Cognitive assemblages: The entangled nature of algorithmic content moderation

Valentine Crosset, Benoît Dupont

Original Research Article

BIG DATA & SOCIETY

Big Data & Society
July–December:1–13

DOI: 10.1177/20539517221143361
journals.sagepub.com/home/bds
SAGE

Cognitive assemblages: The entangled nature of algorithmic content moderation

Valentine Crosset[1] and Benoît Dupont[2]

Abstract

This article examines algorithmic content moderation, using the moderation of violent extremist content as a specific case. In recent years, algorithms have increasingly been mobilized to perform essential moderation functions for online social media platforms such as Facebook, YouTube, and Twitter, including limiting the proliferation of extremist speech. Drawing on Katherine Hayles' concept of "cognitive assemblages" and the Critical Security Studies literature, we show how algorithmic regulation operates within larger assemblages of humans and non-humans to influence the surveillance and regulation of information flows. We argue that the dynamics of algorithmic regulation are more liquid, cobbled together and distributed than it appears. It is characterized by a set of shifting human and machine entities, which mix traditional surveillance methods with more sophisticated tools, and whose linkages and interactions are transient. The processes that enable the consolidation of knowledge about risky profiles and contents are, therefore, collective and distributed among humans and machines. This allows us to argue that the cognitive assemblages involved in content moderation become a cobbled space of preemptive calculation.

Keywords

Algorithms, regulation, content moderation, platforms, social media, cognitive assemblages

Introduction

In the face of violent extremist speech on digital platforms, algorithms have increasingly been promoted as the most effective approach to perform essential moderation functions at scale, embodying a renewed potential for securing flows (Amoore and Raley, 2017; Gillespie, 2020; Gorwa et al., 2020). From a purely technical and computational perspective, algorithms can be defined as mathematical constructs that act on a data set configured to solve a defined problem and accomplish specific tasks (Gillespie, 2014, 2016; Miyazaki, 2012). Social science studies have approached the issue of algorithms from the perspective of powerful and opaque entities that govern, sort, or control our lives (Ziewitz, 2016). Recent narratives suggest that algorithms construct and enact regimes of power and knowledge (Beer, 2009; Lash, 2007). Algorithms "search, collate, sort, categorize, group, match, analyze, profile, model, simulate, visualize and regulate people, processes and places" (Kitchin, 2017: 11) through "automated management" (Dodge and Kitchin, 2007). Algorithms thus claim and express forms of "algorithmic regulation" (Ulbricht and Yeung, 2022; Yeung, 2018) to "influence behavior or manage risk" (Yeung, 2018), resulting in new opaque regimes of population management, control, discrimination, and exclusion (Gillespie 2014; Mittelstadt et al., 2016; Noble, 2018; Pasquale, 2015).

In this paper, we argue that algorithmic content moderation is entangled in hybrid and collaborative modes of cognition, which are oriented towards threat anticipation and preemptive rationalities. We consider that algorithmic moderation operates for regulatory purposes, "to shape, constrain, and coordinate the behavior of others" and "to manage particular kinds of risk" (Ulbricht and Yeung, 2022: 4). While other forms of algorithmic regulation have received significant attention in the social science literature, critical investigations of algorithmic content moderation remain rare compared to the extensive computer science literature (Rieder and Skop, 2021). Some key works have provided excellent insights into algorithmic content moderation, but they focus on the technologies

[1]Médialab Sciences Po, Paris, France
[2]International Centre for Comparative Criminology, Université de Montréal, Montreal, Canada

Corresponding author:
Valentine Crosset, Médialab Sciences Po, 1, Place St Thomas d'Aquin, 75007 Paris, France.
Email: valentine.crosset@sciencespo.fr

involved and their political and ethical issues (Gehl et al., 2017; Gillespie, 2020; Gorwa et al., 2020; Rieder and Skop, 2021). As a result, some questions remain under-studied: how do moderation practices and their algorithmic devices used in the context of the "war on terror" revive and extend surveillance techniques? Amid the apparent proliferation of algorithmic techniques in content moderation, how do these techniques coevolve with the collaboration of humans, giving rise to new modes of cognition to secure data flows?

This article explores the interweaving faculties of algorithmic content moderation and human operators and the implications of this human-algorithmic cognitive assemblage for the surveillance of violent extremist speech. In particular, we focus on the deployment of algorithmic content moderation in the context of online terrorist content. The dissemination of extremist content related to terrorism provides a unique case study for understanding the role of algorithms in regulating complex phenomena, given the heterogeneity of the actants involved. The data in this article originates from a larger research project that involved conducting a digital ethnography of the Islamic State on social media between 2017 and 2019. One of the project's objectives was to explore the regulation of Islamic State speech. During this period, we collected a series of documentary sources that outlined the terrorist content management practices of Facebook, YouTube, and Twitter. This strategy aligns with Kirschenbaum's (2003) proposal to study software as a "product of material environments." In this perspective, technical objects result from papers, technical specifications, reports, etc. For this research, we rely on four discrete sources: platforms' terms of use, transparency reports, blog posts published by digital platforms and statements from platform spokespersons and owners. Drawing on this documentary material, this paper demonstrates the importance of considering the collaborative and relational nature of algorithmic content moderation (Rieder and Skop, 2021). Our empirical work thus illuminates the processes and relationships of human-machine moderation in a more concrete manner.

This paper contributes to the current debate on algorithmic regulation and content moderation by engaging with the rich literature on critical security and posthuman studies. Critical security studies have critically analyzed how algorithms produce security (Amoore, 2009; Amoore and Raley, 2017; Aradau and Blanke, 2018; Bellanova and de Goede, 2022b). This literature helps capture how algorithms are transforming securitization practices and the governance of populations. Scholars have focused particularly on the functioning of security algorithms and how they produce new forms of political authority (Amoore and Raley, 2017). They have paid attention to algorithms' generative capacities based on analyzing different forms of data (de Goede and Sullivan, 2016; Wilcox, 2017). This literature also describes how algorithmic security decisions are made under conditions of uncertainty about a future threat (Anderson, 2010; Amoore, 2013; Aradau and van Munster, 2007; de Goede, 2012). Therefore, algorithms have been analyzed in this literature as a precautionary or preemptive process. Finally, this literature has signaled how the goal of securing in algorithmic security is always emergent, as informed by the available data rather than legal experts (Aradau and Blanke, 2015; Bellanova and de Goede, 2022b). Although algorithms were already a focus of the content moderation literature, existing contributions say little about how algorithms and content moderation became a *security practice* (Bellanova and de Goede, 2022a). Moderation has mostly been treated as a censorship issue and less as part of a broader set of surveillance and security practices. Conversely, the security and algorithmic regulation literature has yet to incorporate content moderation practices into its framework.

