JAMIE HINE: Good morning. On behalf of my colleagues at the Federal Trade Commission, I am happy to welcome you to our 6th annual PrivacyCon. My name is Jamie Hine. I'm in the Division of-- an attorney in the Division of Privacy and Identity Protection.

And along with my co-presenter, Lerone Banks, we're happy to bring you PrivacyCon for 6th year. A few details before we get started. We're happy to have you join the webcast. Our agenda is available on the PrivacyCon page with links to all the presented research as well as the biographies of all the moderators and presenters today.

There's also a webcast link on ftc.gov page as well as the event page if you happen to get disconnected where you're joining us sometime in the afternoon. Following PrivacyCon, we will make all the presentations available online. Usually takes about 2 weeks, but we archive all of the presentations today. You'll be able to go back and watch them again.

We'll also provide a transcript that you'll be able to read along and see all of the great presentations again today. After what seems like a lifetime assumes, we all realize that technology happens. So we ask for your patience today.

We have a technology team that's here to address any issues. But if you have specific problems, feel free to email us at privacycon@ftc.gov, and we'll try and help you out as quickly as possible. We welcome questions for our audience.

PrivacyCon is a participatory event, and what that means is if you have any questions for any of our panelists today, send them to privacycon@ftc.gov. We'll have somebody watching that email address and we'll send your questions to the moderators on the various panels. In addition, we'll be live tweeting today's event. So be sure to use hashtag FTC and hashtag PrivacyCon21.

You can also ask questions through the live tweets, and we'll make sure that those also get passed on to our panelists. I want to thank all of our researchers and panelists for all the work that they've done today. We have 19 different presentations.

And while all the people that are presenting the research today are fantastic, we're so excited for all the work that they've done, they represent a lot of research teams. There are hundreds of people that work with the 19 presenters today to bring you all of the work that we have. Please go to our website. You'll see all the papers there. You can read the papers and eventually, we'll put up the presentations.

We hope that you engage with those people. Most of all the contact information is in all the papers we want you to engage with people. So if you're not able to do so today, please contact them in the future and create a dialogue.

So today's program wouldn't be possible without a very large number of people. I just want to say a couple of quick thank yous. All of the panelists today, I'd like to thank Devin Willis, Lerone Banks, Danielle Estrada, Miles Plant, Linda Kopp, Manmeet Dhindsa, and Christina Yeung. I want to thank the FTC event and media teams, I want to thank all the people that will be helping with the Livestream today, including our Office of Public Affairs.

We also appreciate help from people like June Chang, in the Division of Consumer Business Education, Brianna James, Julia, [? Gruenwald ?] Henderson, and there are so many other people. But there are two special thank yous that I want to make. The first is for Leah Hebron. She's a former paralegal and soon to be law student.

And Alex Iglesias. They're the two people that make everything happen today. I want to say a special thank you to both of them. And without further ado, it's my pleasure to introduce Commissioner Rebecca Kelly Slaughter.

REBECCA SLAUGHTER: Excuse me. Thank you, Jamie, and thank you so much to everybody who worked so hard to put on today's event. Good morning. I am Commissioner Rebecca Kelly Slaughter, and on behalf of my colleagues at the Federal Trade Commission and my fellow commissioners, I'm so pleased to welcome you to PrivacyCon 2021.

Thanks again to you all for attending virtually again this year. While circumstances have been and continue to be challenging for all of us, I'm so glad we have the tools to convene so many distinguished advocates, researchers, technologists, and academics from around the country and across the globe. Our second virtual conference is a good occasion to note just how much the pandemic has accelerated our reliance on digital services.

People turning to online platforms and marketplaces for everything from socializing to soap, the necessity of moving so much of our lives online has also highlighted challenges in the digital marketplace and the serious issues data driven business models pose to our privacy, autonomy, and society at large. You have a full menu ahead of you today. And I want to open the buffet with a little food for thought onto your topics.

First, I know we are here today for privacy. But I would like to challenge everyone to reject privacy as the animating framework for the important issues being discussed at today's conference and among thought leaders generally with respect to our data driven economy. Today's agenda addresses the algorithmic bias, issues around consent, misinformation during the pandemic, and special concerns related to kids and teens as well as more conventional privacy concepts.

The FTC has been working on all these fronts and has recently issued guidance on algorithmic bias and unfairness and work to reevaluate outdated and deceptive consent. Frameworks around dark patterns. These issues go way beyond privacy as it is traditionally conceived. The broad agenda this PrivacyCon reflects a growing understanding the data issues with which both the Commission and society at large are concerned have moved past the narrow framework of who has access to your personal data. This emerging understanding is why I prefer the term data abuses to the narrower language of privacy.

