violent extremist narratives with positive messages of social inclusion.
3. Counter-narratives, to "directly deconstruct, discredit and demystify" the narrative of violent extremists.[16]

The success of a communication campaign depends on multiple factors, including the message itself, the medium of delivery and the messenger of that message. To be successful a message needs to be conveyed by a "credible messenger" in the eyes of the target audience. In terms of CT-CVE communications, the target audience is unlikely to perceive government sponsored messengers as a "credible messenger."[17] This has prompted counter-narrative campaigns that are delivered by community and civil society organizations (CSOs). However, such campaigns are often technically and financially supported by governments. This presents a dilemma for CSOs, to be transparent and open about where their funding and support comes from. Otherwise they risk being branded part of the state apparatus and hence not credible. Or to not disclose their sources of funding, in the aim of presenting themselves as a more credible messenger.[18] Highlighting government support may undermine the CSOs efforts, however, hiding sources of funding if they were to emerge at a later stage may prove far more damaging in undermining trust long-term.

One-to-One Online Engagement

Another approach directly targets known extremists, drawing them into one-to-one conversations online. Away from peer pressure and group dynamics, this creates the space to confront the extremists about their beliefs and lead them to question their acquired worldview.[19] It is a skilled job that "often involves former extremists, religious scholars and other credible messengers joining online forums under a pseudonym, building up relationships with individual members and through sustained engagement drawing them into discussions about their extremist views."[20]

A recent example is the Online Civil Courage initiative (OCCI), run by Facebook and the Institute of Strategic Dialogue (ISD).[21] There have been pilot counter-speech projects in the U.K., France and Germany, which involved identifying those with extremist views online and reaching out to them, with the offer to engage in dialogue. The objective behind one-to-one engagements is to start a conversation such that they hope can lead to the individual consenting to themselves being referred to wider support structures such as mental health providers or a face-to-face meeting with a trained intervention provider. As ISD notes "[T]his simple engagement provides a unique window of opportunity to open trusted communications with individuals least likely to be known to frontline services or agencies and who are engaging with the most prolific extremist content online."[22]

Education and Media Literacy

Another approach is education focused, where media literacy is intended to empower "users about where they are, who they are, and how best to act in relation to online challenges."[23] Online extremism can be treated as a child safety issue, much the same way as children are educated about the dangers of online grooming from pedophiles and warned not to give away personal information to strangers online. Literacy training on extremism could include warnings about becoming involved with violent extremism and awareness of potential grooming behavior by extremists.[24] Another approach sees education as a tool that can be used to develop the resilience of children by fostering

debate and providing information on societal context. This can include promoting the values of citizenship and diversity, an understanding of history and power relations in society and religious literacy, among others.[25] A final aspect is media literacy, which gives students the skills to be able to independently critically evaluate media coverage: to question the sources of stories and be aware to conspiracy stories, historical revisionism and misinformation. In the age of fake news, media literacy is a skill that has become ever more important.[26]

Section 2: An Ethical Analysis of Policy Responses to Extremist Propaganda Online

Drawing from human rights, consequentialism and virtue ethics, this section presents a set of different ethical concerns that are raised across a set of policy areas. First is the issue of extremist content. Second, who has moral authority for decision making in this space? The third concern is the role that Artificial Intelligence (AI) plays in decision making. Fourth, issues of manipulation or brainwashing. Fifth, how and where does privacy fit in these responses? Finally, we look at what role efficacy plays in our assessments of these policies.

What Counts as Extremist Content?

At its core, disruption is focused on controlling the dissemination of extremist content online to prevent the spread of extremist ideas and practices. We locate these ethical concerns as they tie to debates about free belief, free speech, and free public communication.[27] Those arguments hold that in liberal democracies we protect free belief, speech, and public communication, while recognizing limits on these freedoms.[28] This raises the first ethical challenge for disruption of extremist content online – how do we decide what is 'extremist' content?

One way is simply to censor content, for example IS beheading videos, or any instructional material that describes how to carry out violent activity. The problem is that implementing these policies for online material can be complex. In liberal democracies, free public communication of beliefs and ideas are ethical principles that are generally adhered to,[29] meaning that any instance where an effort to disrupt online content targets the wrong material is a potential violation of free public communication. For example, if censoring of news organization's broadcasts were to occur, the government could be said to be interfering with the freedom of the press. Second, a well edited piece of extremist content could quite easily produce shocking or offensive material by describing or suggesting graphic material, but not directly broadcast that material.

The IRU has argued that nonviolent terrorist propaganda is a potent radicalizing force that is often overlooked.[30] With that in mind, we are faced with the challenge of how to respond to propaganda that itself does not have or use violent imagery or content. For instance, considering instructional material, should sites or pages that direct an audience to other sites be censored, even if they host no detailed instructions themselves? And how ought we respond to sites in which a set of instructions have been broken down and spread across three different hosts? In this scenario, no portion of an

instruction video itself details the violent activity. So, should each host site be taken down, though none of them singly hosts violent instructions?

In attempt to offer answers to these questions, we draw from a range of ethical theories, such as those found in human rights, consequentialism, and virtue ethics. Using a human rights approach, if extremist content posted online is demonstrably offensive then the content showing the beheading could be disrupted. Showing a victim's beheading is disturbing and disrespectful to the victim and their loved ones. This approach is centered on the argument that certain speech acts can be constrained if they are deemed offensive.[31] This human rights account can be founded on 'recognition respect' and consists "in a disposition to weigh appropriately in one's deliberations some feature of the thing in question and to act accordingly".[32] In virtue of being a human, we all ought to be granted recognition respect and see that causing significant offense is a moral problem. If someone has been a victim of a particularly horrendous moral transgression, particular speech acts about that act that shows significant disrespect, could potentially be constrained or disrupted. Of course, nuance is needed here – a journalistic news article that described the moral transgression would likely not be justifiably constrained, whereas a public commentator who reveled in the details of the transgression and said that the victim deserved it could justifiably be constrained. As explained by Weckert, "[W]e can say that what is wrong with giving offense in general is that it is showing a lack of respect for others and that it may cause them to lose some of their self-respect."[33]