Furthermore, the debates on algorithms within the content moderation literature have, to a certain extent, been isolated from posthuman and feminist analyses. Generally speaking, our approach considers algorithmic content moderation as a heterogeneous set of entities, some of which are social and others technical (Latour, 2005). The literature on content moderation has already highlighted the role of algorithmic and human interventions and the changing nature of power relations between governments, Internet companies, and civil society (Bellanova and de Goede, 2022a, Gillespie, 2018; Roberts, 2019). Hence, to develop our critical approach, it is not enough to point out that algorithms are entangled in human and non-human assemblages. Beyond a better understanding of the relational nature of human and non-human assemblages, it is vital to grasp the features of the relationships that solidify them (Allen, 2011; Aradau and Blanke, 2015). This article adopts the notion of "cognitive assemblages" developed by Hayles (2017) in her work on human and technical cognizers to describe the entanglement of algorithms with other entities and surveillance faculties. From Katherine Hayles' perspective, cognition is a much broader process than thought and consciousness. Katherine Hayles defines cognition as a "process that interprets information within contexts that connect it with meaning" (p. 22). This definition enables Katherine Hayles to make a distinction between cognizers and noncognizers. That is, humans and all forms of biological life, as well as many technical systems on the one side; and material processes and inanimate objects on the other. For her, "the crucial distinguishing characteristics of cognition that separate it from these underlying processes are choice and decision, and thus possibilities for interpretation and meaning" (p. 28). Hayles developed the idea of "cognitive assemblages" to describe "an arrangement of systems, subsystems, and individual actors through which information flows, effecting transformations through the interpretive activities of cognizers operating upon the flows" (2017: 118). However, for the author, these assemblages remain unpredictable due to the

interconnection of humans and technical systems, "the cognitive decisions of each affect [ing] the others" (Hayles, 2017: 118). In Hayles' perspective, algorithms, as nonconscious cognition, are thus thought of as an entangled and collaborative enterprise in which analog and digital forms of computation and cognition cohabit. By putting the surveillance studies literature in conversation with Katherine Hayles' analyses, we show how the cognitive assemblages mobilized for the moderation of extremist content respond to a terrorist risk management rationale, oriented toward anticipatory goals aimed at preemptively disrupting potential terrorist groups (Massumi, 2015).

Ultimately, this article offers a critical reading of algorithmic regulation. Therefore, our contribution to the current debate on algorithmic regulation is to show how the governance dynamics of algorithms are entangled in hybrid and collaborative modes of cognition. This is consistent with Munn's approach (2018) to understanding algorithms as an ecology. Algorithms are not only executed sequentially, line-by-line, but distributed throughout the "ecology's diverse array of heterogeneous actors and agents and executed asynchronously" (Munn, 2018). Considering the heterogeneity of algorithmic regulation, it is possible to demonstrate that the governance of information flows results in a set of mutations and movements, mixing traditional surveillance methods with more sophisticated approaches, whose linkages are not characterized by permanence and stability. Some have expressed concerns about algorithms' apparent power and agential capacity in controlling our lives (Beer, 2009; Gillespie, 2014; Introna and Wood, 2004; Lash, 2007). We do not wish to dismiss this approach, but on the contrary, to take it further, establishing an ontological connection between how materiality, experts, and non-experts operate through the interoperability, continuity, and relationality of human and non-human skills to control and exclude unwanted informational flows.

This article proceeds as follows: the first section traces the context and issues related to the online governance of violent extremist content. The second section describes the use of algorithms to enforce platform norms and policies and how it transforms risk management practices. Then, the third section describes how algorithms are embedded in an extensive network of cognizers, who collectively and in a distributed manner manage information flows. We conclude by arguing that the cognitive assemblage involved in content moderation becomes a cobbled space of preemptive calculation. We finally suggest some future directions for research.

Governing extremist speech in the war on terror

US Senator Joseph Liberman made one of the first calls to regulate terrorist-related content when he asked YouTube to remove videos posted by Al-Qaeda and other Islamist organizations in 2008.[1] While YouTube partially honored his request, the company also stated that it would not remove content sympathetic to those organizations that did not contain explicitly violent and hateful content, as it did not violate community guidelines. YouTube also recalled that the platform "encourages free speech and defends everyone's right to express unpopular points of view" (2008). Therefore, social media platform owners needed to acknowledge the dangers that certain forms of expression can induce. Still, the values of free speech in American constitutional law were unquestionably reflected in the private rules they initially developed (Klonick, 2017). This implied a principle of minimal intervention in content moderation, even concerning terrorist organizations.

After the attacks in Paris (2015), Brussels (2016), and Christchurch (2019), governments increased their efforts to enroll digital platforms in the fight against terrorism (Highman and Nakashima, 2015). Public authorities in countries that have suffered terrorist attacks have expressed strong concerns about technologies that allow terrorist groups to disseminate their propaganda, teach operational skills, recruit and radicalize new members and raise funds (Highman and Nakashima, 2015). In this context, they expected platforms to remove extremist content online more aggressively. European authorities implemented new legislation on illegal content[2] to hold platforms legally responsible for what their users post. In the European case, these initiatives generally provide for heavy civil and criminal penalties if digital platforms do not promptly remove extremists and hate speech. For example, the NetzDG law in Germany requires large digital platforms to remove or block access to "manifestly illegal" content within 24 h of receiving a complaint. In the same vein, the European Commission proposed in 2018 a regulation "on the prevention of the dissemination of terrorist content online," whereby platforms are required to block or remove content within one hour of receiving a removal order issued by a court from one of the Member States.

This approach involves pushing platforms to do more to ensure online security. Several platforms have agreed to review their content policies and implement proactive measures regarding the removal of terrorist material. These normative changes are reflected in the evolution of the language used by major platform owners. Twitter, which has long maintained one of the most tolerant free speech policies, stated on its blog in 2017 that "making Twitter a safer place is our main goal" and that "there is no place on Twitter for violent organizations, including terrorist organizations, violent extremist groups, or individuals who affiliate with and promote their illicit activities" (Twitter, 2019). There are many other examples. Victoria Grand, director of policy strategy at Google, stated in an interview that "the goal […] is how to strike a balance

between allowing people to discuss and access information about ISIS, but also not becoming the distribution channel for their propaganda" (as cited in Highman and Nakashima, 2015). On the other hand, Facebook's head of policy, Monika Bickert, stipulated that "we want to make sure that we keep our community safe and that we are not a propaganda tool" (as cited in Highman and Nakashima, 2015).

Moderation takes on a new meaning: fighting terrorist threats. This is part of a larger effort to counter radicalization, a phenomenon understood by public authorities as potentially catastrophic and dispersed (de Goede and Simon, 2012). A growing body of research has examined the culture of preemption that underlies the governance of terrorism, which has been widely deployed in the wake of 9/11 within the "war on terror" framework (Amoore, 2006, 2009; Aradau and van Munster, 2007; de Goede et al., 2014). Rather than being confined to the future, the logic of preemption aims to "act on multiple potential futures that are made exploitable (or liquid) in the present" (de Goede et al., 2014: 413). In the critical security literature, it is now become commonplace to assert that the preemption of unknown threats—and those recognized as unknowable, requires security interventions that can eliminate the threat before it has emerged (Aradau and van Munster, 2007, 2012; Massumi, 2007, 2015). In other words, preemption refers to a mode of calculation that is oriented toward potential trends rather than focused on possibilities that have actual existence (Massumi, 2007, 2015). As a result, the security practices of digital platforms aim to "anticipate an uncertain future" (Amoore, 2009: 55). Digital platforms must thus relinquish their role as content hosts and develop more active strategies to exclude users that are deemed extremist by public authorities. For this, "companies identify, select, search and interpret suspicious datasets" (Bellanova and de Goede, 2022a: 2). The implementation of such strategies implies the mobilization of a network of cognizers that includes companies, engineers, algorithms, users, terrorism experts, etc., which are articulated through modes of coordination that include community rules and policies, machine learning algorithms, and content flagging features offered to users.

