Governing the Social Media City

April 20, 2021

- Hello, everyone, welcome. Good morning, good afternoon, good evening wherever you all are joining us today for the governing of the social media city, integrity design, approaches to challenging digital community problems. Sahar is here with us and he will be doing a quick introduction but I wanted to first come in to say hello and to let y'all know that I will be here to be the respondent to Sahar's presentation and we'll also be doing Q&A. So please pop in that your questions, thoughts into the Q&A feature and we'll make sure we get to those as well. I'm so thrilled to be here today. Sahar and I crossed paths when he started as a Berkman Klein fellow here at Harvard and we've had many discussions around the role of technology, the role of social media, the role of organizations, the complexities of organizations and power and how, you know, the role of individuals versus leadership and what that looks like in tech. Was really thrilled and he asked if I would be a respondant to the presentation today. So I think we're really lucky to hear from someone who has worked at Facebook and has worked in tech and really has spent a lot of time thinking about what we really should be doing as our next steps. So with that, Sahar I'll turn it over to you and then we will have our discussion afterwards.

- Thanks, Kathy. You honor me with your words and your presence and thank you everyone for being here. I realized that watching someone else talk for a while on Zoom, isn't necessarily the most fun thing. So you honor me with your presence. And now onto the show. So the title of his talk is Designing a Better City, the Clash of Content Moderation and Integrity. And I think it's important for me to talk a little bit about where I'm coming from in all this. So my name is Sahar Massachi, I was born in Israel to my parents who both fled Iran as children fleeing for their lives to the only country that would take them. When I was two years old, we left for the United States because of war and violence in the Middle East. And I grew up in Rochester which is where I physically am now visiting my parents. When this experience really shaped me growing up and it helped me think about just being really critical of the police state and critical of surveillance and unaccountable power. And that's really been with me this, like my entire adult life. When I was in college, my friends were often shocked to realize I was a computer science major because I had such a political presence on campus to give you a sense of it, and I feel like my entire adult life has been sort of navigating between being someone that enjoys and knows how to write code and so on and being someone who cares deeply about social issues and social movements. And this really came to a head when I started working at Facebook. I started in early 2016 on a growth team, so that's just normal, you know, getting people to use the platform, to stay on stuff like that. And pretty early on, I realized that there was a civic engagement team and that seemed perfect and I demanded that I join and joined as soon as the company would let me. And it really was something special. To join the team you had to swear to the civic oath that said I will do what's right for people not what's right for the company. And we did such wholesome things as helping people learn how to vote, when to vote, register to vote, things like that. And also between elections, helping people talk to their

elected officials. It kind of felt like working for a nonprofit inside a corporation and it was a really happy time. Then something interesting happened after the 2016 election and all the Russia, IRA scandals, where we started this team that we called election integrity which became known as civic integrity after a while. And suddenly we started working on things like preventing all the bad things and attacks on democracy, sort of fighting political misinformation and hate speech and stuff like that. It was a totally new thing for me and for many of us and it was startlingly intellectually productive. And over time we realized that we stopped fighting or paying attention so much to the surface level problems as much as the tectonic plates of what was going on in social media and not just, you know, this instinct of what Facebook is, but sort of a Facebook or a social media organization and what that looks like. There's a lot more red tape, there was a lot more interest from the company what we were doing because we were working on such core issues. And I felt like there was less room to think big thoughts and do big things and less freedom to just sort of think about what we were doing instead of narrowly specializing on a specific thing. And eventually I left and I'm very happy to be at Berkman to sort of think big thoughts and think about, again, those tectonic problems. And one of those problems is something I wanna talk about with you today. So here we go. Working at social media feels like building a new city and the city is kind of amazing. Millions of people can do things their parents never dreamed of. They can live together, play together, learn together. The city is a marvel. The city has also rotten. Raw sewage runs in the streets. Once in a while mass hysteria takes hold, citizen denounces citizen, friendships are irrevocably broken. People are gradually sorting themselves into armed conclaves within different neighborhoods. What do we do about it? Pro tip: my job was to help answer that question. And let's deep dive, let's dig into it. A really useful framing question is why is this new city so hard to govern? And it's useful to think of online cities versus physical cities. Physical cities don't have the problems of, you know, if you hold a conference, people might show up and just, you know, draw swastikas all over it. Real cities don't have the problem of if you have clubs, there's a high chance I'll turn into propaganda machines. Real cities just have a lot of problems but they don't have kinds of problems. And so the question is what does the physical city have that online doesn't? And a big answer is physical limits. So I think one way to think about it is like this. We have built civilization that has a set of rules and norms to deal with these kinds of problems. Again, that's hoaxes, misinformation, hate speech, harassment, things like that. And these rules and norms work but they assume that you can't teleport, you can't clone yourself. You know, just physical limits. Online we have superpowers, those break those assumptions. So rules and norms no longer really apply. They have nothing to latch on to. Eventually we're gonna come up with better rules and norms. But until we do, I argue that we need to redesign our online spaces to have some more limits so that more closely represents the real world or resembles the real world so that our existing rules and norms have somebody to work with. And a few examples might be useful. So in a physical city, you're limited by your vocal chords and the city of social media you can have simultaneous conversations with many people at once. You can post in 400 groups an hour and you can teleport all over the world to have conversations with people in very different places at the same time. In a physical city, if you are caught and denounced and known to be a liar, assuming identity is really hard. In the city of social media, you just need to sign up for a new account, and it takes about five minutes. In physical city you always know you're talking to a human being, in the city of social media you could be talking to a bot. You

