SPEECH LIMITS IN PUBLIC LIFE:
AT THE INTERSECTION OF FREE SPEECH AND HATE

SPONSORED BY	The Institute for Humane Studies at George Mason University and the University of Delaware's Office of the President, Office of the Provost, Office of the Vice Provost for Diversity, the Class of '55 Ethics Endowment Fund, College of Arts & Sciences Dean's Office, and the Department of Communications.
CO-SPONSORED BY	Heterodox Academy and the University of Delaware's Center for Political Communication
SESSION 4:	Are more laws necessary for responding to hate speech?
	Introduction by Jenny Lambe, Associate Professor of Communication and event organizer.
PANELIST:	Sara Wachter-Boettcher, author – Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech

Transcript of Event

Date: March 15, 2019 Place: Embassy Suites, Newark, Delaware

MODERATOR: Co-host of the podcast called *Strong Feelings About Living Your Best Feminist Life at Work*, and her most recent book *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech* was named one of the best tech books of the year by Wired, and one of the top business books of the year by Fast Company. And that's also available for sale in that room although I think they are leaving shortly after lunchtime so grab it now if you want. Sara is also the co-author with Eric Myer of *Design for Real Life*, a book about creating products and interfaces that are more inclusive and compassionate, and the author of Content Everywhere, a book about creating flexible mobile-ready content. Sara speaks about design, tech and digital publishing at conferences around the world and consults with startups, Fortune 100 companies and academic institutions. Her work has been featured in the Washington Post, Slate, The Guardian, Salon, Courts, McSweeney's and more. Please welcome Sara Wachter-Boettcher.

AUDIENCE: [Applause.]

MS. WACHTER-BOETTCHER: Hello. Thank you all for being here today. I'm really excited to, to give this talk, um, at a conference that has been all about free speech, hate speech, and a lot of very confusing questions in between. I'm going to talk about some things that I think started to come up in the last panel where I'm going to really be digging into what's happening within tech companies themselves and what are some of the underlying causes of some of the problems that we're seeing online from these companies. Now, a lot of what I'm going to talk about is going to be directly related to hate speech and free speech issues, um, but some of what I'm going to talk about is going to

get into some other areas. So, I think you're going to find there's a lot of fascinating threads when we really start to look at what's happening within tech. As somebody who's worked with tech and design for a long time, I started noticing this a few years ago and once I started noticing I couldn't stop paying attention. So, I want to start out today by talking a little bit about this example. Um, this is a screen shot of a video that was on YouTube Kids for quite some time. So, this is a knockoff Peppa Pig cartoon and at the end of 2017 a writer named James Bridle uncovered this creepy under belly of content that was being targeted at kids on YouTube. So, in the regular Peppa Pig cartoon Peppa Pig goes to the dentist. Now, this might not be super surprising if you've ever seen children's programming but in the regular cartoon she's scared and then she goes to the dentist and it's fine. In the knockoff Peppa Pig cartoon Peppa Pig goes to the dentist and she's scared and then the actual visit morphs into this graphic and violent torture scene. So, this was being targeted at children on the YouTube Kids app. And it wasn't just this one knockoff Peppa Pig video, or it wasn't knockoff Peppa Pig videos in general that were a problem, it was actually thousands and thousands and thousands of videos that James Bridle uncovered as part of this issue. Now, they were being produced and added to YouTube and tagged with what he called a keyword salad. So basically, they would stuff them full of all of these keywords to tag them about what the video was about. And they would use keywords that were things that they knew resonated with kids, so things like characters or types of games or types of videos kids like, like things that had to do with like matching. So, they stuff them with keywords and then those would be auto played to your kids based on similarity to content that

they've seen in the past. So, for example, let's say your kid starts out watching a standard Peppa Pig cartoon from the official channel. Let's say you've given them an iPad on a long car trip or at a restaurant. You just need 30 minutes. I don't have kids, but I hear that's a very real thing.

AUDIENCE: [Laughter.]

MS. WACHTER-BOETTCHER: You just need 30 minutes. Well, they start out watching that regular Peppa Pig cartoon. And then very quickly based off of the similarity in content they are sent deeper and deeper into a weird and creepy and dark universe. A universe that I know none of you would want a child to be put into. Right? And this was happening over and over with cartoons, it was happening with live action videos, it was happening with all kinds of online content. And it had been happening for a long time. James Bridle called this out and there became sort of this uproar about it and finally YouTube was like, okay, okay, okay we're going to clean this up off the platform. Right? And they went through a process of trying to remove all of this stuff from their platform. But here's the thing, this content may YouTube a lot of money. It's very profitable to target kids with creepy violent content because as long as the kids are engaged, you're making money. And if fact, of course, it's not just what's happening with children, right? We see this all over on a platform like YouTube. Content that garners engagement makes money. And so, you can see it in something like this. So, this is a screen shot from the interview that James Damore, if you remember him, he was the guy who wrote the Google memo. The Google memo is the memo where he argued that scientifically women were just less suited for technical work than men. And he used a bunch of references in his work. And so, if you were

reading it casually and you weren't paying a lot of attention you would read his work and you would think well, hum, maybe he has a point. He sure seems to write like he knows something. He sure has some footnotes. Footnotes are not a signifier of knowing anything except knowing how to create footnotes in Microsoft Word.

