Solving antisemitic hate speech in social media through a global approach to local action

By Dr Andre Oboler, Online Hate Prevention Institute and La Trobe University Law School

Abstract / Introduction

In 2008 the term "Antisemitism 2.0" was coined to describe the normalisation of antisemitism in society through the use of social media.[1] In the past decade the impact of social media in daily life has grown dramatically as has its use as a medium for hate speech.[2] Antisemitism remains one of the most common forms of hate speech in social media along with the rise in anti-Muslim hate speech following the rise of Daesh (ISIS), the resulting refugee crisis and the rise in global terrorism. Other groups in society are also targeted with misogyny, homophobia and racism against Indigenous peoples making headlines around the world. The Jews have again been the canary in the coal mine with efforts to tackle Antisemitism 2.0 leading the way in the broader response to what has become known as Hate 2.0.[3]

The first problem in tackling antisemitism 2.0 is being able to identify antisemitic content in social media in an efficient and effective manner so it can be empirically measured. This problem was identified as a key challenge at the 2009 Global Forum for Combating Antisemitism and a solution involving crowd sourcing of reports and automated verification was presented to a meeting of the Online Antisemitism Working Group of the Global Forum in 2011, the software was presented at the 2013 meeting and formally endorsed after a draft report based on the first 2,024 reported items was circulated at the 2015 meeting.[4] The final report was released on Holocaust Memorial Day in 2016.[5]

The new technical solution allows the problem to be redefined as a quality of service challenge where the level of hate must be constantly measured and kept below a threshold of unacceptability.[6] As was foreshadowed in 2010, if platforms failed to keep the level of hate low enough, governments would step in with regulation.[7] This occurred in 2016 in Germany and the European Union with agreements

[1] A. Oboler, "Online Antisemitism 2.0. 'Social Antisemitism on the Social Web'," *Jerusalem Center for Public Affairs Post-Holocaust and Antisemitism Series*, No. 67 (April 2008)
[2] M. Wendling, "2015: The year that angry won the internet," *BBC News*, December 30, 2015, http://www.bbc.com/news/blogs-trending-35111707
[3] A. Oboler, *Aboriginal Memes and Online Hate* (Online Hate Prevention Institute, 2012)
[4] A. Oboler, *Measuring the Hate: The State of Antisemitism in Social Media* (Online Hate Prevention Institute, 2016)
[5] Ibid.
[6] A. Oboler and K. Connelly. 2014. "Hate Speech: a Quality of Service Challenge". In *Proceedings of the IEEE Conference on e-Learning, e-Services and e-Management*, 117 – 121.
[7] A. Oboler, "Time to Regulate Internet Hate with a New Approach?" *Internet Law Bulletin* 13, no. 6 (2010); A. Oboler, "A legal model for government intervention to combat online hate," *Internet Law Bulletin* 14, no. 2 (2011)

between companies and governments,[8] then in 2017 Germany passed regulatory laws targeting non-compliance.[9]

The solution to antisemitism in social media has two parts. The first is a global effort to create transparency and accountability through a sharing of real-time data about hate speech in social media. The second part is local action in response to this data which is in keeping with the values and norms of each society. For example: criminal sanctions for posters of hate speech; penalties for social media platforms; counter speech exposing hate speech; counter speech promoting alternative positive narratives; education; campaigns targeting: hate promoters, social media platforms or advertisers.

The danger of Antisemitism 2.0

Antisemitism 2.0 is "the use of online social networking and content collaboration to share demonization, conspiracy theories, Holocaust denial, and classical antisemitic motifs with a view to creating social acceptability for such content".[10] The paper describing the new phenomena was pre-released at the Global Forum for Combating Antisemitism in Jerusalem in February 2008, while that week's *New York Jewish Week* carried a page 1 story warning that the phenomena was "potentially more hazardous than the relatively straightforward smear campaigns and petitions of yesteryear".[11] Facebook at this time had just turned four, was slightly less popular than MySpace, and boasted around 100 million users.[12]

As social media's influence continued to grow, the danger of Antisemitism 2.0 was further explained in a hearing before the Italian Parliament's sub-committee on antisemitism. The hearing heard that, "the danger is not so much that people might read content inspired by anti-Semitism, but rather that they may be induced to accept it as a valid point of view, a fact of life, or something with which one may or may not agree, but not something whose dissemination one should oppose. This is where the risk lies. Some people will feel affected by it and will want to do something against anti-Semitism, but others will remain passive and consider it normal, humdrum, legitimate. And this gives rise to a culture in which hatred, racism and antisocial behaviour are able to spread, posing huge risks to law and order and to security".[13] It is not just the online world that is threatened but the values of society as a whole.