could be talking, if you're talking 100 people on the city social media, how can you tell the difference between these are 100 different distinct people who are arguing with me and this is just one person with 100 sock puppets arguing with me. Your sense of what is going on and how seriously to take is really messed with. And there's more, in the city of social media you can be a Macedonian teen pretending to be a Midwestern rancher. It is hard to do that in the physical city, it is trivial to do it online. And lastly, there's also just wisdom that we have that we somehow have a blind spot when it comes to online. So in the physical city, you know that if you threw 1000 people in a high school gym, you would not have a good conversation. You know that you need to have smaller groups, maybe put them in classrooms, you can have a more structured conversation just because a thousand people can't have a good conversation. Somehow we throw people into groups online of 10,000 people and just assume that a conversation will happen well without those kinds of structures. Okay, so this is the Tubi shore slide. We still want people to be able to teleport, clone, disguise et cetera, in the magical internet world. The internet is wonderful, I grew up on the internet. I love it. The point here is we want to impose some limits rather than no limits. And I think that's really important. So let's go back to talking about the city. The physical city deals with problems using a lot of different tools and sees problems not just problems of incarceration and these are bad people that we need to lock up but also public health problems or educational problems, or, you know, it's not that people who run red lights or speed are necessarily bad people, as much as we need to assume they exist and build an environment that mitigates the damage. But what happens online? Online, we've been hiring more and more content cops. And I'm talking about cops this week or this year or this decade, it brings in a lot of baggage and we can talk about it. So, you know, let's note that and be a little more specific about what we're talking about. What happens right now is that there's a sort of, seems like an implicit belief in Congress, by the firms and a lot of conversations online, that what we want to deal with all the problems online are hire more people to just enforce Facebook law or Twitter law or YouTube law against saying the wrong things. There isn't really a good mechanism for appeal. The people who enforce the law are kind of like judge Dredd types, you know, there are sort of a self-contained investigators, arresters judges and jury, and it's just not great. It also won't work. We just cannot hire enough content cops to get us out of this mess. So quickly, what are content cops? We call them content moderators and they are underpaid and over-traumatized. They, if you wanted to get into it, there's three different kinds of them. You have the policy, people who decide the rules, you have ops teams that oversee this army of people and then you have these moderators implementing these rules. And what they do is just, you know, look at a piece of content, thumbs up, thumbs down, move on. It doesn't seem like a very intrinsically fun sort of job. It actually doesn't have a lot of power and it's kind of just an assembly line. And one thing to notice is that every few months you'll see a rollout of new rules that these content cops enforce from the policy teams and these new rules are just constantly adding more edge cases, maybe epicycles for adding more epicycles for those of you who catch the reference. We're just making a more and more complex system and it just doesn't work. We'll never have enough cops and we'll never agree on what exactly censorship is. Luckily there's an alternative. Hooray, and that is integrity design. And integrity design has a lot of names. Sometimes people call it fighting abuse, sometimes people call it integrity work systems design, trust and safety fits in this kind of, and it's roots comes from instead of fighting bad words, fighting bad actions. And it