Using the consequentialist approach, if the extremist content is likely to result in acts of violence, either in actually producing physical harm or in inciting violence, then the speech should be disrupted. As per a consequentialist reasoning, rather than an individual's rights being the chief moral concern, the bad consequences of a speech can justify constraints on or disruptions to that speech act. As U.S. Justice Oliver Wendall Holmes famously articulated in *Schenck v. United States*, "The most stringent protection of free speech would not protect a man falsely shouting fire in a theater and causing a panic."[34] In terms of extremist communications – production, distribution and "owning" terrorist material/manuals etc. could be justifiably constrained given worry that it could likely result in physical harms. As with offense under the human rights approach, a fundamental point to note here is that the justification for constraints on speech related to extremism requires the harm it may cause to be significant and likely.[35]

Using the virtue ethics approach, decisions hinge on what repeated actions or behaviors by individuals reveal about their character.[36] When applied to online behavior, judgements based on virtue ethics rely on observation of patterns or repeated instances of posting and hosting extremist material.[37] For example, if a particular website is repeatedly putting up disrespectful and/or potentially harmful material, then the takedowns could extend from specific content to the website overall. Alex Jones, for example, was banned from a series of online communications platforms through 2018, because of his patterns of behavior that were deliberately spreading misinformation.[38] Similarly, in mid-2019, 8Chan – a website used frequently by right wing extremists – was significantly disrupted when its service provider, Cloudflare, terminated the site as a customer following a spate of postings on 8Chan linked to extremist acts of violence.[39]

Combining the principles inherent in the human rights, consequentialist, and virtue ethics approaches, in situations of *significant* offense, high chances of *significant harm*, and/or when a person persistently communicates extremist views thus displaying a *significant* pattern of extremist behavior, there may be justification for disrupting online speech.[40] This, however, is not necessarily universally applicable – each case would still need to be assessed on its particulars. That said, in liberal democratic societies, people have a *prima facie* claim to noninterference in their communications.

Moral Authority for Decision Making

This leads us to the second ethical issue – who has the authority to decide what extremist content is and what the appropriate responses should be? Free speech debates are not so much about what is said, but about who has the authority to decide what is said.[41] The online environment presents particular challenges here because the majority of services and infrastructure are privately owned and run. The moral challenge here is, in part, where the moral authority of the "banning institution" comes from. With respect to governments, we can present an argument about the social contract and security, whereby "the first duty of government is the security of its people."[42] The basic argument is that citizens cede certain rights to their government, and as a result it has a responsibility to provide security to its citizens.[43] Taking down disrespectful and harmful content is part of that responsibility.

What of private institutions like Facebook, Google, Twitter and so on? Their moral authority and accompanying responsibilities derive from the type of institution they are.[44] For instance, if they are a media organization equivalent to the press, then they have a duty to report important news to the world, which would also mean that they are constrained by issues of journalistic professionalism. If they are, instead, more like public infrastructure – like that of a road system or energy system – then they may have to constrain their responsibility to shareholders and profits by reference to public safety and extremist content that poses a public safety threat would likely be justifiably disrupted. This leads us to an open question about what public safety is, and what risks are significant enough to warrant institutional responses. A sensible discussion of moral authority and actions like disruption of online extremist material needs to include a discussion of what public safety means and why it is important.

There is an important counterargument to the notion of private social media institutions being like the press or public infrastructure – as they are not a public good, and definitely not publicly elected like a government, they remain simply private institutions.[45] On a notion of corporate social responsibility, we may hold that these private institutions do indeed have a set of moral responsibilities beyond following the law and shareholder returns. Corporate social responsibility, is "typically understood as actions by businesses that are (i) not legally required, and (ii) intended to benefit parties other than the corporation (where benefits to the corporation are understood in terms of return on equity, return on assets, or some other measure of financial performance)."[46] A deeper argument found in Seumas Miller's work is that certain social institutions provide key services which are of moral importance. As a result of these services, those institutions have moral responsibilities above and beyond the legal and shareholder

responsibilities.[47] It is beyond the scope of this article to argue for this point here. Rather, we suggest that social media institutions may in fact have duties beyond seeing them simply as private companies. Importantly, given the active role that social media institutions are playing in the fights against extremism, we can suggest that they also see that they have some duties above and beyond adhering to the law and maximizing shareholder returns.

AI and Decision Making

A third concern arises with the means of disruption. Many social media companies are now using artificial intelligence (AI) to identify problematic content and suspicious accounts and shut them down, sometimes without direct human involvement.[48] There are two particular issues associated with the use of AI in this way. First, should AI even be performing this role?[49] From a consequentialist viewpoint, the means would likely be largely irrelevant. The primary concern for a consequentialist would likely be if the AI is "getting it right." If harmful content is taken down, with greater accuracy than a human operator, then AI would be the preferred option. On the other hand, we might hold that significant moral questions require the decision maker to have the moral autonomy to make "moral decisions." It would likely follow that AI lacks such moral autonomy and so ought not be making these decisions. There is a potential compromise position, which is in practice how the main social media companies employ AI. AI serves the role of flagging extremist content or suspicious accounts and once flagged, a human operator steps in and makes the ultimate decision. In this way, AI is being used as an aid in human decision making, seeking to increase efficiency while ensuring morally significant decisions are made by people with moral autonomy.

Transparencies

Considerations associated with AI and decision-making are closely related to another ethical issue – transparency. One chief practice of ensuring, assuring, and improving best practice decision-making is scrutability – can we look to the AI decision making processes to see how decisions were actually made, and if so, is that process in line with the relevant laws or rules? The problem is that certain forms of machine learning lack scrutability – there is no audit trail detailing how the AI made decisions.[50] Perhaps this alone is not a major issue, but if we significantly value free public communication as a functional necessity for free speech, and AI is directly disrupting online material without a human in or on the loop, then this may present a significant moral concern.

Liberal democracies distinguish themselves from non-democratic cultures in part by reference to fair and just legal systems. If someone has their speech constrained online, then they may have a right of appeal. But this appeal process is founded on the notion of transparency – one must know that their speech act was constrained or interfered with, the reason why their speech act was constrained, and how they can appeal such decisions. If the AI being used is inscrutable, then there is a failing in fair and just processes. Similarly, if humans are involved, but that decision-making lacks transparency in

either the action or the appeals process, then we have a diminution or failing of fairness and justness.