The Alt-Right in the United States is a manifestation of Antisemitism 2.0. It began in parts of Reddit and 4Chan,[14] as an "obscure, largely online subculture" before entering the "very center of American

[8] "European Union agreement with Social Media Platforms on tackling Hate Speech," Online Hate Prevention Institute, last modified May 31, 2016, http://ohpi.org.au/european-union-agreement-with-social-media-platforms-on-hate-speech/

[9] M. Connellan, "Germany holds social media companies to account for hate speech," *SBS News*, April 6, 2017.

[10] A. Oboler, "Online antisemitism 2.0."

[11] T. Snyder, "Anti-Semitism 2.0 Going Largely Unchallenged," *The New York Jewish Week*, February 20, 2008, 1. Also online at: http://jewishweek.timesofisrael.com/anti-semitism-2-0-going-largely-unchallenged/

[12] M. Arlington, "Facebook Now Nearly Twice The Size Of MySpace Worldwide," *Tech Crunch*, January 22, 2009, https://techcrunch.com/2009/01/22/facebook-now-nearly-twice-the-size-of-myspace-worldwide/.

[13] "Presentation of the Final Document of the Sub-Committee of Inquiry into Antisemitism", Chamber of Deputies (Italy), October 17, 2011, 304.

[14] "Alt Right," Know Your Meme, accessed September 1, 2017, http://knowyourmeme.com/memes/cultures/alt-right

politics".[15] Rolling Stone describes it as a white supremacy movement linked by a "contempt for mainstream liberals, feminists, 'social justice warriors' and immigrants", [16] but overlooked the most common target: the Jews.

The antisemitism of the Alt-Right and the subculture it emerged from can be seen in /pol/'s efforts to mainstream the antisemitism meme of the Jew.[17] It can be seen the dedication of this sub-culture to white nationalism with antisemitism 'at its theoretical core'.[18] It can be seen the promotion of the idea of Jews as grand manipulators trying to destroy the white race.[19] DNA testing followed up by posts to prove one's 'whiteness' have taken off within these groups.[20] In the Alt-Right, antisemitism can be seen in Richard Spencer's call to "Hail Trump, hail our people, hail victory!" followed by a Nazi salute at the "DeploraBall".[21] It can be seen in the meme of Pepe the Frog as a Nazi with Trump's face, spread virally by the Alt-Right and even by Trump's son.[22] It can be seen in the resulting widespread use of Pepe as a symbol by the Alt-Right and the ADL's response listing it as a hate symbol.[23] It can also be seen in the Alt-Right's (((echos))) targeting Jews on Twitter.[24]

The Alt-Right and those in the sub-culture from which it grew are not just promotors of antisemitism. They are deliberate promotors of the normalization of antisemitic messages in social through social media. They are deliberate promoters and accelerators of Antisemitism 2.0. Within the sub-culture the brainwashing of people to accept the groups conspiracy theories is known as "red-pilling", named after the choice given to Neo in The Matrix movie where he had to choose between a blue pill that would put him back to sleep in an artificial reality, or the red pill which would break him out of this controlled

---

[15] G. Michael, "The seeds of the alt-right, America's emergent right-wing populist movement," *The Conversation*, November 23 2016, https://theconversation.com/the-seeds-of-the-alt-right-americas-emergent-right-wing-populist-movement-69036.