AUDIENCE: [Laughter.]

MS. WACHTER-BOETTCHER: Um, but he had footnotes, right? So, okay, he's got research and so you might think well I want to hear what this is all about. Well, when he went out and started doing interviews the first place he went and did interviews was with Jordan Peterson, I'm sure some of you know who Jordan Peterson is. For those who don't, Jordan Peterson is a right wing academic from Canada, from Toronto. And, um, here's somethings that he has become very well known for. He believes that "the idea that women were oppressed in our history is an appalling theory;" Islamophobia is "a word created by fascists and used by cowards to manipulate morons;" white privilege is "a Marxist lie;" and he refuses to call people by the pronouns that they request to be called by in his classes which is how he originally got a lot of notoriety. Now I will leave it to you to decide how you feel about Jordan Peterson. You can probably guess how I feel about Jordan Peterson. But, Peterson interviews James Damore, right? And so more than half a million people watch this right away. And here's the thing, if you watch that interview you don't just watch that interview because within a few clicks of related content you can go from a Peterson lecture to a video titled something like How Savage Are Blacks in America & Why Is Everyone Afraid to Discuss It? And if you watch that video you are watching explicit racist anti-black content. And that is what happens

over and over again when you have a YouTube algorithm that is prizing, showing you things that are similar to what you liked in the past and in fact more like what you watched in the past. So, it trends toward extremism. And you can see it over and over again. And you've heard it talked about maybe a little bit in terms of things like radicalization of let's say, um, young Muslim men for the Islamic state. You probably haven't heard as much about it when it comes to the radicalization of young white men toward right-wing causes. But it seems it's also part of the radicalization, um, for the white terrorists in New Zealand just today. So, Zeynep Tufekci, she's a, a digital sociologist and she talks about this a bit and she says, you know, as we click and click, we're carried along by the exciting sensation that uncovering more secrets, deeper truths. YouTube leads viewers down a rabbit hole of extremism while Google racks up the ad sales. And that is one of the realities with online platforms is that we have created an online system where for so many of these companies that we rely on they have a model that's really engagement at all costs. I have a very simple example for this that isn't really about, um, hate speech so much but I think it crystalizes it in the simplest way and I love it so much for that. This is an example of a notification that a woman named Sally Rooney, she's actually an Irish novelist; her book's out now and it's great. But, um, but before her book could come out, she was just some random person on Twitter like many of us, and on Tumblr like many of us, Facebook like many of us, right? And so, she had the Tumblr app and one day she got this notification on her home screen of her phone. Beep, beep. Neo Nazis is here. And she looked at it and she was like excuse me? Why am I getting this on my home screen? And she actually – I emailed her about this and we talked a little bit

about it and she told me that, um, she spent awhile like digging through everything, all of her settings in Tumblr because she was like, oh my God, have I accidentally subscribed to Neo Nazi channels? Like, what is going on? Why is Tumblr sending me this? She got this because she'd actually read some posts about the rise of neo Fascism in the United States, something probably many of us have read some posts about. And so, Tumblr decided that she was really interested in all of the latest Neo Nazi content. And so, they decided they were going to push Neo Nazi content to her phone. And the thing about this is that of course nobody sat down, right, there was nobody at Tumblr, there's no writer there who sat down with their keyboard and was like beep, beep, Neo Nazis is here. What they're using is a text string. So, they have this whole set of text strings, right, that they can use for notifications and the idea is that it can be like fun and zany. And so, people shared other examples of the same kind of thing where they got this same message but with a different label in it. So, the, it's like beep, beep mental illness is here was another one that was pretty funny. Less funny if you suffer from mental illness, I guess, and you happen to get that post on a very bad day. But, after this happened, the reason I bring this up is that this is kind of funny but then after it happened I sort of paid attention to the conversation and what I found was actually the head writer at Tumblr tweeting about it, responding to people talking about it and he said, you know, we talked about getting rid of it but it performs kind of great. Hum. And I think that that sums up so much of what you see happening in the tech industry. We thought about not doing this objectively bad thing, but you know it makes us money. And, what you see over and over and over again is this very singular focus, this very narrow

focus. And when you focus so much on that one thing, on something like engagement, you can do all kinds of harms off on the side and never pay any attention to it. So, I started paying attention to this many years ago in kind of a related but different arena when, um, this particularly terrible thing happened to my friend Eric Myer. Him and I actually ended up writing that book *Design for Real Life* together and this was sort of one the very early reasons we did. So, what happened was at the end of 2014 he got onto Facebook, like many of us do, and he was planning to check in on peoples Christmases, right? Because it was Christmas Eve and he thought that he would be seeing some good old Christmas content. Instead he saw a new thing and it was called Year in Review. Now, what Year in Review did is it would take all of your content you had posted over the past year and it would decide what was the most popular content and then it would take all of that popular content and package it up for you into an album and then it would put that album in front of you and be like, here look at your awesome year. Right?

