Received: 3 June 2023 | Accepted: 26 October 2023

DOI: 10.1002/poi3.381

RESEARCH ARTICLE

P&I Policy & Internet | PSO | WILEY

Moderating manipulation: Demystifying extremist tactics for gaming the (regulatory) system

Ashley A. Mattheis[1] | **Ashton Kingdon**[2]

[1]Hillary Rodham Clinton School of Law, Cyber Threats Research Centre, Swansea University, Swansea, UK

[2]Department of Economic, Social, and Political Sciences, University of Southampton, Southampton, England

Correspondence

Ashley A. Mattheis, Hillary Rodham Clinton School of Law, Cyber Threats Research Centre, Swansea University, Swansea, UK.
Email: ashley.mattheis@swansea.ac.uk

Abstract

Due to its ease of scalability and broad applicability, the use of artificial intelligence (AI) and machine learning in platform management has gained prominence. This has led to widespread debates about the use of deplatforming as the default tool for repeated or severe violations of terms or service. But technologically deterministic approaches are not infallible and can be predictable based on their actions. This opens the door for manipulation of media content and technological affordances to become tactical options for actors seeking to subvert regulation. Existing discussions often neglect the topic of manipulation of content, algorithms, or platform affordances as a primary aspect of the strategies used by extremists in relation to the difficulties of moderation from a policy perspective. This study argues that it is essential to understand how extremists and conspiracy theorists use manipulation tactics to 'game' the current policy, regulatory and legislative systems of content moderation. Developing approaches that attend to manipulation as a strategy and focus on platform and context-specific tactics will generate more effective policies, platform rules, AI developments and moderation procedures. This study analyses and demystifies three primary tactics, which the authors categorize as numerology, borderlands and merchandising, regularly used by extremists online in their strategies to 'game' content moderation. We provide case examples from a variety of ideologies including far-right, QAnon and male supremacism to highlight the

tactics rather than ideological nature of such manipulation. We conclude with a discussion of how demystification processes could be incorporated into content moderation settings. This study contributes new insights about evasion tactics to the content moderation discussion and expands current understanding of how platforms can develop sociotechnical remedial measures.

KEYWORDS

artificial intelligence (AI), content moderation, digital culture, extremism, machine learning, manipulation tactics, propaganda

INTRODUCTION

Artificial intelligence (AI) particularly in the form of machine learning has become increasingly influential, leading to rapid social and technological change as it becomes embedded within many aspects of society. However, the potential transformation of humankind by AI confers not only benefits, but also risks encompassing privacy, autonomy, dignity and human safety (Hall & Pesenti, 2017). The regulation of AI has become an urgent priority, with countries around the world proposing legislation aimed at promoting the responsible and safe application of AI to limit the harms it can pose.

Holistic (2023) outlines the fact that different initiatives want to regulate the same technology, yet diverge in how they actually define AI. They surveyed some of the most widespread AI definitions offered by contrasting regulatory initiatives and bodies outlining their commonalities and differences. This study will incorporate the definition laid out by Hall and Pesenti (2017) in the UK Government report, *Growing the Artificial Intelligence Industry in the UK*, as 'a set of advances general purpose digital technologies that enable machines to do highly complex tasks effectively'. This was adapted from The Engineering and Physical Science Research Council, which uses the following description: "Artificial Intelligence technologies aim to reproduce or surpass abilities (in computational systems) that would require 'intelligence' if humans were to perform them. These include learning and adaptation; sensory understanding and interaction; reasoning and planning; optimization of procedures and parameters; autonomy; creativity; and extracting knowledge and predictions from large, diverse digital data" (UK Research and Innovation, 2022). This study uses 'Artificial Intelligence' as an umbrella term to cover a set of complementary techniques that have developed from statistics, computer science and cognitive psychology. While recognizing distinctions between specific technologies and terms (e.g., AI vs. machine learning, machine learning vs. deep learning), it is useful to see these technologies as a group, when considering how to support their development and use (Hall & Pesenti, 2017).

Due to its ease of scalability and broad applicability, the use of AI in platform management has gained prominence. This has led to widespread debates about the use of deplatforming as a response by social media companies as the default tool for repeated or severe violations of terms of service (Gunton, 2022; Jhaver et al., 2021; Klinenberg, 2022). Content moderation practices and developments have predominantly focused on technological solutions, given the vast scale of content and interactions that need to be monitored, and the relentlessness of violations (Gillespie, 2020). However, technological solutions have not been as effective as industry, researchers and policymakers would have hoped. Issues with these solutions include technical limitations (Church & Liberman, 2021),

for example, language-based 'big data' technologies, which consistently have limited capacities for success in non-English language contexts (Knight, 2016), and common-sense language processing (Sankhe, 2022). They include business limitations, for example the need to 'tweak' algorithms to ensure that platforms maintain profits from networked circulation and advertising which often reduces the efficacy of moderation of hateful/extreme content (Hagey & Horowitz, 2021). Further, they include user limitations, that is the use of human creativity to subvert rules, processes and to create practices to achieve desired outcomes. For example, the splicing of 10 s of 'dead air' or alternate video to the beginning of Christchurch attack livestream video copies to trick algorithmic scrubbers to surpass content bans of the mass murder footage (Knight, 2019).

In any discussion of content moderation, parameters around content labelling are important. Much of the type of content we are discussing would likely fall into the 'awful, but lawful' or 'borderline content' categories in relation to policy. This means that platform rules, rather than governmental regulation are the guiding factor in removal. At issue in many such instances is the amorphous character of 'extremist' content or 'extremism' online. Another major factor in content moderation policies and how they are applied are the terms of service of the platforms themselves, which are directed by each platform type and business needs. And, lastly, there are differences in how content moderation streams are generated and prioritized (e.g., user flagging, automated flagging, moderator flagging and so on). To clarify for the purposes of our research, extremism and extremist content are identified following Berger's (2018) definition, namely content that positions 'an ingroup's survival as inseparable from the need for hostile action against an outgroup'. As this study will show, much extremist content is therefore not moderated, particularly if propagandists come up with ways to 'hide in plain sight' as we describe below. We draw on our specializations in Criminology and Communication to engage with an underexplored problem associated with AI in relation to its use in digital content moderation by specifically focusing on human creativity as an arena for users' forcing limits on AI content moderation. This underexplored issue is urgently in need of attention in both practice and policy to improve the efficacy of moderation and reduce the success of extremist media efforts. Here, human creativity—that is the tendency of people to create ways of doing things to achieve their desired aims by working with the tools (or systems affordances) available to them—poses a user-driven recurrent limitation to the use of AI-based content moderation practices. Thus, the question animating this study is: What nontechnological strategies do extremists use to circulate their preferred content and propaganda in light of technological (AI) content moderation practices?

To address the above, we situate our discussion within the wider context of AI content moderation practices, and then describe and provide three short case examples of strategies we have seen extremists use over time in our own research to circumvent restrictions. These three strategies which we have categorized as (1) numerology, or the use of numbers and coding to bypass bans and rules; (2) borderlands, or the infiltration of ingroup eligible online communities (such as #TradWife online spaces) to both normalize content and generate plausible deniability; and (3) merchandising, or the marketing, influence-based promotion and sales of products to spread ideology and fund operations. These three strategies are essentially developed around using regular capacities of socially networked digital platforms aimed at information and product consumption. As such, they do not represent technological advances, but rather creative applications of available affordances and practices. Further, crucially, they are used by extremist actors and digital influencers to evade and circumvent, that is manipulate, different types of content moderation. We follow the short case examples with a discussion and recommendations for addressing the human side of the content moderation equation. This study contributes to the content moderation discussion by providing new insights into how bad faith actors are circumventing filtering strategies. In doing so, the study also expands current understanding

of how social media platforms can develop sociotechnical remedial measures and move away from the inordinate focus on solely technical modes of content moderation.