Terrorist content and algorithms

Classifying patterns

For some years, private and public actors have prioritized technology tools, specifically machine learning algorithms, to exclude terrorist content from online platforms. In this regard, Zuckerberg (2017) declared, "artificial intelligence can help provide a better approach. […] This is technically difficult as it requires building AI that can read and understand the news, but we need to work on this to help fight terrorism worldwide." In his first appearance before Congress in April 2018, Mark Zuckerberg (2018) stipulated that these technologies were still in their infancy but that in the coming years, AI will become more powerful to ensure the protection and security of the Internet. In particular, the Facebook CEO hopes for an improvement in AI's ability to distinguish linguistic nuances. Algorithms are therefore expected to enable proactive control over circulating flows of online content. Facebook (2018a) argued: "by using technology like machine learning, artificial intelligence and computer vision, we can proactively detect more bad actors and take action more quickly." Thus, the deployment of computational methods to filter and govern violent extremist content has become one of the prominent features in managing information flows (Facebook, 2017a, 2017b, 2018a; Twitter, 2016; YouTube, 2017a).

Algorithms are increasingly capable of analyzing different types of data (images, texts, videos, audio recordings), making it possible to ensure a wider coercive dimension in the moderation of content. Using these technologies to generate and quantify risky profiles and contents aims to ensure the "good" management of informational flows by excluding harmful elements. The exercise of control is therefore implemented by a "security device" that moves away from the disciplinary techniques exercised in a closed environment and draws inspiration from regulatory techniques focused on risk factors and population management (Foucault, 1978). To control the fluidity of informational flows, algorithms monitor content flows associated with the figure of an "enemy within" (Amoore, 2009). These techniques inevitably create a dichotomy between "good" and "bad" circulation. From then on, the issue at stake for the algorithmic technologies of digital platforms focuses on regulating deviance in information flows.

Currently, large platforms mainly use two types of automated tools in their content moderation (Gorwa et al., 2020). The first is image matching algorithms, which compare new posts to an existing content base. To remove content, platforms use the so-called "hash" technology. If a user uploads a new terrorist video or image, the system checks whether the image matches a known terrorist photo or video. Secondly, platforms have invested in content detection and classification technologies using machine learning algorithms. In particular, they are experimenting with using natural language processing to predict whether a text advocates terrorist propaganda (De Smedt et al., 2018; Schmidt and Wiegand, 2017). These algorithms learn to detect terrorist publications from previously deleted al-Qaeda and Islamic State terrorist propaganda texts. The instructions, this time, will no longer be explicitly programmed by an engineer but will be generated by the machine itself, which learns based on labeled data provided to it (Burrell, 2016; McQuillan, 2015).

Although algorithms are often considered logical series, machine learning algorithms redefine this representation. Amoore (2020) explains that machine learning algorithms

are essentially characterized by the relations among functions and less by the series of steps in a calculation. One of the specificities of machine learning algorithms is their capacity to modify themselves "in and through their non-linear iterative relations to input data" (Amoore, 2020: 11). The consequence of this second approach is that the underlying logic of the algorithm becomes incomprehensible and opaque to its designers and any human observer (Burrell, 2016). This constitutes a significant difference in the management of informational flows. While image matching algorithms require a manual process of collecting and curating specific terrorist images and videos to match, machine learning algorithms "involve inducing generalizations about features of many examples from a given category into which unknown examples may be classified (e.g. terrorist images in general)" (Gorwa et al., 2020: 5). These features can amount to thousands or even millions of variables for some of the most complex models, making human interpretation impossible. Therefore, machine learning algorithms establish probabilities and correlations of what constitutes terrorist speech based on the interferential relationships between information fields in a large volume of random data.

Furthermore, content is not the only indicator digital platforms use to maximize the potential for detecting jihadist accounts. Another way is to use algorithms that detect "terrorist clusters." For example, Facebook (2017a) uses "signals like whether an account is friends with a high number of accounts that have been disabled for terrorism, or whether an account shares the same attributes as a disabled account." In addition, digital platforms have implemented algorithms that identify "repeat offender accounts." This term, which refers directly to criminological language, concerns fraudulent Internet users who repeatedly create false accounts following their suspension. In short, algorithmic interventions are varied and are deployed on different levels; that of the profile, the contents, the relationships or the uses.

Faced with the multiplicity of algorithmic uses in the moderation of extremist content and their capacity to evolve, we cannot rely on a single representation of the algorithm. On the contrary, the algorithmic agents involved in content moderation are highly diversified and specialized. These algorithms constitute a complex "algorithmic ecosystem" (Lange, 2016; Parisi, 2015), with a strong capacity for surveillance and social sorting. Automation plays a more significant role in accommodating this proliferation of algorithmic applications. Moderation techniques now depend on learning, open-ended, adaptive response capabilities rather than on invariant rules according to pre-ordained ideas. Whereas algorithmic systems tended to start with a fixed set of criteria for the threat or target, machine learning algorithms instead "abductively generates threats and targets via pattern recognition in large volumes of data" (Amoore and Raley, 2017: 6). Consequently, machine learning algorithms involve that the security objective is not often clearly articulated or pre-specified (Amoore, 2013; Bellanova and de Goede, 2022b). These generative and abductive processes underlying learning algorithms will necessarily be characterized by continuous variation and uncertainty (Aradau and Blanke, 2017; Parisi, 2013). Thus, the abductive logic of many of these algorithms contrasts with deductive reasoning, "so that they are closer to the experimental processes of learning and verifying through available data" (Amoore and Raley, 2017: 6).

Rather than collecting evidence, this algorithmic form of governability reinforces preemption in the management of terrorist risk (Bellanova and de Goede, 2022a, 2022b). Platform owners have embraced probabilistic association rules for their security needs. The idea is to locate regularities in vast and disparate data patterns to establish suspicious content and profiles from increasingly autonomous methods (Munn, 2017), making content moderation inherently probabilistic (Ananny, 2016, 2020). Indeed, algorithms make probable associations by measuring the interval between one existing piece of data and another (Aradau and Blanke, 2017; Parisi, 2015). By connecting probabilistic associations, algorithms perform "anticipatory actions" against potentially radicalized users who could launch terrorist attacks. The critical point is that probabilistic knowledge based on algorithmic outputs becomes a security device. The algorithmic moderation reinforces a geography of suspicion (Amoore, 2009).

Data infrastructure

To be effective, algorithms must access large databases maintained by platforms. Contemporary algorithms, primarily those based on machine learning algorithms, rely on the largest possible cache of indiscriminate data (Gillespie, 2016; Parisi, 2019).[3] As Parisi (2019) states: "machine languages use the data environment to select, evaluate, rank match and reconfigure information according to the social use of data" (p.102). For Munn (2017), algorithms thus realize an ideological performance where data volume, variety and velocity are considered representative of a certain reality. These stocks of "digital footprints" appear to constitute a "generalized digital behaviorism" that is used to illuminate sets of relations between past behaviors (Cardon, 2015; Rouvroy, 2013). The algorithm is not only based on the user's activity history. It also works on the digital traces of those who have performed the same actions as him (Cardon, 2015). This mechanism is generally referred to as "collaborative filtering" in technical jargon. The data will then be transformed into trends by algorithms to draw deductions and predictions (McQuillan, 2015). Thus, the traces of Internet users will be constantly quantified and translated into predictive or non-predictive scenarios. By having a global view of Internet users, that is, their communications and actions,

both visible and silent, the algorithm acts as a powerful magnifying lens that categorizes certain users and contents as terrorists. The main feature of these non-conscious forms of automated cognition is, therefore, their ability to aggregate data and sort it semi-autonomously into categories while drawing conclusions through a surveillance infrastructure that most users never directly encounter (Ananny, 2016; Zuboff, 2019).

