Assembly Project Fellowship Showcase

May 17, 2021

- Welcome, everybody. Let's see. I can still see people filing in. Oh, right. Why don't we get started? Hello, everybody. I'm Jonathan Zittrain and I'm so pleased to be welcoming all of you to our session today where we will hear from the participants in the 2020/2021 Berkman Klein Center Assembly Project fellowships. This marks, I think the fifth year, that we've done the Assembly program. The idea was to bring participants together from academia, industry, government, civil society, and across all sorts of disciplines to work on problems arising from technology that touched the public interest and that are really hard, and that might call upon multiple sectors or parties or types of thinking to make progress on. And they're every bit as pressing as they are difficult. And the original theory behind the program encouraged by our colleague Jordi Weinstock was that if you had just the right amount of a little bit of structure, but not too much, and managed to yank people out of their day-to-day wherever they might be working in a day job or occupation, and kind of put people in a new configuration and load them into a particle accelerator, figuratively speaking, and try to inspire some collisions, that good particles might result. And that has borne out over the five years of this program where we've touched on in different theme years issues of privacy and security, ethics and governance of artificial intelligence, more recently disinformation on online platforms. To be sure not that we've solved all this stuff but it's just been fascinating to see people who consider themselves already somewhat expert or steeped in one of these sets of problems where the relevant fields try to address them and to see folks of that sort encountering some of the people and projects that have coalesced through this program and say, oh, I'm thinking of a new way of looking at that stuff. And for each of these, over the course of the five years, we've had more than 75 fellows. And out of that together, there's been I think just under 20 projects. And every year, one or two of those projects has continued beyond the life of the program whether a proof of concept or beyond trying to continue addressing the problem that they originally identified. And I think a number of people as you'll hear from them today have said that the program can change their own professional trajectories and ideals about what they might wanna work on. But this pandemic here, instead of picking an entirely new topic and starting sort of from ground zero, we invited five teams who were continuing elements of work they'd done in earlier iterations of the fellowship to come back for further support. We matched them with advisers on the Harvard campus, created kind of a new sort of all-star cohort among everybody, we're able to facilitate some small grants and offer some opportunities to share their work for feedback with other relevant players, some of whom might be able to help implement some of the ideas they've had. So today, we'll be hearing from those five projects. And before we do that, I just want to first give a huge thanks, not only, only as I already mentioned to Jordi Weinstock who kind of had the initial vision for this program, but to Zenzele Best, Hilary Ross, David O'Brien, Oumou Ly, and Sarah Newman who have really helped pull all this together and create just that right amount of structure and inspiration and creativity to get things really cooking over the course of this project and this initiative. I'm also extremely grateful to the advisors both in past years and most recently in this year to include on these projects Kade Crockford, Mary Gray, James Mickens and Margo Seltzer, who have each spent their time offering honest thoughts and suggestions on what you're about to see tonight. So grateful you could be here, and that will be, this is by way of warning, recording this for posterity. And I think now back to you, Hilary, again with thanks to introduce our five teams. I think there's probably a banner already up. Oh, yes. That's the thing, the project fellows. There's the five things you'll hear about today. And then I think it was probably already up there, the opportunity to put in questions in the Q and A tool. And then at the very end of each of the five presentations, we'll have time for possibly advancing some of those questions to the teams. So thank you all. And thanks to our Assembly fellows. I'm just so pleased at how much you've really brought it this year and at the work you're doing on questions that really matter. Hilary, over to you.

- Thanks, thank you so much, Jonathan. So I'm Hilary, I'm a senior program manager at Berkman Klein for the Assembly program. I'm really thrilled to be introducing these five projects that are working to improve the state of disinformation and privacy and the governance of AI technologies. As Jonathan said, all of these projects started during Assembly fellowships over the last five years. They've been working on their own and then they've rejoined this year to dig deeper and further some piece of their project. So, tonight they'll be sharing about their work broadly and about what they've done over the last few months. So before we jump in, just a huge thank you to our staff team as Jonathan said, to Jonathan himself, to the fellows, our advisors and the Assembly alumni community with a particular shout out to Tanay Jaeel and Michaela Lee for their peer advising with project teams this spring. And as Jonathan said, the flow of tonight's event over the next hour, these five teams will be sharing about their work with brief intros in between for me, then we'll shift after their presentations to Q and A in the same order that they're presenting up top. And that Q and A will be moderated by Jonathan. We'll be collecting questions as we go. So please, please share your questions through the Q and A tool as you have them, and I will remind you to do that as we go also. So with that, I'm excited to get going and to introduce the AI Blindspot project. AI Blindspot came together during Assembly 2019, working on the ethics and governance of AI. The team is Ania Calderon, Hong Qu, Dan Taber, and Jeff Wen. And we asked each project team you might've seen in the promotion for this event to create a riddle for their project. And theirs is, we all have it, it looks like a blur. If we don't spot it, harms will occur. So I'll hand it over to Dan from here to untangle that.