LITERARY OVERVIEW

The use of AI to automatically remove extremist propaganda directly hinders dissemination, online and offline mobilization, trolling and other networked harassment, and thus reduces extremists' ability to achieve their main goals (Jhaver et al., 2021). As propagandists conventionally need to operate on mainstream platforms to normalize their ideas and have them resonate with wider audiences, the removal of content directly from the mainstream, makes it harder for narratives to appear acceptable, with a consequent decrease in recruitment (Amarasingam et al., 2021; Squire, 2023). Deplatforming also disrupts extremist planning and organization, particularly on sites such as Facebook, Twitter and Instagram, which seamlessly offer both propaganda potential (through public posts), and confidential communications (through private chats), for those trying to recruit (Kingdon & Ylitalo-James, 2023). However, research has shown that a consequence of stringent filtering systems and censorship is that those who have been deplatformed from mainstream sites may migrate to private chats and encrypted sites like Telegram (Bloom et al., 2019), which, although often smaller and offering extremists a reduced audience, are perceived by users as more secretive through their encoding, making it far more difficult for public monitoring of misuse and emerging threats. Concern has also arisen with respect to the possibility of subjective, disparate and potentially biased enforcement of content and creators by different platforms resulting from the absence of internationally agreed upon definitions of what constitutes terrorism and violent extremism (Keller & Leerssen, 2020). Moreover, transparency regarding the tools and approaches used by social media companies for content moderation is highly limited and the absence of details regarding their implementation makes evaluating their effectiveness difficult (United Nations, 2021).

Currently, much research related to extremism and content moderation focuses on technicalities, specific platforms or other technologically deterministic frames (Gillespie, 2018). The desire for technological solutions is high for a variety of reasons, not the least of which is the large scale and rapidly changing pace of the problem. The current state of practice in regard to such technological solutions is highly complex with possible solutions dictated by each specific platform or technology (Gorwa et al., 2020). Technical forms of content moderation include automated filtering, machine learning and sentiment analysis to detect problematic content. In some cases, such content is automatically removed (filtered or matched against lists of problem content) and some are predictively detected content that may be reviewed by human content moderators (Gillespie, 2020; Gorwa et al., 2020). In addition to deplatforming (content or creator removal from the platform), platforms leverage different forms of automated solutions to curate and moderate content such as downgrading the content (algorithmic suppression) (Karizat et al., 2021), or by redirecting users to other content (Moonshot, 2023). Another approach often used is a technique known as 'shadow-banning' which refers to content being either deleted entirely or having its visibility significantly limited, without the user that published the content being made aware (United Nations, 2021). Platforms including Facebook, Twitter, Google and Microsoft make use of image and video matching using 'hashing' technology to shorten an output of any length into a fixed string of text—the hash—which operates like a digital fingerprint. This input can then be compared to other files to find duplicates and prevent them from being shared, thus enabling the identification of identical harmful content in real-time at high scale (United Nations, 2021). In the aftermath of the Christchurch attack in 2019, the Global Internet Forum for Counter Terrorism, (2023)

created a shared industry database of hashes of terrorist propaganda used by its members, but it is not available to nonmember platforms. Their stated aim was to support the coordinated takedown of such content across platforms, while adhering to data privacy and retention policies.

Currently, social media companies employ their largely bespoke content filtering models with limited success in preventing the spread of extremist (or other disallowed) content. This inability to manage dangerous or illegal material when paired with reduced transparency about industry practices and procedures exemplifies a crucial ethical problem. That is, the increasing application of algorithms in informing decision-making across a variety of sectors (e.g., finance, legal and judicial systems, and communication systems) with material effects on daily life. This situation has prompted demands for algorithmic accountability and the development of responsible AI (Busuioc, 2020; Ugwudike, 2022). Bias, unfairness and lack of transparency and accountability in AI systems, and the potential misuse of predictive models for decision-making have raised concerns about the ethical impact and unintended consequences of new technologies for society across every sector where data-driven innovation is taking place (Ayling & Chapman, 2022). Concerns surrounding AI can be grouped epistemically (the probabilistic nature of insights, the inherent inscrutability of 'black box' algorithms and the fallibility of the data used for training an input). There are also normative apprehensions regarding the fairness of decisional outcomes, the increasing surveillance and profiling of individuals, and the erosion of informational privacy. Algorithmic systems likewise create problems of accountability and moral responsibility, where it is unclear which moral agent in the process bears (or shares) responsibility for outcomes from the system (Ayling & Chapman, 2022). Without a detailed knowledge of the key factors providing the basis for algorithmic determination, it is impossible to know the extent to which social media companies are engaging in practices that could be considered unethical, deceptive, or discriminatory (Allcott & Gentzkow, 2017).

What is known about how technological moderation happens is information supplied by the companies themselves and is therefore proscribed by the company's own public-facing communication strategies. For example, Facebook says that it relies on machine learning to prioritize which content needs to be reviewed first; posts that violate the company's policies are flagged by users or machine learning filters, which include everything from spam to hate speech and content that glorifies violence. Since 2020, the company has elected to deal with clear-cut cases by automatically removing posts or blocking accounts. Only content that does not obviously violate the company's policy is reviewed by a human content moderator (United Nations, 2021). Other platforms like Twitter and YouTube rely on AI to quickly remove comments that violate the companies' rules. It is clear from these examples that content moderation AI models are trained to filter out specific content that fits company defined criteria. However, in their current form, these types of AI models suffer from inherent limitations (one of which, human creativity, is discussed in this study). For instance, machine learning models trained to find content from one terrorist organization, may not work for another because of language and stylistic differences in their propaganda (United Nations, 2021).

Thus, there are multiple problems with these types of AI content moderation approaches. Along with previously noted language limitations, automated systems, even machine learning systems, are not particularly good at understanding human meaning and context, either machine learning or sentiment analysis. Human communicative forms like irony, sarcasm and joking are highly complex and difficult for machines to manage (Gillespie, 2020). Further, crucially, what counts as 'problematic' content is dictated by platforms themselves. This creates a basis for content removal that can vary with business priorities and in response to socio-political concerns, both are nontechnical factors (Gorwa et al., 2020). Thus, content removal is not consistent across platforms and it may or may not

occur in accordance with stated platform terms and conditions with little transparency about such decisions. This means that content moderation practices are often highly contested and can easily become politically weaponized in culture wars framings of 'free speech online' used commonly by radical, far and extremist right actors.

Work focused on the human side of this problem explores the negative effects of the current frameworks of content moderation and policies, the most well-known of which is the broad range of research on 'echo chambers', and 'filter bubbles' (Kitchens et al., 2020). Although these concepts have become widely used both inside and outside of academia, with abundant research focused on them, extant literature leaves a variety of gaps and problems including issues with comparable data, replicability and consistency (Kitchens et al., 2020). Here, again, the issue of different platforms with different technological affordances impacts the study outcomes. Moreover, many such studies focus narrowly on users' specific information diets and behaviours on single platforms across limited topics of discussion. Broader studies of digital communication show that overall digital access to content increases information diversity among users rather than reducing information diversity (Pearson, 2021). Furthermore, rather than consistently speaking to likeminded thinkers or ideological companions, tools and features released on many platforms (e.g., lists, groups and tools to segment follower bases) actually seek to help users prevent 'context collapse'—a case where there is divergence between a user's imagined and actual audience—on their platforms (Pearson, 2021). Recent research also explores information operations—ranging from those conducted by individual actors to state actors— through a framework of human practice looking at how manipulation of platforms takes up specific repertoires of tactics, techniques and procedures that can be used to establish narrative control and shift public perspectives (Mattheis et al., 2022). This reframing is meant to illuminate a larger view of the practices and processes used to manipulate content and evade detection. It is essential to explore these issues in their full scope and complexity—as sociotechnical, that is as an interconnected human and technological problem—to successfully advance discussions of policy and practice related to better content moderation practices and maintaining a balance of safety and free speech in digital contexts.