Being categorized as a terrorist thus results from the processing of a set of signals. This power of enunciation no longer belongs exclusively to governments but derives from an aggregation of data by algorithms designed and maintained by private companies (Cheney-Lippold, 2018). By categorizing an individual from a set of digital traces, the algorithm exercises a narrative power over this individual. Cheney-Lippold summarizes this exceptionally well: "we are narrated when our data is algorithmically spoken for" (2018: 39). In the case of detection algorithms, they "say" something in mathematical language about the user; they assign him a legitimate or illegitimate user posture. In other words, algorithmic devices translate "raw" data that "historicize" the user into supposed objectivity constructed by the platforms' rules and policies. These operations show that, in the end, technical devices are powerful tools that condense users' past and present.

This algorithmic power is further enhanced when companies pool their database, as evidenced by the Global Internet Forum to Counter Terrorism (GIFCT) (Twitter, 2017). Facebook, Microsoft, Twitter, and YouTube created this partnership in 2016. The partnership aims to undermine the ability of terrorists to disseminate their content by developing effective technological means and sharing best practices. The power of this partnership—and its most controversial character—is reflected in their joint database. This opaque database aims to share *hashes* of terrorist images and videos to stop their distribution. York (2021) reminds us that when one company tags an image or video as terrorist or violently extremist according to its policies and removes it, any other company using this database will also automatically remove that same content without "seeing" what it contains. While the GIFCT initially prioritized Islamic State and Al-Qaeda propaganda, the GIFCT decided to expand its database following the Christchurch Call to Action. In particular, GIFCT will target attacker manifestos—often shared by supporters of white supremacist attacks and other publications linking to neo-Nazi and white supremacist groups.

Propelled by this large-scale analytical capacity, invisible to humans, computer algorithms are assigned a superior normative capacity because of their unrivalled ability to report "terrorist signals" and to find the most rational course of action. Moreover, this power is essentially exercised by the Internet giants, which have a monopoly on the colossal databases necessary for the effectiveness of machine learning algorithms. Insensitive to the potential biases that the processing of this data could reproduce, their algorithms are shielded by incredible opacity, forming new invisible regimes of population management. For technology companies, algorithms have a "disposition to objectivity" (Hillis et al., 2013: 37) and are neutral, politically speaking. This posture reflects a rational destiny at its peak, ignoring that this could merely be an "opinion embedded in mathematics" (O'Neil, 2016).

Machine + human: The rise of cognitive assemblages

The transparency reports from digital platforms make one unequivocal statement: algorithms report more content than humans. Automated techniques have allowed companies to remove more content faster, finding the majority of content themselves. Regarding terrorist content, the results obtained, thanks to the advancement of automated tools, showed a success rate of over 90% for YouTube and 99.7% for Facebook. These algorithms sort through a massive amount of content that needs to be reviewed daily. At the end of 2020, Facebook had more than 2.85 billion active users. In the second quarter of 2021, Facebook (2021) had removed more than 7.1 million items labeled as terrorist. YouTube (2021), where users can upload as much as 500 h of video every minute, suspended more than 6,278,771 videos between April 2021 and June 2021, including 431,355 for inciting violence or violent extremism and 116,215,629 comments. From July to December 2020, Twitter (2021), for its part, deleted 3.8 million Tweets that violated the platform's rules. Action was taken on 58,750 accounts labeled as terrorist. These figures only reflect cases for which action was taken. As the former head of security at Facebook, Alex Stamos, reminds us, the total number of decisions, including content for which no sanctions were taken, is much higher.[4]

While these numbers are impressive, what is considered "terrorist" by platforms remains fuzzy. Platforms do not always clearly define the groups they label as extremist, and when they do, they follow the qualifications of the US government. Facebook says it refers to lists issued by the US government (Foreign Terrorist Organizations or Specially Designated Global Terrorists) to define terrorist entities. These lists focus mainly on foreign organizations and are primarily associated with jihadist groups. Other extremist acts related to hate and right-wing extremism will belong to the "organized hate" category. Companies have thus focused their automated detection systems on "terrorist groups that pose the greatest threat globally, in the real world and online" (Facebook, 2017a), namely the Islamic State and al-Qaeda. In this case, platforms are highly selective in the choice of groups they decide to ban. By reducing terrorism to jihadist groups, platforms reinforce the myths surrounding Islam and terrorism.

Implicitly, these impressive capacities to secure information flows refer to the traditional dualism that opposes machines and humans (Latour, 2005). But the human-machine duopoly is hardly tenable if we understand algorithms as sociotechnical assemblages composed of heterogeneous entities that are always uncertain and provisional (DeLanda, 2006; Latour, 2005). STS researchers have shown that algorithms are embedded in sets of associations (Beer, 2009; Gillespie, 2016; Neyland and Möllers, 2017; Seaver, 2013) and are inevitably loaded with values (Mittelstadt et al., 2016). From this perspective, algorithms are an assemblage "of institutionally situated computational code, human practices, and normative logic that creates, sustains, and signifies relationships among people and data through minimally observable, semi-autonomous action" (Ananny, 2016: 99). While algorithms are typically touted as a powerful and autonomous force by digital platforms, they still need engineers, experts, NGOs, governments, civil society, moderators, etc., to operate. Therefore, we see three core ways algorithms are entangled with human decision-making to form emerging cognitive assemblages: through moderators, online communities and experts.

Entangled with moderators

Surprisingly, while technology companies advertise the deployment of powerful detection algorithms, they also have considerably increased the size of their security and safety teams (Facebook, 2017a; YouTube, 2017b). In 2018, Facebook doubled the number of people working on these teams to 20,000 employees, including 7500 content reviewers. In 2019, this number increased again to reach 35,000. As for Google, it employed 10,000 moderators in 2017. A YouTube statement published in 2018 mentioned that "deploying machine learning actually means more people reviewing content, not fewer. Our systems rely on human review to assess whether content violates our policies." Algorithmic automation is, therefore, not supposed to work alone but to interact with the input of human moderators. However, this labor is characterized by precariousness and heavy psychological consequences (Roberts, 2019). Furthermore, this work is often delocalized and subcontracted in Global South countries, creating significant "digital labor" flows between companies in Western countries and developing countries (Graham et al., 2017).

Human expertise remains essential in two types of areas. First, algorithms can only be assembled into stable structures through design and training sequences that depend on humans. As we have just seen with the increase in the number of safety team employees, automation has not so much replaced humans as multiplied the need for their intervention. Automated machines are the ones that need humans the most. Second, algorithms can be poor regulators in more complex contexts. If companies regularly publicize the performance of their algorithms, they rarely talk about their weaknesses and error rates[5]. They thus elude a central question: What if algorithms were wrong? In Syria, activists and journalists who use social media to document possible war crimes and abuses by jihadist groups have regularly seen their accounts suspended by mistake. These errors were further reinforced in the first months of the Covid-19 pandemic when moderation was primarily delegated to algorithms (Scott and Kayli, 2020). Moreover, algorithm recommendation systems can alter the fight against terrorism. Based on a study on YouTube, Schmitt and its collaborators (2018) show that counter-messages are closely linked to extremist content. They observe that automated algorithms impact the interrelatedness of counter-messages and extremist content. Faced with a congruence of themes and specific keywords (e.g. "jihad"), the counter-messages disseminated on platforms put users at risk of being exposed to extremist content.

