Combatting Terrorist Content: The Social Media Challenge (11/5/19)

00:00:26 Noah Rauch: Good evening, everyone. My name is Noah Rauch. I'm the senior vice president for education and public programs here at the 9/11 Memorial & Museum. It is my pleasure to welcome you to tonight's program, "Combatting Terrorist Content: The Social Media Challenge." As always, I'd like to extend a special welcome to our museum members and to those tuning in to our live web broadcast at 911memorial.org/live.

00:00:51 The near-ubiquity of social media platforms have brought about deep questions about their role and their ability in stemming the spread of misinformation, violence, and extremist content. During a speech at Georgetown University last month, Mark Zuckerberg explained that Facebook's focus has been to "give people voices and bring people together." "To err on the side of free expression." And implicit in this idea is that good ideas, good arguments, good discourse, will drown out bad ones. This program will look to examine the efficacy of this idea across platforms, in a world where what constitutes extremist content and misinformation is contested, and the scale of social media platforms and the content uploaded to them is unprecedented.

00:01:31 Every day, Twitter sees about 500 million tweets, 350 million photos are uploaded to Facebook, and 82 years of content are uploaded to YouTube. How do you actively and appropriately monitor free expression in the face of this, especially as the content is evolving, often live and often contested? To discuss these questions, and others, I'm delighted to welcome David Tessler and Joan Donovan.

David Tessler is a public policy manager on the dangerous organizations team at Facebook. He joined the company in May 2019 after almost 20

years of public service in the U.S. federal government. Most recently, Mr. Tessler was at the State Department, where he held a variety of positions, including acting sanctions coordinator and deputy director of policy planning.

00:02:19 Mr. Tessler also worked for seven years at the Treasury Department, where he focused on sanctions policy and counter-terrorist financing. Mr. Tessler began his career in the Foreign Service and was posted at Bucharest, Baghdad, and to the U.S. Mission to the United Nations in New York.

Dr. Joan Donovan is director and lead researcher of the Technology and Social Change Research Project at the Shorenstein Center on Media, Politics, and Public Policy at the Harvard Kennedy School. This project is leading the development of 100 case studies and training of over 100 practitioners, including journalists, researchers, and community organizations, to enhance the field of critical internet studies as a means to better identify, understand, and combat media manipulation around the world.

00:03:01 Dr. Donovan's research and teaching interests are focused on media manipulation tactics and techniques, effects of disinformation campaigns, and adversarial media movements that target journalists.

As you can see, we are in for a wonderful conversation this evening. And we thank you both for joining us this evening to share your insights and your expertise. So with that, please join me in welcoming David Tessler and Dr. Joan Donovan in conversation with executive vice president and deputy director for museum programs, Clifford Chanin.

(applause)

00:03:38 (no audio)

(laughter)

00:03:52 (laughter)

(no audio)

(no audio)

(no audio)

00:04:32 (no audio)

00:04:52 David Tessler: ...so big, that those kind of concrete definitions help us in, in trying to find the content, the violating content, the extremist and violent content, and taking it-- take it down as soon as possible. So, we try to do it really concretely, and we find the more concrete we can get, the more nuanced we can get, the better that we can bring down the content at-- as quickly as we can, as quickly as possible.

Clifford Chanin: Joan?

00:05:20 Joan Donovan: Uh, yeah. So, as a researcher, I come at this problem in probably a different way in the sense that I have to look for content across the web in all platforms. I have to look at what constitutes a media manipulation or a disinformation campaign. So, when we talk about media, we talk about media as an artifact of communication, right?

So, everything that you post online is a piece of media, and it's a remnant of a thought or an event or-- and the capacity for the online environment to structure media and hold on to it and archive it and keep it for long

periods of time actually changes our relationships to what we might call "the real world" or "I.R.L."

00:06:06 And so, we study media, and then, when we study media manipulation, we're looking at, well, what are the deceptive ways in which people are using low tech, or even no tech, to deceive broad publics? To pretend to be bigger and more impactful than they are? For the last ten years, I have been a researcher of online social networks, social movements, online social movements, and internet communication.

00:06:34 And so, when people started to realize that there was a problem online, it's because it had reached such an enormous scale. There had been so many people good at deceiving using media, and using social media, in particular, that we had a problem that really had outgrown any potential first-stop solutions.

00:06:56 And so, as we define media manipulation, and then disinformation in particular, we often run into what might be classified as terrorist content in the sense that it is torture, or beheadings, or messages meant to demonize different groups of people. And then our job as researchers is, try to find that evidence across every platform.

00:07:22 So, you can imagine, it's a really-- every day is a bad day at work, right? And so, for us, we're always trying to think about, well, is it an image, how do we do this image search, how do we find this on different platforms? And then our role is to try our best, either if journalists are asking us about the production of some kind of content, to help journalists get to the bottom of things and source materials appropriately, or our role sometime is to actually help the platform company see the problem differently than they already do. And so, in looking at it, though, over the last ten years, the scale is just enormous now, in terms of how big some of these disinformation campaigns can get.

00:08:06 Clifford Chanin: So, I want to disaggregate that a little bit, because the disinformation campaigns cross over into certain gray areas and the

politics of different countries. But if we start with the terrorist materials, the violent materials, is that an easier fix in terms of being able to identify it and being able to keep it from circulating? David, do you want to start that one?

00:08:30 David Tessler: Sure. So we use technology, we use people, and we use partnerships to try to get at that terrorist content that's trying to come, or is, has made it on platform. When it comes to Al-Qaeda and ISIS content, we don't actually wait for it to come to Facebook. We go out and we try to find that content, whether it's videos or images, we bring it to Facebook, and we put it into our, into our content-matching banks, into, into our systems, our screening systems, so that we can try to catch that material, that content, before you even successfully can upload it.

00:09:11 Clifford Chanin: Give us some examples of how it's identifiable beforehand.

David Tessler: So, for example, we will go out and we will find ISIS imagery, whether it's a logo or a photograph, some kind of image or a video. We will then put it into our system, into our screening system, and then, if you try to upload that logo or that, that video, our content-matching system will catch it before you're even able to upload it and will stop you.