One of the main concerns with social media's increased activity in regulating online communications is that these companies' decision-making processes lack transparency. Twitter, for instance, long held to a notion of protecting free public communication. In 2017-2018, they faced increased public and political pressures around the world to be more active in removing extreme content, and deleting accounts associated with extremists. However, this process lacked transparency – by some accounts, these decisions were made by Twitter CEO Jack Dorsey directly, raising concerns about capricious decision making and the abuse of discretion.[51] In particular, it is important to have clarity over how social media companies define terms such as "terrorism" and "extremism" in their terms of service, definitions that are at the heart of content removal, but which often vary significantly between platforms. Finally, clear and transparent appeals process by which users can challenge decisions are needed.[52]

This point becomes even more ethically complex as social media companies are increasingly taking action to limit the visibility of some extremist or problematic postings that come close to but do not break the platforms terms of service. In these cases, an account posting and/or hosting extreme views and extremist content is not shut down, nor is their content necessarily removed. Instead, their postings' accessibility or visibility is restricted by the host provider, by actions such as removal or restrictions on recommendation, searchability or engagement functions, the application of warnings and prevention of monetization[53]. While those who want to view the content will still likely find it, the content is no longer as accessible to a wider audience. This is not direct interference in free speech, rather it is a diminution of free public communication. The practice is usually resorted to in circumstances in which the account or content does not breach the platforms terms of service but comes sufficiently close to breaching those terms that the platform felt the need to act. Related to transparency, however, if the terms of service have not been broken, how and why was the decision made? These questions return us to the earlier discussions of what counts as extremism, and who has the authority to make decisions about what extremist content is.

A final point on the issue of transparency relates to redirection, counter-narratives, one-to-one and broader education strategies – should users be told that they are subject to a counter extremism response? Should counter-narratives be explicitly noted as counter-narratives. If a state agency is funding a counter-narrative strategy, should that state's involvement be openly and explicitly noted? When engaging in one-to-one discussions, does the "counselor" need to let the target know who they are and what they are doing? Finally, and related to counter-narratives, do any broad education-based strategies need to explicitly note if a state agency is funding their work?

One key aspect here is oversight – transparency does not necessarily only mean "open to the public," it can also mean "open to scrutiny." Because of this, any responses to online extremism must contain some measures of oversight, scrutiny, and/or accountability. Though the state's activity may be secret or covert, liberal democratic values still hold that such actions require some form of scrutability and oversight. Practices like warranting, FISA Courts in the U.S., Ministerial accountability in the Westminster system, must meet the fair and just principles of liberal democracies whilst protecting the

integrity of particular tactics and operations. As one of the authors has argued elsewhere,[54] if any of these responses to extremist propaganda are justified by reference to national security, then equivalent national security oversight and accountability processes should apply. Moreover, a parallel set of processes ought to apply to non-state institutions engaged in these practices. Other aspects arising from transparency are best discussed in terms of efficacy, discussed below.

Manipulation and Brainwashing

Given that a number of responses to extremist propaganda center on changing people's motivations and beliefs, "brainwashing" is of further ethical concern. Transparency highlights the worry that, should the manipulation be known, the targets will no longer be vulnerable to it. Furthermore, there is the fundamental concern of whether such manipulation ought to be permitted in the first place. Deradicalization efforts can be traced back to movements in the 1980s that sought to deprogram cult members and neo-Nazis. Current initiatives, including counter-messaging, one-to-one redirection efforts and broader education programs raise the ethical concern that counter-propaganda is manipulation and brainwashing. In order to ethically assess these worries, we need to answer at least two questions – should counter-propaganda programs seek to change a target's heart and mind? And if so, who has the moral authority to do this?

This then leads us to the final ethical challenge around manipulation and brainwashing – what is the state's role?[55] On the one hand, as mentioned, the state has a duty to stop people engaged in, and perhaps planning to engage in, malicious activities. On the other hand, however, liberal democracies distinguish themselves from other forms of government, in part, by reference to the liberty that their citizens have. On the face of it, if a state seeks to change people's motivations and beliefs, then they are either on a slippery slope to authoritarianism or perhaps already there. Perhaps here, the motivations and means become operative. If the state's motivation to intervene is in reference to national security concerns, this implies that the target of the intervention is a threat to the state in some way. If, instead, the state's motivation to intervene is based on safety concerns – that the target of the intervention is a risk to themselves, specifically – this implies that the target of the intervention is being treated with due respect and care. The means of intervention matter. If the state actively supports relevant community groups, friends, and families to be the primary methods of intervention, the state's efforts are more likely to be effective in garnering the trust and support of relevant communities.[56]

Privacy

As with constraints on free speech and free public communication of ideas, a right to privacy is not absolute. If someone repeatedly places extremely offensive and/or dangerous material online, they can not simply argue "you cannot interfere with this because this is private."[57] For the easy cases, again, like free speech and public communication online, disruption is easy. The notion of privacy becomes more complex in cases where "pre-extremist"[58] information is analyzed and an individual is identified for being

"potentially at risk" of extremism. In this case, we are making assumptions about a person's nonpublic beliefs and motivations. Modern information analysis claims to be able to read people's minds.[59] These internal mind states are obviously deeply intimate and personal. Notions of privacy, therefore, would suggest some significant constraints on the sorts of information gathered, who has access to it, and what is done with it, particularly in cases where an extremist act has yet to occur. In dealing with these cases, the virtue ethics approach becomes particularly useful – if a person has a habit or pattern of extremist behavior online then we have a greater justification for closer surveillance, to see if they are moving from "mere" extremism into something more offensive or dangerous.

In liberal democracies, there is typically an assumption of a right to privacy. Moreover, in liberal democracies, privacy is also understood as a constraint on government overreach – the government cannot violate the private spaces of people. Privacy violations can potentially be justified in situations where there is significant reason for the state to believe that an individual is engaged in preparatory activities that might result in rights violations or harms to others, and they therefore forfeit their general right to privacy. Here we see a set of considerations about human rights, consequences and a person's habits playing a role similar to those points covered in the free speech section.