borrows a lot from computer security. You can think of it as the practice of I have this big system I need to understand how it works. There are bugs in how the system operates and loopholes, people are gonna attack the system by abusing those loopholes and they're gonna attack it not by hacking something or doing a SQL injection or a buffer overflow, they're gonna hack it so that they can spam more or so that they can get away with hate speech or whatever it is. And definition of our work could be systematically solving the online harms that uses inflict on each other. So ethics could be solving the harms that companies inflict on their users. Security could be solving the online harms that outside parties inflict on your users. And integrity borrows and has things to say to both those things but it's a specific new kind of work that's happening. And people seem to talk a lot about content moderation. Sometimes they think they're being very smart and they talk about feed ranking and algorithms. But when I'm here to talk to you about today is integrity design and somewhat algorithms. So there are a lot of strategies of doing integrity and one of the biggest ones we found is bringing back some real world physics into online interactions. There's so many ideas about how to do this. Oh my gosh, I'm only gonna talk about a few because of time, but it's a really rich vein to tap. So here's a few. You're gonna hear a lot of people talk about friction and different people mean different things by it. It really is a design principle that undergirds a lot of this and the idea is make things that could be bad, hard, be harder to do as people do them more often. And the cost could be things like a lag time or making the text in gray rather than black. There's a lot of ideas but this is a sort of an underlying idea here. Let's get more specific. A Bar Mitzvah is a thing in the Jewish tradition where a child becomes an adult. And this involves typically a lot of study, a lot of hard work, a ritualistic thing, and then you have a party and we can borrow from this idea to deal with a specific problem which is cloning. There is a problem of people having thousands of fake accounts. We want people to have a few multiple accounts just so that they can have some alternate identities because the internet is wonderful. And it's hard sometimes to be able to tell what is a fake account and what's not. The strategy here is just make the costs higher, jump through some hoops before you get blessed as an adult account. Maybe you just have to wait some time, maybe you have to do to demonstrate some pro-social behavior. We just wanna create the cost of a new account that has access to the features that could get abused to be higher. And crucially, we expect attackers to jump through these hoops. We just think that it would be possible to have three fake accounts, a thousand though it would just be a lot harder. And that's a key principle here for trying to mitigate risk not get rid of it and we're trying to make doing the wrong thing harder so that it's not impossible. Another version of this is speed bumps. So online the worst harms almost always come from power users and that is people who post a lot. We could give special scrutiny on the people who post a lot. We could throw cops at them to sort of pay attention to what they're doing. And honestly, that's better than what we do now which is mostly ignore them. But you can also solve this with design, just add speed bumps. If people are doing risky behavior like posting in many groups at once, this is like speeding. If you speed too much, you're locked out of using the feature that you're speeding with for a while. This is another example of just easy design inspired by physical limits. So this gives us a really fun definition, spam is behavior of people who take advantage of this lack of physics and spam and spam fighting is the other part of the pillar to what I would consider to be really just like how you fight these problems in a non-throw tons of cops of the problems kind of way. So here's a me. So as an industry, we know how to fight spam. Your inbox right now