00:09:41 We also, though, have what we call machine learning classifiers. So, if you try to upload an image that isn't quite the ISIS image that we have in our system but it's similar, you might be able to upload it because the-- that first line of defense hasn't, hasn't caught you. But we then have these machine learning classifiers. Those classifiers look at content that you're uploading and try to figure out if it's similar to something that we have in our bank, something similar that we have in our systems.

00:10:16 So, there are all kinds of different contextual clues we have as to why something might be similar to a piece of content that we've already identified as being ISIS or Al-Qaeda. So those, that machine learning classifier system-- and we constantly try to improve it, add more

information to the algorithm so it gets better and better-- that system, those classifiers, will then search, search for content that is similar enough that it will either, if it's sure, if it's confident that it's so similar that it's pretty sure that it's ISIS or Al-Qaeda, it will automatically delete it.

00:10:57 If it, if the classifiers think, "Well, it's, I'm pretty sure"-- "I" being the computer-- "I'm pretty sure, but I'm not 100% sure," it will send it to a human reviewer, and then, and a human reviewer will look at it. So, we use-- we use technology, we then use human reviewers, and we have, we have 15,000 human reviewers who are looking at this content and trying to see if it is, in fact, a match, a violating, a piece of violating content.

00:11:28 Um, and then we use partnerships with other tech companies, um, where we collaborate really intensely on trying to make sure that we're all working and swimming in the same direction, all working in the same direction, to find this content and get it off our platforms quickly. So, we have all of those layers of defense.

00:11:50 Um, I think, when it comes to Al-Qaeda and ISIS content, um, we're making really good progress, but this is, this is a battle that, that is gonna be constant, right? That the terrorists are going to continually try to outmaneuver us, to find new ways to get, to get around our systems, to upload the content, because they want to be on, on our platform and on other platforms.

And so we have to constantly be trying to stay one step ahead of them, to improve our technology, improve our algorithms, and to understand their, their intent, their aims, and their goals, so that we can figure out how best to protect, protect our users, protect our space.

00:12:38 Clifford Chanin: Joan, in that context, I mean, you have this extraordinary machine capacity, if you will. You also have the human evaluative capacity. But, nonetheless, it doesn't always hit the mark. So, what are the issues as you see them with this review process? And, I guess the

question that I would follow for both of you is, you know, is some degree of failure to catch these sorts of things built into the system itself?

00:13:09 Joan Donovan: I think we have to take a bit of a step back and think about, well, what are the features of Facebook that incentivizes terrorists to want to be on that platform in the first place? And part of it has to do with access to all of you, and every user, and a broad public, and very few barriers to getting into your feed.

00:13:34 And in fact, you know, my head started exploding when I was thinking about, when Facebook rolled out advertising technology, because no longer different groups stuck in echo chambers and known networks, but then they can pay to target specific people-- often vulnerable people.

00:13:54 And so, the features of Facebook itself are incentive to create a media wing of your militia, for example, and so we don't just look at, um, stuff related to known terrorist networks, but we look at burgeoning movements and how much time they invest in their social media channels. Because it's about reaching out.

00:14:20 And we group them loosely into two different categories on our team. Who are the messengers? That is, who is trying to carry an ideology and use their social media platforms to espouse a particular set of beliefs? And then there's a special group that we look very intensely at, is mobilizers. These are people that are doing active recruitment. How are they trying to use that messaging to bring people in, to set up group chats, and to also bring them in to other spaces offline where they can do more of the radicalization work?

00:14:58 And so, it's a really... it's a really different process once you start to realize that the technology itself is facilitating quite a bit of this connectivity. And by design, of course-- Facebook is about connectivity. But I think if we roll back the clock ten years and we thought about, well, what kind of connectivity is good for society, we might have thought about the product differently. We might have said, "Okay, different kinds of connectivity are gonna be important. Groups are gonna be important,

but maybe they shouldn't be massive. Maybe the connectivity across, you know, a couple million people is a different kind of problem than the connectivity across a group of, say, 30 or 40 people."

00:15:46 And so, with that massive connectivity comes a whole slew of different problems that are functions of the product itself. And so, from our perspective as researchers, we're constantly trying to struggle against, yes, of course, you have A.I., yes, of course, it works when it has a very particular set of parameters. But we're dealing with human actors that are incentivized to create workarounds and are very good at that, as well.

00:16:15 And so, when you say this is a forever problem, I agree with you. But also, I have to think about, well, is the forever problem baked into the structure of the internet itself, or is it amplified by the products and the platforms that terrorists are using? And in some instances, certain features of different platforms are a significant incentive for them to continue to try to spread their message or to mobilize different groups of people.

00:16:45 Clifford Chanin: Yeah, but it does seem like all of these platforms, not just Facebook, are intending to provide more communication and more connection to its users. And so, you know, if that really is the model, in some sense, playing defense is automatically going to put you in a position of having to catch up all along the way.

00:17:06 David Tessler: Well, we're also playing offense. And I'd like to just give two examples of where our size has been effective in, in trying to counter terrorism. One is the awful-- I don't know if you remember-- but the awful attack in Nigeria years ago, where Boko Haram kidnapped hundreds of Nigerian girls. There was a movement started in Nigeria, Bring Back Our Girls. It was a small movement. It was grassroots. They actually used Facebook primarily to spread the movement globally to the point where, you know, then the first lady, Michelle Obama, was, was involved in holding a sign that said, "#BringBackOurGirls."

00:17:49 That's an example of where our size, where our connectivity, is really working to help promote these, these really important efforts to counter terrorist activity or actions or, in this case, this awful, awful Boko Haram kidnapping. One other example is, and I mentioned, we have the technology, we have people, and we have partnerships. We started partnering with a nonprofit called EdVenture in 2015, to give a challenge globally to university students around the world to try to come up with social media campaigns to counter violent, extremist, hateful speech.

00:18:34 Um, so, 2015, not that long ago. We have this campaign, this program is now in 75 countries. It's worked, I think, with over 6,500 students and helped to generate 600 of these anti-extremist campaigns that, through Facebook, has reached 200 million people. So these are examples of where the size and where the connectivity is really essential in playing offense against the terrorist, hateful, and extremist content.

