CHAPTER 7

Artificial Intelligence for Peace: An Early Warning System for Mass Violence

Michael Yankoski, William Theisen, Ernesto Verdeja, and Walter J. Scheirer

Introduction

Pundits are increasingly raising concerns about the dangers that advanced artificial intelligence (AI) systems may pose to human peace and safety. Elon Musk (Clifford 2018) warned that AI has the potential to be more

M. Yankoski (✉) · W. Theisen · W. J. Scheirer
Department of Computer Science and Engineering, University of Notre Dame, South Bend, IN, USA
e-mail: Michael.G.Yankoski.2@nd.edu

W. Theisen
e-mail: wtheisen@nd.edu

W. J. Scheirer
e-mail: walter.scheirer@nd.edu

E. Verdeja
Kroc Institute for International Peace Studies and Department of Political Science, University of Notre Dame, South Bend, IN, USA
e-mail: Ernesto.Verdeja.1@nd.edu

T. Keskin and R. D. Kiggins (eds.), *Towards an International Political Economy of Artificial Intelligence*, International Political Economy Series,
https://doi.org/10.1007/978-3-030-74420-5_7

dangerous than nuclear weapons. Stephen Hawking (BBC 2014) worried that AI could mean the end of the human species. A recent Special Issue of the Bulletin of the Atomic Scientists included several warnings about the coming AI Arms Race (Roff 2019). Indeed, many of the chapters in this volume are similarly concerned with mitigating the negative implications of advanced AI.

We agree that caution is warranted regarding the rapid development in the field of AI, but we also believe that some AI research trajectories may be employed toward positive ends. In this chapter, we introduce one current research trajectory that combines AI and social scientific research on political violence in order to contribute to practical conflict prevention. AI systems are capable of significantly enhancing the work of peacebuilders in specific but important ways, for artificial intelligence provides unique tools for identifying and analyzing emergent trends and threats within massive volumes of real-time data on the Internet, well beyond the capacities of most existing political violence early warning systems.

This chapter discusses a novel project that brings together computer and social scientists using artificial intelligence to advance current atrocity and political instability early warning capabilities. We focus on how the spread of disinformation, rumors, and lies on social media—essentially, hate propaganda—in already unstable political contexts may function as early warning indicators of imminent large-scale violence. Researchers have long argued that hate propaganda legitimizes violence against vulnerable groups (Chirot and McCauley 2010; Sémelin 2005; Kiernan 2003; Koonz 2003; Weitz 2003), but in our current social media landscape, where harmful and manipulative political content circulate more rapidly and widely than ever before, the dangers are especially acute. This is evident, for instance, in the lead up to the Indonesian elections in 2019, where Instagram and Twitter were filled with conspiratorial allegations about treasonous politicians who had to be prevented from winning at the ballot box by any means, including through terror, threats, and killings (Suhartono 2019; BBC 2019). In Myanmar, ongoing atrocities against the Rohingya minority group have been fueled by nationalist memes on Twitter and Facebook accusing the Rohingya of being dangerous foreigners, an existential threat to the integrity and survival of the country that must be eradicated (Azeem 2018; BSR 2020). In these cases and many others, social media has played a defining role in perpetuating dehumanization and facilitating violence.

Our focus in this project is on advancing what peace studies scholars often refer to as "negative peace"—the absence of armed conflict and direct violations of bodily integrity, such as killings, assaults, and torture—by contributing to early warning modeling and analyses of political violence.[1] As such, we present a model of computational forensic analysis of digital images in social media to help identify where and when unstable societies may tip into large-scale violence, by providing journalists and prevention practitioners—that is, policymakers, analysts, and human rights advocates in the atrocity prevention community—with data-rich, theoretically robust assessments of processes of violence escalation in near "real time." This type of triage mechanism is central to ensuring timely and effective preventive responses on the ground. Our project is a work in progress, but we believe it suggests a path forward that can also benefit from contributions from the broader scholarly community.

We realize that developing the political will to prevent or stop violence is crucial and extremely difficult (Lupel and Verdeja 2013; Weiss 2016), but providing more accurate and actionable information on conflict escalation can aid prevention work significantly. Our unique focus on digital images is driven by the new ways in which people communicate on the Internet, which are no longer just text-based. Given the enormity of social media data produced daily and the need for timely and accurate analysis, AI systems offer unique capabilities for enhancing and even transforming the work of practitioners tasked with anticipating and responding to violence.

The chapter is dedicated to outlining the computational dimensions of our project. We proceed in several steps. First, we outline the current state of risk assessment and early warning research and practice, and situate our project within this field. We then introduce specific problems of disinformation: namely, how social media is increasingly used to sow fear and distrust in already fragile communities and further legitimize violence against vulnerable groups. This kind of disinformation is an important indicator of likely violence in the near- or mid-term, and thus needs to be actively monitored if early warning analyses are to be more focused, accurate, and actionable. Despite their importance, it is exceedingly difficult to understand the spread and impact of coordinated disinformation campaigns in real time using currently available tools. We then discuss our project in detail, first by outlining the types of entities we analyze—social media memes—and then by sketching the overall computational model of analysis. We then turn to some further areas of development, and

finally explore the ethical and policy implications of AI work in atrocity prevention.

Risk Assessment and Early Warning: What We Know, What We Need to Know

There is a long history of systematic attempts to anticipate the outbreak of large-scale political violence, going back at least to the 1950s when the superpowers sought to model the likelihood of nuclear war through a variety of simulations (Edwards 1997; Poundstone 1992). Since the international community's failure to prevent genocides in Bosnia-Herzegovina and Rwanda in the 1990s, governments and human rights advocates have devoted more attention to forecasting political instability and mass violence, often working closely with scholars to develop rigorous and evidence-based approaches to prevention work.[2] Today there are numerous early warning and risk assessments initiatives focusing on the main drivers and signs of impending violence and also to informing atrocity prevention (Waller 2016).

We now have a sophisticated understanding of conditions that elevate risk of violence, but systematic and generalized models of short-term patterns of violence onset are less well developed. It is exceedingly difficult to interpret fast-moving political violence dynamics. Our project contributes to these early warning efforts to understand real-time violence escalation through an analysis of the circulation of digital images on social media that can encourage and legitimize harm against vulnerable minorities. The project primarily focuses on countries already at high risk of violence where the timing or onset of violence is difficult to know. This is ultimately about forecasting the likelihood of violence, not about providing a causal theory of violence. The distinction between these is pivotal. Much like sharp pains in the left torso may indicate an imminent heart attack without being its cause, political violence forecasting is concerned with assessing the probability of future violations, rather than causally explaining their occurrence after the fact.