[16] C. Skutsch, "The History of White Supremacy in America," *Rolling Stone*, August 19, 2017, http://www.rollingstone.com/politics/features/the-history-of-white-supremacy-in-america-w498334

[17] A. Oboler, *The Antisemitic Meme of the Jew* (Online Hate Prevention Institute, 2014), http://ohpi.org.au/the-antisemitic-meme-of-the-jew/

[18] E.K. Ward, "Skin in the Game: How Antisemitism animates white nationalism," *The Public Eye*, Summer 2017, http://www.politicalresearch.org/2017/06/29/skin-in-the-game-how-antisemitism-animates-white-nationalism/

[19] This idea of war between the Jews and the white race is the origin of the "Gas the Kikes, Race War Now" slogan common on these forums. The idea of the Jews as the enemy emerges from Nazi literature, including their use of the Protocols of the Elders of Zion, and continues to be portrayed in modern online memes and websites e.g. that of Holocaust denier Andrew Carrington Hitchcock, see http://andrewcarringtonhitchcock.com/jewish-genocide-of-the-white-race-case-closed/

[20] E. Reeve, "White nonsense: Alt-right trolls are arguing over genetic tests they think 'prove' their whiteness," *Vice News*, October 9, 2016, https://news.vice.com/story/alt-right-trolls-are-getting-23andme-genetic-tests-to-prove-their-whiteness

[21] D. Lombroso and Y. Appelbaum, "'Hail Trump!': White Nationalists Salute the President-Elect," *The Atlantic*, November 21, 2016, https://www.theatlantic.com/politics/archive/2016/11/richard-spencer-speech-npi/508379/

[22] A. Ohlheiser, "Why Pepe the Frog's Nazi phase doesn't worry his creator," *The Washington Post*, September 14, 2016, https://www.washingtonpost.com/news/the-intersect/wp/2016/09/14/why-pepe-the-frogs-nazi-phase-doesnt-worry-his-creator/?utm_term=.afc8b7441709

[23] S. Begley, "Anti-Defamation League Declares Pepe the Frog a Hate Symbol," *Time*, September 28, 2016, http://time.com/4510849/pepe-the-frog-adl-hate-symbol/

[24] "Triple parentheses echo," Know Your Meme, accessed September 1, 2017, http://knowyourmeme.com/memes/triple-parentheses-echo

environment and let him see the real world.[25] It is the use of social media to red-pill the public, opening their eyes to narratives that demonize Jews, propagate conspiracy theories, promote Holocaust denial and spread classical antisemitic motifs in mainstream online spaces drives Antisemitism 2.0 forward.

Take for example the triple parentheses identifying Jews on Twitter. Identifying Jews online is not new. The infamous JewWatch website is one of the most well-known and oldest antisemitic websites on the internet and is built on this concept. What the triple parentheses adds is embedding the identification of Jews into the fabric of the Twittersphere. The markers would appear in the Twitter feeds of otherwise regular conversations and normalize the singling out of Jews.

The Alt-Right and others in this sub-culture formulate plans to manipulate the mainstream media, social media, and online culture to spread their narratives – any "escape" of a meme or narrative into the mainstream is seen as victory.[26] The aim is to have the sub-culture's narratives embedded as part of the fabric of the online world and daily life. This leaves individual with the "choice" of becoming red-pilled or continuing life as what the sub-culture derogatorily call "normies".

Following the election of President Trump and the emergency of the Alt-Right as a public force, members of the sub-culture were encouraged to red-pill their family enlarging the support base.[27] The Alt-Right marches, such as that in Charlottesville, further promote the message of white supremacy as a normal part of politics which people should accept. This message was reinforced by President Trump's comment, "I think there is blame on both sides," after the violence at the Alt-Right march in Charlottesville.[28] What started with an effort to normalize antisemitism in the online world has in 2017 shifted to an effort to normalize it on the streets of America. It's not just in America either, when a senior Google representative tells a UK Home Affairs Select Committee that a YouTube video titled "Jews admit organising white genocide" and featuring former KKK Grand Wizard David Duke is "did not cross the line into hate speech" and therefore remains online (with over 91,000 views),[29] that too helps to normalize antisemitism online and in society.

Antisemitism has been, and remains, the canary in the coal mine for society. It is through the prism of the fight against antisemitism that that both the new manifestations of hate and new efforts to tackle it emerge. We must continue to specifically tackle antisemitism 2.0 even as we simultaneously use what we learn to also tackle the wider problem of hate 2.0 affecting other groups in society. The creation of social acceptability for racism, [30] religious vilification (particularly against Muslims),[31] misogyny,

---

[25] A. Marwick and B. Lewis, "The Online Radicalization We're Not Talking About," *NY Magazine*, May 18, 2017, http://nymag.com/selectall/2017/05/the-online-radicalization-were-not-talking-about.html

[26] Ibid.