Words matter, and data abuses reflects the fact that rampant corporate data collection, sharing, and exploitation harms consumers, workers, and competition in ways that go well beyond more traditional or libertarian privacy concerns. We must examine a wide variety of data abuses including questions of racial bias, civil rights, and economic exclusion, considering practices that undermine personal autonomy and dignity and re-evaluating damaging and dangerous business models and market practices. In addition to examining these practices, we need to consider what to do about the problems we find in the markets.

And so the second challenge I would like to issue today is the following. Can we move away from outdated notice and consent models to govern questions surrounding personal data and instead turn our focus to the underlying business structures and incentives that are anchored in indiscriminate collection and application of personal data to fuel data driven business models such as behavioral advertising? It is this underlying incentive structure that has caused so many of the harms and privacy risks we're here to discuss today.

Rather than focusing on opt in versus opt out and whether privacy policies are clear enough, I believe we should be discussing the concept of data minimization, a principle that would ensure companies can collect only the information necessary to provide consumers with the service on offer and use the data they collect only to provide that service. That minimization could be coupled with further use, purpose, sharing, and security requirements to ensure that the information companies can permissively collect isn't then used to build tools or services that imperil people's civil rights, economic opportunities, or personal autonomy. Corporate self-dealing is also a serious problem in the data ecosystem. And as long as key digital markets are controlled by just a few large data hungry online platforms, both consumers and prospective entrants are at their mercy.

The commission has a shared concern about many of these practices, and I've heard the call from members of the public at our two open meetings for us to take decisive action against these abuses. This moment of renewed energy at the FTC offers a window of time to catalyze meaningful changes in the markets and ensure that the data economy actually works for people, not just the largest corporate players. And of course, unchecked data collection is not just a consumer protection issue. It is also a competition issue.

The enormous amounts of data incumbents have collected gives them a profound advantage when competing against new entrants or seeking to enter new product markets themselves. We absolutely must look at these issues holistically rather than myopically viewing them through the lens of either competition or consumer protection. I believe that the FTC has an obligation to use all the tools in its toolbox to address these issues.

Simply challenging the application of abusive data practices on a case by case basis isn't likely to bring the systemic change we need to see in the market. The FTC has benefited greatly from workshops and conferences like this one, and I hope participants and observers of today's conference help us chart a path forward so we can build a more fair and just future together. Thanks again to everyone from the FTC that made today's event possible, especially our agency speakers and moderators.

And to our attendees, I'm personally grateful for your work and hope to hear more from your at the commission. And it is now my honor and pleasure to introduce Erie Meyer, the FTC Chief Technologist for additional opening remarks.

ERIE MEYER: Thank you, Commissioner Slaughter, and to everyone who put on today's event, especially Jamie Hine and Lerone Banks and to our wonderful presenters. Thank you for today, and all the work that led to today. My name is Erie Meyer, and I'm the Federal Trade Commission's Chief Technologist and an advisor to FTC Chair, Lina Khan, and I'm so glad to be here.

I was living in Central Ohio during the height of the financial crisis. I watched families I've known my whole life lose their homes because of the government's failure to rein in devastating industry abuses. And it kind of broke my brain. I had seen myself as just a tech person before.

But now, I just didn't want to design multivariate tests to improve ad conversion rates. I wanted to make sure that even the big guys had to follow the law and to ensure my neighbors were treated like human beings. Before we kick off today's event, I want to share a few places where the market should expect some changes and how the FTC will approach its work when it comes to protecting the public from the misuse and abuse of data.

The Khan Commission approach to data is not through a narrow lens of consumer protection. Data abuses do not happen in a vacuum. They're fed by incentives, among them beating out competitors. So with that broader view, you can expect key changes in our work.

We're going to make sure that data abusers face consequences for their wrongdoing and provide real help for affected individuals. When a firm breaks the law or worse breaks the law over and over and over, regulators like the Federal Trade Commission need to design and impose remedies that actually fix things. And fixing things doesn't mean simply making a disclosure longer or a one-time fine bigger.

It means making sure the firm cannot and will not benefit from ill-gotten data, including against their competitors. It means making sure that the rest of the industry is deterred from engaging in similar wrongdoing. It might mean that we need to look at restructuring business incentives or even corporate structure. And it means making sure that the people who are targeted or hurt are able to understand what happened to them and to get help, actual help.

And what does this look like in practice? It looks like companies who break the law having to not just disgorge data and money, but algorithms that were juiced by ill-gotten data. Companies that sacrifice security in service of speed being subject to bans just like abusive debt collectors. People getting the dignity of specific and clear, but most importantly usable information about what happened to them and where they can connect with straight answers about what's next. It turns out that paperwork and a fine, no matter how large, does not seem to fix the fundamental problem.