probably has very little spam in it. That's not in the spam box. This is a more or less solved problem and we can learn a lot from spam fighting in general. For example, a simpler system gives you less attack surface, pay attention to incentives. We've talked about a lot of this. But the thing I really wanna focus on right now is this last bullet. If your current reach comes from spam, you don't own it. It's theft, not an entitlement. So we've just gone over a lot of ways to fix a problem. We've talked about being inspired by physical limits. We've gestured at spam fighting, we can talk more about that. And the question is, why aren't we doing this already? If this is so obvious, why isn't it implemented? And one of those reasons is the stakeholder trap. So let's say that we are a firm and we're an integrity team inside a firm. And our job is to find a loopholes in the rules that allow people to spam. If we crack down on those loopholes, by definition there are some people on the platform who are benefiting from spammy behavior, who will be upset. It is impossible to crack down on these loopholes and to fix these bugs and to stop spam-like tactics without affecting the people who are currently benefiting from them. If you allow them to cow you, you will fail. And this is an important concept. And that brings me to sort of the end of the talk which is integrity design can do a lot. It's a beautiful alternative to throwing tons of content cops at the problem but it can't solve political problems. And this brings me to what I like to call the wire problem. At the end of the day, platforms need ethics to do the right thing. They need to enforce the rules fairly, they need to knock you off to spammers and they need to actually do the work. We can have amazing investigations, we can have amazing ideas for how to fix the problems online but they only take you too far, so far if you're afraid to bust the perps. And I'll close with integrity is the superior alternative to censorship. We know how to do it and we know the sort of problems that arise when trying to implement it. Here's a sort of sense of what a firm could do to solve the problem. What we haven't talked about is what the public and governments can do about it and a bunch of other things. But I hope this gives you a sense of what it might feel like from the inside and sort of how one could fix the problem. if one took this path. The internet is too precious to flood with cops. And as a coder, I wanna take a step back and and talk about a meta argument here, which is the integrity professional. So this whole presentation you just saw is the product of an integrity worker, that's me. And integrity is emerging skillset in a professional role in thinking like this and fighting these problems. And it's really fun for me because finally I get to be at the intersection of working on technology, thinking about society, thinking about policy and a lot of people are working on it. Online in Congress, even from the firms themselves, we see a lot of arguments about free speech forever. No, we need to be on the Nazis. This just seems like a little bit of a distraction to me. Let's talk about feed ranking, let's talk about algorithms, but even more than that, let's talk about integrity design and how you could do, how you could get at these problems in a more systematic way that doesn't have these problems of censorship at them. There are integrity professionals working on these inside of companies and companies are a terrain of struggle. So please know that this is happening, people are having these conversations all the time. And one thing that really saddens me is that there's this really, like important conversation happening inside of firms where they have access to research and A/B testing and trying out things and a whole nomenclature but the public doesn't really have access to it. The public has their own concerns that aren't permeating the firms as much. And these two things seem, I think the public could really learn a lot from just the experiences of people on the inside who've studied these dynamics. So my personal project is we're thinking

about can we gather integrity professionals and give them a voice? This isn't the only analysis, there a lot of caveats I didn't have time to get into and there are conflicting concerns. This isn't the only way to look at the problem. But it is a way, it is an important way. And to my mind, we need more of it. And that's it. Thanks for coming to this thing. Kathy I cede the floor to you. Done.

- Great, Thank you. Hi everyone, as a reminder to go ahead and put in your questions. thank you to Lance and Reuben for going through those questions and helped me find them as well. Sahar, thank you for sharing your background, your experience at Facebook, your research and thinking around this space and tying it all together in ways that many of us can think about, dissect, question, understand. I'll share some thoughts, some reinforcements, maybe a few challenges and some questions, and then we'll open the floor for so many others as well. So we both are computer scientists who found her way to the Berkman Center to really think about a lot of these topics. There are a few things you raised in the beginning that really had me thinking, right. You mentioned that you were inside a company and part of part of the values was to do what's right for people and not the company. And there we've seen so many of these instances in companies as well, including places I've worked. So I spent about 15 years in the private sector in big tech companies before moving on to about four years in government at the White House then later on into academia. And I actually run, I'm responsible for computing. At Mozilla and we see, you know, these different value sets in companies do as little harm as possible, do what's right for people, et cetera. And it looks like somewhere along the way that can break down. And you mentioned some of that, right, when other priorities start to take over, or when you have to look at like the tectonic plates and now all these other issues maybe are seeming less important. So one of the questions to raise to think about more is really what happens when that happens? How do we think about even if a company has a set of values how do we make sure that's incorporated into all the products we build? You know, Karen Hao had that piece of investigative journalism looking into the inside of Facebook and there are others who have now spoken about their experience there. So what do we do when it looks like there are teams, like you mentioned, you know, in your final closing thoughts, there are integrity professionals who are doing this work, are they effective? If not, what are some of the obstacles that they have to overcome and what are our roles, either as users of the product or if we were to go into those companies, how do we really think about how do we empower people like you who were there because you're no longer there? And this also goes to another piece you had me think about too, was, you know, all these people who are empowered in companies and I certainly know people who are still there but we also now have seen professionals who've been fired, who think about integrity and ethics. And so how do we both trust companies when they are trying to do this work and also if either we go into these organizations or again, are continued users, how do we think through what that looks like? How are these integrity professionals empowered to do their jobs? Another thing I really, really pulled out was the city analogy. Major kudos to that, analogies are always so hard because we all come to the table with our own assumptions of the thing you're doing an analogy to. And you really were able to carry that through, one piece that really struck me, specifically was that piece about how in cities, not only do you have the policing and cop factor but you have social workers and libraries and parks and sanitation, and so many other tools that are at the disposal