Joan Donovan: I just...

Clifford Chanin: Yeah.

00:19:10 Joan Donovan: I want to just pick up a little bit on that, because the activists that are incentivized to use the platform in order to spread those messages, you know, it's not just a function of Facebook that they provided, you know, a conduit to an international audience.

Those activists took a real risk. I mean, a serious risk in developing content, and, you know, having some of their real identities, um, laid bare to the public. And so, it's not the case that there's just this kind of, like, clean break where you can say, "Well, activists use this platform in order to spread awareness about this other thing that we know is terrible and is also, exists on our platform," like Boko Haram has used Facebook in very particular kinds of ways, as well.

00:19:59 But activists, many of whom I talk to, are very afraid of governments using different technologies that you've built in order to, you know, harness the power of a crowd to be retargeted and to be surveilled. And so it's not a clean distinction to say, activists, on the one hand, benefit, and terrorists, on the other hand, are, you know, somehow at odds with this.

00:20:28 There's an interactive effect that we have to consider related to the risk that activists take on when they decide that they're gonna use platforms in order to combat terrorism or police brutality. And those risks are not necessarily an unmitigated good.

Clifford Chanin: So the platform itself, you know, is, of course, highly visible. We were talking before, Joan, and I'd ask you to go back into a point that we discussed before the program, where the algorithms, you said, that underlie these programs assume good-faith communication on everybody's part.

Joan Donovan: Mm-hmm.

00:21:06 Clifford Chanin: That they're built along the assumption that everybody sort of wants to have this common end. And there is no real conflict outcome that's intended by these communications, which is obviously not the case for the various extremists who use it. But tell us more about that, because I think it's a really interesting point.

00:21:25 Joan Donovan: Yeah, so when you're programming an algorithm, you actually rely on what are called natural-language-processing algorithms. And ultimately, algorithms, if they are to think, right? And I'm not gonna try to humanize them, but algorithms make the assumption that it is connecting good-faith actors.

So, you post something on Twitter, and an algorithm then takes that information and rebroadcasts it to the rest of the network. And the

algorithm believes, in a certain sense, that what you're trying to spread is intentionally well-thought-out and is meant to be received by this broad audience.

00:22:10 Other algorithms, like on Facebook, that sort that content for the news feed, say, "Okay if you've written, 'Congratulations,' 'Happy birthday,' 'I'm having a baby,'" it will take those set of keywords and say, "This is a good-faith conversation and I want to make sure everybody knows that so-and-so just had a kid." Right?

00:22:33 And algorithms make sense in that way. But when you're trying to deceive, the use of keywords, the use of different language patterns, hiding your identity, all of those things are techniques meant to trick an algorithm. For instance, tonight, go home and write into your Facebook news feed a set of keywords around "congratulations," "it's a boy," you will get more eyes on that post than any other posts that you've done, because those are the kinds of happy words that the algorithm responds to.

00:23:10 And, of course, if you're making posts, and I'm sure some of us have gone through tough times, you make a post that says, you know, "I'm pretty depressed today," those go nowhere, and you're, like, "Where's my family? Where are people in this?" Right? And that's because depressive posts don't go far, and the algorithm determines that. And so, when we're talking about manipulation tactics, there's a key set of behaviors that my team really tries to look for in terms of, where is the lie? Where are they hiding different pieces of content? And how are they getting through and across and around these things?

00:23:47 And of course, you see it quite a bit, I'm sure, in misnamed Facebook groups that are trying to not be known for what it is that they're actually talking about. And so we study those techniques, because that algorithmic manipulation is something that these groups have grown very good at over the last, I would say, four years. They've really honed what is called "search engine optimization techniques" in order to remain present and effective on platforms and largely undetected.

00:24:18 Clifford Chanin: So, David, your efforts are, of course, to head them off at the pass, if you will, to understand the ways in which they are trying to get around the controls and the restraints that you have into the system. I wonder, um, in terms of what Joan has said, does that resonate with the experience you're-- you face in trying to uncover the things that are not obvious?

00:24:39 David Tessler: So, it's certainly difficult, right? This is a-- this is a mammoth undertaking, but, you know, we have the commitment to do the best we can. And I think in many respects, we're doing very well. In the last two years, we've taken down 26 million pieces of Al-Qaeda and ISIS content because of our, our technology, both our content-matching and our machine-learning. 99% of that, of those 26 million pieces of content, were taken down before any user reported it.

00:25:14 So, it's not to say that—I agree with you. It's an enormous problem. And, and... terrorists are going to continue to try to get around our screening. They're going to try to figure out ways to get around what we're trying to do to protect the space. But we're going to continue to try to get better. We're going to continue to try to make the technology better.

00:25:41 Um, we're working right now with the Met Police in London on an initiative. One of the tragic things about Christchurch... The whole thing was, of course, horrific. One of the tragic things about it was that our technology was not accustomed to screening and finding that kind of awful content from a first-person perspective. Um...

Clifford Chanin: Meaning the shooter has a camera.

David Tessler: Meaning the shooter, correct.

Clifford Chanin: And is broadcasting whatever it is he or she is doing at that moment.

00:26:16 David Tessler: Right, right. So immediately after that, we, we worked to try to improve that. And we recently announced, with the Met Police in London, a program where they are going to, in training, uh, with body cams, take video of themselves shooting so that we can load those images, those videos, into our screening system to get better at being able to screen from the first-person perspective.

00:26:41 That's just one example, though, of how we have to constantly look at, what's the, what is the technology we have now? What's the capability we have now? And where do we have to get better? So, it is, it is... It is a constant, it is a constant effort to do better, both on the technology side and on the human side, right?

00:27:01 We have to understand better what they're trying to do. Different terrorist groups sometimes behave differently, right? ISIS, from its inception, has been very interested in broadcasting out to the widest group of people to try to... to try to build the widest following. Al-Qaeda has always been a little bit different than that in trying to, um, I wouldn't say bespoke, but trying to curate their message and their followers a little bit more carefully.

00:27:34 So, different terrorist groups act differently. We have to understand that so that we can understand how they're going to try to get around our screening systems. So, this is-- this is part of, part of our commitment, and part of our work.