The second privacy issue centers on limiting the power and influence of states by limiting where and when it can access its citizens data (and perhaps non-citizens as well). In this political sense, privacy is seen as the opposite of government intrusion. In other words, "[p]rotecting privacy involves reducing the extent to which individuals, institutions, and the government can encroach on people's lives."[60] Continuing this political frame, privacy might be thought of as an instrumental good, something necessary for democratic freedom. As before, the easy cases are easy, privacy is either trumped by national security duties, or the extremist has forfeited privacy rights by repeatedly producing and communicating information that is significantly offensive and/or dangerous. Where things become complex again is if the activities online are "merely" extreme, not yet verging into significantly offensive and dangerous.

This poses a deep problem for liberal democratic states. Such states have a commitment to the idea that people are free to believe what they will, and the state should be constrained in what it can do in the lives of its citizens. Liberal democracies distinguish themselves from authoritarian or theological societies by recognizing that their citizens will have differing conceptions of what is "good": they are at their core, pluralistic. For example, "[R]adical beliefs and extremist attitudes are not necessarily illegal, nor are they inherently negative … Martin Luther King and Gandhi [were] radicals of benevolent intent. It is not altogether uncommon for several individuals to, at some point in their life, hold views or opinions that may be considered extreme. In a majority of these cases, violence or any other problematic manifestations of these beliefs will not occur."[61] Put simply, a person in a liberal democracy can think whatever they want, including an extremist. What matters is not the belief, per se, but how the beliefs play a role in the person's behavior.[62]

This issue becomes particularly challenging when considering one-to-one efforts, counter-narratives and education. The problem here is the combination of concerns

about manipulation and brainwashing with the legitimate role of government. It is not simply that a person is being exposed to ideas and practices that seek to change their mind, but that the government is involved. For liberal democracies, conscious government efforts to change people's hearts and minds about core beliefs through deception and exploitation is deeply problematic. While complex, a two-part response to this challenge is needed. First, any agency involved in such efforts needs to be distinct from, and independent of, traditional security related agencies like the police, intelligence, or military. Social services or healthcare providers or, perhaps independent non-government organizations are more suited to carrying out such initiatives. Second, as early as it is possible, any such efforts would need to be made transparent. That is, if the effort is only going to be successful through deception, then it needs to be rethought and redesigned. The point here is that a liberal democracy must be constrained to direct interventions where the public offensiveness and danger posed by those beliefs are significant. In liberal democracies, the individuals can maintain whatever beliefs they want.

Efficacy

A fundamental issue for any policies in an ethically complex terrain is whether such policies are effective or not. Responses to extremist propaganda inherently intrude on or violate free public communication, privacy etc., and/or can lead to a set of undesirable consequences. As such, in a liberal democratic context, any intrusions or negative consequences from responses to extremist content online must be justified by the responses actually working. Efficacy is core to any ethical framework.

The problem is that evidence of success is problematic. As noted by Harris-Hogan, Barrelle, et al., "while evaluation of CVE policies and measures is crucial to improving knowledge and establishing best practices, there is currently no consensus on standardized CVE evaluation practice.[63] Despite this, "counter-radicalization policies in countries like Denmark, Germany, the Netherlands, the USA and the U.K. are required to be 'evidence-based' which suggests that evaluators should apply rigorous empirical methodology and measurement techniques. However, it is often unclear what this evidence should consist of and how it should be gathered … Also, there is to date no consensus on indicators of successful de-radicalization."[64]

As there are a complex set of different factors involved in the pathways to violent extremism and disengagement in carrying out an attack, the success of one program can prevent another program from being successful.[65] Moreover, if the individuals are no longer interested or engaged in the violent extremism, there is literally no evidence of their potential future activity. That is, the absence of evidence itself is evidence of the program's success. But, at the same time, an absence of evidence is absolutely no evidence of success. Parallel to this, there are different ways to consider success, which complicate and confound easy measures of the efficacy of intervention programs. A first issue to recognize is whether permanence is a good measure of a program's success. That is, while it seems ideal that a person is no longer involved in violent extremism as the result of a program, is a program a failure if the person becomes re-radicalized and/or reengages in VE ten years after a program's end? Underpinning this is the problem of evidence and counterfactuals: It is hard to get convincing evidence that the subject

would've otherwise been engaged in violent extremism had they not been involved in a program, and evidence of this becomes impossible the more general a program is.

Section 3: Toward an Ethical Framework for Countering Extremist Propaganda Online

Assuming then that there is some confidence that the interventions are successful, there are five elements that we argue should comprise the ethical parameters for countering extremist propaganda online: Free speech, moral authority of the decision maker, transparency, privacy and efficacy. In the paragraphs below, we set out the ethical foundations for the limitations that these parameters should place around online CVE.

Free Speech

As noted, liberal democracies typically consider free speech a foundational value; the presumption is against interventions that impose constraints on free speech. However, there are conditions where free speech can legitimately be constrained – in situations of significant offense, high chances of significant harm and/or when a person persistently communicates extremist views, these public communications might legitimately be disrupted. To be explicit, mere "extremism" alone is not a justification for disruption. Interference in speech acts can be justified if the disruptor can show that the speech acts display *significant* disrespect for individuals, can cause *significant* harm and/or there is a *significant* pattern of extremist behavior.

Such a response is deliberately pluralistic. It does not look for a single moral foundation to explain every answer to questions of free speech. Instead, it seeks to identify a set of values to help in decision making. On this pluralistic approach "a complete account will need to appeal to several foundational theories, each one of which is able to explain the basis of *some* of the normative factors, but no one of which explains all of them."[66] Efforts to constrain or disrupt the problematic speech act would be guided by human rights considerations of respect, consequentialist concerns about the harms of allowing the speech act to remain undisrupted and character aspects drawn from virtue ethics about how frequent and persistent the problematic speech acts are. All of these, as stated, turn on the significance of the problematic speech act. We note here that the notion of significance is vague and undefined, however, it is a common aspect to free speech discussions. Furthermore, the pluralistic approach gives us a basis for public discussion about what counts as significantly offensive, significantly harmful or a pattern of significant extremist behavior.