of the city to really ensure that the citizens are taken care of. So it's not just like architect which is important. We're gonna build the city and we're gonna run away and we'll forget about everything else. There are so many other factors involved. So how do we get there with tech? We actually the Berkman Klein Center had an honor and expertise talk about 1 1/2 ago. And I think tech to some extent is starting to get there. But just the other day there was, you know, a panel talk around tech's role in society. And again, we came back to this idea that in tech there's a lot of hubris and those with engineering computer science, computing backgrounds, and those with deep expertise in society and humanity and politics and economics, aren't always at the table on equal footing. We have the research now, right, of people like Dr. Safiya Noble with algorithms of oppression and Sarah T. Roberts who wrote "Behind the Screen" and was in the film Cleaners and Joy Buolamwini, and Dr. Ruha Benjamin, and so many others, who have such a depth of experience in these other parts of the city. And so what does it take to now bring those other parts of the city, whether it's integrity design or something else, into our online virtual city so that we now also have a version of social work and sanitation and these other aspects that you bring up in the real life city. I thought that was such an interesting analogy to bring up. And, oh, and then you also mentioned this idea of a mature adult and you talked about how do we think about having these either speed bumpers, and people have questions on those too, speed bumps or other criteria in place before someone can be a mature adult or someone can advance the next level, et cetera. With my product hat on, I generally, I think often about edge cases and who's left out when positions of power in things like speed bumps or criteria and that's more a question than anything with an answer but more the assuming, there are speed bumps and what happens to people who might not fit your criteria, yes, the goal is to perhaps catch the bots and the people who are standing, but what if in the way we're now also removing access to people that could have used the platform? One, I think one scenario that some of the companies have had to grapple with is the real name scenario, right. Well, to ensure integrity, we are gonna require real names on YouTube accounts and Facebook accounts, et cetera or at least this is a more Google thing. I shouldn't, I don't know if she knows with Facebook with this but we're gonna required real names on some of our accounts. And immediately, as you know having real names also causes a lot of problems for certain people who can be in danger if that happen. So what are some of the scenarios where if there are requirements to be a quote, mature adult to be on the platform, now we're also leaving out so many other people. So thank you for bringing that point up for us to think about and discuss as well. And then finally, I think another piece before I turn it over to the audience is to also really think through this idea of the redesign on online spaces. There's such, it's such an interesting framing and there's so much interesting work also done in this space. Our own, we had a fellow here a few years ago, Johanne Chan who's a designer by trade and talks a lot about the thoughtfulness that has to go into designing both physical spaces but also online spaces and how they tie together. So how do we also bring those voices into the room as we design these spaces? I know there's a new, there's New Public by Civic Signals that's also thinking through, you know, how do we just design more meaningful spaces online or like in the digital world that really takes care of people? And it would be interesting to see what other resources that exist, that both I didn't mention and that we can also look into as we think about this as well. And let's see, oh, and then one other thought that came up too was this idea of the comparison to computer security and how that's been around for a little bit. And we respond to

computer security 'cause we have all this scaffolding in place, right. If someone breaches our system, if someone at this point, although it's imperfect we are constantly still hearing cases about security issues, you have to safeguard your accounts. But to some extent, companies prioritize having either an SRE team or dev ops team or some kind of SRE stands or site reliability engineer or security team but the moment your system goes down there's a team that swarms to take care of it. Part of that because that's viewed as like quote, real technical problem and your users can't get access to your product. And there's like a whole team that's built out in this world. So what is the equivalent of that, expanding on what you've already said about integrity design to have a whole team that now swarms on the same thing, the site doesn't go down in the same way, but it is going down, the city is decaying with more sewage, to use your analogy. So what are ways we can prioritize things to think of this just the same as if a system went down? Someone pointed, you know, to the point the other day that well, when a system goes down like that in like the in like security sense, the companies actually lose revenue. So that's a motivating factor when it just slowly rots away in different settings that's harder to see. So how do we think about those metrics or how do we think about the outcomes of that in such a way that the companies really are actionable? So that's such a great point to compare to security and would love also thoughts from folks in general on what that might look like. And with that, yeah, I think I'll turn it over back to you to Sahar to either say a few words or turn it over to the audience. You have 32 open questions in Q&A so we'll try to get to as many as that as possible. And then there are about 14 just love notes and comments and thoughts on the presentation.