Current research has identified a host of general conditions that elevate a country's likelihood of future violence. The first condition is a *history of unpunished violence against vulnerable minorities* (Harff 2003; Goldsmith et al. 2013). Prior impunity legitimizes future violence because potential perpetrators know they face little or no sanction. *Severe political instability*, especially armed conflict, is another major risk factor (Midlarsky

2005; Goldstone et al. 2010). In countries experiencing ongoing and profound crises, such as war, insurgencies, coups, or violent changes of political control, leaders are much more likely to rely on increasingly harsh repressive measures to eliminate perceived threats and remain in power. Armed conflict, whether an international or civil war, is among the strongest predictors of future atrocities against civilians. A third factor is the espousal of a *radical ideology* by government leaders and/or armed challengers that systematically dehumanizes others (Robinson 2018; Weitz 2003). Such extreme ideologies—whether religious, ethnic, racial, or authoritarian variants of left- or right-wing ideologies—justify the use of increasingly repressive measures against vulnerable civilians. Related to this is ongoing *state-led discrimination*, including the denial of basic civil and political rights as well as movement restrictions, which are highly correlated with atrocities (Fein 2007; Goldstone et al. 2010). Finally, *government regime* matters; authoritarian regimes are more likely than democracies to engage in violence. Semi-democratic regimes, with limited political contestation and some opposition political movements but weak rule of law and unaccountable political leaders, are also more prone to violence than robust democracies (Stewart 2013). The greater the presence of these factors, the greater risk a country has of experiencing mass violence. However, these factors are largely static—they are useful for providing general risk assessments (low, mid, or high risk) of violence, but the factors do not fluctuate much over time. They tell us the relative likelihood of future violence, but do not provide precise insights into *when* a high-risk situation may devolve into overt killings and atrocities.

Much harder to pinpoint in real time are the short- and mid-term events and processes that are indicators of the shift from high-risk conditions into actual violence, or what is normally known as early warning (Heldt 2012). Broadly, there are three clusters of early warning indicators (Verdeja 2016). First, there are *dangerous symbolic moments or discourses* that significantly dehumanize already vulnerable populations, or reinforce deep identity cleavages between groups. This includes rallies and commemorations of divisive events or the spread of hate propaganda. Second is an uptick in *state repression*, such as moving security forces to places with vulnerable populations, stripping those populations of legal rights, attacks against prominent minority or opposition leaders and their followers, or widespread civilian arrests. Finally, *political and security crises* challenging incumbent political leaders are important indicators, and these can include new or resumed armed conflict between the

state and rebels, rapid changes in government leadership, or the spread of confrontational protests. Unforseen exogenous shocks like natural disasters or neighboring conflict spillover can also trigger violence by challenging the ability of political leaders to maintain control. In many instances, several early warning indicators will occur simultaneously or in clusters.

Despite these existing indicators, there is room to improve and expand early warning capabilities. The primary concern involves limitations in overall availability and quality of information. Many conflict scenarios are extremely difficult and dangerous to access physically, and thus we must rely on the information provided by relatively limited numbers of people in the field, whether journalists, aid workers, local residents, government officials, or displaced civilians. The reliability of these sources varies, but even when sources are dependable, it can be exceedingly hard to know how representative information is, especially when mobility is limited. For instance, does a journalist's reporting in a small area reflect the entire region or the hardest hit areas of a country? If not, what is missing? How can we compensate for these limitations?

In short, many existing early warning models can expand their source material to systematically tap into much richer social media streams, which can inform how violence may be escalating or if a situation is on the cusp of escalation. To be clear: social media streams do not represent a more factually accurate portrayal of what is occurring. Many politicized memes, for instance, are misleading or outright lies and their provenance (i.e., origin and mode of creation) can be hard to identify, as we discuss below. Indeed, social media is itself a new battleground in contemporary conflicts, as armed actors frequently distort and misrepresent the actions and motives of their enemies as a justification for violence. However, politicized social media gives a stream of real-time data that our system is capable of analyzing in order to identify trends that signify increased potential for outbreaks of political violence. Thus, integrating social media analysis into early warning evaluations may significantly enhance conflict prevention and intervention work. In order to understand what this entails, the next section discusses what a political meme is and then presents the main features of our AI project.

What Is a Political Meme?

If the new media landscape is a battleground, then one must study how contemporary political movements communicate on the Internet in order to begin to formulate a response. Unlike in the past, where most people were largely consumers of professionally curated media content, anyone can now make and disseminate their own political messages to an audience of millions on the Internet. The widespread availability of powerful image editing tools has democratized digital content creation, allowing users with basic computer skills and time to produce custom images and videos. This content most often takes the form of a *meme*. Memes are cultural artifacts that evolve and spread like a biological organism, but are completely external to biology. On the Internet, memes consist of images, often humorous, that adhere to a set genre, which acts as a guideline for the creation of new instances. But these images are often more than just jokes. Memes have served as the impetus for political actions in movements as diverse as the Arab Spring (York 2012), Occupy Wall Street (Know Your Meme 2020), and the Black Lives Matter (Leach and Allen 2017) movements (Fig. 7.1). And they are now a significant resource

Fig. 7.1 A selection of political memes from the past decade, all exemplifying cultural remixing. Left: A meme that is critical of the Syrian regime's use of poison gas, in the style of the iconic Obama "Hope" poster. Center: a meme associated with the UC Davis Pepper Spray Incident during the Occupy Wall Street Protests. Right: A Black Lives Matter meme where the raised fist is composed of the names of police victims. This meme also includes the movement's signature hashtag

for monitoring the pulse of an election, responses toward international security incidents, or the viewpoints surrounding a domestic controversy.

The popularity of memes continues to grow, and so does the scope of political messaging attached to them. Cases where political memes prefigured violence in some form are not difficult to come by on social media. Our own work has uncovered cases across the globe where memes have been used to spread antisocial messages from discriminatory stereotypes to outright calls for violence. In Brazil, we found memes of right-wing President Jair Bolsonaro depicted as an action hero, ready to take on the country's drug traffickers (left panel of Fig. 7.2). This coincides with an escalation of the Brazilian drug war and violence against the poor, in which state security forces have been responsible for over a third of the violent deaths reported in Rio De Janeiro (Santoro 2019). In India, we witnessed misogynistic memes featuring a cheerful Prime Minister Narendra Modi overlaid with text containing degrading messages against women (center panel of Fig. 7.2), while the government continues to undermine legal protections for women (Human Rights Watch 2018). In Indonesia, we discovered images where a hammer and sickle were superimposed on the prayer mats of Muslim worshipers, meant to insinuate that they are crypto-communists (right panel of Fig. 7.2). These images were found shortly before the 2019 presidential election, which ended with violent street protests in Jakarta (Suhartono 2019; BBC 2019). The list goes on.

Fig. 7.2 A selection of political memes with disturbing messaging. Left: Brazilian President Jair Bolsonaro depicted as an action hero, ready to take on Brazil's drug traffickers. Center: A misogynistic meme featuring Indian Prime Minister Narendra Modi. Right: Hammer and sickle superimposed on the prayer mats of Islamic worshipers in Indonesia

If AI is to be deployed as a mechanism to watch for political memes that may be used to incite large-scale violence in high-risk contexts, we require a rigorous definition of a political meme to operationalize the algorithms. Definitions like the one given above referring to evolution and organism-like propagation are typically attributed to the biologist Richard Dawkins (2016). However, Dawkins' thinking on the meme, whether intentional or not, borrowed liberally from the notion of intertextuality in literary theory: the shaping of a text's meaning by other texts. Julia Kristeva, and Mikhail Bakhtin before her, suggested that the novel reworking and retransmission of information is fundamental to human communication (Kristeva 1986). In the Kristevan mode, intertextuality applies to all semiotic systems, including digital images and video. Intertextuality is a more useful framing when assessing content on the Internet, which can be collected and analyzed via automatic means to identify *intertexts*, those specific points of correspondence between artifacts (Forstall and Scheirer 2019). While other researchers have sought to define the meme in general terms (Shifman 2014), we are not interested in all memes for an early warning system for violence. Thus we offer the following definition for a political meme: a multimedia intertext meant to engage an in-group and/or antagonize an out-group.