Clifford Chanin: Let me ask, obviously, the major platforms have both more users and more resources to try, anyway, to regulate what goes on in those sites. But it does seem that the use of social media, in one form or another, is now part of the planning of these kind of attacks?

00:28:05 You think of the synagogue shooting in Pittsburgh. The most recent arrest yesterday, the plan to blow up a synagogue in Colorado. Christchurch being a perfect example. The attacks on the mosques, where this was

supposed to be broadcast through a webcam live, as it was happening. Is there a risk, or is depriving them of the major platforms, like Facebook and whatever else might be used, that doesn't necessarily cut out what their options might be.

00:28:33 It may limit their audience, but there are other sites they could go to, other channels and platforms, which are perhaps not avowedly going to support these kinds of actions, but are much less concerned about limiting the impact. Is that fair to say, Joan?

00:28:51 Joan Donovan: I think it's fair to say that there are a range of services that allow for live broadcast, and baked into the design of live broadcast is this, I think, woefully misguided assumption that everyone should have access to live broadcast instantaneously, without really had to... really having to do much. What we see is a pattern emerging where violent acts are used in order to shape the way in which news and other media organizations, um, talk about whatever terrorists, whatever the terrorist is carrying out.

00:29:37 And so, for instance, if you look at the chain of events around Christchurch: it gets posted, the live stream, everything is set up, it's been tested, and it's... He posts the link along with a manifesto, and not just a manifesto in one place, but many places for this manifesto, so that it's very difficult to take down, right?

00:30:01 So you can imagine, you have to get in touch with Scribd and DocumentCloud and all these other places in which this manifesto spreads. But there's really only one place you can get the live stream, and this person is able to stream this 20-minute video. And then there's, what, from what I've been told, over a million attempts to re-upload that video to Facebook within the next 24 hours.

00:30:29 So there's an entire group of people that are behind this shooter that is kind of like a, an adversarial movement online that is intent on keeping that contact-- that content in play as journalists try to discover what's happening, as everyday people try to discover what's happening, you

know, so people start using different keywords. They use the name of the shooter.

00:30:55 They use-- they try to go find that video that was uploaded to Facebook Live. They head over to different platforms. But ultimately, my point is to say that the infrastructure that they choose, they choose it because it's stable. They don't tend to go to these platforms where the technology is, is wonky or doesn't always work or, you know, flickers in and out of signal.

00:31:19 Especially a platform like Facebook, which is global in its footprint, makes... is really enticing as a tool in order to do this kind of, this kind of live broadcast. And it's, you know, to the credit of Facebook, it is remarkably stable infrastructure. It pretty much works every time, which is a really, like, great thing. But in this instance, that stability, the background around how many times this person was able to inspire other people to spread this manifesto, as well as to re-upload this, this video, that makes me rethink this feature.

00:32:03 It makes me rethink broadcast. It makes me think, well, what did we miss when we first designed these systems, to say, well, how would you earn that kind of attention? How would you earn the privilege of doing a broadcast? Um, and what do we need to have-- either FCC regulations or something that would put guardrails in that would help us prevent the next tragedy?

00:32:28 David Tessler: So, just, just to get back to your original question. Facebook is, was, is very cognizant of our need to work together and collaborate to try to counter the terrorist threat. And so, in 2017, we launched, together with Microsoft and with YouTube and with Twitter, this Global Internet Forum to Counter Terrorism.

00:32:54 We call it the GIFCT, and that... Those, that core of four companies has grown substantially, include a lot of other tech companies, both big and small. Our mission we boiled down into three kind of pillars. One is to prevent, one is to respond, and one is to learn. On the prevention side,

we've worked together, hand-in-glove, in developing a bank of what we call hashes. And a hash is a digital fingerprint of a video or, or an image. So, even if you try to manipulate it a little bit along the edges, that fingerprint is an easy way for us to identify that as a piece of terrorist content.

00:33:40 So, we developed this technology at Facebook. We shared it. It is now... it's open. We've shared the technology. We have collectively in the GIFCT, we have now collected 200,000 distinct hashes. So 200,000 distinct fingerprints of, of terrorist content that can then be used by any member of this hash consortium in the GIFCT to better prevent this kind of terrorist content from, from being uploaded onto their platforms. So that's on the prevention side.

00:34:17 On the respond side-- and this is also following the call to action, the Christchurch call to action-- we developed what we call a content incident protocol, which means that, in the event of a real-world crisis, an offline crisis, like, for example, what recently happened in Halle, Germany, we have a protocol among members of the GIFCT to immediately share whatever information we have as to what might be going on online.

00:34:50 So we have an offline crisis, or a terrorist attack, in this case, what's happening online? And how do we as the, as the members of the GIFCT, make sure that whatever we see, we make sure everyone else in the GIFCT sees, so that we can prevent things from going viral? Amazon is a member of the GIFCT, and if you remember, the terrorist in Halle tried to-- tried-- to use Twitch, which is, which is Amazon's live-streaming service, to live-stream the terrorist attack.

00:35:26 So, Amazon is a member. We were able to very quickly share information in that respect. So, these types of, of collaborative efforts are key. On the learning side, and this is, this gets to, I think, the core of your question, we want to make sure that the resources and the experience that we have as a larger tech company, we're able to share with smaller tech companies that may not have the same kind of resources.

00:35:52 At the same time, sometimes smaller companies have a very, very important perspective or experience that we benefit from. So it's a two-way street. And we work with an NGO called Tech Against Terrorism to put on workshops. This year, I think we've put on about a dozen workshops and reached, you know, over, I think, 100 or 120 different tech companies around the world to try to make sure that the knowledge that we have, and experience we have, and technology that we've been able to develop as a larger company, we're sharing, so that these smaller companies can use it and protect themselves, too.

00:36:31 Because we know that as we get better... And, and I take your point, there is a benefit that Al-Qaeda and ISIS will always want to try to get back on Facebook. But we also know that as we get better, we are... There's a risk that terrorists go to smaller platforms where they think it's a more permissive environment, and we want to help those smaller tech companies protect themselves.