- Wow. Well, thank you, Kathy and thank you everyone else. These are such good questions. Oh my gosh, I feel like many of them could be a whole hour-long discussion themselves. In order to honor both your very good questions and all the people in the audience, I'm gonna try like a very quick, like churn through them. But, you know, with the caveat that like we could spend a lot of time talking about it. So I hope I'm more gesturing towards an answer than being seen as giving the answer. So how do we help people inside the companies? Whistleblower laws are really important. So Facebook, Google, to some extent, Amazon, they are surveillance companies. Being a worker at a surveillance company is different than being a worker anywhere else because the company surveys you. This means that we are seeing, one reason we see much fewer whistleblowing ethics accounts from companies like Google and Facebook is that their workers are terrified because they know just the vast reach of those companies have into like their lives. It's much easier to whistle blow something like Boeing because they're not on your phone. And I think legislation or something is really important in dealing with that. Furthermore, when it comes to the sort of like firms as a field of struggle, there's a really long conversation we can have about metrics, you know, different firms work differently. I can only speak to the firms that I'm familiar with. And I think this is a good point to say, like, I come from Facebook, therefore my talk is colored by Facebook but it is important to give them credit where it is due. To my understanding They hire a lot more integrity workers than anyone else, they've done a lot of work. There is a lot of like interesting good thinking happening in the firm. And though I have my frustration with them, at the very least, they're hiring a lot of people to work on it and it's important to say, like, if I had come from YouTube, you know, you can imagine this being like a bit more of a YouTube flavor talk. So as far as I know, it's all about

metrics. When someone runs an A/B test, what are the top five metrics they see? They're not integrity metrics, they're growth metrics. And that really means that you have one small team fighting against the tide and all these other teams in the company being incentivized to do things that often like make the problem worse by accident, right. They're just pushing up their metrics. And that's huge. Your point about this all sounds like a bunch of engineers is really well taken. In my experience, We were actually, if we were dominated by any role it was researchers, specifically people with PhDs in the social sciences. And that was a really fun, important time. You can imagine companies hiring people in the social sciences much more but you can also imagine companies creating better channels for talking with people of different backgrounds but you could also imagine us building a scaffolding for that to happen without waiting for like the largess of the dictators of these companies. You know, the dictators is a charged word. The CEO's who own more than 50% of the voting shares to decide they want to do it. And you also could imagine dealing with this problem through an antitrust lens or through a co-op lens or through a public utility lens. And that's also kind of beyond the scope of this but I think that we've started gesturing towards questions of like power. And let's say society decides they want something, what does that mean? Next page. The thing about mature adults, Bar Mitzvahs, real names, the real name policy, I think is also a good one. All your points are good ones. And I think the big thing I'd say here is again, it's like about costs versus locking people out and escalating costs as people do a thing that is likely abusable as opposed to high costs upfront. And I think that's the best way you can sort of deal with this problem in a very abstract, three senetnce way but it is a concern. The meaningful spaces thing is important, right? The internet archive is amazing, Craigslist, Wikipedia, they're all part of this like, I don't think they could exist today because the internet has changed so much and we definitely need more public spaces. Like I think the new signals, ELI pairs or people are talking about creating sort of oases off the dominant firms to have public space and that's important and in part an inspiration to what I'm talking about which is inside these commercial firms and commercialized spaces, can we create, can we think about designing those as well as alternatives or compliments? And lastly, I'll talk about the security team and I'm really glad you asked. I'm gonna share a slide that's in my appendix that I care a lot about, and that is, wait a go, boop, boop. What about the cops, right? What about all the content moderators who we like and care for? And the answer is, I think have them be more like Sherlock Holmes. You can imagine people responding to attacks by looking at, well, not just what was the content that this person did that we don't like, but are they part of a ring of people? Are they acting in a coordinated fashion? What is the post-mortem here? What loopholes did they use? And then writing up a report and then handing it to the rest of the company. I'd say that the content moderators are among the people who have a really distinct sense of where are the emerging attacks, tactically and also strategically like what's happening? What are people saying? What are our current processes missing? And to my understanding, they're just not being listened to right now, they're not really given that agency to sort of like report. And that's one of the many things that you could change and that's like gesturing at a sort of post-mortem. Okay, that might've been too long. I'll try and be even more brief.