AUDIENCE: [Laughter.]

MS. WACHTER-BOETTCHER: And so, the most popular photo that Eric had posted all year, the most popular thing Facebook was to find was this picture of his daughter Rebecca. It's the picture that he posted on her sixth birthday when she died of an aggressive brain cancer. It was the worst year of his life, the worst day of the worst year of his life but it was the content that it was deemed the most popular by the platform because a lot of people commented on it. So, this, in this particular example, right, this was awful for him at a very personal level and this blew up. This went viral. It was all over the news. He got apologies from the produce manager who had worked on it. A friend of

mine was actually the content strategist, so the person who was working on like the little bits of microcopy, like, um, hey, Eric, here's what your year looked like. Um, my friend worked on this and she felt awful. I'm sure she still continues to feel awful about it, right? Nobody wanted this to happen. But here's the thing, you look at an example like that and you think well, we really missed the boat. Woops. Right? Like, most people had good years. We didn't really think about the people who had bad years. Oopsies. But the thing is if you look at a company like Facebook, they keep doing the same thing over and over again. I'm going to explain this one to you because this one is [[indiscernible] and it has an extra little layer to it, and it might not make sense at first. So, in 2017, so we're talking years later, right, they're still packaging up past stuff and putting it back in front of you because of engagement. In this particular example, what had happened is a little different. So, a journalist named Olivia Solon, who writes about tech, happens to get things like rape threats in her email. Pretty common if you are a journalist and a woman and its even more common if you are a journalist and a woman and you write about tech. So, she had taken a screen shot of one of those rape threats that she had received, and she had posted that to her Instagram account because she had wanted to let people know, like, this is the kind of stuff that I get. Now once again, you could imagine that her posting that on Instagram got a lot of engagement. People wanted to talk with her about this photo she had posted. So, what Facebook did – you know, Facebook owns Instagram – so what Facebook did is they said, hum, that's a popular photo; I know, we want more Facebook users to also use Instagram so what we're going to do is we are going to take that photo that she posted and use it in an ad for Instagram on

Facebook and then display that ad to all of her friends. Don't you want to download the Instagram app? You too could get screen shots of rape threats. You see this over and over again, and Facebook is in a lot of ways sort of like the nucleus of a lot of these problems because here's some places where sort of this obsession with engagement, this obsession with getting you to click more and more and more can go really wrong. So, this is a brief story coming out of Germany. Um, so last summer a study came out, um, that was centered on what was going on on Facebook in Germany. And if you look at this town Altena, Germany, um, it's a small town and they had taken in a lot of refugees. When they took in refugees originally there was a lot of pro-refugee stuff happening in the town, right? So, they had a lot of events to welcome them, donate goods to them, make sure that these people were set up effectively in the community. But then in 2015 there was a young firefighter trainee named Dirk Denkhaus who showed up there at this house on the screen which was a refugee group home and he tried to light it on fire. Now, he was not known for having extreme views, he had never done anything violent before. Everybody was shocked, everybody was so surprised that he would do something like this. But it turns out that he was doing something that a lot of us have done. He was using Facebook. He was using Facebook a lot. And so, in this study there were two researchers, Karsten Mueller and Carlo Schwarz and what they did is they went and looked at all of the data they had over a two-year period about every anti-refugee attack in Germany. So, there was 300, oh I'm sorry, 3,335 data points they looked at. So, 3,335 refugee attacks. And they looked at every variable they could come up with. They looked at the wealth of those communities, they looked at the demographics of

those communities, they looked at do those communities support far right politics, do they have a lot of newspaper sales, do they have a large percentage of refugees, did they have a history of hate crime before the refugees had entered the community. Do they have a lot of political protests? Everything they could come up with and they came up with one thing, one factor. It wasn't political leanings; it wasn't income or education or anything like that. It was Facebook usage. They found that over and over again if there was a town where Facebook usage was higher than average those towns reliably experienced more attacks on refugees. And that was true in big or small places. That was true in affluent places or poorer places. Wherever per person Facebook use was one standard deviation above the national average attacks on refugees increased by about 50 percent. And the reason that they gave for this, so in their findings what they found was that once again it comes back to an algorithm designed to optimize how much content you're going to look at. And it turns out the way that you optimize how much content you're going to look at is to show you more and more extreme content. So, they found that content that was, um, ah, eliciting negative primal emotions like anger or fear performed best. So, it proliferates. So that is what happened. So that's how even in a place that is relatively pro-refugee in person, in real life, can start having a really strong anti-refugee, um, sense online. It doesn't match up with how people really think because that is what proliferates. And the thing about Facebook is that there are so many examples like this, from the little tiny things, the, like, why are you taking this screen shot of a rape threat and displaying it to my friends to the really big things like refugee attacks. Over and over again. I have like 700 headlines about stuff like this