Data abuse is not just an issue of privacy. It's a matter of civil rights and national security. People from communities whose rights and safety are constantly threatened can tell you this isn't just about someone knowing what you've looked up online. The US Department of Housing and Urban Development recently charged Facebook with violating the Fair Housing Act. The Department of Justice charged a Zoom executive alleging that his actions led to people being able to use Zoom's data to track down an intimidate family members of people who use the platform to discuss the Tiananmen Square massacre.

There's been a 2,920% increase in reports of identity theft via government benefits this year. So what this means, for example, is when a bad actor applies for something like unemployment benefits using personal information gleaned from a data breach from one of these firms. I sincerely hope none of you heard people you love cry this year because they lost their jobs in the pandemic and struggled to access benefits.

But I did, and I want to reaffirm that the recklessness of firms that think they can get away with not keeping their promises about protecting your data and hope you'll think it's a scandal rather than a system, is resulting in families-- real families not having enough money to buy food. A pandemic has only sharpened the view of what's happening to our country's resilience because of these data disasters. In addition, we're moving away from a legalistic approach to addressing data abuses and towards a more rigorous approach. This means we'll be approaching investigations with an interdisciplinary lens including privacy engineers and designers, financial analysts and product managers, and yes, technologists. But this won't happen overnight.

And [INAUDIBLE] already begun to assemble a team to make some of these shifts when it comes to fashioning more effective remedies for law breakers, understanding the full range of harms and sharpening our analytical approach. I'm pleased that Stephanie Nguyen has joined the agency as deputy CTO and Presidential Innovation fellows [INAUDIBLE] and Vivian Lee are working to drive many of these efforts. One of the pioneers of privacy engineering, Leah Kissner, once told me the way to tell whether or not a privacy fix would actually be meaningful was that if a lawyer could do it alone, it wasn't actually going to change anything.

Lee is right. If a company can come into compliance by papering over questionable conduct, it's not actually changing the facts on the ground. So to all the other people who might be tired of working on designing multivariate tests to improve ad conversion rates, come help us change the facts on the ground, we're hiring. In closing, I want to join my colleagues in thanking the incredible team that put on this event and to welcome you all here today to our discussion as the FTC charts a new approach to policing data abuses on our economy. And now to panel 1.

[MUSIC PLAYING]

DEVIN WILLIS: Good morning, everyone or good evening, depending on your location. I'm Devin Willis, an attorney in the Division of Privacy and Identity Protection of the Federal Trade Commission. I would like to welcome you to the first panel of PrivacyCon 2021 entitled "Algorithms."

As you may recall, last year's PrivacyCon had an interesting panel discussing the existence of bias in particular artificial intelligence algorithms. This year, we are fortunate to have three panelists who will present very fascinating research on auditing, machine learning algorithms for bias. First, we will hear from Basileal Imana from the University of Southern California on his paper Auditing for Discrimination and Algorithms delivering Job ads, which presents a method for auditing ad delivery algorithms.

Next, Hongyan Chan from the National University of Singapore will present her paper entitled On the Privacy Risks of Algorithmic Fairness. Finally, Martin Strobel, also from the National University of Singapore, will conclude the presentation portion of our panel discussing his paper entitled On the Privacy Risk of Model Explanations, which study tools used to provide algorithmic explainability. More detailed bios of all our panelists and links to the research papers are available on our PrivacyCon 2021 website at ftc.gov.

After we conclude the presentation portion of our session, we will have a question and answer period where we will be able to take some questions from audience as time permits. So if you have questions, please email them to privacycon@ftc.gov or send Twitter to hashtag PrivacyCon2021. Without further ado, I would like to turn to Basi to start us off.

BASILEAL IMANA: Thank you, Devin for the introduction. My name is Basi, and I'm a PhD student at University of Southern California. In this talk, I'll talk about our work on Auditing for Discrimination and Algorithms Delivering Job Ads. This is joint work done in collaboration with my PhD advisors Professor Aleksandra Korolova and Professor John Heidemann.

Next slide. Targeted advertising has become ubiquitous in recent years and it's one of the ways that people access opportunities such as employment or education credit and housing. Therefore, externally auditing the role that these algorithms play in shaping society is important to ensure that the ads are being delivered in a fair way and also that they're being compliant with applicable laws, for example, in regulated domains such as employment.

Next slide. So I'll start off with an example. Let's say I'd like to hire a software engineer, so I create a digital ad. And I want to target exclusively. So I create a gender balance audience, so with 50% females and 50% male, and then I run this ad on Facebook.