This allows us to address one of the limitations now widely accepted by machine learning algorithms researchers and companies: machine learning systems find understanding the context of speech particularly challenging. Facebook (2017a) explains in one of its releases that:

> AI can't catch everything. Figuring out what support terrorism and what does not isn't always straightforward, and algorithms are not yet as good as people when it comes to understanding this kind of context. A photo of an armed man waving an ISIS flag might be propaganda or recruiting material, but could be an image in a news story. Some of the most effective criticisms of brutal groups like ISIS utilize the group's own propaganda against it. To understand more nuanced cases, we need human expertise.

At the current stage of development, several observations can be made about machine learning algorithms. First, humans are better at understanding complexity and context than algorithms. While they can help flag questionable content more quickly, a total delegation of moderation to automated systems would generate disproportionate risks of censorship. On the other hand, the challenge for developers is to develop an "integrative technology" that works across different media types. When an algorithm is programmed for a particular sequence, it has little room to adapt to another data set.

> In figuring out what's effective, we face the challenges that any company faces in developing technology that can work across different types of media. For instance, a solution that works for photos will not necessarily help with videos or text (Facebook, 2017a).

The usefulness of machine learning algorithms thus remains limited in this particular context. Companies explain that a system designed to search content from one terrorist group does not work on other groups due to differences in language and propaganda styles (Facebook, 2017b). For example, while Facebook offers its interface in 111 languages, its algorithms can only detect hate speech in 30 languages and terrorist propaganda in 19 languages (Flick and Paresh, 2019).

Meanwhile, while not always easy to outline, the boundaries between humans and algorithms tend to harden. While humans tasked with moderating terrorist content and detection algorithms work simultaneously, the temporal gap at which they operate is widening (Hayles, 2017; Lange, 2016; MacKenzie, 2019). The sophistication of detection algorithms and the dramatic increase in the performance of database management tools have allowed for faster content detection than any human can achieve. Algorithms analyze content and make decisions in milliseconds, which creates "a realm of autonomy for technical agency" (Hayles, 2017: 142). This technical temporality fits the speed and scale of a globalized infrastructure where billions of users constantly post vast amounts of content. However, it does not fit with the need for rationalities that are qualitatively different from the purely correlational logic of algorithms.

Entangled with online communities

Although algorithms do most reporting, companies also seek to enlist their user communities in this process. This role is usually outlined in the platforms' security policies and rules, where the following statements can be found: "we rely on members of the YouTube community to report content they find inappropriate"; "we [Facebook] ask people to share content responsibly and to let us know when they see something that may violate our Community Standards." The mechanisms seeking to establish social order on platforms thus involve a "diffusion of responsibility" (Garland, 2012), where the community is incentivized to monitor information flows. Platforms frame this surveillance work as valuable and necessary for their proper functioning. Alongside algorithms, users apply their own sense of what regulation should entail, complying with the platforms' injunctions. They do not shape the social order but conform anonymously to the "reporting work" prescribed by platforms. Thus, users play an essential role, "propelled by both civic responsibility and self-interest" (Marx, 2013: 58).

As a result, some communities have made it their primary mission to denounce jihadist accounts. Faced with an increase in jihadist content and a string of terrorist attacks in the West, users and hacktivist groups such as Anonymous have embraced the responsibility and moral duty to expose terrorist accounts to companies. The deadly outcome of the attacks contributed to creating an environment in which "lateral surveillance" (Andrejevic, 2004) flourished. These "vigilante users" have become the self-appointed spokespersons for the online fight against terrorism. They act to regain control of the technical infrastructure, which in their view, has been wrongly coopted by terrorist groups and their followers. The message of these Internet users is clear: the Web cannot be used for terrorist propaganda. To restore the balance of the technical infrastructure, these users use discursive and technical means of surveillance and denunciation to expel undesirable users. These Internet users embody, in a way, the successful enrollment that enables an increase in the reporting of banned content. More than the emergence of new trends in algorithmic regulation, it is also the persistence of traditional rationalities that should be analyzed. This approach is supported by Bonelli and Ragazzi (2014), who remind us of the enduring importance of old data collection and processing methods, despite the use of sophisticated technologies in terrorism prediction. In doing so, the security measures deployed by platforms actively configure the conditions and goals of the "new" and the "old" (Bonelli and Ragazzi, 2014; de Goede et al., 2014; Hoijtink, 2014). Viewed in this light, as de Geode and his colleagues (2014) point out, the "new" and the "old" bond with, merge, and contest each other.

More concretely and in a more utilitarian way, platforms need the community to make up for the shortfalls of their detection algorithms, particularly for all contents with a complex context and those posted live. While platforms can filter most photos and videos that have already been suspended (and therefore have been "fingerprinted") through their detection algorithms, such an approach is not possible for the live videos that can be streamed on some platforms. Facebook, for example, introduced a live video capacity in 2016. In 2016, Larossi Abballa used Facebook Live to claim responsibility for the murder of two police officers at their home in Magnanville (France), shooting his video on the spot shortly after his deed. The video lasted 13 min, was watched live by 98 people and was taken down 11 h after it was broadcast. More recently, Facebook Live was activated during the Christchurch shooting. Viewed by 4000 people, it took 29 min for the video to get its first flag and be removed. The algorithm was blindsided, as it was confronted with data it had never been trained on, even though the company has developed technologies that can spot certain themes or images in live videos and block them immediately (see Klonick, 2019), at least in theory. While automated systems develop intelligent skills through rapid, non-conscious, and non-hierarchical decision orders, their interventions are only effective if the present looks like the past (Amoore, 2020; Parisi, 2019). Non-conscious forms of automated cognition require that new terrorist materials be similar and analogous to old terrorist materials. In this

context, where abductive reasoning is used to elaborate hypotheses, knowledge about terrorist content always remains incomplete or unaware of emerging trends. This kind of cognitive elaboration reshapes how extremist speech is produced and consumed.

Entangled with experts

Furthermore, to enhance their moderation tools' effectiveness and expand their expertise, platforms have intensified partnerships with other technology companies, governments, civil society groups, academics, and NGOs (Facebook, 2017a, 2018b; Twitter, 2016; YouTube, 2017a, 2018). The objective is to foster shared learning about terrorism and propaganda mechanisms related to the Islamic State and al-Qaeda. Non-conscious forms of automated cognition become entangled with a sprawling network of security analysts and experts in counter-radicalization. For example, several organizations specializing in terrorism or cybersecurity can flag pages, profiles, and groups on these platforms. They can also send companies photo and video files associated with the Islamic State and Al-Qaeda, which the companies then feed to their algorithms to check for existing matches or use to prevent future uploads. The companies can also grant a small group of vetted partners some reporting "privileges." Take the example of YouTube's "Trusted Flagger" program. It "provides powerful tools for users, government agencies and non-governmental organizations (NGOs) to report content that violates the Community Policy (YouTube, n.d.)." Trusted Flaggers are selected by YouTube. The selection criteria are that they regularly report content with a high-reliability rate. The platform continuously evaluates the Trusted Flaggers' skills. YouTube reserves the right to withdraw the status in cases where the Trusted Flagger regularly reports content that does not violate the platform's community rules. Algorithms outperform trusted Flaggers in terms of reporting rates. For instance, between April 2021 and June 2021, algorithms detected 5,927,201 pieces of content compared to 54,339 pieces for Trusted Flaggers. If algorithmic content moderation does not replace human-based reasoning, it is important to underline that the humans involved in moderation "operate according to an ontology of big data analysis" (Heath-Kelly, 2017: 37). This interweaving of human and machine infrastructures favors a preemptive form of filtering. Consequently, this assemblage of experts participates in the inductive elaboration of threats, where the user is carried in posthuman terms "as part material body, part informational flow" (Heath-Kelly, 2017: 36). Following Amoore (2020), we can argue that this complex cognitive assemblage becomes a new space of "calculative reasoning," where new and old surveillance rationalities are intertwined.