00:36:54 Clifford Chanin: Let me switch a little bit from... I mean, I think it's obvious that no one would want... Or that the line would be clear in terms of terrorist violence and the desire to keep that from being disseminated. But let's cross over to hate speech, where the lines are not always as clear, where there are contextual issues that may make some distinctions difficult for, whether it's machine-learning or individuals reading it, to sort out.

00:37:21 Um, so putting aside, okay, we don't, we don't... We have ways, presumably more effectively, of keeping live shootings off of these platforms. But hate speech and misrepresentation in various ways is in some ways a more serious issue because the effects are longer and it's harder to sort out. So how can you make the distinction, each of you, in terms of what the challenges of dealing with the hate speech issue might be as compared to-- what seems to me, anyway-- to be a much more obvious example of the kind of thing you could identify and keep off the platforms? Assuming you are aware of what's going on? Joan?

00:38:02 Joan Donovan: Um, yeah, the hate speech issue has been something that... So, globally, there are different rules. So in Germany, you get a different experience of the internet than you get here in the U.S., because the U.S. sort of toggles towards 1A, the first amendment, and, and freedom of expression is paramount. Whereas in Germany, um, you can't use certain terminology. There's a kind of dictionary of slurs and whatnot that platform companies do abide by.

00:38:34 And there's more stringent restrictions on particular forms of, for instance, Holocaust denial. But platforms really have to build a system that is global because it's very easy to circumvent and pretend you're in one country or in another. And for the U.S. question around hate speech, there's a really great organization-- well, it's an organization of organizations that have been offering this change-the-terms look at policy, where they're... Civil society organizations are trying to get different platforms to take up the same hate speech policies, so that if YouTube takes something down based on hate speech, and that group has also served that link on their Twitter account, it's just clear that these policies are going to be abided by across platforms.

00:39:28 From our perspective, as researchers, we saw not... maybe not an increase in the amount of hate speech in 2016 and 2017. But we definitely saw an increase in the focus on hateful groups, and in particular, the rise of an American white supremacist movement that was going under the moniker of the Alt Right. And I just, I felt like Chicken Little, running around, being, like, "It's happening. This is dangerous. There's something going on."

00:40:02 And, you know, the Unite the Right rally happens. And, you know, for all of the research, and all of the ways in which we tried to get hate speech on the table at platform companies, it just seemed to fall flat. And then the day after, literally, the day after, platform companies started taking down massive amounts of hate speech on their platforms. They were removing accounts, they were de-platforming people. Even Uber was, like...(laughing): "Nazis can't get a ride." OkCupid was, like, "Nazis can't get a date," right?

(laughter)

00:40:36 Joan Donovan: It was just on and on and on. And so you saw this moment in tech where, up until that day, you know, April... You know, I mean, August, you know, 11-- "We don't see it." August 12-- "It's not our fault." August 13-- "We're going to do something about it," right? And that progression of events I've written quite a bit about-- I have a recent article out about content moderation-- introduced a whole new field of play.

00:41:05 Which is to say that platform companies are now interested in thinking about hate speech, thinking about its impacts, but also understand the question of youth radicalization differently than they had done before. Platform companies were very proud of the work they had done with different social movements. And, you know, Twitter had the three, you know, black fists for Black Lives Matter up, and would argue, you know, that they were instrumental in these social movements. And then when you say, "But the, the, white supremacists," they just go, "I don't see it. I don't know what you're talking about. It's not even here," right?

00:41:44 Meanwhile, they've got... it's just, you know, people are running rampant. And even Facebook took down... The day before the Unite the Right rally, they took down the event page for the Unite the Right. And so in this moment, now we're starting to see tech companies try to get in line and try to work together in a flock to develop content moderation around hate speech. And they're at a crossroads, because they actually don't know what it is that's unique about the American Neo-Nazi movement and the Alt Right in order to create those rules, right?

00:42:22 And so I would love to see a way in which platform companies could come together and work with Change the Terms to get more uniform policy around hate speech so that we can avoid, um, a growing, you know, radicalization of essentially, you know, white male American youth. And I know that you guys do a lot of great work with Life After Hate. So I'm, I'm interested to see how that progresses, though, because for so long, I was saying there was a problem, and a lot of people just,

you know, didn't see their culpability. Especially, uh, you know, people at platform companies who were building these tools.

00:43:05 Clifford Chanin: Yeah, but is it more difficult to make these distinctions and to, um... find these exclusive practices and, and eliminate this kind of problem of hate speech, if it's not really about, you know, live videos of violence?

David Tessler: So I think it's important to go back to the principle, and in the introduction, Mark Zuckerberg's talk at Georgetown was, was referenced. If we're, if we're trying to give people voice, allow people to connect and build community, they have to feel safe. And so this goes-- that's the principle. This goes to people, users on platform, feeling safe.

00:43:43 And I mentioned the community standards before. We have published kind of our definition of a hate group and our definition of hate speech. If, if it meets those definitions, if it is speech that attacks a person for... (stammering): As, as... for a protected characteristic like race or gender, sexual orientation, we will take it down, because that's the kind of content that makes people feel unsafe.

00:44:13 Our definition of, of terrorism is... has nothing to do with ideology. It's not ideology-specific, it's behavior-specific. If you meet that definition of terrorism, no matter what your ideology is, we will, we will prohibit you from the platform, and we will prohibit any praise, support, or representation of you on the platform.

00:44:35 So, in that respect, um, it's, you know, our mission is clear. Is it, is hate speech sometimes very contextual or culturally contextual? Yes. Do we have challenges with linguistics? Yes, we do. You know, I think...

Clifford Chanin: Can you give me a couple of examples of what you mean by challenges of linguistics?

00:44:58 David Tessler: Well, I mean, I think... A good but very tragic example is what happened in Burma. There was a lot of hate speech going on in Burma, including on our platform, at the time. And this is something we've talked about as a company quite a bit, and we have worked to rectify. We didn't have enough Burmese speakers. So, so it... hate speech in that regard is complex, but our policies are very clear about it.