Given this definition, how can we operationalize it within the context of today's AI capabilities to give an algorithm what it needs to automatically assess any potential threats of violence that political memes might pose? First, the source of the meme can be scrutinized. Where the meme was found on social media, and who posted it, can be diagnostic. For instance, if the source is known to be political, the meme might be as well. But the source is not essential for an observer to understand the message a meme conveys. More importantly, the content of the meme is composed of visual and textual cues that deliver the message. In many instances, decisions are made based purely on the visual style of an image. For example, does it look like something we have seen before, which is known to be political in some regard? If so, then an intertextual association has been made. If it is new, is there something in the visual content or text, if present, that gives us a clue as to whether or not it is political? Such clues can be the presence of political figures in the image, places associated with political history, symbols associated with political or religious groups, or objects with some political significance. Finally, not all political memes are something to worry about; thus, we need to separate the innocuous from the dangerous. This is done by establishing semantic links between

the visual content and text, as well as by assessing the sentiment of those elements. The ability to recognize all of this involves pattern recognition—an area where AI systems excel, because they can be trained to recognize all of the aforementioned elements. The challenge is in making the system understand their relevance in a manner that is consistent with human observers.

Technologies for an Early Warning System for Violence

If visual content found on social media is the object of focus for our early warning system, then what, exactly, are the technical requirements for understanding it? This task is fundamentally different from more traditional forms of early warning, where indicators are established using quantitative variables from the social science literature (e.g., datasets on protests, armed conflict onsets, coups, etc.), that can be processed using statistical techniques from data science and used to make general predictions. Here we bring to bear new methods from the areas of computer vision and media forensics, as well as established best practices from high performance computing and web and social media studies. We do this in order to build a comprehensive system that proceeds from data ingestion to making predictions about content that may contain messaging that seems intended to incite large-scale violence. In general, there are three basic required components of such a system (Fig. 7.3): (1) the data ingestion platform; (2) the AI analysis engine; and (3) the user interface (UI). In this section, we will detail these technical aspects of the system.

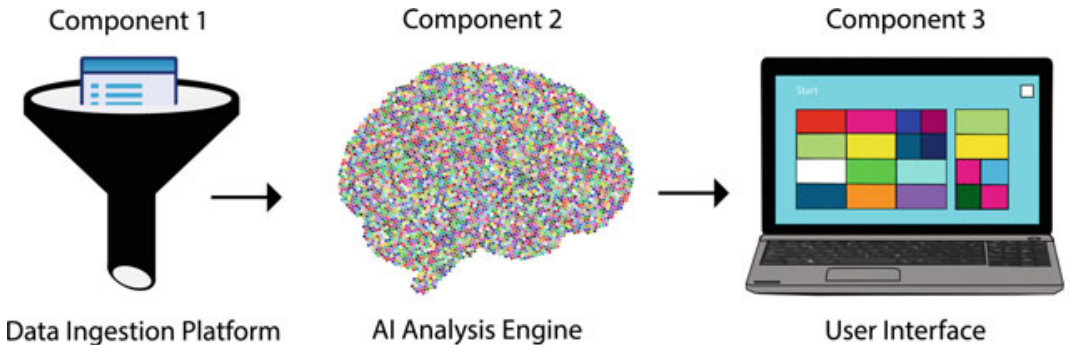

Fig. 7.3 The processing pipeline of the proposed early warning system for large-scale violence, composed of three basic required technology components

Data ingestion is the most important part of the process. If we do not look in the right place to begin with, then the early warning system will not be useful. Given the consolidation of the Internet over the past decade (itself a contributing factor to the overall problem we are studying) into a handful of popular social networks (Internet Society 2019), data targeting is, in some sense, straightforward. However, as one dives into these social networks, their vast internal complexity becomes apparent, with many communities and subcommunities existing in complex webs on platforms like Reddit, 4chan, Twitter, Facebook, and Instagram. So, where to look? Social media platforms have always striven to add structuring elements to their posts, in order for users to better identify relevant topics and authors. Examples of these structuring elements include user-defined hashtags (the "#" symbol followed by a plaintext string) that are used as meta-data to tag posts with custom topics on most social networks, as well as account names (often indicated by the "@" symbol). However, the reliability of these elements can be questionable, and some platforms go out of their way to render them useless (most notably 4chan, where nearly every user is anonymous). Media objects such as images and videos also form structured elements in posts, and for our purposes, posts containing these media types are of primary interest. Accordingly, our attention will be on visual content for the rest of this chapter. Note though that it is possible to bring comments, shared or reposted content, and "likes" into the analysis as well.

Further, when it comes to the wholesale data harvesting that is necessary to watch the Internet in a meaningful way in real time, there is a diversity of access control mechanisms put in place that restrict direct access to the data. Some sites make this process easy, while others make it very difficult. For instance, Twitter provides an Application Program Interface (API) (Twitter 2020) that makes all of its data very accessible to automated data ingestion. The main motivation for this is the development of third-party apps that make creative use of the data that appears on Twitter, but there is also some sympathy for academic work that looks at the social aspects of information exchange on the platform. Other sites, like Facebook, are far more closed, leading to the need for specialized content crawlers that can search for relevant content without the assistance of an open API. Rate limiting procedures put in place on social networks are another confounding element, and a potentially dangerous roadblock for any early warning system that needs timely access to data. An example of this is Facebook, which uses dynamic links to content,

which change on a regular basis. These problems can be mitigated with a distributed system design, whereby many instances of the data ingestion component run from different parts of the Internet.

A practical constraint of data ingestion is the scale of the data an early warning system must consider in order to be effective. In a targeted operation, where a particular region and/or set of actors is being monitored, we can operate over data on the order of millions of images within a period of weeks. This is the current upper-bound for media forensics being conducted at the academic level (NIST 2018). Beyond this, the cost of data storage and the time required to collect the data becomes prohibitively expensive. However, there is a need for the system to scale to the order of billions of images per day (Eveleth 2015) for a truly comprehensive real-time capability. The ultimate goal is to be able to watch all of social media for emerging trends. In order to accomplish this goal, partnerships will need to be established between non-governmental organizations (NGOs), social media companies, and academics to provide access to the data at the source. The reluctance of social media companies to allow outsiders access to their internal data repositories is a serious stumbling block in this regard. We envision that such partnerships will lead to the transfer of early warning techniques that can be run internally at a social network when data access is problematic.