[27] M. Pearl, "How to Tell if Your Alt-Right Relative Is Trying to Redpill You at Thanksgiving," *Vice*, November 24, 2016, https://www.vice.com/en_au/article/nnk3bm/how-to-tell-if-your-alt-right-relative-is-trying-to-redpill-you-at-thanksgiving

[28] D. Merica, "Trump says both sides to blame amid Charlottesville backlash," *CNN*, August 16, 2017, http://edition.cnn.com/2017/08/15/politics/trump-charlottesville-delay/index.html

[29] S. Oryszczuk, "Google chief: Far right video accusing Jews of 'organising genocide', isn't hate speech," *Times of Israel*, March 15, 2017, http://jewishnews.timesofisrael.com/google-chief-far-right-video-accusing-jews-of-organising-genocide-isnt-hate-speech/

[30] Oboler, *Aboriginal Memes*.

[31] A. Oboler, "The normalisation of Islamophobia through social media: Facebook," in *Islamophobia in Cyberspace: Hate Crimes Go Viral*, Imran Awan (Routledge, 2016).

homophobia and other forms of hate weaken society and makes the fight against antisemitism that much harder.

Accountability of platforms and people

At the 2009 Global Forum for Combating Antisemitism the working group on antisemitism on the Internet and in the media identified the lack of metrics for measuring antisemitism in social media as a major challenge. The challenge remained open and was reaffirmed at a 2011 meeting of the working group and then in a report released at the 2013 Global Forum.[32] It noted the "lack of metrics on: a. The number of problem items in specific platforms e.g. reported groups in Facebook, reported Videos on YouTube; b. The number of items resolved on specific platforms e.g. groups shut down, videos removed, complaints reviewed or dismissed;… d. The time delay between something being reported and action being taken in a specific platform".[33]

The reluctance of social media platforms to tackle antisemitic was clear. Facebook, for example, refusal to recognise Holocaust denial as a form of hate speech and therefore as a breach of its community standards.[34] This was confirmed to the Global Forum's working group in a 2011 letter that stated in part, "the mere statement of denying the Holocaust is not a violation of our policies. We recognize people's right to be factually wrong about historical events."[35] It was readily demonstrated that even obvious cases of antisemitism were being rejected when they were reported to Facebook, for example the picture of Anne Frank with the words "What's what Burning? Oh, it's my family" written across it.[36] Further work looking at 47 antisemitic Facebook pages showed how many remained online despite numerous reports.[37] The report led to formal complaints through the Australian Human Rights Commission in which a mediated solution, involving the removal of the listed content (at least for Australian users) and a commitment to remove any identical content uploaded in the future was given by Facebook. The problem is not limited to Facebook, indeed later research has shown Facebook's response, while still far from acceptable, to the be the most effective of the major social media platforms.

While isolated examples and small samples demonstrate the problem, without detailed metrics there is no transparency and as a result there is no accountability. Social media platforms have been largely self-regulated. Inside the United States, hate speech enjoys first amendment protection, meaning laws seeking to restrict it would be deemed unconstitutional. Outside the United States, the platforms argue they are mere carriers and it is the users who should be prosecuted if illegal hate speech is uploaded.

[32] A. Oboler and D. Matas, "Online Antisemitism: A systematic review of the problem, the response and the need for change," Israeli Ministry of Foreign Affairs, 2013, http://mfa.gov.il/MFA/AboutTheMinistry/Conferences-Seminars/GFCA2013/Pages/Online-Antisemitism-A-systematic-review.aspx
[33] Ibid.
[34] A. Oboler, "Facebook, Holocaust Denial and Antisemitism 2.0," *Jerusalem Center for Public Affairs*, August 27, 2009, http://jcpa.org/article/facebook-holocaust-denial-and-anti-semitism-2-0/
[35] The full text of the letter can be seen in: Oboler and Matas, "Online Antisemitism", 50.
[36] "Facebook Fails Review", Online Hate Prevention Institute, September 14, 2012, http://ohpi.org.au/facebook-fails-review/
[37] A. Oboler, *Recognizing Hate Speech: Antisemitism on Facebook* (Online Hate Prevention Institute, 2013), http://ohpi.org.au/recognizing-hate-speech-antisemitism-on-facebook/

This argument is problematic when the platforms decide what content is promoted to whom and profit from the existence of the content.