00:45:29 What we intend to do is very clear about... is very clear. You know, Joan mentioned Life After Hate, is a group that we work with, a nonprofit we work with, in our efforts to try to redirect people. So if you, in the U.S., you search some hateful terms, Facebook will offer you—a little window will pop up-- and will offer you and say, "Hey, you're searching for this really bad term. If you want the resource... a resource to, to talk about what you're feeling, and, and try to leave hate, here's a resource."

00:46:10 And, and so we redirect to this group Life After Hate, which is an amazing group of people who themselves have come from, from extremist pasts, and so know firsthand what it's like both to be in that milieu, but then also to find a path to leave it. And we've expanded this program now also in Australia and in Indonesia.

00:46:35 So, you know, we recognize that we don't have the credible voice. We don't have the experience of being... having a past of being in an extremist group. But what we can do is, we can partner with the people who do have credible voices in society, civil society organizations who know the ground truths in different parts of the world, and who, and who have the credibility.

00:47:00 We can use our platform in the size and scale of it and partner with the people who have the voice, and together in that partnership, I think we can, we can really make a difference. And we're seeing, we're seeing that now, and expanding to two other, two other countries, and, and perhaps we'll go, we'll go beyond that.

Clifford Chanin: Joan, we were talking before about the ways in which someone can get through the filters by changing the language and the

references. You see various ways of punctuating, or italicizing in some sense, as a signal to your very narrow audience that, "You know what I mean by this." Can you talk a little bit more about those kinds of attempts?

00:47:41 Joan Donovan: Yeah, so we see in a lot of these hate group conversations, or in chats, where, if they know that the algorithm is moderating for a very specific dictionary-style term, they'll make a change. And then part of that reactive adaptability is, it's... it's viral in and of itself.

So we've seen when, for instance, YouTube Chat was removing any use of the word "Jew," they changed to "juice," right? And so they're able to go around. And of course, you can't ban the word "juice," and so, on and on. And there's ways in which pop culture figures into this.

00:48:29 So, for instance, if you're gonna make a slang... You know, use, use popular culture... pop culture references in order to denigrate different populations, we'll see that pop up. So during the time when "Black Panther" was a really popular movie out, we saw, instead of using racial derogatory terms, they would just say, "Wakandan," right?

00:48:53 And so the way in which these communities derive a bunch of their legitimacy has to do with their ability to kind of stay on these platforms and remain productive. But then also, you know, signal back to their audiences that, "While we're on this platform, we have to play with, by these rules. But you can pay for the podcast over here where there are no rules," for instance.

00:49:23 And so part of the ways in which they hide things has to do with playing against the terms of service on very specific platforms, and then having, you know, more disgusting conversations in other places. And that, I think, also, is difficult for the policies that you have, which is to say that there's no real policy that says, "Well, if you do this hateful thing over here, and we know that it's you, you can't do your reputation-brand

management over on our platform to ensure that everybody knows that your podcast is live at 9:00 p.m. tonight."

00:50:01 And so there's ways in which they play those platforms off of one another, and then manipulate as they learn the new terrain. And so I think, as we're thinking about media manipulation and disinformation, and how hate groups use it—as well as terrorist groups-- we also have to start thinking about, what is the terrain here, and where is it that this stuff is, is popping up?

00:50:26 And are there ways in which, once we've validated certain identities, to say, "Yeah, we are not gonna expose our customers to this," right? The platforms, you know, ultimately, you might use the word community, I might use the word customer. But that is to say that if you are providing a service, you should also make sure that within that service, it's not just hard to stumble upon these things, but that I shouldn't be able to be targeted or harassed, you know, in such a way that, that the platform, you know, pretends like they don't see it. And so I would love to see some kind of cross-platform plan for dealing with something that is a much bigger problem than any one platform faces alone.

00:51:15 David Tessler: Well, we actually do have a designation process at Facebook, both for hate groups, for terrorist groups, and for criminal organizations. And part of our... part of our criteria is looking off-platform at their behavior. If, if a hate group, for example, is displaying violative behavior, behavior that would get them designated off-platform, that's enough for us.

00:51:40 We can designate a hate group based on a holistic look at what they're doing. It's not confined to just what they're doing on Facebook. So... and the same is also true for terrorist groups as for criminal organizations. If we see them meeting the criteria for our... for our designation, no matter where it is, if the criteria are severe, the behavior is severe enough, then the threshold is met, we'll designate them, which will mean, even if on Facebook their behavior is benign, they will be prohibited from being on Facebook.

00:52:13 And it goes beyond just themselves being prohibited, but you won't be able to praise them, you won't be able to support them, or represent them on Facebook. So it's, you know, the designation... The designation tool is, is pretty potent. And we have a process in place where we, where we do that research and try to find these, these groups that we just want to expunge completely from the platform.

Joan Donovan: Yeah, and, I'm... Sorry.

Clifford Chanin: No, no, that's what you're here for.

00:52:43 Joan Donovan: I'm interested though, in that, in that, that moment where you designate a group. So, for instance, I was reading an article from BuzzFeed. Craig Silverman and Jane Lytvynenko are very good at looking at Facebook as a platform. And so there was a designation of a group, a hate group in Canada-- I think it was the Soldiers of Odin-- where, a few days later, after being designated as a hate group, they went out as reporters to see if that stuff was still there, and it was still easily searchable and, and retrievable. They didn't really have to go very far. And so, from my perspective, I'm wondering, then, policies-- yes. Like, enforcement-- how, though?

00:53:29 David Tessler: So... so... I don't know that particular case, so I can't speak to it. But I think, in general, part of the designation process is working with the implementation teams, the enforcement teams, to make sure that they have the, the information they need.

For example, the prominent members, the founders, the leaders of the, of the organization-- they have all the, as much information as, as possible, so that when they go out to enforce the designation, they have what they need to take down... to take down the presence. And then look for the praise and the support and that sort of thing.

00:54:11 But with the, with the size of the platform, it's not always easy. We're, we're not perfect, but the, the system, the process, is one that is, that is robust. And so it's, it's always a question of, how can we do better? How can we implement better? And that may be... that may be a case of that.

Clifford Chanin: Good, I'm gonna see if we have a question or two from audience. Please wait until you get a microphone. Gonna... yes, this lady right here in the front. And Harmony will have a microphone for you in a second.