- No, thank you for all those terrifying pieces. I wanna use your slide as also a way to plug both again, the behind the team work with Sarah T. Roberts and Mary Gray's Ghost Work who really

get into both content moderation, but also the whole ecosystem of all the people with much less power than the big tech companies who make tech run and how we have to think about that as well. Oh, and I also made note of a clarifying point, for those of you who don't know what A/B testing is, it's when companies often say we have option A, option B, we're going to test out different options to see what people use better and even the premise of A/B testing has now been well-researched as sometimes also being problematic in how companies use that to basically research people. So I wanted to clarify that in case anyone didn't know. And okay, let's dive into some of the 31 of these, and they've been, some of these have been plus one. I also wanna recognize that the plus one system in the Q&A, just can also be imperfect. Those questions that have been around longer tend to also have to have more additions. So we'll start with Nick and then I'll jump over to Anna. And so Nick says I'm most struck by the comment that it's impossible to a meaningful way, or sorry have a meaningful conversation, every one with 1000 people. So how does integrity deal with that? And surely not a matter of more than just speed bumps and nudges and how could it be dismantling certain spaces completely. That actually were quite a few questions about speed bumps in general, as well in the Q&A.

- Thank you, Nick, thanks for being here. So to clarify a lot of what I'm talking about, you know, I'm evangelizing integrity as a discipline but also my own specific ideas within it. And there are other ideas within integrity that are not these. And I might have, you know, I think integrity people might disagree with some of it, et cetera. So that's important to note. My sense of an integrity inspired way of thinking about this 1000 people in the group problem is thinking hard as a product about what a group is meant for and building it. So that might mean more reputational tools so that people who are like heavily involved, people in a group are marked rather than someone who's dropping by for the first time. That could mean if a group is trying to make a decision than having a specific affordance for it in the technology. If a group is trying to have a long running discussion, maybe instead of having comments on posts which seems to be the prevailing way of discussion, maybe it's, you know, you build something inside that lets you have like speaker for, speaker against Robert rules. I'm not really sure, I can't really think off the top of my head about this. I think the answer flippantly is just think about what kind of discussion people are having inside the group and build structures that allow that discussion to happen in a way that isn't just commenting on a link. And I hope that makes sense.

- Great, thank you. Anna, I'm curious whether integrity design is just about changing technical affordances or whether it's also considered as broader support and compensation for the social labor that folks like volunteer moderators do. There are plenty of social workers who aren't taught.

- Yeah. Well, integrity design is kind of a term that I'm fighting to be used. So it could mean anything you want it to mean. I think that the, every sort of discipline maybe has like a universalizing or like empire building component to it. So as a partisan of integrity, I'm incentivized to say, yeah, that's part of what we're talking about too. It's all part of it. But being a little less flippant, yeah, I think it really matters, right. So Reddit has a volunteer moderator structure which both has problems of unpaid labor in terms of uncompensated is, you know, you're doing work for the firms, but also has problems in terms of unpaid labor means the

people who are moderating these giant subreddits are often not held accountable or have to find a different way to monetize, there are all kinds of problems there. If you think of it as a systems problem and like this is a weak point in our system, then I think there's like a really strong integrity case we made for we need to start worrying about it.

- Thank you, Sahar. Yeah, thank you for bringing it up what happens when there's mass moderation outside unpaid moderation as well, the power structures that really come into that as well. I'm gonna combine two questions by, let's see, I think it's Sandra and Kelly 'cause they both get to the idea of incentives and profit. So is there a really a way to get large firms to pursue a goal other than profit maximization to adopt something like integrity design and then Kelly's similar was, you know, a major difference between a city and a platform like Facebook is that Facebook is for-profit and not a government entity. So really can't afford profit platform ever have the right incentives to build a public space that really works for everyone. I'm gonna add there that really what struck me going from private sector to government was that we didn't have a luxury to not design for everyone whether it's a digital piece or a policy, even people still do. But yeah, we don't have edge cases, it's everyone. So how do you think about that?