and this is like a very small selection of them. You see it over and over and over again. And so, what I think what we haven't talked enough about is what is the responsibility, what is the culpability of tech platforms in this discussion? Because for a long time Facebook has been talking about how, you know, they just didn't think about it in advance. I think we talked about that a little bit in the last session, right? But like we didn't mean to –whoops. And you hear that in this statement from last year. This is when the Cambridge Analytica scandal broke and they said, you know, what you heard from Mark was, um, you know we're an idealistic and optimistic company and for a decade we focused on all the good that we could do. But we didn't focus enough on preventing abuse and thinking through how people could use these tools to do harm as well. That's again the same thing, right? You focus narrowly on that one positive thing and you don't see all the negative consequences. The thing about this though is that I don't really buy it. When you think about the fact that you have started to see this pattern emerge over and over and over and over again. You can go back to 2003 when a FaceMash, which is what Facebook originally started out as, was scraping Harvard student's images without their knowledge or consent and then asking users to rate their hotness. That was the origin story of Facebook by the way. If any of you don't know that that's what they did. That's what Mark decided was cool at first. One of the things that he, ah, wanted to do at that point was also compare your photo to pictures of barn animals. So, just so you know that is the original reason that we have Facebook today.

AUDIENCE: [Laughter.]

MS. WACHTER-BOETTCHER: And back then he said, you know what, that's now

how I meant for things to go and I apologize for any harm done as a neglect, my, my neglect to consider how quickly the site would spread and its consequences. So, for how many years do we let huge companies with tons of money get away with saying like whoops I didn't realize there might be consequences. At this point I think what we really have to say is these are not oversights. What we're talking about are choices. These are choices that are being made over and over and over again and they are being made for one very specific reason. In Silicon Valley you will hear the term hockey stick over and over again. Unfortunately, it never seems to die, and the hockey stick conversation is all about growth. We need our growth to look like a hockey stick. It needs to jump up right? And every single time we have this conversation about hockey stick growth, every single time we say that is what has to happen then we allow ourselves all kinds of harms because they're never as important as maintaining that growth. And did you know Facebook made $55 billion in 2018 off of ad revenue. That's billion with a B. And, interesting detail, that's after they had a huge amount of scandals; that's after they had a ton of bad press and that's up from about $40 billion the year before. It is very profitable to do exactly what Facebook is doing. And they even admit it. This is a memo that leaked last year but it came from back in 2016 so it just leaked last year from one of the senior VP's who has been there like since the beginning – Andrew Bosworth. Boz. And, ah, he said you know the ugly truth is we believe in connecting people so deeply that anything that allows us to connect more people more often is de facto good. That's why all the work we do in growth is justified; all the questionable contact importing practices; all the subtle language that helps people stay searchable by friends.

All of it. What he's saying is, like, we're continually nudging you to not change your privacy settings to be more secure. We're continually nudging you to, ah, import your contacts, meaning like spam your friends. Like, we're nudging you, nudging you, nudging you and guess what, everything we do is de facto good as long as it helps us to connect more people. If that is the mentality you have at the very core of your business, then you're not just vulnerable to abuse. What you're actually doing is optimizing your business model for abuse. And that's what we're really seeing. Right? We are seeing tech companies that have created business models that are optimized for this. And where they routinely over and over again will choose the thing that gives them money over the thing that could actually be a safer or better community. And it's led us to a place where I look at something like this which is a post from last week from Mark Zuckerberg where he said, like, Facebook cares about privacy now. I look at this and I say bullshit.

AUDIENCE: [Laughter.]

MS. WACHTER-BOETTCHER: I would like to see a hell of a lot more evidence, excuse me, um, than what I've seen so far that they actually care about this because I have not seen it. I have not seen it in how you think about your business model and I have not seen it in how you think about your users. I've not seen it in anything that you've done since 2003 when you were comparing people to barn animals. So many of these problems I think also really come back to origin stories. Speaking of Facebook in 2003, the origin story of where a tech company comes from can tell us a lot about the choices that it makes. And I think this came up in the last panel too, the origin stories of a huge percentage of tech companies, almost all of them, are a bunch of young white

guys sitting around a room. Or maybe one brilliant lone white guy genius sitting behind a computer. You can see that at something like Twitter which back in the day was actually going to be called Status. This is an early sketch that Jack Dorsey has posted of the thing that would become Twitter. And when we were first talking about Twitter in those early days and they were sketching things out and trying to figure out what they were going to make they had this idea. This is the origin story of Twitter. They wanted it to be these live from the road updates. He imagined it as being kind of like updating your AIMS status, that's your instant messenger status, um, from wherever you are and sharing that with the world. And he also talked about how he was really fascinated with how cities functioned, and he was fascinated with things like how bike messengers and delivery drivers move throw cities and call step out over CB radios. Now, when you think about who is that that he's talking about randomly driving around a city delivering stuff and calling stuff out over CB radio? That's not going to be reflective of the people who actually ended up using Twitter, right? And then there's also the core model that it's built off of, sort of the underlying mental model of how it works. So, when you think about it, prior social networks, things like, um, Myspace, or Friendster, or Facebook, they are all kind of built off of an idea of relationships. Right? So, you have like friendships with people. That's kind of the point. Now, the core, the thing that's at the center of Twitter is not really you, it's the update, the update itself is the center. So, there's sort of a, a distancing or an alienation from the user, right? Because you're more interacting with the tweet than you are with the person in a lot of ways. So, the update is the core thing. And, you go from a model that tends to be reciprocal, right, right, like I want to be your friend, I ask