Given existing constraints, we have found an effective combination for data ingestion to involve partnering with local experts who can provide relevant hashtags and accounts to review, thus mitigating the need to monitor everything on the Internet at once. While this approach may miss quite a bit of potentially threatening content, targeted sampling is still effective. For example, in our work on the 2019 Indonesian Presidential Election (Yankoski et al. 2020), we partnered with the local fact checking organization Cekfakta (2020). By using local volunteers throughout Indonesia, Cekfakta is able to identify social media sources of concern. However, human monitoring of even a limited number of sources proved incapable of keeping up with the pace of content creation. From just 26 hashtags and eight users on Twitter and Instagram, we harvested over two million images for analysis (Theisen et al. 2020)—a staggering number that exceeds the human capacity to find patterns in unorganized data.

The artificial intelligence component of the early warning system is designed to automate the analysis of the large collections of media content assembled at the data ingestion stage. A key to success in this

regard is the use of state-of-the-art artificial perception methods. In the debates surrounding the use of artificial intelligence technology, there is a pervasive misunderstanding of the difference between perception and cognition. As we discussed in the beginning of this chapter, fears over AGI, or artificial general intelligence, have led to accusations that all AI technology represents an existential threat to humanity. Should AGI ever appear, it would embody a set of cognitive models meant to mimic the conscious mental actions of knowledge acquisition and reasoning in the human mind. This type of technology does not exist at the time of writing, and we are skeptical that it will emerge in the foreseeable future. The complexities of the human brain as a system are beyond the current understanding of science. We do not possess a model of computation for the brain, nor do we have explanatory models for complex phenomena such as conscious thought (Marcus and Davis 2019).

Where AI has made inroads into modeling competencies of the brain is in the sensory systems (e.g., audition, olfaction, vision). Perception is the ability to take information from a sensory system and make decisions over it. This process mostly unfolds in an unconscious manner, but embodies a set off complex pattern recognition behaviors. The most studied and best modeled sensory modality is vision. The fields of computer vision and machine learning have taken direct inspiration from experimental observations in neuroscience and psychology (Goodfellow et al. 2016), leading to features (i.e., descriptions of the data) and classifiers (i.e., models that make decisions over features) that are, in some cases, at human- or super human-level performance (RichardWebster et al. 2018). These technologies can be used safely and effectively in the appropriate context. For instance, computer vision can be used to determine which images out of a large collection are similar in overall appearance, identify specific objects within images, and match common objects across images.

Concern over manually and automatically generated fake content has driven advances in media forensics—a field within computer science that borrows heavily from the fields of computer vision and machine learning. For a violence early warning system, we need a way to characterize the images such that (1) from an initial collection, they can be placed into distinct genres, (2) new images can be placed into known genres or new genres where appropriate, and (3) semantic understanding of the visual content can be extracted, so that threatening messages can be identified. Select techniques from media forensics give us a path forward for each of these requirements.

For establishing connections between images, *image provenance analysis* provides a powerful framework (Moreira et al. 2018). Work in image manipulation detection has shown that it is possible to estimate, through image processing and computer vision techniques, the types and parameters of transformations that have been applied to the content of individual images to obtain new versions of those images (Rocha et al. 2011). Given a large corpus of images and a query image (i.e., an image we would like to use to find other related images), a useful further step is to retrieve the set of original images whose content is present in the query image, as well as the detailed sequences of transformations that yield the query image given the original images. This is known as image provenance analysis in the media forensics literature. The entire process is performed in an automated unsupervised manner, requiring no human intervention. Such a process can be used to trace the evolution of memes and other content, which is a piece of what we need for the early warning system. This process can also be used for fact checking and authorship verification. In general, provenance analysis consists of an image retrieval step followed by a graph building step that provides a temporal ordering of the images. In place of the latter, we suggest that an image clustering step is more useful for early warning analysis. We will, in broad strokes, explain how each of these steps works below.

In order to find related images, each must first be indexed based on features that describe the style and content of the images, but in a compact way that reduces the amount of space needed to store the data. Such a representation of the data can be generated using techniques from the area of *content-based image retrieval*, which addresses the problem of matching millions of images based on visual appearance. In our prototype early warning system, we build an index of all images based on local features, instead of the entire global appearance of the images. This strategy allows us to find matching images based on small localized objects that they share. For distinct meme genres of diverse visual appearance, finding just one small shared object could be the link for establishing a valid relationship between images. This gives our technique excellent recall abilities over large collections of images. In order to scale to millions of images, we make use of SURF features (Bay et al. 2008) that can be computed quickly and stored efficiently. The index is an Inverted File (IVF) index trained via Optimized Product Quantization (OPQ) (Ge et al. 2013).

After the index is built, it can be used to find related images through a querying process. In a manner similar to what is done in traditional image provenance analysis, query images are chosen and are matched against the images in the index to return the closest matches (the number returned is a user-defined parameter). The choice of query images could be random (i.e., randomly sampled from all of the available images), or determined through the use of image manipulation detectors that can identify suspicious images. Our prototype system has defined a scoring system relying on the quality of matches between individual objects in images, based on the correspondence between the pre-calculated features in the index with the query.

The matching process for image retrieval results in collections of ranked lists (Fig. 7.4). This is somewhat useful, but what is ideally needed here is a data clustering approach that depicts how each image is visually situated with respect to other related images. Further, each cluster should represent a distinct genre of content that is evident to the human

Fig. 7.4 The output of the provenance filtering process to find related images in a large collection for three different meme genres from Indonesia. Each row depicts the best matches to a query image (the left-most image in these examples) in sorted order, where images ideally share some aspect of visual appearance. Scores reflect the quality of matches between individual objects in images. At the very end of the sorted list, we expect the weakest match, and the very low scores reflect that. These ranks form the input to the clustering step, which presents a better arrangement for human analysis

observer. In a meme context, this means that memes that humans have labeled (e.g., "Action Hero Bolsonaro," "Misogynistic Modi") should be found in a single coherent cluster. With respect to other content, derivatives of an ad for instance should also be grouped together in a similar way. Our prototype system uses a spectral clustering algorithm to produce the final output of the system (Yu and Shi 2003).

An important question is, how well does this prototype early warning system work? We validated the system on the dataset of two million images from the 2019 Indonesian Presidential Election that is referenced above. This case study is particularly salient for this work, in that the results of the election led to violent episodes in Jakarta, some of which were stoked by content found on social media (BBC 2019). Most critically, our validation sought to verify that the prototype system was able to detect useful meme genres from millions of images, as well as verify that the images contained within a genre are meaningful to human observers. In total, the system discovered 7,691 content genres out of the pool of roughly two million images. Some examples are shown in Fig. 7.5. Each of these genres was checked by a human observer to assess visual coherence and to tag the genre with a label that described its content. Roughly 75% of the images were placed into human interpretable clusters. Further, controlled human perception experiments were also conducted to verify that the genres were not simply the product of random chance. These showed that not only were human observers adept at perceiving a pattern in most presented clusters, but also that a majority of the detected genres had a cohesive-enough theme that was identifiable even in the presence

Fig. 7.5 Selections from three different content genres from the total pool of 7,691 discovered by the prototype system

of an impostor image. In other words, the AI system is performing accurate pattern recognition and organizing data at a scale that was once unthinkable to human analysts.