Then the outcome I get is that more fraction of females see the ads. Now, the question is why is there such a gender skew in the outcome even though I targeted a gender balanced ad audience? One reason might be just because there are more females by the time the ad was being run. But there might be also other confounding factors.

Prior work looked at this question and control for such confounding factors and showed that it's only the role that Facebook's ad algorithms play in deciding who sees a particular ad as the cause for such skewed outcome. What this prior work did not look at is the role that qualifications might play in the outcome.

So if we go back to this example and look at what fraction of males and females in the audience are actually qualified for the job being advertised and use that to interpret the outcome, we can see that this cue can be explained by the differences in qualification between males and females. So this is the main insight in our work, and looking at qualification is also important from the legal perspective because companies are able to use it as a legal justification against claims of discrimination. So we want to rule out qualification from companies from being able to use that to rule out audit findings that show discrimination.

So based on this [INAUDIBLE] main contributions are first to give a new method for auditing for discrimination in the delivery of job ads. Our method is the first look on for potential differences in qualification. Then we take this method and we study ad delivery on two prominent ad platforms-- LinkedIn and Facebook and we find results that show discriminatory ad delivery by gender in the case of Facebook, whereas we find no such evidence in LinkedIn case.

Next slide. So how we account for qualification is the main part of our contribution. So I'll focus on that and just talk. But there is more in the paper.

So the main challenge with accounting for qualification is that as external auditors, we don't have access to user profile data that would let us directly control for qualification in the audience that we target when we run ads. So what we do is we rely on an indirect approach. So what we do is we find a pair of job positions with two conditions. First, they must have similar qualification requirement, and second, there must be a de facto gender skew in the real world. So known gender imbalance for the [INAUDIBLE] position.

To give an example, if we take delivery driver jobs at Domino's versus Instacart, one is a pizza delivery company, the other one's a grocery delivery company, there's data that shows that 98% of Domino's drivers are male, whereas the majority of the Instacart drivers are females even though both jobs have very similar qualification requirement. So what we do is we take such pair of jobs and we run ads for them at the same time and we look at the outcome. So we look out whether there is a relative difference in how this pair of ads are delivered by gender.

And because we control for qualification and also other confounding factors, if we see a difference, we hypothesize that is due to the platforms until the algorithm propagating the existing skew. So in this example case, we would expect the Instacart ad to be shown to more females if the platform is indeed perpetuating the existing skew. Next slide.

So we take this methodology and we register as advertisers on LinkedIn and Facebook, and we run ads. And to show one of our results for the Instacart and Domino's example, so if you look at the graph on the slide, the x-axis on the plot is the fraction of the females the given ad was shown to and the y-axis is the two platforms that we study-- Facebook and LinkedIn. And for each platform, we have a pair of ads, one for Instacart and Domino's And on the right hand side, we have the result of the statistical test that we applied to test whether the skew is indeed significant.

So if you look at the top row, you can see that the Instacart ad is shown to higher fraction of females, which is consistent with the direction of the de facto skew. On the other hand on the bottom row, we can see that there is no statistically significant difference between the Instacart and the Domino's ad. Overall, what this results shows is this skew on Facebook because we control for qualification as not just skill, but also discriminatory in the legal sense, and the role that Facebook's ad delivery algorithms are playing might be contributing to this discriminatory outcome. And this result is just one instance of our experiments. We repeat this on multiple audiences on different-- and also on different job categories and we find some of the results.

Next slide. Next, we looked at whether the skew that we see on Facebook's case is merely due to Facebook optimizing for click through rate or engagement, which is something an advertiser might actually be interested in. So what we do is we looked at the advertising objectives that advertisers can choose when creating an ad, and we compare two kind of objective.

The first one is reach objective whose aim is to show the ads to as many people as possible in the targeted audience. And other objective is conversion, which shows the ad to people who are more likely we can convert, for example, apply for the job being advertised. So while we were interested in how to run the ads with both objectives and compare and see whether an advertiser can reach a more wider audience by using a reach objective.

If you look at the plot on this slide, it's a similar graph, but in this case, the y-axis shows the reach and conversion case. Both are run on the Facebook platform. And you can see that in both cases, the delivery is skewed including the reach case, which shows that Facebook's algorithms ad delivery even if the advertiser uses the reach objective to try and reach a more wider audience. Again for this experiment, we reproduce this results on more job categories and audience.

So in light of this result to discuss some implicate policy implications of our work, what are technical evidence has shown is that the role that ad delivery algorithms play in ad delivery is important in that regulation needs to take this into consideration. And the questions that we would like policymakers to engage with are first is this technical evidence sufficient to enact new policies that will mandate our platforms to change how the ad delivery algorithms work? In the past, legal challenges and also civil rights audits have pushed our platforms to change, how their ad targeting works. And so we hope to see similar action in the context of ad delivery.