- Thank you. I'm excited to kick us off by presenting the AI blindspot which is a tool for advancing equity in AI systems. I'm presenting this, you can go to the next slide. I'm presenting this on behalf of Ania, Jeff and Hong. And the four of us have been working on a blindspot for the past two years with the mission of dissolving barriers between those who build AI systems and those who don't. Particularly those who advocate for particularly civil society actors who advocate for more equitable AI systems. If you go to, sorry, can you go to the next slide. I'll start by introducing the AI Blindspot Framework and what it represents. And then I'll discuss how we take an art-based storytelling approach to change the narrative around how these AI systems are used. And then dive deeper into a specific case study of tenant screening algorithms, and show how we're using the AI Blindspot to improve how these algorithms are utilized. You can go to the next slide. So here's the AI Blindspot Framework. From the very beginning, the framework was designed to mirror the steps that a data scientist would typically go through when they're designing a model or AI system. There are four stages to the framework of planning, building, deploying and monitoring a system. And these represent a typical data science workflow. And through each stage, we identify oversights where unconscious biases can give rise to structural inequalities. You go to the next slide. Part of our mission is to, as I said, is to dissolve barriers between those who build systems and those who don't. And in an effort to do those, we always strive to create accessible content. Two years ago, we designed a deck of cards which you've used in interactive workshops at different conferences, such as MozFest. And the latest iteration of AI Blindspot is that cute guy you see the right and we have actually call Al. We go to the next slide. Al is a child like being with a human body and a gear for a head. And Al represent the idea that just like children are a reflection of the values of the people and environment around them, AI reflects the values and priorities of those who create them. We don't believe that AI and technology are inherently bad but when they're implemented, they often represent the values and systems of those who create them. And as a result, they often reflect the structural inequalities that exist in our society. You go to the next slide. This encourages people to ask the question, what kind of AI are we raising? If you go to the next slide. In the past few months, we've transformed the original AI Blindspot icons into these illustrated scenarios. And these illustrations represent how the blindspots would play out if AI were a person. We plan to create animated versions of these illustrations, which you could use in social media and educational campaigns. And we're already using them in interactive workshops including one that Hong is doing in two days at Stanford as part of their tech and racial equity conference. If you go to the next slide. The workshop- worksheet you see here is an example of an activity participants will do where they identify what is Al, what type of AI system? It could be a law enforcement or candidate screening or robot assisted surgery or anything at all. And then identify who are the stakeholders who are designing Al, who's harming, being harmed by Al and what types of values and systems is Al reinforcing. And then most importantly, how can you change that story? How can you create a more equitable version of Al? And I'll give an example of that by going into tenant screening, if you go to the next slide. Just to give some background, tenant-screening algorithms are a billion dollar industry. 90% of landlords say they use some type of algorithm to screen potential tenants. And these algorithms are almost entirely unregulated. They take access to data from people's credit scores and employment history and criminal history, but really virtually, anything they want because they're unregulated. And many critics have argued that these algorithms disproportionately harm black and Latino communities. So one thing we're trying to do is, excuse me, is to, excuse me, let me rewind a bit. Not only does this deny an individual housing but it can create a feedback loop where because an individual has now been denied, they have that on their record. And it can also, which affects their likelihood of getting housing in the future. And it can also affect their likelihood of getting employment because the lack of housing. And this feedback loop can persist for generations particularly as the federal eviction moratorium is lifted. If you go to the next slide, this goes back to the worksheet that we, that I showed a second ago. This is an example of what participants would work through in the workshop where they identify the stakeholders who are influencing analysis design, not just those who use the algorithms but other organizations like Consumer Data Industry Association and related companies like those who do background checks and identify communities who were harmed and what values and systems are being reinforced throughout such as society's history of housing discrimination. And then how you can change that narrative. One thing, a lot of communities are trying to change the narrative, but one thing we often hear is the community struggle where they don't really know how these algorithms work. So they don't know what changes to call for. So another thing we're doing is designing a case study in which we walk through each blindspot and show how it manifests itself in the form of tenant screening algorithms and what changes you can call for. I'll show you a couple of examples in the next two slides. If you go to the next one. Success criteria refers to the idea that an AI system's metrics for success determine whether it's, a specific metrics determines whether it's successful or not. But when you try to optimize for one thing, you're often gonna give rise to harm because you're ignoring other priorities. Just like in the case of Al. You know, you can evaluate whether the hose is successful based on whether he can water a plant but then you're ignoring the fact that he's not putting out a burning building. In the case of a tenant-screening algorithms, if you go to the next slide, these algorithms are likely rewarded for minimizing false positives. This means that you want to make sure that, companies wanna make sure that when they recommend a tenant, it really is a quality tenant. But that means you're making a statement, that you're willing to make a mistake and potentially screen out those who actually are qualified. And in doing so, when you make that mistake, you're most likely to mislabel those who come from disadvantaged backgrounds who really are qualified yet are being denied housing. So a specific solution in this case is to not just require some type of audit, which is very vague and really could look like anything, but specifically require an audit that requires companies to measure whether false, negatives are being equally distributed across communities. If you go to the next slide, another blind spot is explainability. And this represents the idea that algorithms are a complex maze of decisions. And it's very difficult to understand why the decisions are made. Even to those who design them, the algorithms are often a complete black box. But the people who design these algorithms have a responsibility to be able to justify decisions that are made, particularly high stake decisions that affect individuals' wellbeing including their access to housing. If you go to the next slide. Applicants who are denied housing, currently don't have any way of knowing why they were denied except for knowing that their credit report was involved. So they have no mechanism to understand the reason and that's they can't really contest decisions that are made. And this is a problem, particularly because these algorithms are vulnerable to inaccurate data that can lead to false recommendations. So a solution here is to give people the right to understand why recommendations are made so that they can correct any inaccuracies and contest any decisions. So that's another solution that we're working, I give you two examples of working through all 11 blankspots and then looking at what they look like in the case of tenant screening and what solutions we can recommend. And we're doing this in collaboration with the National Fair Housing Alliance. If you go the next slide. Tenant-screening is just one example, but you can visit our website to, where you can access our workshop materials if you're interested in designing your own workshop for different areas or connecting with us to potentially do a case study on a different area. We purposely designed AI Blindspot to be very, to be able to adapt it to any type of AI system. And I'm quickly running out of time, but I wanna go to the last slide just to quickly acknowledge, And Also Too as our collaborators. Their design justice have been great collaborators in creating this newest version of Al. Also Kade Crockford, our advisor, who's given us excellent guidance through Assembly, as well as Vinhcent Le, Tawana Petty and Serena Oduro who participated in user testing as part of this latest version. Last slide is thank you.

- Wonderful, thank you so much, Dan. And thank you to the AI Blindspot team. It's really great to hear about your work with civil society organizations and broadly. So now I'm really pleased to introduce the Clean Insights Project which came together during the first year of Assembly in 2017, working on privacy and security. Their project's riddle is, what takes what you need but nothing more, to give the knowledge you want, without gathering a hoard? And Nathan Freitas leads the project, and I'll turn it over to him to share more.

- Excellent, the next slide please. And next, so we're going through... Oh, we had an agenda, sorry. Today on our agenda, we'll be talking through the problem of analytics. A lot of you might know, think of analytics in the context of the thing you'd put into your website, but in fact, sort of measurement of data and signals is everywhere. And then we'll talk about how we've tackled a more private solution. What we're doing, working with developers and designers today around consent and user experience, and then how we've moved the project forward through Assembly Next. So at Assembly 2017, I came into this cohort as a independent open source privacy preserving kind of focused developer. It's the work I do with my team at Guardian Project. And on my team where people from Apple and Google and Square, and people that love data. But they listened to my struggle and the things that I was uniquely challenged with. Next. So at Guardian Project, we have been working for over 10 years on building privacy, preserving software. We work with projects like Tor and Signal and building encrypted databases and messaging apps in human rights and humanitarian context. But we don't add analytics. We don't measure our apps because there was nothing that we trusted. And this is a big struggle that we had at the time and still do today. Next. You might've heard recently in the news about Audacity or if any of you are bloggers or podcasters use this software. Well, they were acquired and they're kind of in the same boat that we are at Guardian Project. They have this great open source community and support but they don't know how many users they have and what features need to be improved or where people are struggling. So their new owner after they got acquired, just dumped in Google Analytics and Yandex Metrica and proceeded to piss everyone off Next. You might've also heard in the news about Apple and Facebook and others fighting around the new moves that Apple has made around locking down app tracking. And really what we've seen there is that most analytics and kind of things around usability are tied up with advertising identifiers. And everything's kind of a big mess. If you actually just wanna responsibly learn how to help your users and improve their experience, it's very hard to find out the right toolkit to do that. Next. This expands all the way to smart cities and contact tracing and health systems and traffic and everything in our lives that's been connected and instrumented. So this is really an existential problem in our lives where everything is constantly monitoring us. And the way that these things are being implemented is often very privacy invasive. Next. On top of all this is that the consent experience for the user is confusing and hard to understand and is full of bias and blindspots as you'll hear about today. And that consent and all of this data can often be weaponized and turned back on to users in unexpected ways. So a big issue there. Next. So, next, please. So we have from the 2017 cohort, I've been shepherding this project along and expanding and building a new team around this who's been working for the last year and it's been an amazing process. So we've built a team that has a wide variety of skills. Next. And we held a really fun symposium last year online but brought together a number of different interested developers and designers and data scientists to think through how to, what people should know about data, how they should use it, what developers need and started building from there. Next. We also did a number of interviews and reports with Open Source tool teams who need this code like we do across kind of internet freedom and human rights spaces and published a report on that. Next. Fundamentally, the idea of Clean Insights is that yes, data is amazing, but if you just hold on to all of the data, it becomes a toxic asset, and there's a lot of liability there. What we wanna do is separate out the true knowledge and retain privacy.