In addition to the AI algorithms, we must also consider how to make the information such a system generates accessible to end users. It is very likely that most of the target users of such a system will not be comfortable interacting with the command line of a computer system. Thus a user interface (UI) layer must be developed and tested, as well as an alert system that will be capable of notifying the correct people with the correct information when an imminent threat begins to trend. It is conceivable that policymakers would require one set of information from this system, while civil society actors would require a different set, and reporters still another set. Working closely with these distinct communities of users is a critical aspect of ensuring this system's utility as an early warning system.

Major considerations remain for the design of a UI that is accessible to users who are not computer scientists. As can be seen in the figures included in this section, the current UI is minimalist by design and can be further developed. At this point in our development, the UI primarily consists of a web interface that presents a list of clusters to users, which can be selected and visualized (Fig. 7.6). We envision the next phase of UI development to not only present the user with genres of memes, but also automatically derived meta-data that describes those genres. Moreover, given automatically derived threat markers from the content (Do the scenes depicted suggest violence? What are the messages found in the text? Does an image share a relationship with content already known to be problematic?), genres of content can be triaged appropriately in order to present the user with the material that requires the most urgent attention.

Limitations of Existing Technologies and a Research Roadmap

Much work still needs to be done so that the prototype system can meaningfully contribute to violence early warning and risk assessment at an operational level. A laundry list of additional features that are still in development includes better data source targeting, data fusion capabilities, the linking of text and image content, natural language understanding, disinformation identification, and the assessment of messaging over time. A good portion of these are related to semantic understanding (i.e., meaning-making)—one growth-edge of artificial intelligence research.

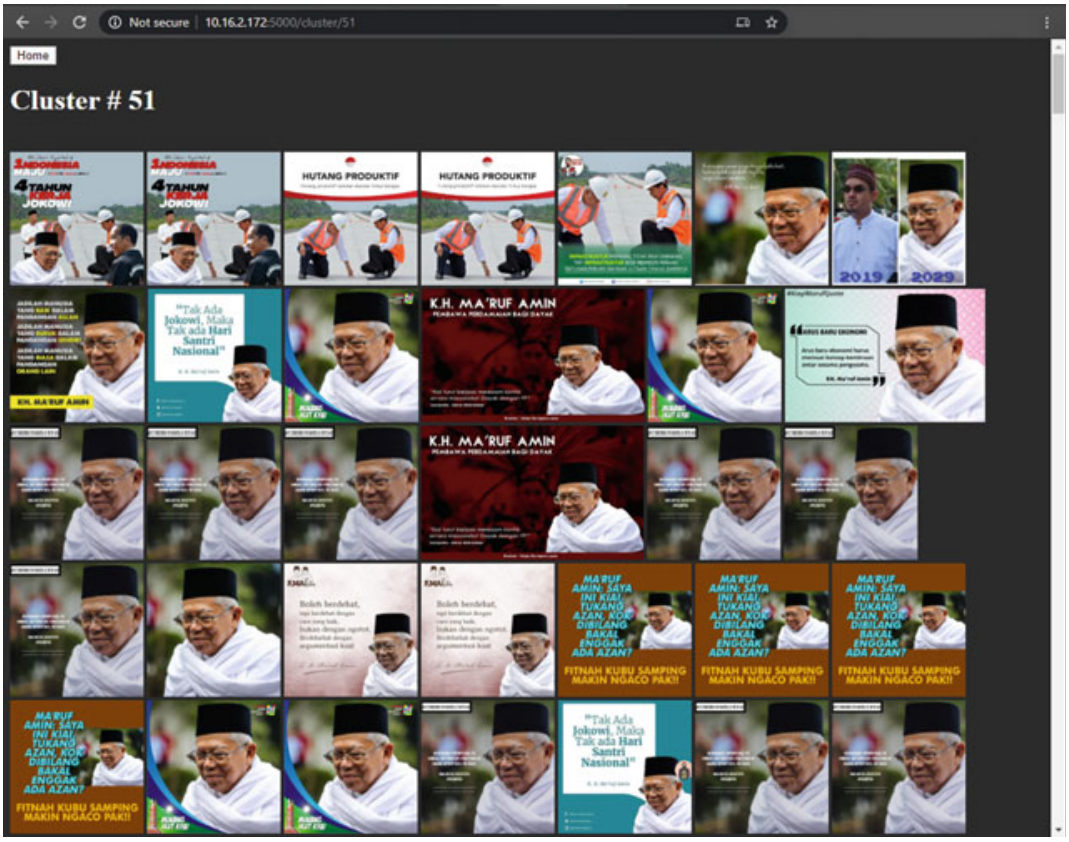

Fig. 7.6 Screenshot of the web-based UI of the prototype system

The high-level goal of our in-process research and development is to design systems that are capable not only of categorizing political memes into particular genres, nor simply of identifying distinct media artifacts that have been manipulated or which may be entirely fake, but which are capable of understanding when these individual items might be indicative of larger trends toward political violence, or when they are being deployed in coordinated ways so as to exploit underlying tensions in particular contexts that are already primed for violence.

As discussed above, current technologies allow for the identification of manipulated media objects in isolation. What has not yet been built are sufficiently robust AI systems that can identify trends occurring across multiple media modalities simultaneously, and which are targeted to incite violence within contexts that have already been identified as volatile or high risk. Consider a scenario wherein a bad actor is deploying a disinformation campaign with the intent to incite violence. In a robust campaign,

we would expect to see thematically resonant media artifacts emerging across modalities in close temporal proximity to one another: a news article here, a photograph there, plus a few videos and some well-sculpted memes designed to solicit response. Of course these media entities would all be shared and liked and cross referenced across multiple platforms and outlets. This is all that is needed for a rapidly spreading disinformation campaign to emerge.

But the ability to identify a targeted campaign across these modalities and platforms *as a single, coordinated campaign* is critical if campaigns intended to incite violence—as well as their individual components—are going to be identified. This task begins by *detecting* that a piece of media has been manipulated or faked but expands into the broader task of semantic analysis and campaign identification across media modalities. Accomplishing this requires mapping the source, flow, spread, and corroboration dynamics as an additional layer on top of the baseline identification of the media object's constituent parts.

The initial assumption here is that manipulated media items are a key signal for our system to find and analyze. Once these manipulated media objects have been detected, the second layer of analysis is *attribution*. Here our main task is to discover traces related to the technological process of creation of the objects which may help us identify when particular collections derive from the same source. This may be possible through simple fingerprinting of media files including EXIF and meta-data embedded in photographs, temporal mapping to identify when particular content is first posted or shared, or even the use of distinct stylistic tendencies in written pieces. Techniques to establish the digital pipeline at the origin of deep fakes and AI-generated media also belong to this category.