And the other question is, are there other additional technical insights or audits that would be useful to help formulate future policies for governing ad policies. In conclusion, in our work, we have shown that ad platforms should change how their ad delivery algorithms work for opportunity ads as such as employment. And we hope that regulators can use both our methodology and findings to inform future policy.

With that, I'll conclude my talk. Our paper and data set can be found at the link shown in the slide. Thank you.

DEVIN WILLIS: Thank you very much, Basi, for that very interesting presentation. Let's move next to Hongyan.

HONGYAN CHANG: Hello. OK, thanks, Devin for our previous introduction. My name is Chang, and I'm a PhD student at National University of Singapore. I'm very happy to be here to give you an overview about our work titled On the Privacy Risks of Algorithmic fairness

This is a joint work with my supervisor Dr. Shokri. Next slide, please. Algorithmic fairness and privacy are essential parts of [INAUDIBLE] machine learning. Most of this work focuses on solving one problem such as designing privacy, preserving the algorithms of fair measurement algorithms.

Time or in real life, fairness and privacy do not exist in isolation. So a deeper understanding of the relationship between fairness and privacy is necessary. So in our paper, we try to find the cost of privacy when we achieve fairness. Next slide, please.

One counterintuitive fact of machine learning is that machine learning models are not neutral. The most impressive example is racial bias [INAUDIBLE] algorithm. It is a popular commercial algorithm used by judges for scoring criminal defendants [INAUDIBLE] committing a crime. It has been shown that the algorithm is biased in favor of White defendants and against the Black defendants.

In the next, let's see why are machine learning models biased. Generally speaking, biased can be introduced via the other training data of the machine learning algorithm. The training data is collected by humans, quite often biased.

This human bias is propagated to the data set and then ultimately to the prediction of the model. The learning algorithm itself may also introduce bias. The model tends to fit the majority group better as the primary concept learning algorithm is to minimize the average loss, which itself favors the majority group.

On the next slide, let me give you an overview about algorithmic fairness. [INAUDIBLE] multiple fairness definitions have been proposed to regulate the prediction behavior of the learning model. We focused on [INAUDIBLE] notion [INAUDIBLE] which would require the model to have similar accuracy across protected groups that are identified based on sensitive attributes such as race, gender.

Most specifically, we see a model is fair when it's two positive rates and two negative rates are similar across protected groups. Accordingly, for measuring the fairness of a model, we compare the gap in the two positive rates and the two negative rates between groups as fairness gap. From now on, let's focus on [INAUDIBLE] as team attributes.

In other words, we'll only focus on the performance differences of two protected groups. Based on this definition, most of the existing algorithms try to find a model that minimizes the average loss while satisfying [INAUDIBLE] of the training data. Let me give you an example to show how they work.

Suppose a University wants to build a linear model for their admission based on the SAT score and JPA of applicants in high school. And there are two populations-- Blue population and yellow population. The University will train model on the historical data, which has the 80% blue people and 20% yellow people as shown in the figures here. Positive and negative means admissions and no admissions respectively.

Next, let's look at the standard model we can learn on this data sets. In this setting, the university will have a model represented by this red line here. The model performs better on the blue people as a result of minimizing the average loss. In this case, the yellow group is the underprivileged group.

Next, to achieve fairness, we may want to use the model represented by this green line. However, using this green can cause privacy issue. Let me explain why.

So implicitly, to achieve fairness, a fair algorithm place this green line right under some yellow points with positive labels. In other words, it increases the influence of the training data from the underprivileged. Originally, for the standard model, the red line [INAUDIBLE] the blue population. So the red line reveals little information about yellow applicants.

Now that the green light is right under some yellow points, so intuitively lack of more information about yellow points in the training data set from the model itself. In our paper, we will formally verify that fairness can cost privacy issues, especially for the underprivileged group. On the next slide, let me clarify what do I mean by privacy.

We used the widely accepted privacy definition differential privacy. Roughly speaking, we see learning algorithm as privacy preserving if whether an individual data was part of the training data sets or not has negative influence on the learning model. So to quantify the privacy, we'll make use of the membership inference tag. The goal of the adversary is to infer whether data point was part of the training data set or not.

So a higher attack accuracy reflects a higher privacy risk. Thus, we use the attack accuracy as a privacy risk in the rest of the presentation. We might know this privacy risk [INAUDIBLE] realistic setting where the adversary only have the access to the protection API of the model. I omit the details about the attack algorithms.