And then that's the actual valuable thing that you're looking for. Next. So we have a number of ways we approach this. And we talk about this on our website in a number of places. But our focus really is just like, take what you need, de-identify the data, de-resolution, de-res it, get consent and really think about it differently as opposed to just sort of wire everything up, suck it all down and then figure it out later. That's not our approach. Next. So we have a way we have toolkits to allow developers of any sort of thing to plug into our tool kits for enabling this. Next. Again, a lot to think about here but we have kind of a different process than if you're a developer who uses tools today where you just sort of activate and forget it. There's actually some thinking and planning to do at the forefront but it leaves you much better off in the end. Next. So all of these tools are available today at cleaninsights.org and on gitlab. So we've actually shipped a lot of work in code and specs and we're really proud that this is starting to be adopted and implemented. Next. If you're a developer and know some JavaScript, this is kind of what it looks like. The interesting thing is that you define kind of these aggregation periods and the start and end dates and these windows of measurement, we have the idea of consent built in and how you can get it from the user much like you might request a permission for use the camera on a phone. And then we have the idea of measuring events and measuring views. And we have ways that those get aggregated and averaged on the device itself before it ever leaves and goes to a server. Next. So videos and more and things that you can learn about on our website, including me and others talking about how to tackle this and a great new blog post on our consent UX, which I'll talk about in a second. Next. So consent, we had to rethink this too. And as you're going back to the Audacity example, the other thing they really failed on it, is they just threw up a big dialogue box with a huge paragraph of text about all the things they're going to measure forever. And they thought that was good enough. And it's not good enough. So we really have some new thinking around how to engage your community, talk about the benefits for them, show them the value of what's being measured and what it's going to look like and allow them to time limit it and participate in different ways, in different times. And so really thinking about again, how you collaborate and co-design and co-measure with your community is important. Next. And I'm so excited we have one of the coolest open source decentralized app stores, probably the only one. F-Droid is now implementing this. And here's a super privacy centric app store that's trying to rival the lockdown systems of Google and Apple. And so they didn't again have any sort of tracking or analytics and measurement. We worked with them to find a way across decentralized app stores, a way to measure and understand where their users can be better served, and everyone's happy about it. Next. So in Assembly, we had some new outcomes. Next. We wanted to add more computational privacy support. So like new techniques beyond kind of the framework we have now. And so we're trying to figure out which one. Next. And we decided with the support of our mentor, Margo Seltzer, that we should add all of them really, that we should basically add all sorts of filters around measuring events and batch measurement filtering and then plug all sorts of things in it and work together with researchers to figure out exactly what is useful and then bring that to market to our developers. So we're really excited that we found a way to sort of not lock ourselves into one system and to find new ways to collaborate moving forward. Next. And adoption awareness is obviously important. Next. And there we also realized we needed a better back-end integration across different infrastructure. So that's really important that we're not locked into our own proprietary closed back-end. And so we're really excited that say, with a civic deployment of this, we could integrate it into something like Tableau or pull into our systems. Next. And lastly, we have a lot of value that we can offer out of the gate to developers, more events and symposiums happening in the future. And finally, next, we're so excited to have our own merch. I drink a lot of coffee to give myself insights and tea, and we're actually gonna have our own coffee and tea that you'll get for free, if you come and participate in any of our events or try to implement Clean Insights yourself. So we'd love to hear from you, stand by for that great benefit. So that's all, and yeah, you can reach me at lots of places and find us at cleaninsights.org. Thank you.

- Thanks, Nathan. I can't wait to get my bag of coffee. Thanks for a great presentation. It's great to hear about your progress and implementation of Clean Insights. For our third presentation, I'm thrilled to introduce the Cloak and Pixel project which definitely turned into a riddle as what hides you so you can be seen. And this team is comprised of three members Gretchen Green, Thom Miano, and Danny Pedraza. They originally came together in 2018, working on the ethics and governance of AI and evolved out of an earlier 2018 project called Equalize. And before I hand it over to them, just a reminder to share your questions, thoughts, comments through the Q and A tool so Jonathan can pose them to the teams after their presentations. And with that over to you team Cloak and Pixel, over to you Tom.

- Thanks, Hilary. We're Cloak and Pixel, brand new name, brand new identity and very loud yellow slides. Next one, please. In 2021, a subset of the Equalize team came back to build on the momentum that we had established and continued iterating on the prototype that we had launched in 2018 in adversarial attack on facial recognition. During the course of the semester, we began to consider additional aspects of the larger problem space of face detection and recognition models. Some other folks had also been working on the problem. And so we felt it was time to look at the problem space from a more broadly representative perspective. And so we wanted to develop additional interventions to advocate for responsible use of FDRs and highlight the risks while others are continuing to deploy the technology. And so defining risks in this space makes them more transparent so that ordinary people are aware of these issues and can make better decisions. That's really the goal of Cloak and Pixel. Next, then over to Thom.

- In our world today, everywhere there are more and more cameras collecting data. With that, there's growing automation and machine learning built on top of this data. This data can reveal highly accurate and detailed information about the individuals captured while simultaneously being fundamentally flawed and susceptible to failure. This is a problem because the level of surveillance going on is often opaque to ordinary people and they have very little control or say in it. Additionally, the information collected is often not in their interests. Finally, both the positive and negative applications impacts and impacts of this surveillance affects different communities disproportionately. Next, please. Well documented though, notably clandestine example of this is Clearview AI. Since 2017 Clearview has amassed over 3 billion photos of individuals by scraping social media sites without the consent of those sites or the users who posted those photos. Among multiple issues with this is that U.S. law enforcement agencies are using Clearview's tools without consistent guidelines, regulation or policy defining appropriate use and without clearly defined quality assurance practices. Next. And so as Thom alluded, this technology affects people disproportionately, and I wanted to highlight an example here. Robert Julian-Borchak Williams, his case may be the first known account of an American being wrongfully arrested based on flawed match from face recognition. This was published by the New York Times. And so it may have been a surprise to Mr.

Williams that he was being arrested since he did not commit the crime, but it is no surprise as we have witnessed this continued debate in the systems used to surveil communities and to identify people for prosecution. Face facial recognition has been used by police forces for more than two decades. And recent studies by MIT, NISD and others have found that this technology while works relatively well on a subset of the population, does not work well on demographics that haven't been included in the datasets. Notably, people of color and females. And so in part because of this lack of diversity, the images used to develop these databases are lacking signal. And that really was the genesis for our project in 2018, Equalize. Next slide, please. We took Equalize established in 2018, the year Thom Gretchen and I met and pushed forward with an additional group of folks, a larger group, was built as a prototype to fight pervasive surveillance. At the time, there was no published work on adversarial attacks to print from facial recognition. And so we wanted to give something to the public. We built a functioning prototype to prep face detection and to really invigorate the missing discussion around the power imbalance that the social platforms have over their users. The tool was successful in reducing the detection and some of the largest vision APIs available. The example shown here on the slide is Equalize confounding Google Cloud reducing the detection rate quite significantly. What also came about was that we wanted to highlight the levels of consent that users could express. And so really wanted to look at what was possible in terms of the control of the data, the discussion and debate, what was desirable and who should control it. Next slide, please. And although we took a technological approach, the Cloak and Pixel iteration now, it's found that the solution isn't just technology. As Heather Burns from the Open Rights Group often says in presentations, if you add digital on top of a thing that is broken, you will have a broken digital thing. The deploying and implementing technology of face detection and recognition has gotten a whole lot easier and will only continue to do so. However, the understanding of its social impact once this technology is available in the world has gotten much harder. And so to that, next slide, and I hand over to Gretchen.

- This semester, we thought not only about the harms of face detection and face recognition but the potential harms of over-reliance on tools like the one that we helped pioneer. So if you start with the harm of face recognition and then imagine a user using an adversarial tech tool to protect themselves, what happens when that tool fails? We looked at example where the face recognition was working, not working like the example Danny brought up and a federally protected witness was found after using an adversarial tag tool that failed and posting their image on social media. Working through that example, you can see that really was the over-reliance that was a problem. And we looked at three areas of the law to see how does society expect toolmakers to interact with and communicate with users to design and manufacture tools like Chainsaws. So if you look at product liability, there's a lot about failure to warn consumer protection. It's about unfair deceptive acts or practices in commerce. Negligence is part of common law about the protection of others against unreasonable risk of harm. All of these come together and say, tell your users about the risk of failure and tell them especially about the kinds of failures that they could not predict. Adversarial attack tools should be expected to fail tomorrow or the next day, even if they work on every single system today, because those systems are constantly changing, they're getting more data, the whole architecture might change. So we presented a poster paper at ICLR, a machine learning conference. And then, next slide. Thinking about how can we position the work that we did this semester? How do we position the work that we did with our teammates in 2018 in a larger area? And there's three main themes that I see. One is thinking about the societal impact of cheap easy access to face detection recognition models which even just in the last few years, barriers to use have plummeted the levers of control and points of intervention. So public communication, user choice, and tools, platform companies as partners for user protection, developing adversarial attack and other technical research and engineering and taking inspiration from and influencing the law. And then the third is it's not just face detection and face recognition, there's the whole question of control of information, data and how it can be analyzed and how it is analyzed and who should have that control. And then back up to number one. And two is sort of what is the impact and what are the levers of control. Next slide, thank you very much

- Wonderful, thank you Cloak and Pixel team. Excited to see your work thrive and evolve. For our fourth presentation, I'm so happy to introduce team Data Nutrition Project which came together during Assembly 2018 working on the ethics and governance of AI. The team has expanded since then and now includes Kasia Chmielinski, Sarah Newman, Matt Taylor, Josh Joseph, Kemi Thomas, and Jess Yurkofsky. And they're going to speak to the riddle. What is something you can't eat that can still be nutritious? And before they go on, I just, a reminder that you can share your questions through the Q and A tool and we will post them to the teams after this round of presentations. And I will hand it over to the team to kick things off.