The third and final goal we aim for beyond detection and attribution is the design and development of *characterization* methods to help us understand *why* a concentrated effort to generate disinformation might be initiated. In this regard, we aim to develop realistic cognitive models to study the effect that media manipulations and the use of fake media have on the users, their intentions (malicious, playful, political), and also their provoked emotions. The ultimate goal is to design an AI early-warning system capable of monitoring both traditional and social media platforms for trending content that may be part of an influence campaign intended to incite violence.

Accomplishing this threefold task of detection, attribution, and characterization across multiple media modalities simultaneously is an even more daunting task than what the prototype system currently accomplishes, given the sheer volume of content created every second online. It is very unlikely that a team of human analysts would be able to sufficiently identify and analyze coordinated disinformation campaigns in real time. The simple fact is that humans are incapable of performing this task across all relevant media modalities at scale and at the speed required to identify a campaign aiming to incite near-term violence. Similarly, at an algorithmic level, there is more demand for computational resources and/or optimizations to improve the algorithms. However, in contrast to human limitations, technology continues to improve in speed and to lower in cost, making this possible in the near term.

Ethical Considerations

Our discussion in this section is limited to the narrow scope of AI designed to function as a violence early warning system such as we have outlined above. Within this narrowed scope, there is an important set of ethical and policy questions that deserves attention throughout the design process of this system. While the space constraints of this chapter prevent a thorough treatment of each area of concern, we first offer a set of four guiding principles for this system, followed by a set of ethical questions for reflection.

The principles are:

Aim: The fundamental aim of the system we propose is to assist efforts to prevent or lessen violence against civilians. Use of this system should only contribute toward that aim.

Transparency: It is crucial that the structure and operation of the system be presented in such a way that the general public can understand what the system entails, and also to allay concerns over the use of AI in this context. Thus, the main aspects of data collection and analysis will be clearly presented and available publicly, although the complete software system will remain proprietary. As explained below, all data sources are public in origin. We do not rely on surveilling private communications.

Accessibility: We will prioritize accessibility for actors whose work has a demonstrable commitment to advancing human rights and protecting civilians. This may include human rights organizations and civil society actors, think tanks and research institutes, journalists, global and

regional governance institutions such as the United Nations, and those parts of the scholarly research community working on peace and conflict, among others. In some instances, this may include certain offices or departments in governments. A structured, transparent committee that includes members with expertise in computational and social sciences, as well as human rights research and practice, should make recommendations on who has access to the system, with a clearly articulated appeals process for those who are denied access upon initial application.

Independence: The principal investigators are committed to evidence-based research for the common good. Thus, we endorse independence of analysis and objective reporting of results.

These principles will help frame our responses to emergent ethical challenges. In addition to the above principles, several pressing questions have emerged from our early development on this system. Questions such as: Is it ethically problematic to have an AI system "listening" in on Internet communications? What safeguards should be required to prevent the possibility of false positives? Is it possible to prevent bad actors from "gaming" the system? Let us address each question in turn.

The idea of an AI used to "listen" for trends—even trends that threaten mass violence—in online communications may seem ethically problematic. Many observers are wary of using AI to monitor digital information, especially private posts, for fear that it can be used to expand the surveillance powers of states or corporations. We share these concerns, but believe we can address these issues for the purposes of this specific project. It is important to emphasize that all of the media instances that our system ingests are publicly available. While other developers, corporations and even national intelligence agencies may seek to develop ways to eavesdrop on private communications, our ingest system simply "scrapes" publicly available communications that can be seen by anyone. The simplest way to distinguish between these is to think about the difference between reading a user's posts on Reddit versus reading that same user's private text messages; we focus on the former types of sources. A legal gray area exists in collections taking place on WhatsApp (Wang 2018) and other messaging services where one must be invited into a group chat to collect information. Nevertheless, from our perspective, we have more than enough public material to sift. There is no need to dig into private

data, especially given the potential harm that can be done to users if they believe a conversation is happening in confidence.

The problem of a false positive is only a concern if we were proposing a system that would be used in isolation rather than as one piece of a robust violence early warning and forecasting model. We do not believe that this system on its own should serve as a single "trigger" for intervention in particular contexts. Rather, we envision this AI early warning system as one facet of the broader early warning forecasting systems already employed. This system would work in tandem with and provide more real-time granular data about what is happening in particular high-risk contexts. The appropriate interventions would then be coordinated by the various stakeholders in a manner that is appropriate to their particular situation, context, and capacities.

It would be prudent to also consider how such a system might be "gamed" for nefarious purposes by bad actors. Manipulation of such a system might range from the harmless hacker to coordinated manipulation attempts made by state actors. Examples of such manipulation of larger systems abound: In early 2020 a performance artist put approximately 100 cell phones into a hand-pulled wagon and walked around the streets of Berlin in order to provide false information to Google Maps. Everywhere he walked Google began reporting a major traffic jam and rerouting traffic around it (Barrett 2020). Consider also the Internet phenomenon known as "swatting," where a hoax "tip" is provided in order to lead a SWAT team to an innocent and unsuspecting person's home (Ellis 2019). It is easy to see that a system designed to detect short-term onset of violence will undoubtedly be targeted by bad actors. A bot army could be deployed to make it *seem* like mass violence is about to erupt in a country, even when there is little to no *real world* movement. While this is an important concern, we emphasize that this system is not meant to stand in isolation but rather is intended to be a component part of a larger early warning system. Social media is a crucial battleground in contemporary conflicts, but it is not the only one; it is important to corroborate the findings of any single early warning indicator system with other types of evidence from other realms of conflict analysis.

Policy Implications

One further question is how the early warning information provided by this system should be employed. From a policy perspective, there are at

least three broad ways in which the data and analysis provided by our AI early warning system might be used: *response*, *shaming*, and *accountability*. The primary contribution of this system is to *response*, by which we mean it can enhance the abilities of conflict prevention practitioners to respond to escalating violence by providing informed, real-time evidence of an emerging threat based on trending campaigns and communications on social media. This is especially important in cases where information is otherwise lacking or limited, such as in places that journalists and human rights monitors are unable to reach due to physical danger. However, one important qualification is needed: Because this system is just one tool within the larger toolkit available to conflict prevention practitioners, it should not be dispositive on its own for substantially coercive response efforts, such as deploying peacekeepers. However, this system will serve to provide more granular and real-time insights and can thus empower conflict prevention practitioners and, where relevant, peacekeeping missions to react swiftly to defuse an incident as it unfolds. Similarly, this system allows election monitors to better gauge the fairness and legitimacy of an election by providing insights into any influence campaigns or intimidation tactics that are being deployed on social media.

Second, this system may also be used to strengthen efforts to publicly *shame* nation-states that deny employing repressive policies, by showing publicly and in near real time how they are in fact endorsing or even committing violence against civilians. The aim of shaming campaigns—which are a key component of much human rights advocacy work—is to change perpetrator behavior by imposing reputational costs that may be transformed into other more, robust costs, such as economic sanctions, among others. The process of publicly shaming a government before the international community has been important in Myanmar, where external monitoring of extremist social media has confirmed what the government has long denied: that there is a widespread campaign to remove and even destroy the Rohingya population, and publicizing this information has placed increased pressure on the government and its allies to lessen the extent of repressive practices (Human Rights Watch 2019). Nevertheless, even if individual actors are difficult to identify because of obfuscated digital content streams, our system's ability to identify a focused campaign may help human rights advocates understand when vulnerable groups are being targeted and provide additional contextual information as this is unfolding, which can aid pressure efforts. To be sure, we do not contend that such public pressures are able to change policies completely, but

shaming has been shown to have some effect on lessening certain forms of repressive behavior, especially where governments or armed challengers have previously agreed to follow existing human rights and humanitarian law (Hafner-Burton 2008).