to be your friend and then you can accept or reject to a model where it is one-sided by default which means I can follow and you don't have to follow me back, you don't have to give permission for me to follow you unless you've locked your account down. But that's not the default behavior. The default behavior is non-reciprocal. Think about that for a second too. I think we've normalized this idea of following people. But if you had asked, I don't know ten random women in the year 2006 when they were talking about Twitter, hey, how do you feel about a bunch of people you don't know following you? Maybe not great. Maybe that would make you a little bit uncomfortable, right? But nobody was thinking about that. We kind of normalized this idea that just following people without relationship with them is fine and there are good things that come out of that kind of openness. But if you're not talking about the potential for harm then you are only seeing half of the story. And then there's this kind of move from like the page to the feed and that we started to see in a lot of places on line as well. But one of the things that that does is it turns a lot of the content into [[indiscernible] that can be amplified a lot more, right? So, if you have a page and you put stuff on your page people have to go to your page. But when you take your stuff and it goes out into the feed then it's sort of like outside of the bounds of your control. And so, your content can be amplified and sent out to communities that you never really intended, or you didn't really realize, or you didn't expect. So again, that is the power of Twitter. It's one of the things that makes Twitter really, really great when it's great and it's one of the things that makes Twitter really, really toxic when it's not. And the issue is that for years nobody wanted to talk about it because every time you would talk about speech on Twitter you would hear the same

thing: we're the free speech wing of the free speech party. That's, that's an actual quote from Dick Costolo. He was the CEO then. That was in 2011. And it wasn't the only time that came out. It wasn't like it came out of his mouth one time, that was said over and over again. There's (sic) quotes of it through 2012, I believe, from a variety of people at Twitter, right. In different roles. That was sort of party line. Free speech wing of the free speech party. That was the party line up through the point when in 2013, just like seven years after the service started, um, Caroline Creata Cortez (phonetic spelling), a British journalist who at the time was trying to get, um, Jane Austin's face onto a pound note, she was targeted by a harassment campaign where she said she was getting 50 rape threats an hour. At that point Twitter decided to finally implement a feature where you could actually, um, report a tweet as abusive. 2013. So, by that point we're getting very close to where a lot of this abuse kind of really started to be noticeable by sort of a broader community where it was being talked about in the media and that's when we get into Gamergate. So, in 2014 Zoe Quinn was the target of Gamergate – one, one of many targets. She was exceptionally targeted by Gamergate. She has like, you know, had to move out of her apartment because she was scared because of all of the threats she was getting to her specific location. Gamergate attempted to destroy her life. And for months and months and months this harassment campaign went on and on, and on and Twitter did almost nothing. They kept doing almost nothing up until a point when, um, Lindy West – Lindy West is a journalist, a writer and, um, her book *Shrill* is great. Um, she eventually quit the service over its failure to take responsibility. At the end of 2014 at the, in the middle of all of this Gamergate she wrote

about how she had reported a tweet as abusive. Now, here's what the tweet was. It's a bit graphic. I'm not going to show it; um, nor have a picture of it. It was an image macro so when you take a, a – you see it all of the time in memes right, you take a picture and put words on top of it – of Thomas the Tank Engine and on it were the words choo choo mother f'er. The rape train's on its way. Next stop you. She reported that as a rape threat because that's a rape threat. It's dressed up in a meme package but that's a rape threat. Twitter told her the tweet didn't violate their policy and it was fine. And that's when she decided she was out. She was done. I can't spend more time on this platform that won't even take this obvious rape threat seriously. So, along the way eventually Twitter kind of started changing its tune. Same person who said free speech party started saying, oh, you know what we suck at dealing with abuse; we've sucked at it for years. We lose core user after core user by not addressing simple trolling issues. I'm ashamed of how poorly we've dealt with this. It's absurd and there's no excuse for it. And you think, finally, we've got some traction here. Finally, they're getting it. Several years too late. They're getting it. But you see they still weren't getting it because still they were very slow to roll out any features that actually help people. And they were still not prioritizing this issue. And we're still a year away from the experience that, um, Jessie talked about in the, the prior talk which is about Leslie Jones being harassed off of Twitter by Milo Yiannopoulos. So, when they finally decided to ban him, that was considered a radical move, right? De-platforming him was a radical move at Twitter at the time. It's still considered a controversial move. They have not done a lot of that. Over and over and over again you have seen problems get bigger and bigger and bigger