- Hey, folks. Can you hear me okay?

- Yes.

- Awesome. Hi, my name is Kasia Chmielinski. I'm the lead for the Data Nutrition Project, which we also called DNP. We're super excited to walk you through what we've been working on. Next slide, please. So we're a team that is actually constantly changing and growing. I just wanna make sure that even though not everyone is speaking today, that you get to see the faces of our lovely team as I will be representing their work along with a few others. Our mission as a team is to empower our data scientists and policymakers with practical tools that improve AI outcomes. And we do this by building things and also talking to folks and we try to be as inclusive and equitable as we can. We really try to walk the talk because we think that's really important. Next slide, please. So today, just to tell you a bit about what we're going to talk about, I'm going to give an overview of our approach and also our impact, and then gonna hand it over to Matt who will walk through some of the improvements we've made on the tool that we call a Label, and then Newman we'll close this out with some of the outreach work that we've been doing and a few exciting announcements. Next slide, please. So the problem that we are trying to address here, actually I love the way that Cloak and Pixel put it. They said, if you add digital on top of a thing that is broken, you will have a broken digital thing with like a giant like smiley face. That is basically what the slide is saying. You know, the problem that we're trying to address here is that artificial intelligence systems that are built on bad data will have bad outcomes. And what I mean by bad is that if you have data that is historically biased or has some issues in completion issues, composition issues, then the model that you build on top of that data is going to recommend, have recommendations that actually exhibit the same issues. So one example, I'll just pull out from the side. These are really recent media examples. There's a lot out there on the right-hand side. Amazon created a hiring tool. It used historical hiring data and historically, women and people of color were not hired as frequently as men, regardless of whether or not they are qualified. And so when Amazon created this tool on that data, it immediately started to say the same things as was in the historical data which means that it was discriminating against various groups, including women. And the other examples on this slide are telling some of the same kinds of stories. So these have real harms and a lot of these models are already out there making decisions. And that's kind of when the problems are identified. So if you go to the next slide, the opportunity that we saw as a team when we were first in Assembly, which is a few years ago now was actually to try to identify the bias before the model is even trained. So currently, like I said, you go all the way through to the end, you deploy the model and only then do you notice that there's something wrong with it. That's problematic for two major reasons, right? The first is that there might already be people who have been harmed. And the second is that it's really expensive to go back and try to retrain that model. And so we said, well, hey, maybe there's something that we can do at the point at which someone grabs a dataset to interrogate the quality of that dataset. If you go to the next slide. The analogy that we ended up using was a nutritional label for datasets. And the same way that I can kind of pick up a can of Coca-Cola and I can look at what's in it and see if it's healthy for me to consume, we wanted to do the same thing for models and data sets. So before a data scientist decides to use a dataset, can they actually look at what's inside the data set through a nutritional label and decide whether it's healthy for their model? And so in 2018, we launched a prototype and a paper. And in 2021, while we were in this fellowship, actually, we ended up launching the newest version of the Label, which focuses a lot more on the use cases. So the intended use of the dataset and less on kind of a generic solution for everything. So in this example here, on the right hand side, this is a melanoma classification challenge dataset. This is a dataset that we worked on with a partner out of Memorial Sloan Kettering. And we said, basically, what are you trying to use this data set to do? And, the most highly, the use case that's mostly used for this dataset is to identify melanoma from images. And so we pulled that use case right out and said, is this dataset about humans? Is this dataset, has it undergone ethical review, quality review? And what are the kinds of harms that could come out of this? What are known issues and mitigation strategies. So that's the most recent version of the Label. If you go to the next slide. Just wanna point out that we've been seeing the impact of our approach, methodology and standard. We're really excited about this. I won't say too much about this but basically we've been having a lot of conversations with large tech giants. A few of them are here but also other organizations, nonprofits, governments in some cases. Our methodology has been used by RAI, which is the Responsible AI Institute in their Responsible AI certification. So we're kind of part of the data standard. And we're also as a standard becoming cited in places like NeurIPS which is a conference where they say, if you're going to submit a paper and it is based on a data set, then you should think about documenting that data set with a data set nutrition label. So really excited to see how our work is getting out there. I'm gonna hand it over at this point to Matt who's gonna walk through the work that we've done during Assembly on the Label itself.

- Thanks, Kasia. So, yeah, so during Assembly, and we've mostly focused our Label improvement work in two domains. One is getting feedback on the current version of the Label and another is work trying to build an ingestion engine. So thing kind of like TurboTax, but for getting information about data sets, to try and automate some parts of the Label creation process, not the whole thing, but just some parts. So when it comes to Label feedback workshops, basically we've been wanting, we did a lot of changes like

Kasia mentioned from the 2018 version of the Label. And we wanted to get feedback on the qualitative direction that the Label was going. So thanks to Assembly we were able to get a lot of focus groups set up, and we're still sifting through a lot of the feedback but some big takeaways we have so far are that, it's really important to prioritize information that helps people build trust when they're trying to find out about a data set and whether to you use it. Also while the qualitative information is super helpful to bring in to the conversation, having some views into the data associated with that qualitative information could really be helpful. On the ingestion engine side, we've been prototyping what questions to ask to help surface some information to add to the label as a starting point since most of our work building labels right now has been working directly with dataset owners and dataset creators. And we've also been doing some research into how to build the technical infrastructure to make an ingestion engine. And I'm gonna pass it over to Kemi to briefly talk about the technical infrastructure research that she's been doing.

- Thank you, Matt. During my research, I wanted to make sure that we use the best tools for the ingestion engine and ensure that it will provide the best options such as maintainability, scalability, and hierarchical storage data for the architecture we wanna build. And I also wanted to alleviate a lot of the unnecessary constraints that some tools spring marks and other technologies contribute to the way we want to develop our code. So through the influence of AI Global's form and their architecture and personal research, I was able to conclude and choose a mixture of forward-facing and eminent technology. And the next steps now is to just use these technologies together, to build our ideal architecture. Now, we can go to Newman to talk about the exciting progress in the children's book and podcast.

- Great, thanks Kemi. And thanks everybody on the DNP team. Not all of us are represented here, but just also a shout out to Jess and Josh who are equally part of the team. And what, the last thing we're gonna share today is our educational work. So alongside the Label work we're doing and the research and the feedback workshops, we've been all along really invested in public education about these issues. So we've hosted AI demystification workshops. We're thinking about public messaging, not dissimilar to the way that in addition to having food nutrition labels on food, you also want to educate the public about the consumption of food and the risks associated with the consumption of certain kinds of foods. So as we have a number of creatives in our group, we've always had this kind of creative bent to our work as well. And so we have a podcast in the works and there's more information coming on that, I'm not gonna share too much today but you will be hearing about that soon. And you can get announcements if you sign up on our website for email updates which are extremely infrequent there about once a year. And then the other thing we're working on is a children's book. It's called, "I'm Not a Tomato." And these are some sketches, they're not final at all from the book in the works. And it's a story about a red round thing that rolls down a magical mountain and about the adventures and exchanges that it has along the way. The stories inspired by how both humans and machines learn and really underscores the importance of diversity in training data. So we're gonna be announcing the launch of that book soon. I would say this summer, early summer, and if you would like to get a notification when that is live and announced, there's a form here for you to fill out and we'll just send you an email blast when that is up. Lastly, and probably most importantly, we want to thank our advisor, Mary Gray, for her support all along the way as well as had other advisors that have been helpful to us including JZ and James Mickens and then huge thanks and props to the Assembly team primarily Hilary and Zenzele for all the terrific work they've done to make this possible in a pandemic. Thank you.