METHODS

The goal of this examination of extremist online strategy is to outline how nontechnological practices can be used to evade and manipulate AI-based content moderation in ways that limit its utility and success in stemming the spread of extremist propaganda and the potential for radicalization. To assess these practices, the authors examined how extreme digital cultures were circulating propaganda images and texts (e.g., posts, books, products) successfully. That is, what types of materials stay online and circulate broadly. To do this, the authors needed to use data from extant data sets that were previously gathered from other projects to assess its utility over time. These original data sets were data collected manually from either extremist, or extreme-adjacent social media and website platforms by the authors.

For this project, the authors analysed their existing corpuses of data to assess patterns of use across them that showed strategies for circulating the content in ways that would avoid detection and removal of the content. Each short case study describes a distinct pattern of use that could be determined by the authors in their assessment of their existing corpuses. The selected short case examples were chosen because they were representative of the identified patterns. Each author used their own method of visual analysis (semiotic or cultural rhetorical criticism) to provide context and explication of the examples for the short case studies.

Underpinning this research, which uses imagery as the primary object of study, are the authors' two methods of visual analyses: (1) the field of semiotics, the study of signs and symbols, with a particular focus placed on the covert symbolism displayed within propaganda; and (2) the field of visual rhetorical criticism, the study of how visual texts, images and symbols make arguments and persuade viewers, with a particular focus on the rhetorical force of propaganda materials. Both semiotics and rhetorical criticism as modes of analysing visual content provide important methods of examination used to extract the obvious and the concealed influences within visual material (Rose, 2012). As our analysis will demonstrate some of the levels of significance were relatively neutral or objective, whereas others were saturated with social meaning and political discourse. The recognition and elucidation of these different meanings involves analysis and decoding, which depends on the nature and knowledge of the researcher's interdisciplinary experience. The images have been carefully selected to demonstrate the ways in which extremists use evasive means, that is how they code, cloak and obscure their ideologies, to circumvent platform restrictions. The ethical considerations that relate to this research have been outlined in full in the ethics application form (ERGO/47657), which received ethical approval at both Faculty and Research Integrity Governance level on 15 March 2019.

SHORT CASE STUDIES: HUMAN CREATIVITY AND MANIPULATING MODERATION

With broadening the scope of the content moderation discussion in mind, it is necessary to discuss the problematics of how human actors can creatively evade, bypass, or 'game' content moderation practices. A well-known example of this practice is hashtag hijacking, a tactic whereby a user (or group of users) attaches a 'trending' hashtag to unrelated content in an effort to get their message to circulate more widely. This was used very effectively by Islamic State who employed hashtags such as #Brazil2014 or #WC2014 to circulate propaganda and increase the visibility of its message on Twitter (Veilleux-Lepage, 2016). This technique has continued to be used in a variety of ways, but most notably by QAnon conspiracy adherents in the summer of 2020 when they hijacked #SavetheChildren to circulate their calls for worldwide street protests in relation to their conspiracy theories regarding child trafficking. The #SavetheChildren hashtag hijack also highlights the material (economic) effects that can be created in this process, as the legitimate Save the Children charity campaign was negatively impacted both online and financially by erroneous calls and website traffic that overwhelmed their capacity for an extended period preventing them from doing their actual work (Rogers, 2020). However, moderating Islamic State or QAnon content in the hashtag hijack is problematic because the hashtag's use is not solely for extremist purposes and legitimate speech would also be removed if all content employing the hashtag was suppressed or taken down. Moreover, using human content moderators to review all the hashtagged material poses problems in terms of the large scale of posts and ensuring the speed of content delivery to end users. Thus, neither technological nor human moderation alone poses a viable solution. These cases show how effective human creativity is when applied to leveraging the capacities of technological affordances (in this case hashtag circulation) embedded in systems.

To explicate more fully ways that extremist actors use manipulative practices online, we identify and describe three additional strategies used to circumvent content moderation. These, we have categorized as numerology (the process of attaching ideological meaning to numbers), borderlands (participation in and leveraging of digital communities that act as a border between normative and extreme cultures online) and merchandising (marketing, influence and sales of extreme ideological products packaged as normative consumer products). These three

strategies, such as hashtag highjacking, are used to circulate extremist content in ways that subvert technological content moderation techniques by making the content difficult for machines to concretely categorize as prohibited. This section begins with examining the case study of numerology.

Numerology

The first strategic practice we describe is 'numerology', the process of attaching meaning to numbers, which has a long and storied tradition in religion, cults and conspiricism (Uscinski & Parent, 2014). Extremists often employ codes to mark their territory and communicate with one another, making numerology an effective approach in concealing motive and evading platform restrictions. Some of the most prominent examples in white supremacist propaganda include the use of the number 14, referring to the 14 words, a popular slogan coined by David Lane, a member of the terrorist group The Order 'We must secure the existence of our people and a future for white children' (Michael, 2009). Similarly, the number 18, referring to the first letter of the alphabet A and the eighth letter H, represents Adolf Hitler, as does the use of 88 standing for HH, to signify 'Heil Hitler'. The code 33 is also used to represent the Klan, because K is the 11th letter of the alphabet and the use of 11 three times numerically makes up 33, thus 33 makes a sensible reference to 'KKK'. The two entries 14 and 88 used in combination, indicate a commitment to both the white supremacy of the 14 words and the ideology of national socialism, this numeric reference is almost uniquely free of alternative meanings, outside of a context where one might refer to the year (Centre for Analysis of the Radical Right, 2020).

These examples help demonstrate that while social media algorithms have been somewhat effective at detecting certain types of hateful and abusive content, there are holes in their defences. Indeed, similar tactics to numerology have been used in word, punctuation and emoji codes to avoid repercussions from social media companies. For example, the coded term 'jogger' was used to reference Ahmaud Arbery (a Black man who was murdered during a racially motivated hate crime while jogging) in extremist online discussions of his murder (Owen, 2020). Another prominent example is the use of multiple paratheses—the echo symbol—a typographical practice used by some antisemites online, typically consisting of three pairs of parenthetical brackets around an individual's name or a phrase such as 'Banker' to indicated that the person is Jewish (Anti-Defamation League, 2023). Similarly, emoji depicting specific images can be used to evade content moderation. Following the UEFA Euro 2020 final wherein three young Black players missed penalty shots triggering an onslaught of racist abuse that made use of the monkey, banana, fried chicken and watermelon emojis as racist symbols that the automated content moderation systems could not detect (Levingston, 2021). The key obstacle for AI models is learning how the same symbols (whether numbers or emoji) can be used in hateful and nonhateful contexts. Teaching an AI model to distinguish between these contexts is paramount in protecting against the considerable harm experienced by victims of online abuse.

Numerology is particularly prevalent within memes, which are successful rhetorical tools that create collectives, while also dividing people through antagonistic methods, which can foster notions of ingroups and outgroups (Daymon, 2020). Thus, although imagery may not necessarily look harmful on the outside, it often contains numerical covert meanings designed to carry violent messages. QAnon, an extreme meta-conspiracy belief system often co-opted by far-right extremists, provides multiple examples of the expansion of numerology as a strategy. Davis (2021) released a report for 'Hope not Hate', which drew attention to a large network of coordinated Telegram channels known as SABMYK (See All Bradley Make Yahweh Known), disguised as pro-QAnon accounts but actually promoting a

new esoteric mythology. Closely linked with the pseudo-religious narratives is the messianic figure Sabmyk—the Orion King, posited as the alleged hero of 'Noah's prophecy' which asserts that the 'ruler of rulers' and 'prophet of prophets', would become conscious on 21 December 2020—the day on which the time of 'Orion' would begin. The SABMYK propaganda below contains the number sequences of 2-2-21 (the day the content was posted) and 4-24-28 (the date of the alleged 'awakening'), along with the written phrases 'Sabmyk woke up' and 'Noah's Prophecy' (Figures 1 and 2).