The 2009 Global Forum for Combating Antisemitism recommended that carrier immunity was "too broad and needs to be limited in the case of antisemitism and other forms of hate", more specific it recommended that "while real time communication may be immune, stored communication e.g. user published content, can be brought to a service providers attention and the provider can then do something about it. Opting not to do something about it after a reasonable time should in all cases open the service provided up to liability."[38] The 2013 report to the Global Forum, which repeated this recommendation and presented the TEMPIS Taxonomy which outlined the different types of online communications defined according to factors such as timing, empowerment of users, moderation, publicness, identity and social impact so that similar types of communication could have the same expectations applied to them regardless of the social media platform being used.[39] A 2010 article warned that, "those who profit from user generated content need to be given the responsibility to take reasonable steps to ensure their platforms have a robust response to the posting of hateful material. The role of government, and the law, is to ensure reasonable steps are indeed taken".[40]

A draft report released at the 2015 Global Forum for Combating Antisemitism for the first time provided a large sample of data about antisemitism in social media. The final report, released on International Holocaust Remembrance Day, January 27th 2016, added statistics about platform responsiveness.[41] Based on a sample of 2024 unique items of antisemitic content from across Facebook, YouTube and Twitter the report divided the content both by platform and across four different categories of antisemitism: traditional antisemitism (49%), New Antisemitism (34%), Holocaust denial (12%) and promoting violence (5%).[42] The percent of content removed by the platforms, after 10 months, varied greatly both by category and within each category, the high was Facebook removing 75% of content promoting violence and the low was YouTube removing just 4% of New Antisemitism.[43] The full spread can be seen in Table 1.

Table 1 Percent of antisemitism removed after 10 months

	Traditional	New Antisemitism	Holocaust Denial	Violence
Facebook	42%	27%	58%	75%
Twitter	25%	20%	20%	14%
YouTube	9%	4%	10%	30%

[38] Oboler and Matas, "Online Antisemitism", 30.
[39] Ibid 5—10.
[40] Oboler, "Time to Regulate", 105.
[41] A. Oboler, *Measuring the Hate: The State of Antisemitism in Social Media* (Online Hate Prevention Institute, 2016), http://ohpi.org.au/measuring-antisemitism/
[42] Ibid p 7.
[43] Ibid.

In 2017, after efforts to resolve the problem of online hate through agreements between the government and the social media companies failed to deliver the desired results, Germany became the first country to legislate liability for platform providers. The new law can result in a fine of up to 500 million euros for platforms that systematically fail to remove obvious breaches of Germany hate speech law within 24 hours.[44] The Germany approach is a positive step forward, and long overdue, but it relies on the ability to track social media hate speech in order to be properly applied. It is also, at present, a blunt tool without differentiation for the different types of communication which may be used online, as discussed in the TEMPIS Taxonomy. Not all hate needs the same priority of response, content which is inciting violence may need a more rapid response than Holocaust denial (for example) while content shared publicly in a form that can go viral may need a more urgent response than the same sort of content sent as a private message to a single person.

What's is that there is a need to track antisemitic content, and content related to other forms of hate and extremism, both at the level of individual items and at the level of summary data showing the content which is impacting society. The challenge first presented at the Global Forum in 2009 is now more urgent than ever.

Approaches to monitoring antisemitism in social media

The gathering of data on online antisemitism can be approached in three ways: expert solicitation, automation through artificial intelligence (AI) or crowd sourcing. Each approach has both advantages and draw backs. The best approach is one which synthesizes the contributions of all three approaches to triangulate a more accurate and complete picture.

Expert solicitation is the oldest approach and involves experts first finding and then assessing examples of antisemitism. The problem is of course the time involved in this task, the limited number of experts, and the fact that using experts for anything but the most viral and high impact cases is simply not an effective use of resources. As an ADL spokesperson told the New York Jewish Week when the issue of Antisemitism 2.0 was first raised back in 2008, "we can't sit here all day monitoring YouTube and Facebook".[45]There is a role for experts, but it needs to be reserved for analysis, investigating new phenomena of antisemitism, providing commentary on high impact cases and perhaps most importantly for training both people and artificial intelligence systems. This includes training for civil society organisations, volunteers, the staff of both social media companies and the companies they outsource relevant functions to, as well as training people in government across the areas of human rights, public policy, law reform and law enforcement.