Moral Authority

The next issue concerns moral authority for decision making. This, we suggest, refers to two-related elements. First is the process by which a given authority has the moral authority to make decisions. Second is the way we understand the institutions and supporting technologies. On the first point, the simplest argument comes from discussions of social contract and rights forfeiture. Individuals forfeit certain rights in order to

receive a set of services and protections, principally discharged by the state. Moral authority, therefore, is dependent on the relations between the forfeiting parties, typically to be understood as citizens, and the parties with special rights and responsibilities, typically to be understood as the state. Any effective framework here would be derived from the given particulars of the social contract arising from the given citizen-state relations. However, as covered, many of the issues of online extremism turn not on direct relations between the state and its citizens but companies and consumers. Here, we can look instead to the processes of informed consent and the legitimacy of terms of service agreements between the company and the consumer. Suffice to say that any legitimacy is based on the consumer having had the opportunity to give meaningful informed consent, and the terms of service being what a reasonable person would consider fair and just.[67]

The second point derived from this is based on the more vexing issue of just what sorts of institutions social media (and the like) actually are. If they are merely a product, then we would suggest that the terms of service agreements would be definitive of the limits of what a person can communicate using that product. However, if we see social media as performing some greater public role, then perhaps these services are not some mere product but a part of public infrastructure. As such, like minimum laws regarding product safety, social media is to be bound by minimum conditions on user safety. If, instead, we see social media as serving a function like that of traditional media, then the rules, codes of conduct and restraints imposed on journalists would likely need to be applied to social media. Designating social media to be "media" brings with it a host of responsibilities. We suggest that governments and relevant companies need to engage the public in a broad discussion about the nature of social media, and what rights and responsibilities are entailed by it.

Transparency

We suggest here that the answers to these questions about transparency largely derive from the moral authority of social media companies – are they morally legitimate actors, with the right to make such decisions? A significant portion of the answer here also comes from what sort of institution social media companies are. As above, we need to ask if these companies are actually getting it right in terms of the decisions and actions they take. For example, one Twitter user had his account suspended after threatening to kill a mosquito.[68] This situation was obviously the result of a false positive. In such situations, the transparency of appeals processes plays a critical role in ensuring and assuring that the correct decisions are being made. A further practical aspect of transparency is the right to easy and timely appeals. That is, if a person has their expressions constrained or if they are banned entirely from a given service, then they should have the right to appeal the service provider's decisions. These appeals should have an appeal process that is easy to locate and carry out. On authority and transparency, we suggest that there needs to be some process linking the decision makers to the citizen or consumer in some way. It would be self-defeating if such appeals processes were closed or "black-boxes." Given the importance of transparency, any appeals would need some form of accountability to allow for and maintain community engagement.

When AI is involved in morally significant decisions, that decision-making process must be scrutable to outside parties. What we suggest here is that responses to extremism online aim to achieve a minimum level of scrutability. This minimum scrutability serves a functional purpose – if the recipient of a negative AI decision was to appeal the decision, a core element of the appeal would explain which communication was offensive and in what way that communication was deemed to be offensive enough to warrant censoring. Moreover, such an appeal process would be more than mere explanation. The recipient should be able to contest the AI decision. Like the case of the user who was suspended for threatening a mosquito,[69] there will likely be cases where the communication was clearly neither extremist nor violent in a meaningful way, and, therefore, did not warrant censure.

Privacy

In terms of privacy, this becomes relevant when we consider how an institution ought to act when an individual has disclosed extremist views in a setting where they may have an expectation of confidentiality. Here, we are thinking more about one-to-one and educational efforts. That is, if the potential extremist becomes engaged in ongoing conversations with an online counselor, or someone who they are of the belief is serving some role in which some confidentiality would be a reasonable expectation, then how ought that deradicalization program deal with that confidential information? This is where transparency is particularly important. In part to establish reasonable expectations – doctors, counselors etc. can break confidentiality if a significant crime has occurred or if they reasonably suspect that their client is likely to do something significantly problematic or dangerous[70].

Efficacy

While this may seem self-evident, we recommend here that any efforts to counter extremist propaganda online be subjected to some assessment of their efficacy. Given that a number of the responses to extremist propaganda involve interventions that impact liberal democratic values like free speech, privacy and so on, we need to ensure that any such responses are likely to work. As a simple proportionality calculation,[71] any rights violations and harms resulting from the efforts would need to be outweighed by the benefits. As we noted earlier though, we have to be aware that measures of efficacy are likely to be vague and perhaps contested. What we suggest here is that such challenges, while very important in a context like counter-terrorism policy,[72] are common to policy development. And while counter extremist propaganda policies are likely to have fuzzy or vague measures of success, given what is at stake when such policies are enacted, we have an ethical responsibility to take the notion of efficacy seriously.

Conclusion

How to respond to the spread of extremist propaganda online is a major challenge facing liberal democracies. It is clear that action is needed to counter and limit the impact

of extremist material online, however the ethical parameters around what action is acceptable within a liberal democracy have so far not been sufficiently addressed. This paper has sought to both set out the ethical justification for existing policy approaches, alongside identifying the ethical parameters in which such policies need to work within in a modern liberal democracy. As a first step toward developing an ethical framework for countering extremist propaganda online, we set out five key parameters on which a framework should be based: Free speech, moral authority of the decision maker, transparency, privacy, and efficacy. These five parameters lay out a clear foundation from which to build a practical ethical framework for delivering an effective policy response to extremism propaganda online, whilst maintaining the core values of liberal democracies.