and during the times when they could have done something about it, during the times when like they were a small platform and they could have made different choices they repeatedly said this, it wasn't a problem. They repeatedly avoided the question. They repeatedly went back to free speech wing of the free speech party and we're just not going to engage with it. Until at some point the problem is really big and it becomes much, much harder to solve. And then you can avoid solving it by just saying [[indiscernible] how are you going to do that at scale, right? So, every time what you have is abdication of responsibility. You have it over and over and over. You have it 2016 saying, you know, many people believe that we haven't done enough. We agree. Okay. Oh, look taking responsibility, right? We're going to invest heavily in improving our tools and enforcement. Okay. But once again harassment goes on and you keep hearing it. March 2018, we love – tell me if this sounds like anything else I talked about earlier – we love instant public global messaging. But we didn't fully predict or understand the real-world negative consequences. Hum. Weird. We acknowledge that now. We aren't proud of how people have taken advantage of our service or our inability to address it fast enough. Okay. You acknowledge it. You're not proud of it. You haven't done it fast enough. Great. So, what are you doing? Well, just last month Kara Swisher hosted, um, an online conversation with Jack. So, it was Jack Dorsey and Kara Swisher going back and forth on Twitter. And she pressed him on this, over and over again. She would ask him about it. And she was like what grade do you give yourself? And he's like C. She's like F. [Chuckle.] And she pressed him on it. What have you done? What have you done? And over and over what he would do, if you go back through that

conversation, he would evade the question. He would be like, well, it's just so hard. Yeah, no shit it's hard. You've been avoiding it for a decade. Your problems don't get easier if you avoid them for a decade. And so, I look at all of this and I just find myself so exhausted. Right?

UNIDENTIFIED: Yes.

MS. WACHTER-BOETTCHER: It's just, like, the same story on loop and I'm like are we doing this again? Do we have to have a whole news cycle about this again? This is the same thing over and over again. And you are still finding over and over that you've got terrible abuse on the platform particularly abuse against women and particularly, particularly abuse against women of color. And then most particularly abuse against black women. And that has been true for years. Black women started telling Twitter that there was an abuse problem in something like 2009, 2010. So, we're still talking about it. So, Amnesty International did this, um, study at the end of 2018 and what they did is they looked at, ah, 778 U.S. and UK women journalists and politicians. So, this is like women who are in, ah, visible positions, right? And they looked at all of their tweets that they were receiving. And they looked at them and they found that about seven percent of them were abusive or problematic but that women of color were 34 percent more likely to be harassed than white women, and that one in ten tweets to black women were abusive compared to one in 15 for white women. So, women in general being abused a lot on the platform and it gets worse and worse, right, [audible exhalation] when you start looking at women of color and then specifically black women. Twitter knows that or at least it should because so many people have told them that. I had a conversation the other day with Feminista Jones (phonetic spelling), she, um,

she's very prominent on social media platforms; she is a black feminist; and, ah, she has, you know, 160,000 followers on Twitter and she's a black woman talking about feminism. You can imagine what she gets. And she told me that Twitter had actually called her and asked her to come to dinner. And I said, Twitter needs to pay you. Like it's not enough to invite a prominent black woman to dinner. You actually need to invest in their expertise. I can bet you that they're paying their programmers with a lot more than a dinner. Why can't they pay somebody who is actually an expert on this issue. And so, again, Amnesty International, right; this is; Amnesty International's getting involved and saying you know what, [audible exhalation] women have long been telling us this and now we can back it up with data, right? Twitter is a place where racism, misogyny and homophobia are allowed to flourish unchecked. I look at this and I say this is called chronic under investment in harm prevention. Right? This is a chronic under investment and it happens over and over and over again. And then there's one other issue I want to talk about briefly here that I think is related to this and kind of digs into some of these questions about these online platforms too. And that's talking about sort of the bias that we're also seeing embedded by tech companies. So, I'm going to talk about an example to give you an idea of what I mean. And this is an example here, um, that comes out of some research done on some Google News articles. So, what's happening right now is you have a lot of people that are building – and, and actually this came up in terms of like automated review of abusive comments to see whether they need base criteria – all of that relies on, ah, machine learning, right? And it relies on this like algorithmic review of content to understand what's it's about. So, there's all kinds of people who are trying

to build all of these systems to be able to evaluate content right? So, let's talk about some of the problems with that. Well, in this particular example, um, what some researchers did is they went, and they took a ton of content from Google News articles. So, they used three million words from Google News articles, and they built this natural language processing tool off of those words. And so, what they found is that it, it would start to do things like complete analogies. So, if you said, you know, Paris is to France as Tokyo is to blank it would know the answer to that is Japan. And that's kind of like a tougher thing for natural language processing to do. That's a kind of logic that takes like a, a little extra special care. It's not the easiest part of natural language processing. So, it's like, oh cool, we can do more advanced stuff with this. And being able to have more advanced natural language processing means we can do more with understanding online texts and maybe we can do something about this stuff at scale. Great. Except some researchers also found that it returned some other interesting results. It also thinks man is to computer programmers, woman is to homemaker.

AUDIENCE: [Laughter.]