If you are interested, you can check our paper for more details. I put the link of the paper at the end of the slides. Great. In the next slide, let me show you the results of the synthetic data first [INAUDIBLE] the real data set.

In this synthetic data sets, we have two protected groups, yellow group and blue group and binary labels positive and negative. So we have four subgroups that are defined based on the protected attributes of the true label. The features for those subgroups are sampled from different two dimensional Gaussian distributions.

The blue group is the majority group, which we are seeing on your network models for both fair and standard models. Here, I show you the results of a training data for two protected groups with positive label. The x-axis is a privacy risk and the y-axis is training accuracy.

The triangle marks represents the results on the standard model and the circle marks represent the results of a fair model. We can see that standard model performs much better on the blue group compared with yellow group. If we choose to use fair model, the yellow group will have a better accuracy.

However, the privacy of risk for the yellow group is also increased. At the same time, it shows that fair model improves the accuracy, but leaks more information about this underrepresented group or under unprivileged group. On the next slide, let's see the trade-off between fairness and privacy.

May vary with distribution of the data. Each [INAUDIBLE] year shows the results for one setting. The x-axis shows the fairness gap of the standard model with respect to equalized odds reflecting the unfairness of the model.

The y-axis is a privacy cost for the underprivileged group. The privacy cost is measured as differences in the privacy risk between standard model fair models. We can see a clear trend that when there are more needs for fairness, privacy cost of achieving fairness it's also higher. We also come back to experiments on multiple real world data sets. On the next slides, I show you the results of the compass data set.

Similarly, we have four samples identified by the risk and true label. We can see that the privacy risks are increased for all those subgroups [INAUDIBLE] achieve fairness as the magnitude of the increase is different for subgroups. So let me conclude my talk in the next slide.

The take away from our empirical result is that group's fairness based on equalizing arrow comes at a cost of privacy. This privacy cost is not distributed evenly across groups. As a result, in practice, if we try to protect the underprivileged groups using fair machine learning algorithms, we must be very careful because it may increase their privacy risk. Thanks for listening, and I'm looking forward to our discussion.

DEVIN WILLIS: Thank you very much, Hongyan. That was a very informative presentation. Martin, we can move on to you.

MARTIN STROBEL: Thanks, Devin. A wonderful morning to everyone. My name is Martin Strobel. I'm a fourth year PhD candidate here at the National University of Singapore.

I'm actually in the same group as Hongyan. And I want to talk about privacy risks that can occur when you try to explain machine learning models which is a work I did together with my supervisor, Reza Shokri. Next slide, please.

So given that this conference is called PrivacyCon, I assume most of you are more familiar with privacy risk than they are with explaining machine learning models. So I want to spend a little bit of time motivating why we want to explain machine learning models. And there are essentially three main arguments for why you want to explain the model.

The first is it gives agency to individuals. So assume we have customer users out there and the decision was made by a machine, and now the person is unhappy with this decision, should happen from time to time, it's impossible for this person to argue against the decision if they don't understand it. And if you have heard one of the horror stories, for example, that the decision might change if you change the spelling of your name or the decision might change if your Main Street instead of Main St in the formula. Then you really want to understand how the decision was made so you can change it.

On a larger level, agencies like the FTC want to go in and audit a model. And if you just look at the big [INAUDIBLE] model, it's really hard to audit it. So you want to be able to explain the model so you can audit it. For example, if you want to look at whether or not the model is fair.

The final argument you might have heard is the right to fairness, and it's more like a philosophical argument that says it's inherently inhumane to be subjected to a black book decision. So even if you don't want to change it or you cannot change your decision, it's still better to know how it was made. And these three are the key arguments for why you want to explain. And let's, on the next slide, look at how not to achieve [INAUDIBLE].

So some people have proposed that you could just really use the entire machine learning model and just dump it out there, and then you're perfectly transparent. The first problem from a privacy perspective is that we already know that if an adversary has access to the parameters of a model, the adversary can learn a lot about the training data. So there's a lot of sensitive information an adversary can obtain as soon as they have access to a model.

There are two more arguments why you not just want to dump the model out there. First, whoever created the model has an interest in it not being released, and second, from a expandability perspective, it's actually a not great. Model machine learning models have like millions of parameters, and just giving that to somebody like a user that doesn't explain anything. They just have a lot of data on their computer, so you don't really achieve expandability. So on the next slide, you see how you actually achieve some expandability.

So a typical framework academia has come up with is instead of trying to explain an entire model, you only want to explain one addiction at a time. So on the bottom, you have this typical simplified machine learning pipeline, you have some data and you train a model, and then the model makes predictions for a user. And on top of this now, you put in an explaining framework.