The recurring insight is that algorithms alone are ineffective and cannot be entirely responsible for preventing and controlling illicit information flows. This strategy of interacting heterogeneous entities erodes the notion of the algorithm as the sole and primary source of control or the human as the only figure of sovereignty (Amoore and Raley, 2017). Instead, this partial and distributed form of control emphasizes the rise of human–machine partnerships in online security practices (Haggerty and Ericson, 2000; Rose and Miller, 1992). This article highlighted an example of productive collaborations emanating from human and technical cognizers, adopting anticipatory goals aimed at preemptively disrupting potential terrorist groups. The security of flows emerges from continuous and complex interactivity between technical, social, and personal multiplicities (Hayles, 2017). These assemblages, furthermore, are never predictable. Their boundaries and linkages cannot be foreknown. Instead, they are defined by a set of movements and reconfigurations, as some users, experts, and NGOs leave the system while others join it or convert to new organizational forms. Similarly, algorithms evolve as contexts change and "new meanings are produced" (Hayles, 2017: 123).

This complicates the widespread "fetishization of algorithms" (Crawford, 2016) when considering the governance of flows. In practice, the algorithmic regulation of information flows is more fluid, liquid, cobbled together and plural than it appears. It is characterized by a set of shifting human and machine entities, which mix traditional methods of surveillance with more sophisticated tools, and whose linkages and interactions are transient. The processes that enable the consolidation of knowledge about risky profiles and contents are therefore distributed and collective among humans and machines. We see, for example, algorithms that can independently generate lists of threats and targets, as well as terrorism experts, civil society workers, and users who can spot signs of terrorist propaganda, requiring possible intervention by platforms.

Conclusion

In this article, we have examined the deployment of algorithms and how they are entangled with other entities to regulate terrorist content. To do so, it was necessary to initiate a more intense conversation between content moderation researchers, on the one hand, and critical security studies and posthuman studies scholars, on the other hand, to analyze how content moderation becomes a security practice. Using algorithms in content moderation has promised to expand the range of control techniques that can adequately secure information flows. This article argues that algorithms are part of a more extensive set of regulation techniques focused on risk factors and population management. Far from becoming an omniscient and omnipresent force, algorithms work in complementarity with older and more traditional forms of intervention. Algorithms cannot succeed without effective partnerships

with human collectives, without the eyes of users and the knowledge of experts necessary to extend the reach of algorithmic surveillance. The expansion of algorithms is embedded in a complex web of moderators, engineers, expert knowledge, and voluntary adherence strategies that provide users with optimized environments where flows are adequately secured. Focusing on a concrete empirical case study showed how this cognitive assemblage becomes a cobbled space of preemptive calculation.

The cognitive assemblages described here in moderation practices encompass human and algorithmic interactions that operate at different reflexive, temporal, and perceptual levels. Therefore, this article challenges the prevailing discourse that presents algorithms as mighty entities. Our emphasis on the plural dynamics of algorithmic regulation offers an alternative narrative to the more dramatic interpretation, where powerful algorithms shape our lives (Beer, 2009; Lash, 2007; Zuboff, 2019). This is not to minimize the manipulative nature of these technologies but to recognize that they act through a "cognitive assemblage" oriented toward rationalities of threat anticipation and that they are not in themselves the primary locus of control. Even if they tend to be overshadowed by algorithms, humans play an essential role in this process. Thus, the solution to manage extensive information flows is never binary but hybrid, fluid and cobbled together. Moreover, the human-machine mix of this assembly is constantly changing, and the importance given to certain actors fluctuates over time. For example, the technological innovations introduced by new algorithms or computing capacities, changes in the structure of the network, the explosive growth of available contents, the skills of humans tasked with the monitoring of large flows, the development of new partnerships involving experts, governments and non-profits, or the innovations of terrorist groups that are submitted to this surveillance, all contribute to the emergence of new configurations whose structure and features remain unpredictable.

Our knowledge of how algorithms cooperate and interact with other human and non-human entities is still fragmented. The concept of "cognitive assemblage" focuses our attention on the distribution and articulation of control and surveillance capacities. The algorithm is no longer understood only through its technical features but also in relational, institutional, contingent, and contextual terms. The rise of this cognitive assemblage for the proactive management of harmful content has four major implications that must be explored further. First, this assemblage illustrates an emerging modality of control that goes beyond liberal individualism and technical rationality. This is not the same as suggesting that knowledge is produced exclusively by a machine or a human. What is central is the collective and distributed process. Human and non-human entities are thus always jointly controlling the flow of information. Second, it illustrates how the hybridity of new risk management practices changes the temporal scale of monitoring and regulatory techniques. Humans and algorithms operate through distinct timelines with qualitatively different rationalities. Third, despite the shifting nature of their boundaries, the risk models underlying these assemblages are based on actuarial frameworks, favoring knowledge about digital traces above knowledge about people. It promotes an epistemological change in the modes of knowledge: a search for correlation rather than a search for causality (Calude and Longo, 2017). As a result, they provide an "intelligence" knowledge that rests on statistical techniques, guided by prediction and action and no longer by understanding. However, as Calude and Longo (2017) point out, too much information tends to be uninformed, implying that most correlations are spurious. Finally, this hybrid form of governance has normative implications, questioning how transparency and oversight mechanisms should be calibrated.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research received funding from Good in Tech research network, under the aegis of the Fondation du Risque in partnership with Institut Mines-Télécom and Sciences Po.

ORCID iD

Valentine Crosset https://orcid.org/0000-0002-5659-1124

Notes

1. See U.S. Senate Committee on Homeland Security and Governmental Affairs, "Lieberman Calls on Google to Take Down Terrorist Content," May 19, 2008, https://www.hsgac.senate.gov/media/majority-media/lieberman-calls-on-google-to-take-down-terrorist-content
2. Unlike the United States, in Europe there is not a strong presumption against restrictions on free speech (for a summary of the legal and philosophical differences between the United States and Europe regarding free speech, see Waldron, 2012).
3. The author indicates that for the algorithm to be minimally reliable or useful, it will need a massive data set on which to train. For example, social networking algorithms rely on a huge number of nodes before they are able to describe or influence an online community. A long period of observation of data streams will also be required for recommendation and prediction algorithms before they can make useful predictions.
4. See Alexander Stamos, Prepared Written Testimony and Statement for the Record of Alexander Stamos, before U.S. House of Representatives Committee on Homeland Security Hearing on "Artificial Intelligence and Counterterrorism:

Possibilites and Limitations," June 25, 2019, https://perma.cc/GRW8-VKPJ.

5. Since 2019, however, several platforms have published in their transparency report statistics about content that was restored after appeal.

References

Allen J (2011) Powerful assemblages? *Area* 43(2): 154–157.

Amoore L (2006) Biometric borders: Governing mobilities in the war on terror. *Political Geography* 25(3): 336–351.

Amoore L (2009) Algorithmic war: Everyday geographies of the war on terror. *Antipode* 41(1): 49–69.

Amoore L (2013) *The Politics of Possibility: Risk and Security Beyond Probability*. Durham, NC: Duke University Press.

Amoore L (2020) *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*. Durham, CN: Duke University Press.