- There's sort of like two questions there, right? Or maybe the question is, are these social media firms bad because they are just as bad as every other corporation or are they worse somehow? And, you know, I think you can make an argument, even that they're better. I would much rather, I dunno, not have like Exxon or Altria in charge of Twitter than Jack Dorsey, I think, I'm not sure. You know, you could imagine a world where the people running these firms are just worse. And I think that there is a large conversation to be had about regulation, anti-trust, structural separation, public options, all which are separate ideas. And I have a lot to say about that. It's a little bit out of scope of this talk just because it'll take a long time to talk about. So maybe the flippant answer is if you are a social democrat or anything to the right of that, I think you would argue that part of the role of the state is changing those incentives, right? Like if you imagine that everything's profit maximizing, internalize the externalities, et cetera, et cetera, and that's all doable. And I guess I'll just state, like we had Standard Oil, we had a bunch of big companies doing bad things in our history and we were able to deal with it. And to some extent, this isn't like a new problem, like pollution isn't a new problem. We just need the political will to like apply the solutions we already have to that problem. That's not 100% correct, but it's like mostly correct.

- Great, thank you. I'm gonna combine a few more. I think are somewhat related. There's just so many. So that folks that wanna learn more from you, Samer and Harvey, I think these are both around like who's doing this? Well, so do you have examples of platforms that are already implementing some of these measures? Are we seeing healthy communities out there? And he shared Wikipedia implementing the child to adults account permissions or adding friction in different kinds of ways. And we also, there's actually a few other questions that asked are the existing platforms capable of building a better city? So what are some good examples and are we capable?

- I think a real platform that I can think of is trying something and I'm not up to date on what every platform is trying to do. I think like Twitter, adding friction to the retweet, where they really want you to like write something when you retweet as opposed to just indicate one is an example of that, as are the ones that you mentioned. And I've heard really good things about for example, Ravelry, and that's really intriguing to me. Nothing jumps to the top of my head and maybe I'll get back to you on that and I'll think about it a little bit with the next few questions.

- Okay, thank you. And let's do one of from about what ethics means. So Jeremy says what is considered to be ethical, varies from person to person. So since you are a fellow at a university, can you talk to us more about the academic philosophical theoretical arguments you use to draw your ethical values from?

- Oh gosh. One thing that always really enraged me working, well, hmm, I think that you can imagine being really frustrating working at a firm is drawn from like a Rawlsian veil of ignorance. So this, you know, Rawls's veil of ignorance doesn't work totally, an all concepts it's been supposedly debunked by some recent thinkers but I think it's a really useful exercise. There are times when, you know, executives or someone set up a set of rules, you say, okay, we're going to apply those rules. And then after you apply those rules, they say we don't like those outcomes, change those outcomes, please. And that just seems like clearly wrong. And I don't know, I feel like philosophies, you know we're all influenced by so many, every sort of formal philosophical school of ethics to me has some flaws. Luckily or unluckily, I don't think that we're in the sort of world where like the really hard ethical questions are the most important, right now we're dealing with ethical questions of the form of, you know, Joel Kaplan is messing with the rules again or, you know, we have a process and then some executive is sidestepping this process because they don't like its results. Like that's kind of the stuff that we're dealing with right now. And I think that's covered in all ethical traditions.

- Yeah, thank you for sharing that, I too have seen that as well. Let's jump to a question, I think also covers a few other questions. Hara Hussein had asked and the session insightful on there, I thank you for asking this one. It's lovely to hear from you Sahar and why he's so passionate about all of this work. And I also appreciate that Sahar said multiple accounts it's a necessary part of the web. In Pakistan where I'm from, women and queer people often have multiple accounts to be able to navigate the web to find information community and express themselves without surveillance by family, conservative groups for the state. I'd love to ask what do you think about progressive feedback loops? For example, users getting more privileges to get involved in governance if they're able to go through all the stages of maturity?

- First, I'll say that I'm a fan of Hara and you all should check out chayn or chayn, I don't know how to spell it, chayn.org, probably. To answer your question, one thing I didn't have time to get into is that like we're talking about a city, but the city is still a dictatorship. And as you know, a good American, I believe in democracy as we all should. And again, kind of out of scope of this talk is sort of how do we bring democracy to the city as well as just good design. And I think that what you're suggesting seems to be like a good building block of that, right. In a

world where people can have multiple accounts, how do you have a vote that isn't tainted? How do you have an ability to like discuss things? I'm told that this is the last question. So my email is integrity@sahar.io and I will paste it in the chat and please feel free to hit me up there. And this doesn't have to be the last, like the end of the conversation. And thank you Berkman for hosting me and Kathy for being such a good MC and questioner and collaborator and everyone for coming.