- Thank you team. Really exciting to see your work and excited to read more about the children's book and share with the children in my life. So we have reached our fifth and final presentation from team Disinfodex which came together last year during Assembly 2020, working on disinformation. This team came together during the pandemic and has continued working throughout the pandemic. So just a big shout out to them for building something in the midst of a difficult situation. And the team is Clem Wolf, Rhona Tarrant, Jenny Fan, Neal Ungerleider, Ashley Tolbert, and Gulsin Harman. And they are going to answer their riddle, where do you go to answer what you don't want to know? And before I hand it over to them, just a quick note that after this we'll shift to Q and A for all the project teams in the order that teams presented. So last, not last chance, but chance to share questions before we jump into Q and A, and then you can keep sharing questions as we go. But get your questions in now. So thank you so much and over to you team Disinfodex.

- Thanks so much, Hilary and thanks for the shout out about organizing a project in the middle of a pandemic. It was certainly interesting. As Hilary mentioned, we came together last year and we have a lot of different backgrounds. So we're from journalism, communications, technology and policy and design. And although it's just me presenting, say it really has been a massive team effort since last February. So I just wanna share it to that. And so the question that we came together to address as a group was, how do we help people working in the disinformation space to better access and analyze all of the publicly available information about disinformation so that they're fully informed and that they can best address the problem? We'll go to the next slide. And this is what we've created. It's called a Disinfodex. It's a database of publicly available information about disinformation campaigns. It currently includes disclosures that are issued by major online platforms, including Facebook Instagram, Twitter, YouTube, Google, and Reddit. And it also has third-party reports that come with those which are from Graphika, DFRLab and Stanford Internet Observatory. And it's really designed for anyone in the disinformation space, people research and influence operations tracking disinformation networks across platforms and generally exploring broader trends. So just to go back really quickly, I just wanna talk about why we created the database 'cause I think it's kind of important. So we initially, as I said, set out to help journalists better cover this information And there were these two major problems that we saw. The first is it's a really new space in journalism. So you're talking about newsrooms pivoting to cover a space that's rapidly evolving, involves new skills, new verification methods, new vocabulary. And the second was that there's a lot of information out there for journalists to navigate. So you've got academic papers, media reports, platform releases, civil society groups. And the information we found was quite disorganized. It's hard to know where to go, where to start. So we decided to focus on that aspect of the problem, putting everything into one place and sort of navigate the disorder about information disorder. So just really quickly, what do we mean by disinformation? It's how we would define it as disinformation is a kind of false hoods that's fabricated or distribution deliberately, presumably with the purpose to do harm. And I suppose in real life, things are a little bit messier. No one has a truth detector, no one can map out all the things that should or should not be in a comprehensive database about disinformation. And as well, as I said, there's so many sources. So we realized we needed to find a place to start and a smaller place to start. And we decided to start then with the actions taken by platforms and those working with them. So this obviously is not the whole story but we feel it's a reliable data set as in the platforms may not tell the whole story but whatever they do publish is vetted by lawyers and it's reliable. And even then though, having found this data set we found that it was actually a little bit messy and hard to navigate. So here's what you need to know about platform disclosures in one minute. So Facebook, Google, YouTube, Twitter, they all release information periodically about what actions they take about disinformation campaigns or influence operations or information operations, they may call them. And these are typically actions based on behavioral signs, they'd noticed. So they might call it inauthentic behavior deceptive practices. The definitions vary across the platforms. And so do the frequency of communication. So some are more frequent than others in their disclosures. And also the formats and events vastly differ across platforms. So things don't emerge in the same way. And then sometimes these operations are attributed to a country or to a government or a specific entity. And sometimes there are very clear targets. This targeted, the U.S. for example, and then sometimes not, sometimes that information is not established. And so that's the platform disclosures. So then beyond that, there are a small group of open-source investigators. And currently we include Graphika, DFRLab and Stanford that have access to the information before these are released. And what they do is they write up their own reports where they might give you more context and show what they were able to establish in addition to the platform. So that's what we currently index. And our first challenge was mapping out all of the data in a way that made sense for readers, make sense of all that to quickly navigate and compare it. So just to kind of dig into the weeds a bit more on this. So there's also kind of an understanding the relationship between the different disclosures and knowing where to find them. So this is one example. There's a Twitter disclosure in April 2nd, 2020. It tells you about a specific network of accounts that was removed by Twitter violating their platform manipulation policy which you may be interested in as someone who tracks this field or wants to understand the strategies behind these influence operations. What you may not know just by looking at it is that it's related to a campaign that Facebook removed and at about the same time related to an Egyptian PR firm. Here by related, we mean that there's at least a strong suspicion that we're talking about the same network, and you may not also be aware that it's related to a prior action taken by Twitter by which we mean it's likely that this prior action that was taken in December, 2019 was from the same network or the same entity. And here's one more action from Twitter as well that may as well be related. This time dating back to September, 2019. So how does this all tie together? Well, one piece of good news is that there is another report that hopefully does this. And that's from Stanford also from April, 2020, that outlines how all of these actions are related. So the database does, is it sort of brings us all together so you can immediately make those connections between them. And I suppose just more quickly analyze the information. Next slide. So here are our goals. We want to find reliable public information. We want to help create connections between different operations. We wanted to make it easy to merge this with other data sets, if you're a researcher and you want to do that. And also just fundamentally answer some basic common questions. For example, how many operations involve country X or show me all the information about operation Y. So before we started the fellowship this year, we had a very basic website online. And we quickly realized that it needed some major upgrades and this is what we've been focused on for the last few months. So there are two views that you're seeing here, the Card View and the Table View, which are now available at disinfodex.org, if you wanna go on to the next slide. And each entry in the database opens up to a detailed network cards which has all the general information such as which platform was affected, which countries were involved as well as all of the related disclosures produced by social media platforms and the third party investigators. And you can also copy links to specific networks which you could not do before this fellowship. And also we worked a lot on and chatted to Jenny for her fantastic work on these on filtering. So by country of origin, you can filter by name of company, by date of disclosure, by entity or individuals involved or by policy violation as well. And if you just wanna go into the next slide. Yeah, so our latest database improvements really helped clean up previous data and allow for filtering of named entities, target origin countries, policy violations which were previously quite difficult to navigate. And so we hope this paves the way for the work visualizing the data, that's kind of our next goal. And also just sort of making the database and website a little bit more accessible, We'll move on to the next slide there. So our next steps, we're updating the website, we're doing some user testing, some more user testing. We're also hoping to reach out to the community, so more journalists who work in this space, researchers, people who work in policy, anyone sort of relates to that. So if that's you, or if it's someone you know, please send them our way or visit the database and let us know what you think. And we're also working on, a few researchers are working with the data at the moment. So we're gonna hopefully showcase that. And we're going to do some more outreach to the platforms and open source investigators to say how we can better collaborate. And then finally something that we've been talking about for awhile, if we were to add new sources what would the new sources be? And we don't have an answer to that question yet. Just wanna run to the next slide there. Yeah, so that's it. To end, we would love to hear from users, as I said, we're still ironing out the kinks on the website. But that being said, if this is useful to your work or if you want to use it and get in touch or just get in touch, please just send us an email. I also want to mention that we're an official partner of Carnegie who've been incredibly supportive and insightful and they've really helped us to get where we are and who we're gonna continue collaborating with. And then also just a massive thanks to Berkman again for inviting us back. It has truly helped us get to a place that we just wouldn't have been otherwise. And thanks to James Mickens as well for his incredible guidance. And yeah, thank you.