Amoore L and Raley R (2017) Securing with algorithms: Knowledge, decision, sovereignty. *Security Dialogue* 48(1): 3–10.

Ananny M (2016) Toward an ethics of algorithms: Convening, observation, probability, and timeliness. *Science, Technology, & Human Values* 41(1): 93–117.

Ananny M (2020) Making up political people: How social media create the ideals, definitions, and probabilities of political speech. *Georgetown Law Technology Review* 4(2): 352–366.

Anderson B (2010) Preemption, precaution, preparedness: Anticipatory action and future geographies. *Progress in Human Geography* 34(6): 777–798.

Andrejevic M (2004) The work of watching one another: Lateral surveillance, risk, and governance. *Surveillance & Society* 2(4), 479–497.

Aradau C and Blanke T (2015) The (Big) data-security assemblage: Knowledge and critique. *Big Data & Society* 2(2): 1–12.

Aradau C and Blanke T (2017) Politics of prediction: Security and the time/space of governmentality in the age of big data. *European Journal of Social Theory* 20(3): 373–391.

Aradau C and Blanke T (2018) Governing others: Anomaly and the algorithmic subject of security. *European Journal of International Security* 3(1): 1–21.

Aradau C and Van Munster R (2007) Governing terrorism through risk: Taking precautions, (un) knowing the future. *European Journal of International Relations* 13(1): 89–115.

Aradau C and Van Munster R (2012) The time/space of preparedness: Anticipating the "next terrorist attack." *Space and Culture* 15(2): 98–109.

Beer D (2009) Power through the algorithm? Participatory web cultures and the technological unconscious. *New Media & Society* 11(6): 985–1002.

Bellanova R and de Goede M (2022a) Co-producing security: Platform content moderation and European security integration. *JCMS: Journal of Common Market Studies* 60(5): 1316–1334.

Bellanova R and de Goede M (2022b) The algorithmic regulation of security: An infrastructural perspective. *Regulation & Governance* 16(1): 102–118.

Bonelli L and Ragazzi F (2014) Low-tech security: Files, notes, and memos as technologies of anticipation. *Security Dialogue* 45(5): 476–493.

Burrell J (2016) How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1): 1–12.

Calude CS and Longo G (2017) The deluge of spurious correlations in big data. *Foundations of Science* 22(3): 595–612.

Cardon D (2015) *À quoi rêvent les algorithmes : Nos vies à l'heure des big data*. Paris: Le Seuil.

Cheney-Lippold J (2018) *We are data: Algorithms and the Making of our Digital Selves*. New York, NY: NYU Press.

Crawford K (2016) Can an algorithm be agonistic? Ten scenes from life in calculated publics. *Science, Technology, & Human Values* 41(1): 77–92.

De Goede M (2012) *Speculative Security: The Politics of Pursuing Terrorist Monies*. Minneapolis, MN: University of Minnesota Press.

De Goede M and Simon S (2012) Governing future radicals in Europe. *Antipode* 45(2): 315–335.

De Goede M and Sullivan G (2016) The politics of security lists. *Environment and Planning D: Society and Space* 34(1): 67–88.

De Goede M, Simon S and Hoijtink M (2014) Performing preemption. *Security Dialogue* 45(5): 411–422.

De Smedt T, De Pauw G and Van Ostaeyen P (2018) Automatic detection of online jihadist hate speech. arXiv preprint arXiv:1803.04596.

DeLanda M (2006) Deleuzian social ontology and assemblage theory. In: Fuglsang M and Meier Sorensen B (eds) *Deleuze and the Social*. Edinburgh: Edinburgh University Press, pp. 250–266.

Dodge M and Kitchin R (2007) The automatic management of drivers and driving spaces. *Geoforum* 38(2): 264–275.

Facebook (2017a, June 15) Hard questions: How we counter terrorism. Facebook Newsroom. https://about.fb.com/news/2017/06/how-we-counter-terrorism/.

Facebook (2017b, November 28) Are we winning the war on terrorism online? Facebook Newsroom. https://about.fb.com/news/2017/11/hard-questions-are-we-winning-the-war-on-terrorism-online/.

Facebook (2018a, July 31) Removing bad actors on facebook. Facebook Newsroom. https://newsroom.fb.com/news/2018/07/removing-bad-actors-on-facebook/.

Facebook (2018b, September 17) Combating hate and extremism. Facebook Newsroom. https://about.fb.com/news/2019/09/combating-hate-and-extremism/.

Facebook (2021) Facebook Transparency report. https://transparency.facebook.com/community-standards-enforcement#terrorist-propaganda [accessed 8 September 2021].

Flick M and Paresh D (2019, April 23) Facebook's flood of languages leaves it struggling to monitor content. Reuters. https://www.reuters.com/article/us-facebook-languages-insight-idUSKCN1RZ0DW.

Foucault M (1978) *Security, Territory, Population. Lectures at the College de France 1977–1978*. London: Palgrave Macmillan.

Garland D (2012) *The Culture of Control: Crime and Social Order in Contemporary Society*. Chicago, IL: University of Chicago Press.

Gehl RW, Moyer-Horner L and Yeo SK (2017) Training computers to see internet pornography: Gender and sexual discrimination in computer vision science. *Television & New Media* 18(6): 529–547.

Gillespie T (2014) The relevance of algorithms. In: Gillespie T, Boczkowski P and Foot K (eds) *Media Technologies: Essays on Communication, Materiality, and Society*. Cambridge, MA: MIT Press, pp. 167–195.

Gillespie T (2016) Algorithms. In: Peters B (ed.) *Digital Keywords: a Vocabulary of Information Society and Culture*. Princeton, NJ: Princeton University Press, pp. 18–31.

Gillespie T (2018) *Custodians of the Internet: Platforms, Content Moderation, and the Information Society and Culture Hidden Decisions that Shape Social Media*. New Haven, CT: Yale University Press.

Gillespie T (2020) Content moderation, AI, and the question of scale. *Big Data & Society* 7(2): 1–5.

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

Graham M, Hjorth I and Lehdonvirta V (2017) Digital labour and development: Impacts of global digital labour platforms and the gig economy on worker livelihoods. *Transfer: European Review of Labour and Research* 23(2): 135–162.

Haggerty KD and Ericson RV (2000) The supervisor assemblage. *The British Journal of Sociology* 51(4): 605–622.

Hayles NK (2017) *Unthought*. Chicago, IL: University of Chicago Press.

Heath-Kelly C (2017) Algorithmic autoimmunity in the NHS: Radicalisation and the clinic. *Security Dialogue* 48(1): 29–45.

Highman S and Nakashima E (2015, July 16) Why the Islamic State Leaves Tech Companies Torn Between Free Speech and Security. Washington Post. *https://www.washingtonpost.com/world/national-security/islamic-states-embrace-of-social-media-puts-tech-companies-in-a-bind/2015/07/15/0e5624c4-169c-11e5-89f3-61410da94eb1_story.html.*

Hillis K, Petit M and Jarrett K (2013) *Google and the Culture of Search*. London & New York: Routledge.

Hoijtink M (2014) Capitalizing on emergence: The "new" civil security market in Europe. *Security Dialogue* 45(5): 458–475.

Introna L and Wood D (2004) Picturing algorithmic surveillance: The politics of facial recognition systems. *Surveillance & Society* 2(2/3): 177–198.