00:54:49 Audience Member: It's just a comment, and it may be ridiculous, but it seems to me that Facebook's desire, business model, is one of expansion, and it puts the public at risk before you have figured out how to solve inevitable problems. And your using Burma-- which, I assume, you meant Myanmar-- as an, as an example, was perfect to me, that you didn't have enough people to understand the speech, the language. So is there a point at which you should draw in your expansion, your business model?

00:55:35 David Tessler: So I mean, I would just go back to our mission and the core mission, which is to give people voice. I think there is a lot of great voice being given, a lot of great connections and communities being built. And where, you know, we are committed to trying to stop the bad, to trying to stop the... Those that are either proclaiming a mission of violence, or engage in violence, from being on the platform.

00:56:08 Joan Donovan: I just wanted to say, I think you hit it right on the nose, which is to say that, you know, from, from what I've been told, you know, that the company has expanded into places, they don't really understand the culture. They don't know how people access the internet, right? And so if the community isn't able to access and use the internet, and you have a situation in which the military is using Facebook as a screen in order to hide their true intentions, it shouldn't be up to reporters and civil society groups to root out that kind of disinformation campaign, right?

00:56:49 But if Facebook had been in the country, in the culture, and was dedicated to the service of the community, and not just deploying a technology, you would have seen it a lot faster. And I know that, because everywhere I went, people were saying, "This is happening. Why isn't anybody doing anything about it?"

And so expansion at all costs is potentially incredibly harmful for societies where authoritarian regimes and military coups are on the table. Because if I pretend to be 40 reporters in a city, and I am completely astroturfing the playing field, and I'm really the government, there's nobody that's gonna hold that government to account.

00:57:39 There's no way in which we're going to get justice for the people who were murdered, right? And so that kind of process isn't just one of, like, Facebook came in and, and deployed a technology, but that Facebook caused a specific kind of amplification of that harm that that government would not have had recourse to.

00:58:04 They wouldn't have been able to do that if it not... had not been for the Facebook tool. And so those are the kinds of difficult questions that research in disinformation comes up... Because it's not like you can say, "Oh, this government is also a terrorist organization," right? Because there's, you know, serious implications for that. But we do see disinformation, and governments really using the ability to hide on platforms as a way to suppress their citizens.

Clifford Chanin: Gentleman here in the front, uh, Harmony?

00:58:43 Audience Member (loudly): So terrorists-- oh, sorry. So terrorists and other organizations upload content-- they try to-- on Facebook and other... and other platforms, and you bring it down as, take it down as fast as you can. But there's certainly a whack-a-mole aspect to that, because, I'm just wondering, if you were to develop a silver bullet and actually take all of it off immediately-- with A.I. or whatever-- wouldn't...

00:59:04 I'm just wondering if some anti-terrorist government organizations would not welcome that, because at least they're following what's going on. And if you force them to go into the dark web, or to, really, encryption, that, aren't some of the governments using the information that's... or following the terrorist organizations through these platforms, and it doesn't help in their efforts at all?

00:59:24 David Tessler: So, I wish there were a silver bullet. There isn't. We are doing everything we can to take all of the content down. You know, we get a lot of it down. There's, like you said, it is... It is a bit of a whack-a-mole game. But as we get better, we're getting better at the game. And so we're not, we're not concerned with, with the hypothetical that you presented. We're, we're concerned with just getting down the content.

00:59:56 Clifford Chanin: I think it is the case, and you see the reporting on the various incidents, including the incident in Colorado yesterday, where contact is made by the FBI undercover through social media, with people who are out there. So it is a means-- whether it's Facebook or not, I don't know in each case-- but there is a means by which, you know, the law enforcement authorities are active on these platforms.

01:00:15 David Tessler: I mean, law enforcement has the ability to make formal requests to us if they see, if they see information and they want to get more information. There's, there's a legal process for that. And we, we comply with those, with that legal process. And if we see something on platform that we think may lead to imminent harm, real-world harm, then we will contact law enforcement if we think that it's, that it's, that it's an imminent threat.

01:00:44 Clifford Chanin: So just reading from "The New York Times": "Mr. Holzer"-- who is the person who planned to blow up the synagogue in Pueblo, Colorado-- "Mr. Holzer was arrested Friday after he was first contacted by an undercover FBI agent in September. It's not clear when investigators began tracking him or how they were first alerted to the posts. Speaking over Facebook Messenger shortly after contact was made, he told the agent he was formerly a member of the KKK and now identified as a skinhead."

01:01:09 So this is clearly, I mean, we know this. This is a law enforcement technique that, you know, they are very adamant about trying to protect for exactly the reasons that you described. Who else had a hand up? All the way in the back, the gentleman, there. Not sure whether Harmony or Ruth will get there first. It's always a race. And Ruth wins.

(laughter)

Joan Donovan: Thanks, Ruth.

01:01:30 Audience Member: Let's pretend you guys just switched jobs.

(laughter)

Joan Donovan: Not for all the money in the world.

(laughter)

Joan Donovan: I just met David, seems like a great guy, love to have him over for dinner, but I'll stick at Harvard Kennedy.

Audience Member: Yeah, okay. And, and so for the professor, what three actions, what three things would you implement if you had his job?

(laughter)

Audience Member: And then, conversely, what three research programs would you start that would help you a few years from now?

Clifford Chanin: This is going to get very confusing. Let's see what happens.

01:02:14 Joan Donovan: Yeah, yeah. I don't know, do we want to popcorn back and forth so that we have... You know, I don't necessarily think about product redesign too much, but I do have an excellent researcher on my team, whose name is Kathy Pham, that does excellent research on product management, and its implications in society. So the first thing I would probably do is begin by implementing a very robust ethics course about technology and ethics in Facebook, and it would be core and it would be mandatory.

01:02:50 Then I would think about a couple of different features on Facebook that are particularly difficult for my kind of research. Which is to say that I would look very closely at Facebook groups, and I would try to ascertain, at what level are Facebook groups harmful? And this has to do with the way in which Facebook groups begin, maybe in one way, shape, or form as an authentic something or other. But then five years down the line, different admins have come in, and they've been repurposed and rebranded.