- Wonderful. Thank you so much, Rhona. Thank you to all the teams and I'm going to pass back to Jonathan to moderate our Q and A.

- Thank you, Hilary. It might well be apocryphal in which case it should probably be in Disinfodex, but Mark Twain was is reputed to have said that everybody's always talking about the weather but nobody ever does anything about it. And I'm just so struck with each of these incredibly tight, incredibly detailed, and I basically say this as a compliment, even though these days, maybe it isn't. Erudite presentations that kind of grapple with very real problems on down in the weeds rather than just with grand pronouncements which of course also have their place, just so exciting to see what you've been cooking up collectively and contemplate what might be able to happen next. So just a huge thanks for the seriousness, the detail orientation and the commitment to building a world that we don't have yet. It's funny that, 25 years into the build out of this digital space, there's still so much up for grabs and both in a positive way and more often portrayed and understood in a negative or invasive way, and what to do about it. So we have a little bit of time to just visit with each of the project groups again with a question. And for our friends from Blindspot AI, I guess this is kind of a two-parter, which is first, do you aspire to kind of more of a product or a service? And by that, I mean, is it you'll drop off these kinds of really helpful, tangible ways that somebody who is about to implement or build an ML system and expose people and their data to it, can actually stop for a moment and have a very cognizant way of kind of walking through it, or is it more like a good housekeeping seal or certification or some kind of thing where a company can say to the world, hey, this independent group, we've used their tool and they've given some form of blessing, however powerful or dangerous that might be? That would be the kind of service piece of it. I'm curious what you're thinking about on that. And the second part only loosely related more of the educational function is from Amelie Sophie Yurkofsky. And she was curious about Blindspot's workshops with civil society organizations. And what kind of grounds you cover? Do you get into tech policy questions such as how to regulate AI more generally and how you balance technical complexity with accessibility in your workshops? I don't know, Dan, if you wanna start with that, yeah.

- Yeah, I guess I can maybe start with the first part. I think certainly when we were getting started, we had ideas about sort of some sort of certification. Yeah, potentially a product we would create that would certify companies saying that they had passed Ai Blindspot assessment. But I think in the course of working on over the last two years, we really found that it resonated more with civil society organizations that were in a position like I said of wanting to advocate for changes but really needing some specific recommendations that they could be advocating for. And I think that that was really a greater need right now. So I think that that's, we've sort of pivoted in that direction sort of away from, maybe certifying companies whom it just wasn't quite resonating with to the same degree, who may need something different, something more like a product to certify them into providing more of a service for the civil society organizations. But it could ultimately turn it into a product benefiting that audience as well. I think developing a product of AI Blindspot is I think maybe a stage we haven't gotten to yet, but potentially when we're invited back, to do Assembly a third time, maybe we could do that.

- What does it take to move a car off the slop today? But that actually perfectly teams up Amelie Sophie's question about, all right, what would I expect to find in a workshop like this?

- I don't know if Hong is able to jump on because he's actually leading the workshop that is being done at Stanford in two days. So he's more directly involved with it. Is he able to join?

- I can speak to that. So on Thursday we're running a workshop that's kind of a beta version of this one called product in the sense that the workshop has a script, has a presentation, has a participate, worksheet for participants to fill out and we're hosting about 30 civil society folks. And having them basically go through this workshop. And if it does work out well, our plan is to, you know it's already open source but our plan is to train the trainers to have as many people be able to run this workshop on their own for other folks, so that we can kind of get out of the way and have a community take the ownership and evolve this workshops. So that's the idea of spreading it to as many people as possible. And regarding the complexity versus kind of the accuracy of the content, I believe the blind spots are pretty technically accurate and robust 'cause we have Jeff and Dan, data scientists who help come up with these blind spots and explain them. So we believe it's very clear but also very, based on the technology and the science. We're happy to get any help I think, if anyone can help us.

- Great, thank you both. And I don't know if there's a way to once again kind of blast into the chat room if somebody wants to reach out to you all, it might've been briefly in the slides. Certainly feel free to do that, but thank you both and thanks to the whole team for this effort. Onto Clean Insights, for which so many interesting things at once, I love how like Nathan, you just like dropped as a parenthetical, the F-Droid store. It's like, yes, it's a decentralized safer marketplace which people wouldn't normally take those two things together. And that was but a parenthetical in the larger presentation. Although one that jumped out to me given the current epic battle between Epic and Google on the antitrust implications of Google's App Store. And I actually bring it up now because that's a great example of when people talk about remedies for problems that exist or what it would look like to build some guard rails when amazingly there are none, and so many of these areas, something like that store is an answer to the question of, well, is there any other way to safely have third parties offer apps for your phone? And you're producing that answer. And similarly for Clean Insights, you're producing a kind of set of tools meant to make it that much easier to undertake the complex work when you're not really getting paid to do it as a company of worrying about your subscribers or your consumers privacy even as you're trying to make the most that you can, that's the insights part of the data that you're collecting. So my question here is two-fold. First, do you think that there could be once you have demonstrated the utility of these tools, I don't know if they include differential privacy. That was a question that Marta Sharma had, to like briefly speak to that. But by demonstrating these tools, do you think that it offers a roadmap to regulators to then insist upon the use of something like that since they exist and it shows that you can or is that kind of too specific? I don't know if the Europeans after many years really regulated cookies. Cookies, specifically, in browsers. And was that kind of a wise move to get that much in the weeds or is that way too specific for a regulatory perspective? And I don't know, of course the team may have different views on that, but I'm curious if you have thoughts on that. And second is that yes, no screen that Apple in its fight with Facebook and others has now offered up in iOS to say, hey, do you really wanna be sharing your info? Would you like to request not to be tracked? Is that screen something that you think Apple should be revising? That if a company were willing to do the sorts of limitations that Clean Insights would have them place on how they collect and use the data, should there be some reward in a differently configured screen from the likes of iOS so that you wouldn't have as many people opting out? And apparently the statistics say a ton of people are maybe unsurprisingly opting out when presented with that screen. So that's what I was curious what you all thought.

- Yeah, the vision we have around the kind of consent engagement process is definitely not this binary one time opt in. I think that's gonna fail. And the numbers are proving that it's failing. And in fact, what that--

- Will succeed depending on your view, hoping people would do it.

- It's wildly succeeding in the right kind of way. I think what's confusing about that, is someone could actually implement Clean Insights today in an iOS app and not be prompted with that because that's specifically around an advertising tracking identifier usage. So Facebook is being a bit disingenuous in that, on that front, but we, it also is going back to F-Droid. F-Droid, actually has pioneered the idea of tracking code being an anti-feature. If you go into the F-Droid for an app, it'll say these are the anti features, anti-features that this app contains and they've worked with Exodus Privacy tracking.

- Like cigarettes saying how much tar they have in them?

- Exactly, they just write it in your face. And so the fact that that also we've said, well, now that you're implementing Clean Insights, how do we talk about this if you're, you've kind of, we've crossed the threshold for you? So I think there's definitely a lot to think about there with that in rewarding, kind of a more nuanced approach and on the front of regulation as well as differential privacy. Today, a lot of what we see with differential privacy are, is implemented, first, you still gather all the data and then you put the data in a private, super secure database and then you implement data differential privacy at the analysis stage when someone wants to query the data. And there are still cases where maybe, it's done in the right way that's got epidemiological studies and that's how you have to implement it. But there's so many cases where know that, the people aren't trained to handle that, there's so much liability, there isn't, as we've seen time and time again these databases get infiltrated and exfiltrated. So I think we are hoping to show that, you can still achieve the outcomes you want and needing to understand if your product is succeeding with a much more constrained way of measuring. And we just need to make it as easy as Google Analytics. So that's been our approach. So, I mean, I think regulation, yeah, should be part of it, though I think there's fore-trained professionals that have a reason to gather all the data. I know the open differential privacy project at Harvard for instance, is operating at that stage and they have a really solid approach. So it's gotta be granular and in the weeds, that's where we're at.