Kirschenbaum MG (2003) Virtuality and VRML: Software studies after Manovich. *Electronic Book Review*, pp. 1–20. https://electronicbookreview.com/essay/virtuality-and-vrml-software-studies-after-manovich/

Kitchin R (2017) Thinking critically about and researching algorithms. *Information, Communication & Society* 20(1): 14–29.

Klonick K (2017) The new governors: The people, rules, and processes governing online speech. *Harv. L. Rev*. 131: 1598.

Klonick K (2019, April 25) Inside the team at Facebook that dealt with the Christchurch Shooting. The New Yorker. *https://www.newyorker.com/news/news-desk/inside-the-team-at-facebook-that-dealt-with-the-christchurch-shooting.*

Lange AC (2016) Organizational ignorance: An ethnographic study of high-frequency trading. *Economy and Society* 45(2): 230–250.

Lash S (2007) Power after hegemony: Cultural studies in mutation? *Theory, Culture & Society* 24(3): 55–78.

Latour B (2005) *Reassembling the social: An introduction to actor-network-theory*. Oxford: Oxford university press.

MacKenzie D (2019) How algorithms interact: Goffman's 'interaction order' in automated trading. *Theory, Culture & Society* 36(2): 39–59.

Marx GT (2013) The public as partner? Technology can make us auxiliaries as well as vigilantes. *IEEE Security & Privacy* 11(5): 56–61.

Massumi B (2007) Potential politics and the primacy of preemption. *Theory & Event* 10(2). DOI: 10.1353/tae.2007.0066.

Massumi B (2015) *Ontopower*. Durham, CN: Duke University Press.

McQuillan D (2015) Algorithmic states of exception. *European Journal of Cultural Studies* 18(4–5): 564–576.

Mittelstadt BD, Allo P, Taddeo M, et al. (2016) The ethics of algorithms: Mapping the debate. *Big Data & Society* 3(2): 1–21.

Miyazaki S (2012) Algorithmics: Understanding micro-temporality in computational cultures. *Computational Culture* 2: 1–20. http://computationalculture.net/algorhythmics-understanding-micro-temporality-in-computational-cultures/

Munn L (2017) Seeing with software: Palantir and the regulation of life. *Studies in Control Societies* 2(1): 1–16.

Munn L (2018) *Ferocious Logics: Unmaking the Algorithm*. Lüneburg: Meson press.

Neyland D and Möllers N (2017) Algorithmic IF… THEN rules and the conditions and consequences of power. *Information, Communication & Society* 20(1): 45–62.

Noble SU (2018) *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York, NY: NYU Press.

O'Neil C (2016) *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy*. New York, NY: Crown Publishing Group.

Parisi L (2013) *Contagious Architecture: Computation, Aesthetics, and Space*. Cambridge, MA: MIT Press.

Parisi L (2015) Instrumental reason, algorithmic capitalism, and the incomputable. In: Pasquinelli M (ed.) *Augmented Intelligence Trauma*. Lüneburg: Meson Press, pp. 125–138.

Parisi L (2019) Critical computation: Digital automata and general artificial thinking. *Theory, Culture & Society* 36(2): 89–121.

Pasquale F (2015) *Black Box Society. The Secret Algorithms that Control the Economy and Information*. Cambridge, MA: Harvard University Press.

Rieder B and Skop Y (2021) The fabrics of machine moderation: Studying the technical, normative, and organizational structure of perspective API. *Big Data & Society* 8(2): 205395172 11046181.

Roberts ST (2019) *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven, CT: Yale University Press.

Rose N and Miller P (1992) Political power beyond the state: Problematics of government. *British Journal of Sociology* 43(2): 172–205.

Rouvroy A (2013) The end(s) of critique: Data-behaviourism vs. due process. In: Hildebrandt M and De Vries K (eds) *Privacy, Due Process and the Computational Turn. Philosophers of Law Meet Philosophers of Technology*. London & New York: Routledge, pp. 143–168.

Schmidt A and Wiegand M (2017) A survey on hate speech detection using natural language processing. In Proceedings of the fifth international workshop on natural language processing for social media, pp. 1–10. https://aclanthology.org/W17-1101/

Schmitt JB, Rieger D, Rutkowski O, et al. (2018) Counter-messages as prevention or promotion of extremism?! the potential role of YouTube: Recommendation algorithms. *Journal of Communication* 68(4): 780–808.

Scott M and Kayli L (2020, October 21) What happened when humans stopped managing social media content. *Politico*, pp. 1–16. https://www.politico.eu/article/facebook-content-moderation-automation/

Seaver N (2013) Knowing algorithms. *Media in Transition* 8: 1–12.

Twitter (2016, February 5) Combating violent extremism. Twitter Blog. https://blog.twitter.com/en_us/a/2016/combating-violent-extremism.html.

Twitter (2017, June 26) Global internet forum to counter terrorism. Twitter Blog. https://blog.twitter.com/en_us/topics/company/2017/Global-Internet-Forum-to-Counter-Terrorism.html.

Twitter (2019, March) Terrorism and violent extremism policy. Help Center. https://help.twitter.com/en/rules-and-policies/violent-groups.

Twitter (2021, July 14) An update to the twitter transparency center. Twitter Blog. https://blog.twitter.com/en_us/topics/company/2021/an-update-to-the-twitter-transparency-center.

Ulbricht L and Yeung K (2022) Algorithmic regulation: A maturing concept for investigating regulation of and through algorithms. *Regulation & Governance* 16(1): 3–22.

Waldron J (2012) *The Harm in Hate Speech*. Cambridge, MA: Harvard University Press.

Wilcox L (2017) Embodying algorithmic war: Gender, race, and the posthuman in drone warfare. *Security Dialogue* 48(1): 11–28.

Yeung K (2018) Algorithmic regulation: A critical interrogation. *Regulation & Governance* 12(4): 505–523.

York JC (2021) *Silicon Values: The Future of Free Speech Under Surveillance Capitalism*. New York, NY: Verso Books.

YouTube (2008, May 19) Dialogue with Sen. Lieberman on terrorism video. Official Blog. https://blog.youtube/news-and-events/dialogue-with-sen-lieberman-on/

YouTube (2017a, August 1) An update on our commitment to fight terror content online. Official Blog. https://youtube.googleblog.com/2017/08/an-update-on-our-commitment-to-fight.html.

YouTube (2017b, December 4) Expanding our work against abuse of our platform. Official Blog. https://youtube.googleblog.com/2017/12/expanding-our-work-against-abuse-of-our.html.

YouTube (2018, April 23) More information, faster removals, more people—an update on what we're doing to enforce YouTube's Community Guidelines. Official Blog. https://youtube.googleblog.com/2018/04/more-information-faster-removals-more.html.

YouTube (2021) Application of the YouTube community rules. Information transparency. https://transparencyreport.google.com/youtube-policy/removals?hl=fr. [accessed May 15, 2021]

YouTube (n.d.) YouTube trusted flagger program. YouTube Support. https://support.google.com/youtube/answer/7554338?hl=fr [accessed 15 May 2019].

Ziewitz M (2016) Governing algorithms: Myth, mess, and methods. *Science, Technology, & Human Values* 41(1): 3–16.

Zuboff S (2019) *The Age of Surveillance Capitalism: The Fight for the Future at the New Frontier of Power*. London: Profile Books.

Zuckerberg M (2017, February 16) Building global community. Facebook Publication. https://www.facebook.com/notes/mark-zuckerberg/building-global-community/10154544292806634/.

Zuckerberg M (2018, April 10) Mark Zuckerberg testifies before Congress. https://www.youtube.com/watch?v=mZaec_mlq9M.