The first image is paying reference to the ninth chapter of the Book of Genesis, versus 21-27—Noah's Prophecy, and specifically the Curse of Ham; according to mythology, having seen his father naked, Ham was cursed to have his black descendants be forever slaves (Whitford, 2016). The entwining of this folklore with religious dogma helped legitimize racial stigmatization as 'scripture', 'gospel' and to many, the 'word of God'. As a result, for hundreds of years, the biblical story of the curse of Ham was used as justification for serfdom, slavery and human bondage of the black races (Goldenberg, 2003). Thus, any

FIGURE 1 Biblical numeric imagery endorsing racism and white supremacy.

FIGURE 2 Messianic numeric imagery linking biblical racism to a new esoteric mythology.

imagery containing this specific code of numbers, while flagging as biblical passages, and therefore not extreme, could be subliminally endorsing racism and white supremacy.

The numerical references on the second image—2-2-21—relate to the notion of the second coming (a messianic prophecy of Christian eschatology regarding the return of Jesus following his ascension into heaven). Both images also contain references to the obscure 'age of Orion', a concept originating in ancient Egyptian mythology that asserted that the souls of the deceased found peace in the constellation of Orion, from where the 'prophets of prophets' would emerge to ignite the 'second coming'. The alleged date detailed in the first images—4-24-28—in which Jupiter and Saturn will align at zero degrees of Aquarius. Supporters of QAnon promote the account that in Vedic astrology, Saturn represents the Lord of Karma, who is said will contribute to the 'Great Awakening'—rhetoric that refers to the reckoning former President Trump was expected to bring an end to what is seen as the satanic cabal that has infiltrated the US government (Amarasingam & Argentino, 2020). SABMYK propaganda is a stunningly clear example of the degree to which conspiracy theorists have utilized numerology—numerical symbols as coding for dangerous content—in their attempt to weave narratives combining occult prophecies and a weaponised historiography to support claims of white power and circumvent take-down restrictions. The study will now turn its attention to our second case study: borderlands.

Borderlands

The second strategic practice we describe, one that has become regularly used by far-right actors online (and offline), we have termed 'borderlands'. This is when extremist ideologues and activists leverage events, narratives, or political hot topics to engage nonextremist populations, or rather extreme-adjacent populations, what Berger (2018) calls 'the eligible ingroup'. These populations often share specific types of ideological affinities (e.g., antifeminism, 'tradition', mixed martial arts, environmental, health or food purity) with extremist cultures that can be used to produce narrative resonances. Borderlands, thus refers to the edges of normative digital cultures that border extremist digital cultures. Importantly, the borderlands shift as extremists work to reposition the 'Overton Window', or acceptable political discourses (Beck, 2010), often using 'culture wars' framing around specific identity or rights-based topics (e.g., anti-Critical Race Theory, antivaccine, antimasking and anti-Trans/LGBTQ movements).

One of the increasingly researched borderland communities for digital cultures of far-right extremism is #TradWife culture online.[1] The moniker 'trad' stands for 'traditional' marking a specific gendered framing of identity and a conservative worldview within the culture. #Trad is popular in 'European/Anglo' digital contexts and non-Western digital contexts including in Hindutva (Hindu nationalist) cultures (Leidig, 2021). #Trad is a reactionary movement rooted in restoring gender roles supposedly as a basis for increasing personal satisfaction—here the notion of 'choice' is used to push against claims that the culture is promoting regressive ideas. More problematically, this line of individualized reasoning is easily transposed to social reasoning positing that satisfaction with daily life is best achieved through restoring a better, more 'natural' (or 'God-given') social order rooted in the nuclear family and a gendered division of labour, both in relation to employment and reproduction (Mattheis, 2021). This 'social' narrative framing of return to a golden age of social order characterized by 'proper' (gender, race, class, citizenship, religious and so on) roles aligns with and is promoted by extreme ideological #Trad participants from far-right extremist cultures such as white supremacy/identitarianism, religious or ethno-nationalism and misogynistic extremism (Mattheis, 2018).

Importantly, the discursive framing of this 'return' to tradition is rooted in postfeminist sensibilities that position women as already equal (postfeminism), so a return to femininity is both acceptable and desirable (McRobbie, 2009). Not surprisingly, women who participate in #Trad culture espouse a virulent antifeminist stance pushing back against what they argue is a culturally sanctioned (woke culture) suppression and belittling of their lifestyle choices (Mattheis, 2021). Such antifeminism is a primary ideological overlap with far-right, typically religious, ethno-nationalist and supremacist cultural ideologies. Antifeminism as a discourse, then, provides the primary border-crossing point between #Trad and far-right extremist, neo-fascist, religious nationalist or White supremacist digital cultures, particularly through female influencers within both digital communities (e.g., influencers such as Wife with a Purpose, the Transformed Wife, Blond in the Belly of the Beast and so on). Narrative and visual arguments within the culture often also pose choices between femininity and feminism as irreconcilable binary opposites using a rhetorical form of good versus evil, which again mimics extreme ingroup versus outgroup framing making it easily co-opted by extremist members (see Figure 3).

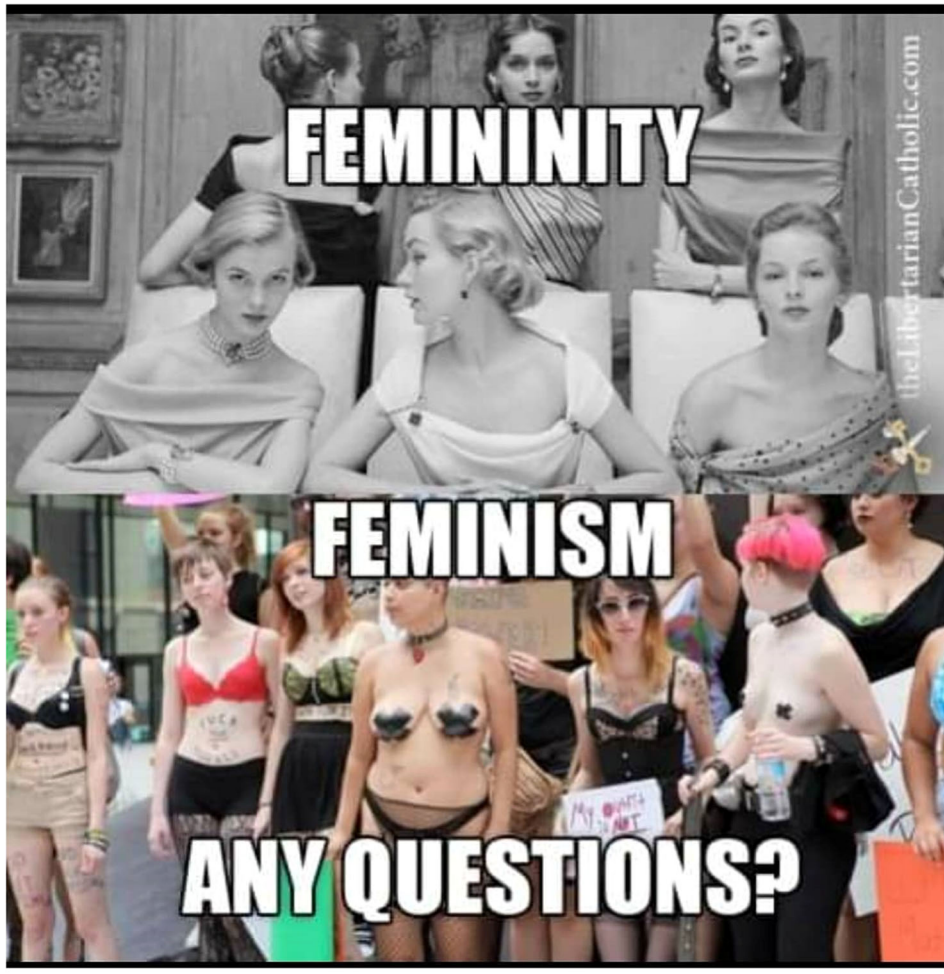

FIGURE 3 Antifeminist Images.