MS. WACHTER-BOETTCHER: And there are a bunch of examples of bias built into the system, right? Over and over and over again. And the problem isn't that the algorithm is like wrong, the problem is that the algorithm is trained on a, on a set of data, right? It's trained on Google News articles. So, the Google News articles reflect the viewpoints of the people who wrote them, they reflect the people who are interviewed, they reflect what makes it into the media. So, it's not true that man is to computer programmer as woman is homemaker in the way that Paris is to France as Tokyo is to Japan. Those are not the same

types of truths but the, is true in the data that they found. In the data the algorithm used those things have the same relationship. So, it was true. And so you can see these kinds of problems cropping up over and over and over again and so many companies are quick to jump on the bandwagon of saying well we need to have machine learning, we need to use these packages in our software and they're buying them up and they're implementing them without thinking about the biases that might be embedded in them. Amazon actually had to talk about this recently because they had to scrap an entire tool. Um, they tried to have a hiring tool that was going to use AI to, like, look at people's resumes. What they found is that what they wanted to do was recruit ideal candidates. Right? And so, what they did is they used ten years of resumes from, um, that Amazon had received and then the outcomes of those resumes. And they used that to train the AI. And so, you know, think okay great. Right? Like, oh, well look at the successful in the past and we'll use that to predict the future. What it did is, it downgraded resumes using the word women. So, if you mentioned that you had gone to a women's college, maybe you'd been involved with a women's group, maybe you were on a women's lacrosse team at some point in your life. All of those things would mean that your resume was downgraded. It would also, um, not really make a difference if you actually knew the coding languages that Amazon had specifically and explicitly stated it wanted. Irrelevant. But it did care if you used particularly aggressive verbs like if you said you executed and captured things in your resume that was a bonus. And the thing is, is that the people who are most likely to use that kind of language are almost all men. And so, they found that they couldn't actually get this to give them results that weren't completely biased against

women. And they had to scrap it. And you can start to see how these kinds of systems playout when it comes to something like speech online with this example from Predictim. So Predictim is a newer piece of software. It was in the news a little bit in I think December because what they wanted to do is, um, help you find the ideal babysitter or nanny, right? And it, it was going do that by using AI to screen candidates. So, what it would do is, um, they want to basically make sure that you find a childcare provider that you trust. So, they went out and they actually asked a lot of, um, "mommy bloggers" – that was the wording they used – um, what they'd want in a babysitter. So, they took this one slice of people and they said what are you looking for in a babysitter. And then they used that to train an algorithm to assess prospects' social media. Interestingly they weren't just assessing prospects social media that you might know about, they were going out and using this, this machine learning system to find social media that was anonymous. So, let's say I have an application out to be a babysitter and it has my email address and all of my information on it and I might expect that maybe they're going to look me up. Right? But I also have my anonymous Twitter account and my anonymous Twitter account doesn't have a photo of me, it's connected to a different email address, isn't using my name, never mentions my location. Their system would still try to go dredge that up and then assess what you might have said. Now, you might say, like, okay that's going to help keep people safer. Maybe it, you might say, like, okay, well, you know, maybe you want to know that your potential babysitter is, you know, trolling people online. Which, sure, maybe you would want to know that. But the thing is, um, they start classifying the data, right? So, they're using a neural network; they're using the neural

network to do that natural language processing to classify the posts. And they're classifying them along lines like is this abusive or polite; is this person positive or negative? And they found some interesting stuff. So, when one writer went through and started, um, trying to run Predictim results on people he knew. And he found that when he tried to run, um, the results on, um, a couple of people; one, a standup comic; and two, his actual babysitter, he got some pretty interesting results. So, for, first of all here, um, we've got his, yeah, actual babysitter and, ah, when he started looking at what Predictim said about her it said that she had some, like, moderate risk for disrespectful attitude. I should note, this is a picture of a black woman. And I think that that comes into account here because there is a strong historic trope of seeing black women as being disrespectful. Okay. Moderate risk for disrespectful. Here are some of the tweets that it found of hers. Didn't put on any makeup but I got that post poop glow.

AUDIENCE: [Laughter.]

MS. WACHTER-BOETTCHER: Our legal system is f 'ing crazy map. Haven't decided if I'm an indigo child or a narcissist. 2018 is the year I stopped talking s—. So, they decided that she was disrespectful and potentially less suited to be his babysitter. Okay. Let's see what they said to his friend who's a standup comic. His friend the standup comic was rated very low risk. Now, looking at what he tweets; no joke, I saw Tom Brady suck off the VP to completion 16 times; forget fake news old fat __ is a pushing __ and watch corporate and alone together tonight. Okay. We celebrate the new Gestapo President Turd, his army of cowardly f-boys, Paul Ryan. Now, what's interesting about these tweets in comparison, now, okay, he's a standup