Notes

1. Financial Times, "Zawahiri Leads Al-Qaeda into Battle for Muslim Hearts and Minds," https://www.ft.com/content/e8bff11a-4d5e-11da-ba44-0000779e2340 (accessed November 4, 2005).
2. Much of the literature looks at the threats and challenges posed by IS and other jihadi extremists. However, given the connections between online materials and the rise of right wing and nationalist violence in recent years, we focus this paper on the broader threat of violent extremism generally, rather than just Jihadi violent extremism.
3. For more on this see Frederick F. Schauer, *Free Speech: A Philosophical Enquiry* (Cambridge: Cambridge University Press, 1982); Stanley Fish, *There's No Such Thing as Free Speech: and it's a Good Thing, Too* (New York, NY: Oxford University Press, 1994); Wojciech Sadurski, *Freedom of Speech and its Limits* (Dordrecht: Kluwer Academic Publishers, 1999); David van Mill, "Freedom of Speech (Stanford Encyclopedia of Philosophy/Summer 2017 Edition)," https://plato.stanford.edu/archives/sum2017/entries/freedom-speech/ (accessed May 1, 2017).
4. We use extremism here in a broad sense and note that there are important distinctions between extremism and violent extremism which we do not address in this paper, and that how these terms are defined are part of an ongoing debate. We expand on these issues later in the paper where we discuss what is extremist content and who should make these decisions.
5. We note here that a number of the principles we discuss are covered by particular laws. While the relation between law and ethics, particularly in complex areas like counter-terrorism, is worthy of analysis, we do not have space to enter into those discussions here. Our approach complements jurisprudence, offering guidance for policy makers and in the application of the relevant laws.
6. We note that there is an apparent lack of evidence of the actual impact of extremist propaganda and the overemphasis of the importance of *online* propaganda and influence. See Kate Ferguson, "Countering Violent Extremism through Media and Communication Strategies: A Review of the Evidence" (Partnership for Conflict, Crime and Security Research, Cambridge, 2016); Alastair Reed, "An Inconvenient Truth: Countering Terrorist Narratives - Fighting a Threat We Do Not Understand - ICCT," https://icct.nl/publication/an-inconvenient-truth-countering-terrorist-narratives-fighting-a-threat-we-do-not-understand/ (accessed July 2, 2018).
7. Ian Brown and Josh Cowls, "Check the Web: Assessing the Ethics and Politics of Policing the Internet for Extremist Material" (VOX-Pol, 2015).
8. While we are focussing our attention to liberal democracies in this paper, following the work of someone like Jonathan Haidt we consider that many of the values we discuss are in fact general Jonathan Haidt, *The Righteous Mind: Why Good People Are Divided by Politics*

and Religion (London: Penguin, 2013). Our recommendations would need to be adapted to make them applicable to non-liberal democratic communities.

9. Europol, "Europol's Internet Referral Unit to Combat Terrorist and Violent Extremist Propaganda," https://www.europol.europa.eu/newsroom/news/europol%E2%80%99s-internet-referral-unit-to-combat-terrorist-and-violent-extremist-propaganda (accessed October 10, 2019).
10. For example in the first quarter of 2018, Facebook took down or added a warning to 1.9 million pieces of VE content– 99% of which Facebook found themselves, see Monika Bickert and Brian Fishman, "Hard Questions: How Effective Is Technology in Keeping Terrorists off Facebook?," *About Facebook* (blog), April 23, 2018, https://about.fb.com/news/2018/04/keeping-terrorists-off-facebook/.
11. Peter R. Neumann, "Options and Strategies for Countering Online Radicalization in the United States," *Studies in Conflict & Terrorism* 36, no. 6 (2013): 431–59, https://doi.org/10.1080/1057610X.2013.784568.
12. Ibid., 443.
13. Kent Walker, "Four Steps We're Taking Today to Fight Terrorism Online," *The Keyword | Google* (blog), June 18, 2017, https://blog.google/around-the-globe/google-europe/four-steps-were-taking-today-fight-online-terror/.
14. The project was initiated by Jigsaw a unit within Google, in collaboration with Moonshot CVE, Quantum Communications, and a team of researchers including Valens Global and Nadia Oweidat. See Google, "The Redirect Method: A Blueprint for Bypassing Extremism," https://redirectmethod.org/downloads/RedirectMethod-FullMethod-PDF.pdf (accessed October 10, 2019); A similar pilot project has been set up by Microsoft on their search engine platform "Bing" in conjunction with the Institute of Strategic Dialogue (ISD). See Microsoft, "Microsoft Partners with Institute for Strategic Dialogue and NGOs to Discourage Online Radicalization to Violence - Microsoft on the Issues," *Microsoft Corporate Blog* (blog), April 18, 2017, https://blogs.microsoft.com/on-the-issues/2017/04/18/microsoft-partners-institute-strategic-dialogue-ngos-discourageonline-radicalization-violence/.
15. Google, "The Redirect Method: A Blueprint for Bypassing Extremism."
16. Rachel Briggs and Sebastien Feve, "Review of Programs to Counter Narratives of Violent Extremism" (Institute for Strategic Dialogue, London, 2013).
17. Patricia Crosby and Assan Ali, "Counter Narratives for Countering Violent Extremism," https://thecommonwealth.org/sites/default/files/inline/ComSec%20CVE%20Counter%20Narratives%20Presentation.pdf (accessed October 10, 2019); Haroro J. Ingram and Alastair Reed, "Lessons from History for Counter-Terrorism Strategic Communications" (The International Centre for Counter-terrorism, The Hague, 2016); Alastair Reed, Haroro J. Ingram, and Joe Whittaker, "Countering Terrorist Narratives" (PE 596.829, Policy Department for Citizens' Rights and Constitutional Affairs, Brussels, 2017).
18. This draws from a distinction between "black" vs. "white" propaganda, in which "white" is government branded and "black" is unbranded or branded under a false flag. For more on this, see Garth Jowett and Victoria O'Donnell, *Propaganda and Persuasion*, (Newbury Park, CA: Sage, 1986); Jason Stanley, *How Propaganda Works*, (Princeton, NJ: Princeton University Press, 2017).
19. The objective is not to "win the argument," as such, but to draw the extremist into sustained debate, and gradually sow the seeds of doubt that leads to them begin to question their beliefs.
20. Briggs and Feve, "Review of Programs to Counter Narratives of Violent Extremism."
21. Institute for Strategic Dialogue, "Online Civil Courage Initiative (OCCI) - ISD A civic response to online hate," https://www.isdglobal.org/programmes/communications-technology/online-civil-courage-initiative-2-2/ (accessed October 10, 2019).
22. Jacob Davey, Jonathan Birdwell, and Rebecca Skellet, "Counter Conversations - A Model for Direct Engagement with Individuals Showing Signs of Radicalisation Online - Executive Summary" (Institute for Strategic Dialogue, London, 2018).