- And finally, just working off of a question Bea Cabello just put in, which was like, this is so cool. How do we help raise awareness, et cetera, certainly curious your answer to Bea's question along with it, what's the receptivity been so far? Are you seeing as a friend or as a buzz stumper by companies that are working with a lot of data?

- I think most companies don't really want or need all the, like everything that something like Google Analytics provides. They just want a few things. There's a few things they need to understand. So they get, we present, we talk to them, they're impressed with like how broad we thought about it from the way we aggregate and average and consent stage. And then they're like, cool, this sounds great. It's a little bit like HTTPS or TLS in the early days. Yes, of course, but it's too hard on our servers. We don't have the budget to XYZ, and then it just happens. And so I think we're not seen as a foe, I think in some ways we alleviate this anxiety, the liability they're worried about and they can reduce the, what is seen as increasingly as surveillance of Google and Yandex and others. So we're definitely like, oh yeah, that's awesome. It totally makes sense. And then when they look at their budget, they're still struggling a little. So we're trying to do as much as we can especially in the civic and humanitarian tech pro bono. And I think it'll just take more time to get folks to adopt us like a good old HTTPS and end-to-end encryption. So we're on that train and hopefully we'll come right along with it.

- Wonderful, thanks so much. If there's more, you're thinking people might do to amplify or assist, please put it into the big chat room and--

- Look out for the coffee, so we'll--

- They can get a tweet storm going and coffee.

- Drink our coffee. It's coming.

- Terrific, great. Well, onward to Cloak and Pixel. And that team, and it's funny, part of my reaction to this extraordinary project and approach is a point that Nathan was just ending on, which was there was a time when the web connections were all un-encrypted because it was just seen as too onerous for servers to pull off encrypting every connection, even Google at one time said, it would just be too hard to do that. And regular Google search when you interacted with google.com for web search was not encrypted. And then at some point, thanks to the good work of people. It wasn't just like the weather, everything did get encrypted. I think we're at like 90 some percent on a random sampling of web links which interestingly mooted the big legal debates in large part about deep packet inspection since there's a lot less you can inspect when the packets are encrypted. And that makes me think about, we're still living in a world that is like un-encrypted web search or un-encrypted website transfers with respect to releasing our photos online. That if at any point, any photo of you has a label attached to it with your name, then any future photo of you whether or not on the web at the time, just a brand new photo extracted from a surveillance camera or from anywhere else, like a picture of you walking down the street or attending a protest can now be associated with your identity. And that's of course the point you all were making with Clearview AI. And I guess then the question back to you would be, how automatic would you like this kind of invisible masking through adversarial attacks used in a productive way here? How automatic would you like it to be? Would you like a standard camera on an iPhone or an Android phone to automatically apply a little bit of this adversarial fuzzing so that wherever the photo goes next, it'll be a lot harder to match it up back to you through these new facial recognitions, ditto for when you post a photo on Facebook and Twitter just as it strips out location data that might normally be captured and embedded, should they be doing this kind of fuzzing? And is your answer affected by the point that Gretchen was making towards the end of the presentation? That this is kind of cat and mouse that broken shield paper is pointing out that it's, you can't really lean too hard on this technique because you never know if there's a way to crack it later. So curious how you all are thinking about a retail use of this versus a wholesale use.

- Yeah, well, I'll definitely let Danny and Gretchen throw in their thoughts here too but I think my first reaction would be to say, I would take it even a step back from what you were just proposing. I'd like it to be automatic in the sense that policy and the perspective of technology companies and anyone building products like this is user privacy first, where you have to, before you can even begin to collect that type of data, you have to get informed consent from the user. Right now, we sort of live in a world where it's, like I don't like the idea that we even have to use an adversarial attack to try and proactively prevent third parties from using machine learning or other processing on the data. I'd rather nothing be, none of that data be collected in that individuals have to provide that data to someone who wants it in the first place. However, that's a very, I think that is an extreme position. Not extreme in the sense that it's necessarily unreasonable but extreme the sense that it's very far from where we're at right now. So to get closer to where we're at, I do believe that I'd like to see it implemented at the platform level. So for example, with Facebook and Clearview as we had talked about harvesting their data, I'd like to see Facebook try and devise more clever ways of protecting that data perhaps by, I mean, there's information that you can only access by logging in and there's information that is only available if you manually make it public yourself. And that's how a lot of platforms have, I think handled that sort of thing. But I think by and large ordinary people don't really expect other people to come and just collect their information. And I think what people also don't expect is for that to be done en masse. And the impacts of when you begin to collect things en masse, how you can actually figure out things about people that even you may not have even told people where you were or who you were with, but by being able to create a social network which is what Facebook has done for third parties to be able to create a sort of network from information like Facebook and other sites online being able to draw this picture about individuals that they haven't offered themselves knowingly, that's where I take issue with it. So I would just, I would like to see more done on the platform level outside of something like a tool like Equalize or some of the other tools where the user has to manually do something themselves. Danny, or Gretchen, do you have anything you wanna say to that that I didn't?

- Yeah, so I think, absolutely, I would like to this sort of security measure much more automated, I think individual consent and the burden on individuals to opt into any kind of security measure that they want especially when they have to go further than like a really simple opt in. But I think the presumption should be that users don't by posting something on social media. For instance, they didn't mean by that that they wanted every possible use by everyone in the world to be fine. And we have a lot of precedents for that in intellectual property licensing. You don't have to give it all away and no one assumes that you gave it all away by allowing one use by allowing a limited license. The fact that adversarial attacks won't always work is no reason to say that they shouldn't be one of the layers of security that we use because you can say this about any cybersecurity issue. Like cyber security is a huge problem. Like we see major problems in the news, like in the last few days, we've seen some major problems in the news and it's exponentially exploding but we don't say. Well, just because whatever you're thinking about using now is probably gonna fail, if not for you, for somebody else, we don't say, well, don't try.

- Yeah, the existence of the locksmiths doesn't mean you shouldn't use locks or a mask might not be 100% percent effective but you still might wanna wear one.

- Right, yeah. Well, if it's gonna be like locking your door, you know, well, it'll make them attack the other guy instead, you know, that's worth something. So this, and, you know, we had some other methods that Thom implemented this semester, the steganography where there's an encoded that you can't see it, a human user won't see it, but it would be easy to uncode or a watermark. So there's various kinds of signals which are then engineering solutions, which if you combine them with either public opinion pressure or law, and then combine that with an automated, things will be done with these marks unless otherwise.

- Yes.

- And society doesn't want you to which is not necessarily the same. Process these things if they have these marks and they will have these marks unless someone specifically said they wanted to give you permission.

- Yes, now it's just genius to offer a technical proof of concept so that it gets out of a stale dichotomy that says, look either people are sharing their photos or they're withholding them. If they share them, this is just part of the cost of it. And this is a way of saying, no, you could share some but not all. And if somebody wants to really go to lengths to get around it, it's showing how much they're trying to override the desire of the person producing the photo. We should probably move on to the next project. Danny, I don't know if there's anything you wanted to just throw in at the very end or.

- No, just saying that the engagement and the conversation aspect of it is also like really an important part of all of our projects, I think, because like all of these issues are things that we should all as citizens be up caring about, sort of like, sort of civic duty, if you will. And that's the only point I was gonna make.

- No, and that nicely connects to a point that LaRay cast made in the Q and A about how to educate technologists and consumers about equity and privacy at scale and to do it as much by showing rather than just telling maybe as a part of that puzzle. And great to mention that function the projects are serving and really all of the projects. Terrific, well onward then to the Data Nutrition Project for which it also feels like such a reminder, back to Nathan's observation about the web pre-encryption that the idea that the state of best practices for even big companies using machine learning on data sets is not to have any labels of this stuff of enduring quality, including if they get the dataset from some public source or from somebody else, it's really like Christmas fruitcakes getting passed around without even an original tin indicating what's in them. And the idea that like that's the state of the art in 2021 is just a mind-blowing, really. So it certainly shows just how much of a role this project could play. And I guess this is kind of pivoting off of Joseph Ben Simon's question, wondering how hungry companies are for this kind of thing? Do they see it as a hassle? It's good enough for government and corporate work, it's really hard to label stuff. Don't, let's just keep going as we go or are they open to it? And should regulators be open to saying, if you're gonna be mucking about with big data sets like this and making recommendations or decisions for you or your customers based on them, maybe you need some basic labeling. So I'm curious, again, kind of similar question, how should this fit into the ecosystem?