comedian and he's not particularly concerned with being proper on Twitter. But these tweets are directly, like, targeted at people, right? Like, he's talking about there are actual people here and he's adding them and he's saying kind of nasty stuff. Now you could say he's funny or not funny or – I actually don't find him very funny. Um, you could say that this is good or bad; it doesn't really matter but what's interesting is that the algorithm thought that he was just fine even though he specifically harasses specific people. And they thought that Quiana the babysitter was moderately disrespectful. So, this writer from Gizmodo, he goes, and he talks to the people behind Predictim and they say well, I guarantee 100 percent there's no bias involved because he says, you know, we don't look at ethnicity. Those aren't even algorithmic inputs. There's no way for us to enter that into the algorithm itself. And I believe that that's probably true. Right? Like I don't think he's lying there. But the problem is that the algorithm starts to learn from patterns that it sees, and you can't actually see what it was learning from. And so, it learns from all kinds of implicit biases and implicit racial biases are all over. So, it goes through and it eats all this text and it comes out the other end reflecting the same biases that society has held for a really long time. And we don't know exactly why its shooting out the answers its shooting out because they won't reveal where they're even getting their data from. Most of these companies won't. And so, before I go, I just want to talk briefly about how do we start thinking about accountability? I mean, that's a very big topic I'm going to barely touch on today. Um, but when we start talking about this, I think what we need to really recognize that is that tech platforms have been responsible for harms for a long time. Ah, Zoe Quinn, the woman who was targeted by

Gamergate so heavily, she said, you know, if you're not asking yourself how could this be used to hurt someone in your design and engineering process you failed. But that's precisely what hasn't been happening for years, and years, and years. And at this point we're so deep into this world, so many business models are dependent on them continuing to do these things over and over again, that we have built a system that if the system has to stay in tact we can't actually solve these problems. We have to do a much better job making sure that tech is pressed to consider consequences and potential harm at every step of its process. Every single thing that is going out there, every single thing that we're using. I mean, I think all of us could stand to just sort of like get more comfortable with the idea that the platforms really have a huge responsibility here and it's a responsibility that they have failed at. None of the solutions are easy or simple. People will say well like, oh, regulation, that's not easy. You know, changing your business model, that's not easy. But fundamentally we have to get comfortable saying that we have gone down a path that is not sustainable. It does not work. It's fundamentally broken. Particularly because we are looking at a world where it's not just that our culture informs tech, although it does, right? So much of what we see in these tech platforms is absolutely the effect of a biased culture, a sexist culture, a racist culture, coming out in tech. But its also that tech is very powerful. Tech informs culture. The technology that we consume plays a big role in how we see the world. And what tech has done over and over again is treat these failures as if they are software bugs. So, with a software bug you log it somewhere, you fix it, you move on with your life. Right? You fix it like you'd fix a typo. But these are not software bugs. These are not things you can just

fix them in one-off ways. Right? Like Twitter can't just go and create one feature to solve one problem. You have to really think about these as systemic patterns, and you need to think about them as systems of action. And that's not something the tech industry has been acculturated to do. Absolutely its been acculturated into this engineering mindset where everything is kind of just a software bug. And so, I want to end with a last quote from, um, oh wait, no I don't. I have a couple of more slides. [Chuckle.] I can't see my slides all from here. Um, okay. So, I want to come back to something that James Damore actually said in that memo I mentioned earlier. He said one of the things he wanted to see tech companies and particularly Google where he worked too, was deemphasize empathy. Being emotionally unengaged helps us better reason about the facts. I call BS on that. I think that one of the biggest problems in technology is that it thinks it's been rational when it hasn't been. It thinks it's been neutral when there isn't a neutral. Its idea of neutral is actually a white male viewpoint. And so what we need to have in the tech industry that is actually much more empathetic, empathetic in a really deep way not just like oh it thinks about feelings but like at its core that it believes that what it's doing has an impact on people and it understands that it is working with people's most intimate issues. We need a tech industry that is thinking about some really difficult questions like whose job even is it to decide what fairness is. Right? Like [[indiscernible] was talking about speech. Whose job is it to decide what speech is okay. It's sort of a really hard question. Also, whose job is it to think about things like historical context. If you are designing something that impacts a certain community and that community has a history of being marginalized or harassed or abused and you

don't have anybody on the team who understands that history whatsoever then I don't think you have any business designing that product. Whose job is it to anticipate unintended consequences? At some level I think that's everybody's job, that's something everybody in an, in an industry needs to know about but I also think, like, where does the buck stop? Right? Like, whose job is on the line if you screw that up, and in the past, it's been nobody's. You can royally screwup and hurt people as long as the hockey stick goes up you keep your job and you can keep your bonus. So, I do want to leave you with one last quote. It's from Anil Dash. He says, you know, most corporate decisions about empathy involve, ah, creating great user experiences but if we make a very friendly and approachable user interface for stripping Americans of their rights, we're complicit. I work with people in the tech industry all the time and I like to tell them how complicit we have been, and I'll include myself in that. How much it's been easier to talk about let's make sure it's like feels delightful to use than actually dealing with the difficult part underneath the surface. And so, I want more of us in every single role that we're in and in every single field that we're in to make sure that these issues aren't staying below the surface; that they're coming to the forefront over and over again. So, thank you so much for having me today.

AUDIENCE: [Applause.]

# # #