#Trad culture, however, also includes nonextremist influencers whose focus on 'traditional' gender roles, submission to heads of households (husbands) and domesticity provide (whether intentionally or not) cover for extremist ideas and narratives by disassociating them with race (or other markers of 'difference'). That is, #Trad discourse is not explicitly racist, or racialized at all (does not use racial term), while simultaneously using imagery of white families, mothers and children, to express notions of 'traditional' social roles. Here the meaning of 'traditional' is thus left amorphous when compared with extremist rhetoric, because narratives assume whiteness as the norm rather than discuss race, heterosexuality, citizenship or other extremist concerns directly. Rather, #Trad discourse uses narratives of 'tradition' tied to implicitly dominant characteristics embedded in national or cultural identity and in Anglophone contexts link 'tradition' with advertising images of happy white families from the 1950s or 'cottage core' imagery of the pioneer (settler colonial) days of yore (Mattheis, 2021). The resultant messaging implicitly promotes notions of a golden white past but provides somewhat plausible deniability when #Trad culture is critiqued as discriminatory and extremist (see Figure 4).

The 'borderlands' strategy has quite a few benefits including normalizing extreme positions by laundering them through quasi normative, but radical communities. For

FIGURE 4 White-centric Tradition Images.

example, using discourses of 'tradition' rather than 'race' to circulate notions of great replacement and civilizational decline (Mattheis, 2021). This strategy and its discursive tactics have contributed to the development of so-called 'pick and mix' ideological fragmentation as messaging is altered to suit a wider variety of extreme-adjacent affinities. Simultaneously, it effects a form of consolidation, broadly speaking by interweaving connections across a wider swath of digital and political networks. Finally, it can provide extremist ideologues continued access to and influence on mainstream platforms through their relationships in borderlands communities. We now turn our focus to the last strategic practice we discuss: merchandising.

Merchandising

The last of the three strategic practices that we outline in this study is 'merchandising'. This strategy is comprised of a series of at least three tactics—'hatejacking' (brand co-optation), influencer marketing and productization (direct sales)—adapted to support extremist causes and enabled by digital practices and norms. Most research focuses on digital radicalization, recruitment and incitement to violence, each of which can be understood as forms of commodifying ideology. Less work focuses on the productization of hate in either offline or online spaces (Miller-Idriss, 2022). To be clear, this strategy makes use of one of the primary intended functions of digital, socially networked technologies: consumer sales.

The first tactic that makes up merchandising is the long-used practice of 'branding', or in this case appropriating existing brands to create extremist group identity and affiliation markers. Benton and Peterka-Benton (2020) use the term 'hatejack' to describe extremist co-optation of nonextremist brands as group identity markers. Brand co-optation is particularly, but not solely, tied to clothing both historically (e.g., skinhead cultures of the 1980s) and contemporarily (e.g., the Proud Boys hijack of Fred Perry shirts). This has the benefit of creating a cohesive identity marker that can be recognized by both other ingroup members and the out-group. Moreover, as Miller-Idriss (2018) has shown in her book *The extreme gone mainstream: Commercialization and far right youth culture in Germany*, extremist use of clothing also intersects with the tactic of direct product sales and the strategy of numerology as a method of evading social taboos and legal bans.

In digital contexts specifically, 'merchandising' is increasingly important because the entanglement of narrative and product consumption has become the essential function of digital and social media. Here, influence is mobilized to create informational 'products' such as podcasts, vlogs and streaming radio shows. Becca Lewis extensively maps what she describes as the 'alternative influence network (AIN)' on YouTube showing how the deeply socially networked group of alt-right influencers commodify hate ideology and coordinate with each other to generate and support audience and channel growth (Lewis, 2018). The AIN has expanded beyond the confines of YouTube and now many influencers maintain operations on multiple streaming sites ostensibly as a protective measure against banning or de-platforming, although this also has the effect of reaching wider audiences by making products available to them on their preferred platforms. The informational products created by this influence network can also be reproductized through their deployment via digital streaming and hosting platforms—particularly newer crypto-enabled block-chain platforms such as Odysee and DLive—which embed mechanisms for monetization through advertising and audience support (Leidig, 2021).

The third tactic of merchandising is the direct sales of products from clothing to accessories and from books to music, and even dietary supplements (see Squirrel & Martiny, 2022), on sales platforms ranging from niche creator sales sites (e.g., Etsy or Redbubble) to massive global sales platforms (e.g., Amazon or Ebay). So, the digital

commodification of hate is big business for both extremists and the online platforms that host their products. Recommender and browser features have a significant influence on the range of content to which audiences are exposed and, because such content is not restricted to clandestine areas of the Dark Web or encrypted platforms, it instinctively appears more legitimate. Slogan and image-based products comprise a range of hate merchandise that is available to suit many ideologies and scores of products from t-shirts and clothing to accessories, art and small wares. Presented frequently as 'humorous', these products are often easily accessible through simple keyword searching on sales platforms even when openly hateful and sellers often use make (or print) to order schemes on smaller platforms (Squirrel & Martiny, 2022). Sometimes products use coded hate (as with numerology) to 'fly under the radar', as well as using personalisable product order formats that also enable the evasion of product delisting (Squirrel & Martiny, 2022).

A case example of this less studied area of merchandising comes from the misogynist extremist cultures of the so-called 'Manosphere'. This digital culture is comprised of four primary ideological variants, Men's Rights Activists (MRAs), Men Going Their Own Way (MGTOWs), Pick Up Artists (PUAs) and misogynist Involuntary Celibates (Incels), which share deeply antifeminist beliefs that men and masculinity are under attack in contemporary society. Each variant differs in its response to this threat, but all pose binary gender relations (man/woman) as an existential zero-sum crisis. These groups have a long history of productizing their beliefs through the writing and publication (often self-publishing) of books that range from ideological tracts, socio-cultural 'exposés', self-help style texts and programs, through to dystopic and utopic fantasies novels (see Figure 5). A particular favourite of one MRA author, a prolific self-publisher, is writing rape fantasies about Anita Sarkeesian an identified 'enemy' of the Manosphere.[2] These texts can be easily found on Amazon through simple keyword searches.

Along with books, the years of developing digital Manosphere culture has generated wider 'cultural' productization of their ideas and taglines, particularly 'red pill' slogan products that enable consumers to participate in the ideology through its commodification of their identification with the Manosphere (see Figure 6).

The range of merchandising tactics and commodities (identity, information, products) indicates that there are more benefits to the strategy than simply increasing monetary value. As noted by Brenton and Petreka-Brenton (2020), Miller-Idriss (2018) and Betuel, 2020,

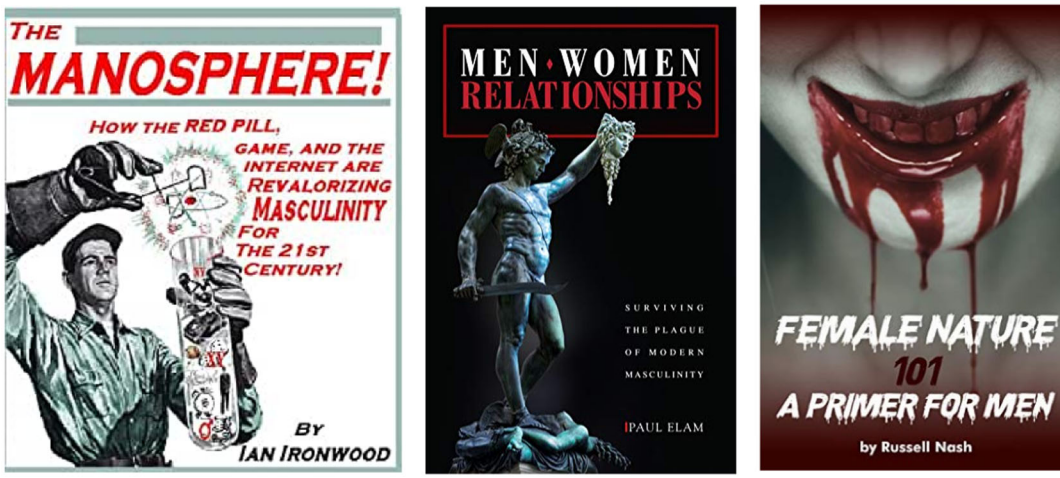

FIGURE 5 Male supremacist ideological texts.