Finally, such a system can contribute to future *accountability* measures. Holding perpetrators accountable after major episodes of violence is often exceedingly difficult, not only because there may be a lack of political will to prosecute, but also because it may be hard to obtain evidence of culpability that meets legal prosecutorial thresholds. While the original authors of media artifacts are often difficult to identify, our system helps map campaigns that are intended to prime populations for violence against vulnerable groups. The additional evidence offered by our AI early warning system can enhance investigation into human rights abuses and war crimes, thus increasing accountability.

In all of these facets—response, shaming, and accountability—our AI early warning system strengthens, rather than replaces, existing conflict prevention policies, initiatives, and programs. We underscore that our purpose is not to cast aside the important accumulated knowledge and expertise of peacebuilding and human rights specialists, but instead to assist their work by providing more fine-tuned analyses of dynamic and shifting conflict situations in near real time.

Conclusion

In this chapter we have described our initial work on an AI early warning system for violence. While there is much that needs to be done in order to maximize the potential for AI technologies to assist in preventing large-scale violence from occurring in distinct conflict contexts, we have demonstrated one aspect of the extraordinary value that niche-specific, targeted AI systems can add within novel domains of application; there are distinct contributions new AI systems and technologies can make to enhancing the capacities and efficacy of experts in distinct fields. Our new collaboration of researchers in peace studies, genocide and conflict studies, and artificial intelligence will continue to push the limits of available technologies, even as it helps scholars of peace studies and human rights practitioners understand more about how online campaigns are capable of sparking and shaping on-the-ground conflict realities.

Acknowledgements This article is based on research sponsored by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL) under agreement number [FA8750-16-2-0173] for the Media Forensics (MediFor) program. Support was also given by USAID under agreement number [7200AA18CA00054].

Notes

1. Scholars frequently distinguish between negative and positive peace, the latter involving the dismantlement of the broad social, political, and economic structures that systematically marginalize people, and the creation of conditions necessary for human flourishing (Galtung 1969). This is the ultimate goal of peacebuilding, of course, but our project focuses on the often pressing and immediate need to prevent and end overt episodes of mass political violence.
2. These include the US government's Political Instability Task Force (PITF); the high-level US Atrocity Early Warning Taskforce; the United Nations (UN) Office on Genocide Prevention and the Responsibility to Protect; various regional efforts by international organizations like the European Union, African Union, and Organization of American States; and, increasingly sophisticated early warning and watch lists by non-governmental organizations. See Verdeja (2016).

References

Azeem, Ibrahim. *The Rohingyas: Inside Myanmar's Genocide*. London: Hurst, 2018.

Barrett, Brian. "An Artist Used 99 Phones to Fake a Google Maps Traffic Jam." *Wired*. February 20, 2020. Accessed March 25, 2020. https://www.wired.com/story/99-phones-fake-google-maps-traffic-jam/.

Bay, Herbert, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. "Speeded-up Robust Features (SURF)." *Computer Vision and Image Understanding* vol. 110, no. 3 (2008): 346–359.

BBC. 2019. "Indonesia Post-Election Protests Leave Six Dead in Jakarta." Accessed March 25, 2020. https://www.bbc.com/news/world-asia-48361782.

BBC. 2014. "Stephen Hawking Warns Artificial Intelligence Could End Mankind." Accessed March 25, 2020. https://www.bbc.com/news/technology-30290540.

BSR. *Human Rights Impact Assessment: Facebook in Myanmar*. October 2018. Accessed March 24, 2020. https://about.fb.com/wp-content/uploads/2018/11/bsr-facebook-myanmar-hria_final.pdf.

Cekfakta. 2020. Accessed March 25, 2020. https://cekfakta.com/.

Chirot, Daniel and Clark McCauley. *Why Not Kill Them All? The Logic and Prevention of Mass Political Murder*. Princeton: Princeton University Press, 2010.

Clifford, Catherine. "Elon Musk: Mark My Words, AI Is More Dangerous Than Nukes." CNBC. March 13, 2018. Accessed March 25, 2020. https://www.cnbc.com/2018/03/13/elon-musk-at-sxsw-a-i-is-more-dangerous-than-nuclear-weapons.html.

Dawkins, Richard. *The Selfish Gene: 40th Anniversary Edition*. Oxford: Oxford University Press, 2016.

Edwards, Paul N. *The Closed World: Computers and the Politics of Discourse in Cold War America*. Cambridge: MIT Press, 1997.

Ellis, Emma Grey. "Swatting Is a Deadly Problem—Here's the Solution." *Wired*. August 8, 2019. Accessed March 25, 2020. https://www.wired.com/story/how-to-stop-swatting-before-it-happens-seattle/.

Eveleth, Rose. 2015. "How Many Photographs of You Are Out There in the World?" Accessed March 25, 2020. https://www.theatlantic.com/technology/archive/2015/11/how-many-photographs-of-you-are-out-there-in-the-world/413389/.

Fein, Helen. *Human Rights and Wrongs*. Boulder, CO: Paradigm Publishers, 2007.

Forstall, Christopher W., and Walter J. Scheirer. *Quantitative Intertextuality*. Cham: Springer, 2019.

Galtung, Johan. "Violence, Peace, and Peace Research." *Journal of Peace Research* vol. 6, no. 3 (1969): 167–191.

Ge, Tiezheng, Kaiming He, Qifa Ke, and Jian Sun. "Optimized Product Quantization for Approximate Nearest Neighbor Search." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2946–2953, 2013.

Goldsmith, Benjamin, Charles Butcher, Arcot Sowmya, Dimitri Semenovich. "Forecasting the Onset of Genocide and Politicide: Annual Out-of-sample Forecasts on a Global Dataset, 1988–2003." *Journal of Peace Research* vol. 50, no. 4 (2013): 437–452.

Goldstone, Jack, Robert H. Bates, David L. Epstein, Tedd Robert Gurr, Michael B. Lustick, Monty G. Marshall, Jay Ulfelder, and Mark Woodward. "A Global Model for Forecasting Political Instability." *American Journal of Political Science* vol. 54, no. 1 (2010): 190–208.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep Learning*. Cambridge, MA: MIT Press, 2016.

Hafner-Burton, Emilie. "Sticks and Stones: Naming and Shaming the Human Rights Enforcement Problem." *International Organization* vol. 62, no. 4. (2008): 689–716.

Harff, Barbara."No Lessons Learned from the Holocaust? Assessing Risks of Genocide and Political Mass Murder Since 1955." *American Political Science Review* vol. 97, no. 1 (2003): 57–73.