Artificial Intelligence is promoted by some as a silver bullet. It involves algorithms that either crawl through the content on social media sites, or with the consent of a platform provider directly access the content on the platforms servers, and then seek to access this content. AI solves the problem of scale by applying raw machine power to the task. Given enough time, it is able to read through all the content is has access to. The problem is in understanding the content. There are two limitations, the first is an inability to process certain content, for example, an AI agent may be limited to parsing text and

[44] M. Connellan, "Germany holds social media companies to account for hate speech," *SBS News*, April 6, 2017, http://www.sbs.com.au/news/article/2017/04/06/germany-holds-social-media-companies-account-hate-speech
[45] Snyder, "Anti-Semitism 2.0".

therefore ignore the vast amount of content in images and videos. The second is understanding the text it reads, processing it and giving it context.

The first problem can in theory be addressed with more complex algorithms which can extract text written into an image, or transcribe sound, these tasks, however, significantly increase the complexity and the cost of the processing. With the huge volume of new content being uploaded all the time these approaches are not practicable. Nor are they likely to be practicable in the future as the quality of the content continues to increase, requiring greater processing power whenever such additional processing power becomes available. What about messages delivered through the images themselves? The Anne Frank meme previously discussed require the ability to recognize Anne Frank, a knowledge of her connection to the Holocaust, then an association between burning, crematoria and the extermination of Jewish families. To further confuse the AI, the statement of her family burning is inaccurate as it was her father who survived and published her diary. For a human with the appropriate background knowledge, the message mocking the victims of the Holocaust is clear. For AI, extracting this message from the many possible messages is beyond what's possible.

The second problem, that of context and understanding, occurs even when the processed content is regular text. Simple AI approaches use keywords to identify hate speech. Imagine running a Google search for "kike" on Facebook (to do this, enter the search in Google as "site:facebook.com kike" without the quotes). The problem with this approach is immediately obvious. There are 1,750,000 results for this found on Google but a quick glance shows that many of them would be false positives. One word is not enough to accurately find such content. What about kike and gas together? The first result is a page called "Kike Gas" which repeatedly posts pictures of an oven with gas cylinders of various sizes, a refill price and a phone number.[46] The page has an address in Puetro Rico and a picture of a real-world gas cylinder storage area with a sign saying Gas Kike. The page appears to be relate to a real business and there is no overt antisemitic content. Searching for kikes (plural) and gas gives better results: a blank Facebook page with the name "Gas the Kikes" claiming to be a university,[47] a post from a group combating antisemitism which quotes antisemitic phrases,[48] a Alt-Right like page dedicated to "Aryan Argentine Memes",[49] a page from Facebook's Help Community titled "I am sick and tired of the antisemitism that is allowed on this site" in which people have posted examples of antisemitism some on which have been linked to from that page for months with no resulting action,[50] and a page titled "Right Wing Death Squad",[51] among many other false positives. A more specific and search for "gas the kikes" gives 594 results from YouTube but none on Facebook (despite the presence of the "gas the kikes" page we know exists).[52] On YouTube itself a search for "gas the kikes" gives 176 results.[53] On Facebook's internal search looking for "gas the kikes" brings up a post with the lyrics of an antisemitic music video with phrases such as "I wanna gas all the kikes until they become zero",[54] another post says

[46] https://www.facebook.com/KIKE-GAS-294941390704685/
[47] https://www.facebook.com/pages/Gas-the-Kikes/920193234669031
[48] https://www.facebook.com/Documenting.Anti.Semitism/posts/1356674271077720:0
[49] https://www.facebook.com/AryanArgieMemes/?nr
[50] https://www.facebook.com/help/community/question/?id=881307688690518
[51] https://www.facebook.com/Right-Wing-Death-Squad-591174651085238/
[52] Search as at 15 October 2017.
[53] Search as at 15 October 2017.
[54] https://www.facebook.com/enix.sho/posts/1491580350918573

"Gas the kikes, race war now!" over a background of rainbow and multi-coloured heart balloons,[55] many others use the phrase to describe what antisemites are saying as they comment on news stories.