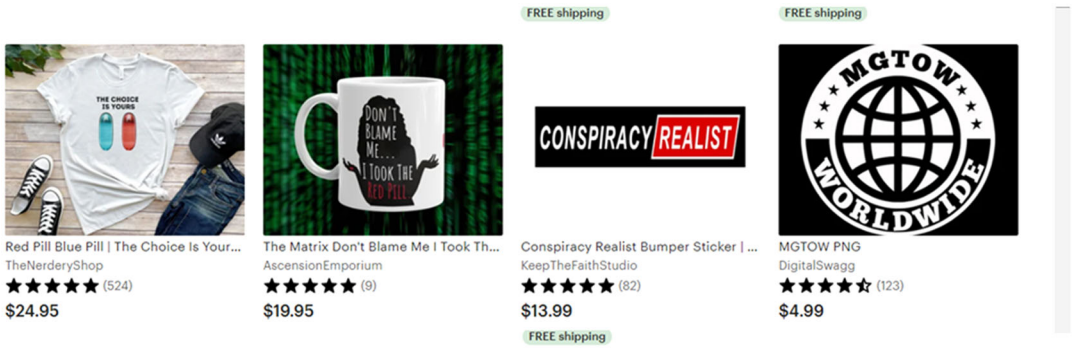

FIGURE 6 Productised extreme ideology.

generating social identification and bypassing taboos or restrictions are concrete benefits of hatejacking and commercialization. Furthermore, engaging with 'borderland' cultures provides pathways for radicalization and recruitment along with commodification, a prime benefit also noted by Miller-Idriss (2022) in her discussion of mixed martial arts as a site of extremist engagement. Along with this, the benefits and impact of influence should not be overlooked. Crucially important to the rise of the alt-right and contemporary resurgence of far-right extremist ideology is the amplifying boost of coordinated, deep networking in the AIN as described by Lewis (2018). Demonetizing (like deplatforming) alone will not address the drivers of merchandising as a strategy. While increased targeting of funding streams is a crucial step, it must be part of a more holistic strategy that addresses the socio-cultural benefits of merchandising.

With these three strategies explicated, we turn to a discussion of ways that researchers, policymakers and technology companies can widen their views of the factors involved in content moderation, as well as the drivers of manipulating moderation. This discussion requires a broader approach to digital technology and practices as a whole, endorsing holistic approaches to moderation.

CONCLUSION AND RECOMMENDATIONS

To better understand how technological content moderation is evaded and manipulated by extremists, more researchers and stakeholders should focus on the nature of digital and social media as sociotechnical systems. It is essential to move away from prioritizing technologically deterministic solutions which only address part of the problem. Human creativity and goal-seeking cannot be addressed successfully through such solutions. For example, user-adapted clips and user-reposting have stymied platforms' attempts to remove all versions, snippets and references drawing on the Christchurch attack livestream, even those with the largest budgets (Konig, 2023). This is a crucial example because the Christchurch livestream video is one of the highest profile pieces of terrorist content—the livestreaming of an attack that killed 51 people at two mosques in Christchurch New Zealand on 15 March 2019—on the web and most of the large platforms have signed onto the Christchurch call as a commitment to taking action.[3] This means most varieties of technological solutions have been used in attempts to permanently remove the Christchurch attack related content. However, because the motivation for circulating, modifying and re-using the content is a human, that is a social problem, technology alone does not provide a comprehensive solution (Mattheis, 2019).

To better understand the human factors undergirding moderation manipulation (and failures), more emphasis needs to be focused on work found at the intersection of Science and Technology Studies and Media Studies, specifically by researchers working in the areas of Extremism, Terrorism and Security studies. These areas of research place an emphasis on digital and social media as sociotechnical systems and human communicative practices. Thus, there is more literature incorporating the social side of the sociotechnical problem.[4] This is not to say that artificial and augmented intelligence systems do not have the potential to make a real difference in the countering violent extremism landscape. They do. However, AI brings with it a wealth of ethical and societal implications to consider with regard to this research (e.g., human enhancement, algorithmic biases, risk of detriment) and in people's lives. What distinguishes AI from traditional sociotechnical systems is the presence of algorithms and technical rules that manage their interaction with the other elements of the system. The way artificial agents learn may not be understandable to human actors, making uncertainty and unpredictability more prevalent in AI. Understanding AI as a sociotechnical system acknowledges that the processes used to develop technology are more than their mathematical and computational constructs. Thus, a sociotechnical approach must consider the values and behaviour modelled from the data sets, the humans who interact with them and the complex organizational factors that go into their commission, design and deployment.

A growing body of evidence also demonstrates that algorithmic systems can propagate racism, classism, sexism, ableism and other intersecting forms of discrimination and forms of oppression that cause real-world harm (Costanza-Chock et al., 2022). Advances in technology and software are rarely inherently bad in themselves, yet that unfortunately does not preclude them from being subverted to ill intent by others, even an unintentional lack of care towards ethical codes and algorithmic accountability can lead to societal and ethical implications. It is therefore urgent from a policy (and practical) perspective to begin viewing the issues of content moderation, especially evasion and manipulation of moderation practices and policies, as driven by human concerns and goals rather than predominantly as technological problems. AI is often constructed from grossly biased and decontextualized information and ideas that can be harmful to the public when turned into automated decision-making systems (Noble, 2018). The current homogeneity of computer scientists and AI experts drawn from dominant groups makes it very difficult to control for and reduce social biases in technology. Moreover, there are few technical solutions to mitigate the problem. Problematically, AI already effects the lives of all social groups, often reifying social harms through differential treatment. As such, it is essential increase the diversity amongst the AI workforce and the stakeholders designing and building these technologies. Thus, more emphasis should be placed on the development of the AI workforce and the training of experts in a socially inclusive way by integrating people from varied backgrounds and experience to ensure better, less biased AI systems with the aim of reducing social biases embedded into the technologies.

Automated content moderation can, or rather should not be fully automated given that false negatives and positives that ban allowable content and miss disallowed content (United Nations, 2021). Moreover, as this study shows, human creativity will always enable actors to find ways around technological content moderation, platform policies and regulations. Thus, human-in-the-loop strategies should be considered best practice. It is also important to explore the sociotechnical challenges of deploying AI technologies on the Web for use in cross-sector ways. Law enforcement, cybersecurity and defence applications of AI often involve decision making that can have significant human impact. Such decisions need support from robust tools and intelligence products, where potential for bias, error and missing data are made clear so that decisions made can be both informed and proportionate. AI is critical to tackle the massive volumes of data from the Web, allowing filtering, summarizing and modelling for use by human analysts and decision makers. However, AI must be deployed with care and needs to be trusted

along with the bias/error of results being understood. Sociotechnical AI systems offer the chance for 'human in the loop' solutions, overcoming some of the problems associated with black box AI. This study therefore emphasizes the need for 'Augmented Intelligence', the combined intelligence of man and machine, when examining the potential impact that computational advances have on preventing online extremism as well as the ways that individuals potentially game the system.

As this study makes clear, the decision-making of AI technology, the people running and developing it, and content moderation practices need oversight. Algorithmic audits are an increasingly popular mechanism for algorithmic accountability. AI audits can help identify whether algorithmic products and systems meet, or fall short of expectations in the areas of bias, effectiveness, transparency, direct impacts on vulnerable communities, data consent, security and access and regulatory compliance (Costanza-Chock et al., 2022). Many large tech companies have established specific auditing teams within their engineering departments. Therefore, in moving forward, it is recommended that companies have mandatory independent AI audits against clearly defined standards applicable to all AI products and operators. It is also crucial for audits to consider the real-world harm caused by the technologies and stakeholder participation, in particular by communities most likely to experience harm from the system, product or tool that is being audited.