An explaining framework in fact is the model, and it potentially interacts with the data and then it provides an explanation to the user. This now opens up potential leakage on the next slide. There are essentially three ways how this model might leak sensitive information. One is already kind of covered by Hongyan. It is the model can leak directly information about training data via its predictions.

However, with the explanation framework, you have two more pipelines kind of. The one that's the explanation in fact that the data directly might leak information, and the explanation also interacts with the model which also might leak information. And in the paper, we mostly focus on how the interaction between an explanation framework and the model might lead to information leak.

On the next slide, you see what I mean when I talk about information leakage or like how you want to quantify the leakage. We use this very similar approach to Hongyan's work where we focus on membership experiments. So given an explanation, can the adversary tell where the data point was in the training set or not?

If you have a background in cryptography, you might be more familiar with a game setting. So you could formulate this as a game and know that adversary would win if we can distinguish two explanations. One, based on a training set that has the data point and one based on a training set that didn't have the data point.

So in the setting, we have an adversary. The adversary has a data record. It gives it to the model, then it gets back with prediction and an explanation. And if the adversary wins the game, he can tell whether or not the training used.

On the next slide, I want to spend a little time showing you how these explanations might look like. We focused in our work mostly on attribute-based explanations, and their approach is to compute the influence of each input feature on the prediction. So if you have a classification task for images, key on the lower left, the classification task is actually figuring out the mood of the phrase and then explanation method might highlight parts of the phrase on top.

Apparently, the eyes were important to predict the mood of the person on the lower side. The eyebrows were actually important to determine the mood according to the explanation method. You also see that there's a lot of flickering, so these explanation methods are far from perfect at the moment.

On more traditional tabular data in explanation method might look like on the right, but let's assume we want to predict whether or not a person gets a loan. And the explanation might say, OK, you would have gotten the loan if your income would be high to say the income was the most important feature. If you want to have a theoretical intuition for how these explanations work, the gradient of the loss with respect to the input is like the canonical example because the gradient tells you how the output changes when you change the input.

So on the next slide, you see the results from our attack. Similar to Hongyan, I'm not going into detail about how the attack works. If you're interested, you find it in the paper.

But we demonstrated on a conceptual level on several data sets. Here, I've shown customer shopping data and hospital data that an adversary with access to the explanations can figure out whether or not a point was used for training and you can do this far better than random guessing. So on the final slide, let me conclude.

What should you take away from my talk? First we have demonstrated that model explanations can leak membership information. And the overall goal seems to be that we want trustworthy machine learning. And white papers out there often say, OK, we need privacy and we need interpretability for trustworthy machinery to work.

And academia has reacted and like there are lots of explanation frameworks out there and new ones are proposed every week, I want to say. So it's very likely that there will be user phasing explainable and machine learning frameworks in the future. Right now, the big companies that give kind of programs to programmers already include explainability tools in their frameworks. So I hope that both developers and regulators have the privacy implications in mind when they are designing explanation methods. Yeah, if you have questions that go beyond [INAUDIBLE], you can reach me at the email address linked to the paper on the bottom of the slide.

DEVIN WILLIS: Thank you very much, Martin, and again to all our other panelists for those very informative presentations. I'm really looking to hearing more about them in our discussion portion, which we will now move on to. Now, I really hope to engage in some great discussion and expanding upon some of the research that you've presented and the implications of such and the work that we all do.

So first, I just wanted to say and ask the question to you, Basi, I mean research has shown the prevalence of bias including unintended discriminatory outcomes and machine learning algorithms used for various purposes. We've seen this from health care to credit to targeted or behavioral advertising decisions as your paper showed. An algorithmic fairness legislation has been proposed in the US and abroad to help mitigate such bias.

Such legislation often calls for more transparency and auditing of algorithms. So sort of to you first, Basi, in light of your research, it seems that you might agree that increased transparency might be useful in achieving fairness and algorithms for online job advertisements. From your technical point of view, I'd be interested in hearing what implications does your research have on such legislation? What would you recommend to increase algorithmic transparency including in job delivery algorithms?

BASILEAL IMANA: Yeah, that's a great question. Yes, transparency is one of the things that we call for in our work. And that can come in different forms.

One of the things that we recommend in our work is that our platforms need to provide additional data and statistics about not just about how ads are delivered, but also what happens at the ad targeting phase and also at the adoption phase. For example, in LinkedIn's case, the platform does not provide breakdown of ad delivery by gender. So we have to rely on a workaround methodology to audit how the ads are delivered by gender. So for example, one way is LinkedIn can provide breakdown of ad delivery by a sensitive attributes.