23. Guy Berger, "Media and Information Literacy: Educational Strategies for the Prevention of Violent Extremism" (speech presented at UNAOC's Media and Information Literacy program, United Nations Headquarters, New York, United States, February 10, 2017).
24. Neumann, "Options and Strategies for Countering Online Radicalization"; European Commission, "Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Supporting the Prevention of Radicalisation Leading to Violent Extremism," https://ec.europa.eu/transparency/regdoc/rep/1/2016/EN/COM-2016-379-F1-EN-MAIN-PART-1.PDF (accessed June 14, 2016).
25. Ratna Ghosh, W.Y. Alice Chan, Ashley Manuel, and Maihemuti Dilimulati, "Can Education Counter Violent Religious Extremism?," *Canadian Foreign Policy Journal* 23, no. 2 (2017): 117-33, https://doi.org/10.1080/11926422.2016.1165713.
26. Neumann, "Options and Strategies for Countering Online Radicalization"; Crosby and Ali, "Counter Narratives for Countering Violent Extremism"; Extremely Together, "Countering Violent Extremism: A Peer-to-Peer Guide by Extremely Together," http://www.extremelytogether-theguide.org/ (accessed October 10, 2019).
27. For a set of discussions around liberalism, free belief, free speech and free communication see Schauer, *Free Speech*; Raphael Cohen-Almagor, *The Boundaries of Liberty and Tolerance: The Struggle against Kahanism in Israel* (Gainesville: University Press of Florida, 1994); Frederick Schauer, "Free Speech On Tuesdays," *Law and Philosophy* 34, no. 2 (2015): 119–40; Anne Showalter, "Resolving the Tension Between Free Speech and Hate Speech: Assessing the Global Convergence Hypothesis," *Duke Journal of Comparative & International Law* 26, no. 2 (2016): 377–415; van Mill, "Freedom of Speech."
28. Schauer, *Free Speech*; van Mill, "Freedom of Speech."
29. Schauer, *Free Speech*.
30. EU Internet Referral Unit, "On the Importance of Taking-down Non-Violent Terrorist Content," *VOX - Pol* (blog), May 8, 2019, https://www.voxpol.eu/on-the-importance-of-taking-down-non-violent-terrorist-content/.
31. van Mill, "Freedom of Speech."
32. Stephen L. Darwall, "Two Kinds of Respect," *Ethics* 88, no. 1 (1977): 36–49, https://doi.org/10.1086/292054.
33. John Weckert, "Giving and Taking Offence in a Global Context," *International Journal of Technology and Human Interaction* 3, no. 3 (2007): 25–35.
34. "Schenck v. United States, 249 U.S. 47 (1919), https://supreme.justia.com/cases/federal/us/249/47/.
35. van Mill, "Freedom of Speech."
36. This is a simplification of virtue ethics, focusing on the *character-as-repetition* aspect of virtue ethics. For more on virtue ethics, see Aristotle and W. David Ross, *The Nicomachean Ethics of Aristotle*, trans. W. David Ross, (London: Oxford University Press, 1972); Rosalind Hursthouse, *On Virtue Ethics* (Oxford: Oxford University Press, 1999); Alasdair C. MacIntyre, *After Virtue: A Study in Moral Theory* (Notre Dame, IN: University of Notre Dame Press, 2007).
37. This terminology of "virtue" draws from the Aristotelian notion of virtue where what matters is a person's character. And this character is created through habit. That is, a person's virtuous character arises from habit, what they do regularly or repeatedly.
38. BBC News, "Twitter Bans Alex Jones and Infowars," https://www.bbc.com/news/world-us-canada-45442417 (accessed September 6, 2018).
39. Josh Taylor and Julia Carrie Wong, "Cloudflare Cuts off Far-Right Message Board 8chan after El Paso Shooting | US News | The Guardian," https://www.theguardian.com/us-news/2019/aug/05/cloudflare-8chan-matthew-prince-terminate-service-cuts-off-far-right-message-board-el-paso-shooting (accessed August 5, 2019); Matthew Prince, "Terminating Service for 8Chan," Cloudflare (blog), August 5, 2019, https://blog.cloudflare.com/terminating-service-for-8chan/.

40. We recognise that the term 'significant' here does a lot of the heavy lifting and is undefined. While we recognise that what counts as significant is likely an open and contested area, what we note is the role that significance plays in free speech constraints. Two influential commentators on hate speech, Jeremy Waldron and David Boonin both "agree that prohibition is acceptable when speech is threatening; they disagree on what counts as a harmful threat. Waldron thinks most forms of racial abuse qualify whereas Boonin is more circumspect. But the disagreement between the two is about what causes harm rather than any major philosophical difference about the appropriate limits on speech. If both agree that a threat constitutes a significant harm, then both will support censorship" van Mill, "Freedom of Speech."
41. Fish, *There's No Such Thing as Free Speech*.
42. Steven Heyman, "The First Duty of Government: Protection, Liberty and the Fourteenth Amendment," *Duke Law Journal* 41 (1991): 507.
43. Thomas Scanlon, *What We Owe to Each Other* (Cambridge, MA: Belknap Press of Harvard University Press, 2000); Ann Cudd, "Contractarianism (Stanford Encyclopedia of Philosophy/Winter 2013 Edition)," https://plato.stanford.edu/archives/win2013/entries/contractarianism/ (accessed October 10, 2019); Fred D'Agostino, Gerald Gaus, and John Thrasher, "Contemporary Approaches to the Social Contract (Stanford Encyclopedia of Philosophy/Winter 2012 Edition)" https://plato.stanford.edu/archives/win2012/entries/contractarianism-contemporary/ (accessed October 10, 2019).
44. Claudia Aradau, "Security That Matters: Critical Infrastructure and Objects of Protection," *Security Dialogue* 41, no. 5 (2010): 491–514, https://doi.org/10.1177/0967010610382687.
45. We thank Kateira Aryaeinejad for raising this point.
46. Jeffrey Moriarty, "Business Ethics (Stanford Encyclopedia of Philosophy/Fall 2017 Edition)," https://plato.stanford.edu/archives/fall2017/entries/ethics-business/ (accessed October 10, 2019).
47. Seumas Miller, *The Moral Foundations of Social Institutions: A Philosophical Study* (Cambridge: Cambridge University Press, 2010).
48. Steve Kirsch, "Identifying Terrorists before They Strike," http://www.skirsch.com/politics/plane/ultimate.htm (accessed October 7, 2001); Emma Woollacott, "The Algorithm That Can Predict Isis's next Move - before They Even Know What It Is," https://www.newstatesman.com/world/middle-east/2015/09/algorithm-can-predict-isis-s-next-move-they-even-know-what-it (accessed September 24, 2015).
49. J. H. Moor, "Are There Decisions Computers Should Never Make," *Nature And System* 1, no. 4 (1979): 217–29.
50. Mark Coeckelbergh, "Artificial Intelligence: Some Ethical Issues and Regulatory Challenges," *Technology and Regulation*, (2019), 31–34, https://doi.org/10.26116/TECHREG.2019.003.
51. Austin Carr, "When Jack Dorsey's Fight Against Twitter Trolls Got Personal," *Fast Company* (blog), April 9, 2018, https://www.fastcompany.com/40549979/when-jack-dorseys-fight-against-twitter-trolls-got-personal.
52. Stuart Macdonald, Daniel Grinnell, Anina Kinzel, and Nuria Lorenzo-Dus, "A Study of Outlinks Contained in Tweets Mentioning 'Rumiyah'" (Global Research Network on Terrorism and Technology: Paper No. 2, Royal United Services Institute for Defence and Security Studies, London, UK, 2019).
53. Walker, "Four Steps We're Taking Today to Fight Terrorism Online"; Reddit Announcements, "Revamping the Quarantine Function: Announcements," https://www.reddit.com/r/announcements/comments/9jf8nh/revamping_the_quarantine_function/ (accessed December 15, 2018); Del Harvey and David Gasca, "Serving Healthy Conversation," https://blog.twitter.com/en_us/topics/product/2018/Serving_Healthy_Conversation.html (accessed May 15, 2018).
54. Adam Henschke and Timothy Legrand, "Counterterrorism Policy in Liberal-Democratic Societies: Locating the Ethical Limits of National Security," *Australian Journal of International Affairs* 71, no. 5 (2017): 544–61, https://doi.org/10.1080/10357718.2017.1342764.
55. We note here the concerns about manipulation of citizens by their governments and the fact that governments do actively intervene in the beliefs and motivations of their citizens.