- I can start and then I'd love to hear what others in the team think 'cause we've all had conversations with different folks. It's a great question. I think it's also changed over time. Our conversations, I think, were more aspirational at the beginning, and I feel like there is now a greater appetite for, call it self-regulation or just like internal standards, especially in large organizations where you have many people who are touching the same datasets and maybe the same models because people are quite frankly worried about risk and liability. And so one way to kind of--

- And we want them to be.

- I think so, I would like them to be. Yeah, as someone who's also a practitioner in an industry, I think that that's a good idea. And I think that having something like a Label is actually a nice tooling solution that doesn't require a huge regulation from outside to say, you have to do it in this particular way, but maybe allows folks to show that they're making, they're trying to provide something along those lines. And so I think that it's definitely a moving industry and a moving ecosystem but I'm definitely seeing at least from the conversations that we've been having increased appetite for something like this. And that's also why we are not so tied to the DNP label as a standard specifically, but also as a methodology or even just as a conversation to say, you might wanna track different things on your data sets internally and on your models internally, but really you should do something. And that's kind of, I mean, you're right. There's just like nothing. So even just saying you should do something, seems to be a good, you'll move it in the right direction but I'd love to hear what others on the team have heard as well 'cause we've all had different conversations.

- Anybody else wanna jump in?

- I agree with Kasia on that. I think that as long as there's an awareness and a demand from the consumer, it should be available for big tech companies 'cause their main priority is profit and if the demand is for what we do, which is the data labels, then I think that it would help a lot. But we also live in an era where we're kind of blind to the data and information. And a lot of it before going into this team, I wasn't even fully aware of all of the practices, unfair practices that were being implemented in these companies. And I think that a lot of the consumers aren't aware as well and it takes a lot of research and actually just being in that industry and following them to actually know that. So I guess the bigger question is like, how do we create that awareness so that the average consumer is also concerned?

- Yeah, it connects right back to Laurie's question around education. Either so they can somewhat self-protect or even more important, what you're suggesting is be aware enough to pressure the companies and those who regulate them to create some atmosphere of like, hey, maybe we should be doing some of the basic stuff that would respect the gravity of the kinds of technology that we're building. Newman, any last thoughts before we move on?

- I would just add that in, I think we have a, I mean there's some very well intentioned people at the big tech companies, of course. And they also, because of their business models have sort of mixed incentives, not the individuals, but the companies themselves have mixed incentives. So I think we have this multi-pronged approach in terms of raising public awareness and having public demand, also doing broader education, including to those who are working in the spaces.

- Terrific.

- And I'll just, oh, sorry!

- Yeah, no real fast, Matt.

- And I'll just add that companies also are in monoliths and pressure can be both external and public as well as internal 'cause a lot of the people we talk to are data scientists. So maybe, I don't know what say, like the executives of a company might want but a lot of the data scientists we talked to really do want this kind of stuff.

- Yes, thank you all very much. I've been informed that there's nobody barging into the Zoom room for the next scheduled use. So we'll be able to, so long as people have the time to take the five more minutes to make sure we check in with this info decks before we adjourn. So speaking of that, over to Disinfodex, and I think a similar curiosity about how much you have found, you're just trying to make the best hay you can out of what the companies already share in their disclosure reports versus trying to blandish them to say a little more, say a little bit more consistently and try to create a true public good cooperating rather than competing in this area, around what they're seeing and when and trying to do it in a way that the public can see, rather than just sort of whatever council of elders there might be that Twitter and Facebook get together to compare notes at the security level to be doing it in a way that everybody can look into. And I'm just curious your sense of the willingness and appetite of the companies to look at something like Disinfodex and say, where have you been all our lives? This is great, we're in, we'll do more.

- I can kick off there. So I suppose, yeah, that's definitely something we've been thinking about a lot. And as we said, it's a really evolving space. So the frequency with which seeds disclosures come out are very widely. And as I said, as well, we're also looking at contacting the platforms ourselves or working more directly with them. So for us really, it's been starting trying to build something on the fly in an area that's constantly evolving. And the platforms themselves are constantly evolving their own disclosing practices as well. To your point as well about the data and the fact that they all kind of different things, you know, might be similar but they all call it different things. So that was a big challenge and something that Jenny helped us out quite a lot with was trying to streamline the data in a way that we can compare it but in a way that didn't distort the underlying data. So I suppose just to answer that, it's definitely something I think the platforms are still working through and it's something that we're still working through as well. We're trying to be flexible in terms of our database and how we enter data. And hopefully at one stage, we'll be able to get to a point where the data can be submitted to us. That's a whole and aspiration for us, but right now it's currently very manual.

- Got it. Yeah, and so it's a really complicated problem. David O'Brien was remarking on how hard this sort of thing was five years ago. And I'm not sure if it's gotten easier or harder since then, probably a vectors in both directions, Jenny.

- Oh, sorry. Yeah, I wasn't just commenting on that, but I think especially the idea of like the concept of a network is something that hasn't been really familiar. Like even from last year, we were thinking about this from a disclosure report perspective. And this year, we really shifted gears to describe more of like a network of activity. And that's something that I think even the platforms are still kind of getting to, and it's something, it's one of our limitations from being able to extract broader ideas like cross-platform, for example. So I think the industry will have to kind of co-evolve with our product as it goes on.

- Got it. Clem, I see you have appeared, I don't know if you wanna say something real quick.

- Oh no, it was just [Crosstalk].

- Other than that, they have done a--

- I know, wonderful.

- Good job in this space, so can't I speak here..

- Thank you, thank you all. Well, I think many of us were struck by the Heather Burns observation, if you added digital to a thing that is broken, you have a digital broken thing. Of course it's faster and cheaper. And it kind of gets to an observation too that was just made, that it may be that, I think this is Newman who was observing this, that isn't just a matter of official corporate policies that we can try to get the companies to subscribe to inform or regulation that would require all the companies to meet a certain baseline level, but the role of kind of bottom up work and people within the companies who are building things that are digital and that may have started out broken and how much they are game to see as within the scope of what they're doing whether or not they've been asked to by their bosses to shine a path towards, hey, I didn't just make it digital, I kind of either fixed it or made the flaws more apparent that we're already already there as a way of respecting the uses to which this will be put and the people's lives that will be touched by it. And thinking about the ways in which there's becoming to be more awareness among tech workers of the power that they may have and the role in which that they might play is a way of seeing latent professionalization of computer science, data science, modeling that the lawyers and the doctors have too long thought that they may be only had a quarter on. And to see these kinds of bottom up solutions be offered and contemplated at all levels of those who might be able to make use of them is extremely gratifying and inspiring. So I just wanna thank you all, both this team, but more generally everybody from the cohort this year for not being tomatoes, for saying we're people, this is technology, that's affecting people and seeing that there are problems and being ready to work on them and shine a path forward for the rest of us. So thank you again for all the work that you've done this year. Thanks again to our staff and advisors for augmenting it. And I really expect that this will just be a waystation on the path towards improvement of our digital world and a model for how we might assess and improve our digital world alongside all of the other models we have to try to put some guardrails and even a direction change on some of this stuff. So thank you all again. Hilary, why don't I turn it over to you in case there's any logistics I have forgotten. But thank you all for a terrific session and for all the work that's gone behind it. Really amazing.

- Yeah, just echoing you, Jonathan. Thank you so much. Thanks everyone for joining us. This is recorded, so it will live on the Event Page on the Berkman Klein site. So if you want to share it with anyone or watch it again through the slides again, it will be there as a resource. Thank you so much and have a good evening or good afternoon, morning, wherever you are.