Text analysis is not much more advanced than a search. If can find specific phrases, or combinations of words and phrases, but it can't really tell if the content is antisemitism or a post discussing antisemitism and seeking to counter it. However well it does, there will always be many false positives. Without human review, the results are likely to be misleading – just as any search is likely to provide some irrelevant results. As can be seen, the results also depend on the search tool. Facebook clearly allows its internal search far more access than Google is able to get. YouTube mean time only searches the titles and descriptions of videos while a Google search for content on YouTube also picks up phrases appearing in the first few comments.

More advanced text analysis can be seen in "Conversation AI" a tool launched by a Google subsidiary in September 2016 and "designed to use machine learning to automatically spot the language of abuse and harassment —with… an accuracy far better than any keyword filter and far faster than any team of human moderators".[56] The tool is still based on text analysis and the use of phrases and the presence or absence of other words to determine if content is antisemitic. While it may be an improvement over simple keyword identification, it is far from a robust solution. Even if it were, 4Chan's /pol/ quickly responded with "Operation Google" which sought to use "Googles" as a code word for African Americans, "Skypes" as code word for Jews, "Yahoos" as code Mexicans and "Skittles" as code for Muslims.[57] The use of code words greatly complicates the process of text analysis, particularly if the code words are regularly changed. In a fight between regulation by Artificial Intelligence and circumvention through human creativity, the computers have little chance of winning. This is especially true when the message seeking to be regulated is an idea which can be rephrased and adapted, rather than a repetition of a known items of content such as song or movie.

The last approach is that of crowd sourcing. This is the method used in the 2016 report previously discussed. Through a custom built advanced online reporting tool called FightAgainstHate.com, members of the public reported and categorized items of antisemitism they found across Facebook, YouTube and Twitter.[58] The volume of data collected is far higher than could be gathered by experts alone, but far lower than what an AI tool could find in an automated search. The main weakness of this system the potential for users, through ignorance or malice, to incorrectly make reports, for example, organized groups could seek to use the tool to report their opponents (be it businesses, sporting teams, schools, political parties etc.) or to report content they disagree agree with but which is not hate.[59] The solution has two parts, one part sees users offering to review content others have reported, and the system determines which items each person gets to insure an independent judgement, and the other involves the limited use of experts to valid items allowing a model of trust to be developed so the

[55] https://www.facebook.com/myron.dawson.965/posts/130188527710071
[56] A. Greenberg, "Inside Google's Internet Justice League and its AI-Powered War on Trolls," *Wired*, September 19, 2016, https://www.wired.com/2016/09/inside-googles-internet-justice-league-ai-powered-war-trolls/
[57] P. Tambutto, "4chan Aims to Fill Google with Racism Following 'War on Trolls'," *Crave*, September 23, 2016, http://www.craveonline.com/design/1125143-4chan-aims-fill-google-racism-following-war-trolls
[58] Oboler, "Measuring the hate".
[59] Oboler and Connelly, "Hate Speech".

system is aware which non-experts tend to agree with the experts judgements.[60] This approaches uses people to identify and review the content, but artificial intelligence approaches to ensure quality.[61]

The best solution would use experts to calibrate a Crowd Sourced system. It would use the crowd sourcing system to review content collected by the AI tools, as occurs with items people have reported. This would give statistics on the level of false positives in the AI system. It would also review how many items in crowd sourced system, gathered initially from human reports, were not found by the AI. This would give data on the AI's blind spots and the degree of false negatives. The three approaches together would allow a triangulation on the real nature of antisemitism in social media.

Transparency through global cooperation

Creating real transparency around the problem of Antisemitism 2.0 requires a global approach. Social media platforms often block content in particular countries rather than globally. Difference can occur between the treatment of content in different languages. A crowd sourced approach fundamentally needs the support not only of a large crowd, but of one seeing social media through the lens of different countries and languages and calibrated to relevant experts. For a truly global picture the tools and methodology also need to be consistent.

The work of the 2016 "Measuring the Hate: the state of antisemitism in social media" report based on crowd sourcing is limited to the English language. The data is not easily accessible to other researchers and was managed by a single organisation. A 2017 report from the World Jewish Congress and Vigo Social Intelligence looks at antisemitism across many countries; [62] it is based on automated text analysis (with some manual review) and skewed towards expressions of antisemitism the software was able to easily identify which results, for example, in over reporting on Twitter and under reporting on YouTube.