ACKNOWLEDGEMENTS

The authors wish to thank Dr. Pamela Ugwudike for providing feedback on earlier versions of this study. The authors gratefully acknowledge the support received from Swansea University's Legal innovation Lab Wales (which is in part funded by the European Regional Development Fund through the Welsh Government).

ORCID

Ashley A. Mattheis http://orcid.org/0000-0002-2919-0712

ENDNOTES

[1] See for example: Christou, 2020; Ebner, 2020; Leidig, 2021; Mattheis, 2018, 2021; Tebaldi, 2021; Veillieux-Lepage et al., 2022.

[2] Anita Sarkeesian is a well-known feminist critic of gaming culture who was identified as an 'enemy' and targeted for violent, sexualized harassment during the first coordinated networked harassment campaign known as #GamerGate, which occurred in 2014. Many digital cultures of far right and male supremacist extremism have connections to #GamerGate cultures and leverage those topics given their shared interests including antifeminism and misogyny, as well as overlapping cultural concerns such as gaming and perfecting abusive online techniques.

[3] See: https://www.christchurchcall.com/our-community/countries-and-states/.

[4] Our research draws heavily from this literature including work mapping the 'Alternative Influence Network' (Lewis, 2018), which highlights the ways that extremists have taken up and become highly successful at using social and digital media, networking and monetization. Notions of 'convergence' (Jenkins, 2006) and 'spreadability' (Jenkins et al., 2013), which explore the interactivity of social and technological aspects and how that produces different meaning making through tech-enabled user-driven practices of participatory mediation. Along with this, research in this vein has focused on the social and cultural impacts of our changed information and media environment including information pollution and amplification (Phillips & Milner, 2021; Phillips, 2018), the ways digital platforms have monetized user response (likes) enables, even promotes, information warfare in efforts to control the 'truth' (Singer & Brooking, 2018) and the power of visual grammars in meme cultures (Milner, 2018).

REFERENCES

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236.

Amarasingam, A., & Argentino, M. A. (2020). *The QAnon conspiracy theory: A security threat in the making?* (Vol. 13). CTC Sentinel. 7.

Amarasingam, A., Maher, S., & Winter, C. (2021). *How Telegram disruption impacts Jihadist platform migration*. Retrieved January 5, 2023, from https://crestresearch.ac.uk/resources/how-telegram-disruption-impacts-jihadist-platform-migration/

Anti-Defamation League. (2023). *Echo*. Retrieved October 15, 2023, from https://www.adl.org/resources/hate-symbol/echo

Ayling, J., & Chapman, A. (2022). Putting AI ethics to work: Are the tools fit for purpose? *AI and Ethics*, *2*(1), 405–429.

Beck, G. (2010). *The Overton window*. Simon and Schuster.

Benton, B., & Peterka-Benton, D. (2020). *Hating in plain sight: The hatejacking of brands by extremist groups* (Vol. 83). Department of Justice Studies Faculty Scholarship and Creative Works. https://digitalcommons.montclair.edu/justice-studies-facpubs/83

Berger, J. M. (2018). *Extremism*. MIT Press.

Betuel, E. (2020). Can masks spread conspiracies: Inside the world of QAnon Merch. *Inverse*. Retrieved November 3, 2020, from https://www.inverse.com/culture/qanon-masks

Bloom, M., Tiflati, H., & Horgan, J. (2019). Navigating ISIS's preferred platform: Telegram. *Terrorism and Political Violence*, *31*(6), 1242–1254.

Busuioc, M. (2020). Accountable artificial intelligence: Holding algorithms to account. *Public Administration Review*, *81*(5), 825–836.

Centre for Analysis of the Radical Right. (2020). *A guide to online radical-right symbols, slogans and slurs*. Retrieved January 1, 2023, from https://www.radicalrightanalysis.com/wp-content/uploads/2020/05/CARR-A-Guide-to-Online-Radical-Right-Symbols-Slogan-and-Slurs.pdf

Christou, M. (2020). #Tradwives: Sexism as a gateway to White supremacy. *Centre for Analysis of the Radical Right Blog*. https://www.radicalrightanalysis.com/2020/03/23/

Church, K., & Liberman, M. (2021). The future of computational linguistics: On beyond alchemy. *Frontiers in Artificial Intelligence*, *4*, 625341. https://www.frontiersin.org/articles/10.3389/frai.2021.625341/full

Costanza-Chock, S., Raji, D., & Buolamwini, J. (2022). Who audits the auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. *ACM Conference on Fairness, Accountability, and Transparency*, 1(1), 1–13. Retrieved April 17, 2023, from https://dl.acm.org/doi/pdf/10.1145/3531146.3533213

Davis, G. (2021). *The Sabmyk network: How a mysterious disinformation network is hijacking QAnon*. Retrieved January 1, 2023, from https://hopenothate.org.uk/2021/02/12/the-sabmyk-network-how-a-mysterious-disinformation-network-is-hijacking-qanon/

Daymon, C. (2020). *LOL extremism: Humour in online extremist content*. Retrieved January 1, 2023, from https://gnet-research.org/2020/10/26/lol-extremism-humour-in-online-extremist-content/

Ebner, J. (2020). Tradwives: Joining the female anti-feminists, *Going dark: The secret social lives of extremists* (pp. 59–78). Bloomsbury.

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gillespie, T. (2020). Content moderation, AI and the question of scale. *Big Data & Society*, *7*(2), 205395172094323.

Global Internet Forum for Counter Terrorism. (2023). *GIFCT's hash-sharing database*. Retrieved October 18, 2023, from https://gifct.org/hsdb/

Goldenberg, D. M. (2003). *The Curse of Ham: Race and slavery in early Judaism, Christianity, and Islam*. Princeton University Press.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 1–15.

Gunton, K. (2022). The use of artificial intelligence in content moderation in countering violent extremism on social media platforms. In R. Montasari (Ed.), *Artificial intelligence and national security* (pp. 69–79). Springer. https://link.springer.com/chapter/10.1007/978-3-031-06709-9_4

Hagey, K., & Horowitz, J. (2021). Facebook tried to make its platform a healthier place. It got angrier instead. *Wall Street Journal*. Retrieved January 4, 2021. https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215

Hall, W., & Pesenti, J. (2017). *Growing the Artificial Intelligence Industry in the UK*. Retrieved January 1, 2023, from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/652097/Growing_the_artificial_intelligence_industry_in_the_UK.pdf

Holistic, A. I. (2023). *Lost in Transl(A)t(I)on: Differing definitions of AI*. Retrieved October 14, 2023, from https://www.holisticai.com/blog/comparing-definitions-of-ai#:~:text=With%20another%20lengthy%20definition%2C%20here,developed%20in%20any%20context%2C%20including%2C

Jenkins, H. (2006). *Convergence culture: Where old and new media collide*. New York University Press.