01:03:26 And usually the only people left in them are really trying to run a significant kind of influence operation. And I've seen a lot of movement pages and groups turn into anti-vaxxer groups, or they turn into ways in which advertising scams happen. And so I'd have a process for looking at and vetting groups, and making sure that what it is they say they are is what they're actually doing.

01:03:53 And this also has to go back to also looking at the pages related to what counts as news on Facebook. Because I can literally make a page on Facebook that says, "Breaking News Now," and it's really just about, like, cute things my cat's doing, right?

(laughter)

Joan Donovan: But that internal designation as news, and, and the self-description, I think, is something that we need to change. I think that's three. I'd do the pages for news, I'd do... I'd take a hard look at groups and say, "Do they represent what they are... what they say they are?" And then I'd bring Kathy in to do a course on, how do you design products ethically?

01:04:29 Clifford Chanin: David? Turning the tables?

David Tessler: Well, I'm very far from being a professor at the Kennedy School. But I guess-- and this is not Facebook-specific, but just, I think, more generally in the tech world-- I would be interested in researching... As, as generally, there is more end-to-end encrypted messaging on the internet, researching what other indicators there are around the message that could help... could help us distinguish benign messaging from, from extremist or violent messaging to help us stop... to help us better stop the violence in a, in a more encrypted world.

01:05:19 So what other indicators are there out there that we would have access to that would be important for us to understand better? That would be one research project. Another would be, I think, in the hate group space, hate groups, many of them are becoming more advanced and more sophisticated in, in how they use social media, and understanding the evolution of hate groups and social media, I think, the direction they're going, would be helpful in order to try to stay one step ahead.

01:06:04 Joan Donovan: I got that.

David Tessler: Those are two. All right, keep on doing that.

Joan Donovan: I got that second one.

David Tessler: Okay, keep on doing it.

Joan Donovan: Yeah, I got that.

Clifford Chanin: Good, we found one place of agreement.

(laughter)

Clifford Chanin: Let's do one more. Let's do one more. Gentleman right there.

01:06:21 Audience Member: Thank you. Actually, with regard to encryption, my question was about that. What exactly is Facebook doing? WhatsApp is Facebook now, and from a research point of view, how exactly are you tackling it? Because it's obviously not as live, not in person, but nonetheless, quite a bit of potent material that is going around WhatsApp, and no one can see it. What's your policy there?

Clifford Chanin: What's up with that?

David Tessler: So...

(laughter)

01:06:47 David Tessler: So privacy is very important. And there's this... there... And this is not, this is not specific to Facebook, but obviously, there's a lot of encrypted messaging already in existence, and a lot of people are using it. As, as you may know, Facebook is moving towards more end-to-end encryption to protect people's privacy-- better protect people's privacy. There is, I think, a natural, a natural push and pull between the privacy advocates and the advocates of counterterrorism.

01:07:26 But there is, there is a lot we can do, I think, and we are working to develop this in the space of indicators. To try to understand if you... Who you are and who are you sending it to, perhaps, is a good indicator. Um... uh... How many messages are you sending?

01:07:51 There are lots of different ways that we are exploring now to try to be as good as we can in continuing to, you know, to work to stop the content, the violent content, from, from being on platform. But it is, it is going to be a challenge. You know, that's... Right? It's gonna be a challenge.

Clifford Chanin: Joan, your thoughts on encryption?

01:08:17 Joan Donovan: Yeah, so in researching in this space, of course, encryption is, is something that we're all really worried about and, and want to understand what the trade-offs are. Because if they do move into encrypted chats, and it's limited to 256 people, the kind of scaling they can do at a very low cost isn't, isn't very good.

01:08:37 They can't really reach out to new audiences. They can't really... unless they get a couple of different phones and phone numbers, which, we've seen different government operatives use WhatsApp in a way that, if you have a lot of resources, you can do these, these massive group texts and things, but you have to have a certain number of telephones and things to be able to push information into the... into encrypted spaces.

01:09:05 The other, of course, the trade-off is privacy, which, you know, one of the things that a lot of people are really worried about is the degree to which Facebook is becoming a sort of global policing apparatus. And how much transparency we're gonna have between when a government asks for things and when Facebook hands things over voluntarily, or how much, you know, information there's gonna be flowing that we don't know about.

01:09:33 But when it comes to the disinformation question, there's excellent research coming out of M.I.T. I believe the researcher's name is Kieran Gramali, who's been looking at... People on Reddit and different places will post open WhatsApp group numbers and say, "Oh, if you're into poetry, join this." (softly): It's usually not poetry, it's pornography.

(laughter)

01:09:55 Joan Donovan: But, you know, "If you're into this, join this group." And within those groups, a certain amount of disinformation also travels. Because if you are someone that's trying to push a mass message, it doesn't always matter to you that it's super-targeted. And so he's devised a way, in his research group, to scrape those public WhatsApp channels, and then get a sense of what is circulating, what amount of it is "poetry," and what amount of it, of course, is disinformation.

01:10:24 And so there are ways of studying this stuff. There are ways of doing it, as well as just good old investigative reporting, where they... reporters will ask their audiences to, "If you see disinformation, take a screenshot of it, and send it to me." And that way, as a reporter, they can track these things. So that was used in Brazil. There was a large project where reporters asked their audiences to give them screenshots, and that seemed to work fairly well.

01:10:57 So I think in the 2020 election moment, we're going to see a lot more tip lines for disinformation that are gonna include this, this ability to take a screenshot and send it over to researchers or, or reporters in order to get a better assessment of what kind of disinformation is gonna flow on the backs of fake campaigns.

Clifford Chanin: Well, I mean, you can see there's a lot more here, but, and we have even gotten to the 2020 election yet.

(laughter)

01:11:24 Clifford Chanin: And, and we will come back to that next year as we go on. Before we close, I know many of you are members, but I'm going to make a plea again. For those of you who are not members, think about joining tonight outside. It will help support our programs here, and we invite you all back whenever you can and as often as you can.

And with that, I think we've had a fascinating discussion, and I ask you to join me in thanking David Tessler and Joan Donovan.

01:11:48 (applause)