Heldt, Birger. "Mass Atrocities Early Warning Systems: Data Gathering, Data Verification, and Other Challenges." *Guiding Principles of the Emerging Architecture Aiming at the Prevention of Genocide, War Crimes, and Crimes Against Humanity*, 2012. http://dx.doi.org/10.2139/ssrn.2028534. Accessed March 25, 2020.

Human Rights Watch. 2018. "India: Events of 2018." Accessed March 25, 2020. https://www.hrw.org/world-report/2019/country-chapters/india.

Human Rights Watch. 2019. "World Report 2018: Myanmar." Accessed March 25, 2020. https://www.hrw.org/world-report/2019/country-chapters/burma.

Internet Society. 2019. "Consolidation in the Internet Economy." Accessed March 25, 2020. https://future.internetsociety.org/2019/consolidation-in-the-internet-economy/.

Kiernan, Ben. "Twentieth-Century Genocides: Underlying Ideological Themes from Armenia to East Timor." In *The Specter of Genocide: Mass Murder in Historical Perspective*. Edited by Robert Gellately and Ben Kiernan. Cambridge: Cambridge University Press, 2003.

Know Your Meme. 2020. "Occupy Wall Street." Accessed March 25, 2020. https://knowyourmeme.com/memes/events/occupy-wall-street.

Koonz, Claudia. *The Nazi Conscience*. Cambridge, MA: Harvard University Press, 2003.

Kristeva, Julia. "Word, Dialogue and Novel." *The Kristeva Reader*. Edited by Toril Moi. Oxford: Basil Blackwell, 1986.

Leach, Colin Wayne, and Aerielle M. Allen. "The Social Psychology of the Black Lives Matter Meme and Movement." *Current Directions in Psychological Science* vol. 26, no. 6 (2017): 543–547.

Lupel, Adam and Ernesto Verdeja. "Developing the Political Will to Respond." In *Responding to Genocide: The Politics of International Action*. Edited by Adam Lupel and Ernesto Verdeja. Boulder, CO: Lynne Rienner, 2013. pp. 241–257.

Marcus, Gary, and Ernest Davis. *Rebooting AI: Building Artificial Intelligence We Can Trust*. New York: Pantheon, 2019.

Midlarsky, Manus. *The Killing Trap: Genocide in the Twentieth Century*. Cambridge: Cambridge University Press, 2005.

Moreira, Daniel, Aparna Bharati, Joel Brogan, Allan Pinto, Michael Parowski, Kevin W. Bowyer, Patrick J. Flynn, Anderson Rocha, and Walter J. Scheirer.

"Image Provenance Analysis at Scale." *IEEE Transactions on Image Processing* vol. 27, no. 12 (2018): 6109–6123.

NIST. 2018. "2018 Medifor Challenge." Accessed March 25, 2020. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=928264.

Poundstone, William. *Prisoner's Dilemma: John Von Neumann, Game Theory and the Puzzle of the Bomb*. New York: Anchor Books, 1992.

Robinson, Geoffrey. *The Killing Season: A History of the Indonesian Massacres, 1965–66*. Princeton: Princeton University Press, 2018.

RichardWebster, Brandon, So Yon Kwon, Christopher Clarizio, Samuel E. Anthony, and Walter J. Scheirer. "Visual Psychophysics for Making Face Recognition Algorithms More Explainable." In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 252–270, 2018.

Rocha, Anderson, Walter Scheirer, Terrance Boult, and Siome Goldenstein. "Vision of the Unseen: Current Trends and Challenges in Digital Image and Video Forensics." *ACM Computing Surveys (CSUR)* vol. 43, no. 4 (2011): 1–42.

Roff, Heather M. "The Frame Problem: The AI "Arms Race" Isn't One." *Bulletin of the Atomic Scientists* vol. 75, no. 3 (2019): 95–98. https://doi.org/10.1080/00963402.2019.1604836.

Santoro, Maurício. 2019. "The Brutal Politics of Brazil's Drug War." Accessed March 25, 2020. https://www.nytimes.com/2019/10/28/opinion/brazi-war-on-poor.html.

Secretary General of the United Nations. *Early Warning Systems*. New York: United Nations, 2006.

Sémelin, Jacques. *Purify and Destroy: The Political Uses of Massacre and Genocide*. London: Hurst & Company, 2005.

Shifman, Limor. *Memes in Digital Culture*. Cambridge: MIT Press, 2014.

Stella, X. Yu, and Jianbo Shi. "Multiclass Spectral Clustering." In *Proceedings of the International Conference on Computer Vision (ICCV)*, p. 313. 2003.

Stewart, Frances. "The Causes of Civil War and Genocide: A Comparison." In *Responding to Genocide: The Politics of International Action*. Edited by Adam Lupel and Ernesto Verdeja. Boulder, CO: Lynne Rienner, 2013. pp. 47–84.

Suhartono, Muktita and Daniel Victor. "Violence Erupts in Indonesia's Capital in Wake of Presidential Election Results." *New York Times*, May 22, 2019. Accessed March 30, 2020. https://www.nytimes.com/2019/05/22/world/asia/indonesia-election-riots.html.

Theisen, William, Joel Brogan, Pamela Bilo Thomas, Daniel Moreira, Pascal Phoa, Tim Weninger, and Walter Scheirer. "Automatic Discovery of Political Meme Genres with Diverse Appearances." *arXiv preprint* arXiv:2001.06122 (2020).

Twitter. 2020. "Developer Documentation." Accessed March 25, 2020. https://developer.twitter.com/en/docs.

Verdeja, Ernesto. "Predicting Genocide and Mass Atrocities." *Genocide Studies and Prevention* vol. 9, no. 3 (2016). http://dx.doi.org/10.5038/1911-9933.9.3.1314.

Waller, James. *Confronting Evil: Engaging Our Responsibility to Protect*. Oxford: Oxford University Press, 2016.

Wang, Shan. "WhatsApp Is a Black Box of Viral Misinformation—But in Brazil, 24 Newsrooms are Teaming Up to Fact-Check It." *Nieman Lab*. August 6, 2018. Accessed March 25, 2020. https://www.niemanlab.org/2018/08/whatsapp-is-a-black-box-of-viral-misinformation-but-in-brazil-24-newsrooms-are-teaming-up-to-fact-check-it/.

Weiss, Thomas G. *What's Wrong with the United Nations and How to Fix It*. London: Polity, 2016.

Weitz, Eric D. *A Century of Genocide: Utopias of Race and Nation*. Princeton: Princeton University Press, 2003.

Yankoski, Michael, Tim Weninger, and Walter Scheirer. "An AI Early Warning System to Monitor Online Disinformation, Stop Violence, and Protect Elections." *Bulletin of the Atomic Scientists* (2020): 1–6. https://thebulletin.org/2020/03/an-ai-early-warning-system-to-monitor-online-disinformation-stop-violence-and-protect-elections/.

York, Jillian. 2012. "Middle East Memes, a Guide." Accessed March 25, 2020. https://www.theguardian.com/commentisfree/2012/apr/20/middle-east-memes-guide.