Jenkins, H., Ford, S., & Green, J. (2013). *Spreadable media: Creating value and meaning in a networked culture*. NYU Press.
Jhaver, S., Boylston, C., Yang, D., & Bruckman, A. (2021). Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter. *Proceedings of the ACM on Human-Computer Interaction*, *5*(1), 1–30.
Karizat, N., Delmonaco, D., Eslami, M., & Andalibi, N. (2021). Algorithmic folk theories and identity: How TikTok users co-produce knowledge of identity and engage in algorithmic resistance. *Proceedings of the ACM on Human-Computer Interaction*, *5*(no. CSCW2), 1–44.
Keller, D., & Leerssen, P. (2020). Facts and where to find them: Empirical research on Internet platforms and content moderation. In N. Persily, & J. A. Tucker (Eds.), *Social media and democracy: The state of the field and prospects for reform* (pp. 220–252). Cambridge University Press.
Kingdon, A., & Ylitalo-James, E. (2023). Using social media to research terrorism and extremism. In L. Frumkin, J. Morrison, & A. Silke (Eds.), *A research agenda for terrorism studies* (pp. 131–143). Edward Elgar.
Kitchens, B., Johnson, S. L., & Grey, P. (2020). Understanding echo chambers and filter bubbles: The impact of social media on diversification and partisan shifts in news consumption. *MIS Quarterly*, *44*(4), 1619–1649. https://doi.org/10.25300/MISQ/2020/16371
Klinenberg, D. (2022). *Does deplatforming work? Unintended consequences of banning far-right content creators*. Retrieved January 4, 2023, from https://ideas.repec.org/p/pri/esocpu/31.html
Knight, W. (2016). AI's language problem. *MIT Technology Review*. Retrieved January 4, 2023, from https://www.technologyreview.com/2016/08/09/158125/ais-language-problem/
Knight, W. (2019). The mass shooting in New Zealand shows how broken social media is. *MIT Technology Review*. https://www.technologyreview.com/2019/03/15/65970/the-mass-shooting-in-new-zealand-shows-how-broken-social-media-is/
Konig, J. (2023). Report charges TikTok failing to moderate extremist content. *Spectrum News NY1, Charter Communications*. https://www.ny1.com/nyc/all-boroughs/politics/2023/03/22/report-charges-tiktok-failing-to-moderate-extremist-content
Leidig, E. (2021). *TradWife, TradLife: Women of the Alt-Right*. Bloomsbury Academic.
Levingston, I. (2021). *Social media fails to curb racist emojis aimed at soccer stars*. Retrieved October 15, 2023, from https://www.bloomberg.com/news/articles/2021-07-15/twitter-facebook-struggle-to-control-racist-use-of-emojis?leadSource=uverify%20wall
Lewis, R. (2018). *Alternative influence networks*. Data and Society. https://datasociety.net/library/alternative-influence/
Mattheis, A. A. (2018). Shieldmaidens of whiteness: (Alt)Maternalism and women recruiting for the far/alt-right. *Journal for Deradicalization*, *17*(2018), 128–161. https://journals.sfu.ca/jd/index.php/jd/article/view/177
Mattheis, A. A. (2019). Manifesto memes: The radical right's new dangerous visual rhetoric. *CARR Insights. Centre for Analysis of the Radical Right*. https://www.radicalrightanalysis.com/2019/09/18/manifesto-memes-the-radical-rights-new-dangerous-visual-rhetorics/
Mattheis, A. A. (2021). # TradCulture: Reproducing whiteness and neo-fascism through gendered discourse online, In S. Hunter, & C. van der Westhuizen (Eds.), *Routledge Handbook of Critical Studies in Whiteness* (pp. 91–101). Routledge.
Mattheis, A. A., Gartenstein-Ross, D., Bills, C., & Koduvayur, V. (2022). Blindsided: A reconceptualization of the role of emerging technologies in shaping the tactics, techniques, and procedures of information operations in the gray zone. *Valen's Global International Strategies and Security*. https://valensglobal.com/blind-sided/
McRobbie, A. (2009). *The aftermath of feminism*. Sage.
Michael, G. (2009). David Lane and the fourteen words. *Totalitarian Movements and Political Religions*, *10*(1), 43–61.
Miller-Idriss, C. (2018). *The extreme gone mainstream: Commercialization and far right youth culture in Germany*. Princeton University Press.
Miller-Idriss, C. (2022). *Hate in the homeland*. Princeton University Press.
Milner, R. M. (2018). *The world made meme: Public conversations and participatory media*. MIT Press.
Moonshot. (2023). *The redirect method*. Retrieved October 15, 2023, from https://moonshotteam.com/the-redirect-method/
Noble, S. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.
Owen, T. (2020). *White supremacists have a disgusting new code for the N-word after Ahmaud Arbery's death*. Retrieved October 15, 2023, from https://www.vice.com/en/article/bv88a5/white-supremacists-have-a-disgusting-new-code-for-the-n-word-after-ahmaud-arberys-death
Pearson, G. (2021). Sources on social media: Information context collapse and volume of content as predictors of source blindness. *New Media and Society*. https://doi.org/10.1177/2F1461444820910505
Phillips, W. (2018). *The oxygen of amplification: Better practices for reporting on extremists, antagonists, and manipulators online*. Data & Society. https://datasociety.net/wp-content/uploads/2018/05/FULLREPORT_Oxygen_of_Amplification_DS.pdf

Phillips, W., & Milner, R. M. (2021). *You are here: A field guide for navigating polarized speech, conspiracy theories, and our polluted media landscape*. MIT Press.

Rogers, K. (2020). Trump said QAnon 'Fights' pedophilia. But the group has made it harder to protect kids. *FiveThirtyEight.com*. https://fivethirtyeight.com/features/qanons-obsession-with-savethechildren-is-making-it-harder-tosave-kids-from-traffickers/

Rose, G. (2012). *Visual methodologies: An introduction to researching with visual methods*. Sage Publications Ltd.

Sankhe, A. (2022). Overcoming the challenges of computational linguistics: An interview with Tom Wolf. *Engati Blog*. Retrieved January 4, 2023, from https://www.engati.com/blog/computational-linguistics

Singer, P. W., & Brooking, E. T. (2018). *LikeWar: The weaponization of social media*. Houghton Mifflin Harcourt.

Squire, M. (2023). *Bad gateway: How deplatforming affects extremist websites*. Retrieved October 18, 2023, from https://www.adl.org/resources/report/bad-gateway-how-deplatforming-affects-extremist-websites

Squirrel, T., & Martiny, C. (2022). *Profiting from hate: Extremist merchandise on Redbubble, Etsy, Teespring, Teerepublic, and Zazzle*. Retrieved January 3, 2023, from https://www.isdglobal.org/wp-content/uploads/2022/12/Profiting-from-Hate-Extremist-Merchandise-on-Redbubble-Etsy-Teespring-Teerepublic-and-Zazzle.pdf

Tebaldi, C. (2021). Make women great again: Women, misogyny and anti-capitalism on the right, *Fast capitalism* (Vol. 18, 1). MAVS Open Press. https://fastcapitalism.journal.library.uta.edu/index.php/fastcapitalism/article/view/421

Ugwudike, P. (2022). AI audits for assessing design logics and building ethical systems: The case of predictive policing algorithms. *AI and Ethics*, *2*(1), 199–208.

UK Research and Innovation. (2022). *Artificial intelligence technologies*. Retrieved October 15, 2023, from https://www.ukri.org/what-we-do/our-main-funds-and-areas-of-support/browse-our-areas-of-investment-and-support/artificial-intelligence-technologies/

United Nations. (2021). *Countering terrorism online with artificial intelligence: An overview for law enforcement and counter-terrorism agencies in South Asia and South-East Asia*. Retrieved October 15, 2023, from https://www.un.org/counterterrorism/sites/www.un.org.counterterrorism/files/countering-terrorism-online-with-ai-uncct-unicri-report-web.pdf

Uscinski, J. E., & Parent, J. M. (2014). *American conspiracy theories*. Oxford University Press.

Veilleux-Lepage, Y. (2016). Paradigmatic shifts in Jihadism in cyberspace: The emerging role of unaffiliated sympathizers in Islamic state's social media strategy. *Journal of Terrorism Research*, *7*(1), 36–51.

Veilleux-Lepage, Y. D., Kisyova, M. E., & Newby, V. F. (2022). Conversations with other (alt-right) women: How do alt-right female influencers narrate a far-right identity? *Journal For Deradicalization*, *31*, 35–72. https://hdl.handle.net/1887/3485420

Whitford, D. M. (2016). *The Curse of Ham in the early modern era*. Routledge.

How to cite this article: Mattheis, A. A., & Kingdon, A. (2023). Moderating manipulation: Demystifying extremist tactics for gaming the (regulatory) system. *Policy & Internet*, 1–20. https://doi.org/10.1002/poi3.381