The next step to global transparency is to enable the FightAgainstHate.com reporting tool to operating in multiple languages and to allow it to be embedded on the websites of many organisations and configured to their needs. A team of five developers have worked through 2017 to enhancing FightAgainstHate.com to meet these requirements. The new version will be released in early 2018 with an invitation for organisations to partner in the project, help translate the tool to their language, and to be among the first to use it.

An additional tool, CSI-CHAT (Crowd Sourced Intelligence - Cyber Hate and Threats), have been developed over the last two years by nine developers and enables organisations to access and work with the data from FightAgainstHate.com. Data can be sorted, searched, classified, annotated and compiled. The tool also produces statistical reports on the data including trend analysis and dataset comparisons.

Organisations will have access to work with the data reported via the gateway on their own website. Each organisation will be able to choose whether they share this data with other organisations, or whether they will release it into the common data pool. Access to additional items from the common data pool will be available in return for reciprocity. As the same item may be reported to many organisations and only the first to release the item to the common data pool will be created with it,

[60] Ibid.
[61] Ibid.
[62] *The Rise of Anti-Semitism on Social Media: Summary of 2016* (World Jewish Congress, 2017), http://vigo.co.il/wp-content/uploads/sites/178/2017/06/AntiSemitismReport2FinP.pdf

there is an inbuilt incentive to pool data. Even where the specific items of data are not shared, the system can account for duplicates in the private data sets of multiple organisations and provide a true picture of the total numbers of items reports and of their nature.

Future work will allow organisations to upload additional datasets gathered through automated approaches and access their coverage for different types of antisemitism. Access to the human collected data, and the support of experts, will automated tools to be further improved.

Accountability through local action

Reports on the hate speech real people have seen and reporting with a particular country, and evidence of any failure of social media companies to appropriately respond, will enable national governments to hold social media platforms accountable. This is this data, from people within the jurisdiction, with commentary from local experts in-line with the values and norms of the society, can support legal schemes like the one created in Germany in 2017 and ensure that they are practically and routinely applied.

Local organisations can also work with the data in a practical way. Knowing the common narratives of antisemitism, including any new narratives or symbols, can assist with the development of responses and counter narratives. Being able to monitor trends can also help to assess the effectiveness of strategies seeking to combat antisemitism. Tracking individual items can allow repeat abuses to be identified and the efforts of law enforcement in this regard supported by civil society. The can also form the basis for education in schools and society more general on responding to antisemitism. Based on greater information campaigns highlighting areas of weakness in the platforms response can be initiated. While the local action may different around the world, with access to a common platform best practices can be created and shared.

Conclusions

The problem of Antisemitism 2.0 is growing globally. The normalisation of hate that started online is increasingly being manifested offline. We can't tackle the problem of antisemitism without paying specific attention to its normalisation through the Antisemitism 2.0 in social media.

To meet the challenge of Antisemitism 2.0 we need data. We need a real-time picture of what's occurring and how its changing. WE need this picture to stretch across countries and across language barriers. We need to draw on experts, artificial intelligence and the resources of the public through crowd sourcing. We need to know not just what is out there, but what people are seeing, what's having an impact.

In addition to the global picture, we need information at the national level to support action against the antisemitism by civil society and governments. From transparency we can create accountability, but creating that accountability is a national responsibility and must take place within the norms and culture of each society.

We must work together and provide, expertise and cooperation needed to bridge the technological gap, the language barriers, and the cultural differences and empower both civil society and governments to tackle the rising global problem of Antisemitism 2.0.

Bio

Dr Andre Oboler is lecture in the Law School at La Trobe University and CEO of the Online Hate Prevention Institute (OHPI). As part of La Trobe's Masters of Cyber Security, Dr Oboler is developing innovative new courses with a focus on cyber-terrorism, international warfare, privacy and surveillance. In his role at OHPI, Dr Oboler develops new methodologies and approaches for monitoring, measuring and responding to online hate. Dr Oboler is a member of Australia's delegation to International Holocaust Remembrance Alliance and co-chair of the Global Forum for Combating Antisemitism's working group on antisemitism on the Internet and in the media. He is a member of the executive of the Jewish Community Council of Victoria, a Councilor on the Executive Council of Australian Jewry and Vice Chair of the IEEE Australia Council. He holds a PhD in Computer Science from Lancaster University (UK) and an LLM(Juris Doctor) from Monash University.