Governments in liberal democracies do routinely engage in 'communication campaigns' that have, in part, the purpose of manipulating their citizens into thinking good things about the government. What we suggest is that transparency plays an important role. While some forms of government propaganda on their own citizens are perhaps inevitable, we can perhaps differentiate between white, grey and black propaganda depending on "an acknowledgement of its source and its accuracy of information" Jowett and O'Donnell, *Propaganda and Persuasion*. Much more needs to be said about these issues, but is beyond the scope of this paper.

56. Allard R. Feddes and Marcello Gallucci, "A Literature Review on Methodology Used in Evaluating Effects of Preventive and De-Radicalisation Interventions," *Journal for Deradicalization*, no. 5 (2015): 1–27; Kurt Braddock and John Horgan, "Towards a Guide for Constructing and Disseminating Counternarratives to Reduce Support for Terrorism," *Studies in Conflict & Terrorism* 39, no. 5 (2016): 381–404, https://doi.org/10.1080/1057610X.2015.1116277; B. Heidi Ellis and Saida Abdi, "Building Community Resilience to Violent Extremism through Genuine Partnerships," *The American Psychologist* 72, no. 3 (2017): 289–300, https://doi.org/10.1037/amp0000065.
57. For more on this point, see Adam Henschke, *Ethics in an Age of Surveillance: Personal Information and Virtual Identities* (Cambridge University Press, 2017).
58. By "pre-extremist" we mean online or other activity that is in itself neither extremist nor violent, but shows tendencies to evolve into extremism or violence.
59. Woollacott, "The Algorithm That Can Predict Isis's next Move – before They Even Know What It Is."
60. Daniel J. Solove, *Understanding Privacy* (Cambridge, MA: Harvard University Press, 2008).
61. Logan Macnair and Richard Frank, "Voices Against Extremism: A Case Study of a Community-Based CVE Counter-Narrative Campaign," *Journal for Deradicalization* 0, no. 10 (2017): 147–74.
62. Henschke, *Ethics in an Age of Surveillance*. 213–14.
63. Shandon Harris-Hogan, Kate Barrelle, and Andrew Zammit, "What Is Countering Violent Extremism? Exploring CVE Policy and Practice in Australia," *Behavioral Sciences of Terrorism and Political Aggression* 8, no. 1 (2016): 6–24, https://doi.org/10.1080/19434472.2015.1104710.
64. Feddes and Gallucci, "A Literature Review on Methodology"
65. Ibid., 309.
66. Shelly Kagan, *Normative Ethics* (Boulder, CO: Westview Press, 1998).
67. It is important to note there that that the terms of service are clear, transparent, and easy to understand. Fine print and overly legal terminology, or even too much terminology inhibits a consumer's ability to actually give informed consent. We thank Kateira Aryaeinejad for this point.
68. Yvette Tan, "A Guy's Twitter Account Got Suspended after He Made a Death Threat—against a Mosquito," *Mashable* (blog), August 30, 2017, https://mashable.com/2017/08/30/twitter-japan-mosquitos-abuse/.
69. Ibid.
70. Paul S. Appelbaum, "Tarasoff and the Clinician: Problems in Fulfilling the Duty to Protect," *The American Journal of Psychiatry* 142, no. 4 (1985): 425–29, https://doi.org/10.1176/ajp.142.4.425.
71. We note here that something like a notion of 'simple proportionality' is in fact quite complex. For more on the complexity of proportionality, see Adam Henschke, "Conceptualising Proportionality And Its Relation to Metadata," in *Intelligence and the Function of Government*, by Daniel Baldino and Rhy Crawley (Melbourne: Melbourne University Press, 2018).
72. For instance, as discussed elsewhere, counter-terrorism policy can be limited by reference to the very values it seeks to protect. See Henschke and Legrand, "Counterterrorism Policy in Liberal-Democratic Societies."

Acknowledgements

The authors are grateful to Kateira Aryaeinejad and CJ O'Connor for reviewing and editing this paper, and for their many insightful comments and suggestions.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant agreement No. 670172); Australia Research Council under the Discovery grant, (DP180103439 *Intelligence And National Security: Ethics, Efficacy And Accountability*); and the Australian Department of Defence under the Strategic Policy Grant, *Countering Foreign Interference And Cyber War Challenges*.

ORCID

Adam Henschke http://orcid.org/0000-0002-2956-0883
Alastair Reed http://orcid.org/0000-0002-9060-5518