In Facebook's case, there are some existing transparency efforts such as a public library API that they made available, which is a good first step. But we don't think it's enough. There is, like I said, like the target advertising pipeline is a complex process with many steps.

Apparently, they provide breakdown by ad delivery, but providing additional statistics about other parts of the pipeline would be useful. And want to add one more point. Another transparency direction that we're thinking about and we hope our platforms would consider is providing auditing interface that auditors can use to query different parts of their algorithms to certify that they're fair enough. So that's one direction we're exploring as well.

DEVIN WILLIS: Thank you. To follow on that, Hongyan and Martin, similar issue. I mean, both of your papers seem to suggest that there may be trade-offs between privacy and algorithmic fairness or methods aimed to achieve fairness such as explainability or other auditing methods. In light of the findings of your research, what implications does your research have on transparency legislation or auditing legislation? Do you think there are ways to achieve algorithmic fairness and protect the privacy of underrepresented communities at the same time, or do you think there might be some inherent trade-offs between privacy and fairness or auditing methods to achieve fairness?

HONGYAN CHANG: I think this is a very interesting question, and the short answer to that is that actually we can't achieve fairness and privacy at the same time. So for example, if we have [INAUDIBLE] classifier that always outputs the same label for all the data points, then the model will be fair, but because it always gave the same output to all the data points, but in different groups. And the model is also privacy preserving because the model is independent of the data.

But I think, more importantly, can achieve fairness and privacy without hurting accuracy of the model. And unfortunately, some theoretical works show that differential privacy can conflict with many group fairness notions including equalized arts. So the results show that if the learning algorithm is satisfied to our differential privacy, which is a very strict differential privacy notion, then the fair models outputted by this algorithm it's a constant classifier.

So in other words, if we want to achieve pure differential privacy group fairness, there's a model accuracy we can guess is no greater than that of a constant clarifier. So when we talk about achieve fairness and privacy at the same time, I think probably the first thing we need to do is relax our privacy notion. For instance, some existing work actually show that relax privacy requirements and then study the differentially private affair always with respect to approximate DP, which is a more relaxed version.

And their empirical results show that differentially private algorithm can output a model that achieve good accuracy and fairness at the same time. So I personally stay quite optimistic. I believe that the accuracy approximate fairness and approximate differential privacy can be achieved at the same time, but it needs further effort from the research communities.

DEVIN WILLIS: Thank you. Martin, do you have any thoughts on that?

MARTIN STROBEL: Yeah, I want to quote on the positive note. So like currently, I don't think there's an inherent [INAUDIBLE] between explainability and privacy. And this is like if you care about the privacy of the training data and you want to explain the model, the only thing you kind of need to guarantee is that your model is private. So if you can guarantee a private model and you explain it only interacts with this model, you would have to have a private explanation. So this is a good note.

On the bad side, in our work we actually found some hints that the privacy risk for minorities might actually be higher than for the majority class. So the privacy risk through explanation for minorities might be higher than majorities, and that's kind of problematic because it's kind of exactly the people you want the explanation for because they are most likely to be the ones who are discriminated. So that's the bad part.

And the other part of bad news is in a follow-up [INAUDIBLE], there's actually some indication that it's harder to explain private models. So how do we make a model private? You introduce noise.

And this noise kind of in a very wishy-washy level is counterintuitive. So if you introduce noise, that makes the model more complicated and now it's harder to understand. So yeah, I don't think there's a fundamental trade-off, but there's still work to do to get it done.

DEVIN WILLIS: Thank you for that. Going back to you, Basi. Your paper demonstrates a method for algorithmic auditing, specifically for gender discrimination in job advertisement ad delivery when platforms aren't transparent about their algorithms. Do you think similar methods could be used to audit for other forms of discrimination such as race or sexual orientation or could they be used for other machine learning algorithms such as those used to serve ads for other types of things such as credit or housing or even algorithms that are used to assist in credit or housing decisions? And if so, do you think there will be any limitations for using such methods?

BASILEAL IMANA: Yeah, that's a great question. The short answer is yes. We'll be able to explain this methodology.

DEVIN WILLIS: I think, did Basi freeze? OK, we will maybe go to the next question, and hopefully we can hear from him on that because I'd really be interested.

BASILEAL IMANA: Am I back?

DEVIN WILLIS: Yes, you're back. You froze for a moment.

BASILEAL IMANA: OK, great. Yeah, so we use a voter data from North Carolina to create an audience where we know both the gender and the location because LinkedIn provides breakdown by location, and we use that as a proxy to calculate the gender breakdown. Similarly, the data set also includes other fields like race and age. So we would be able to use those two audit for discrimination by race and age.

But the caveat is that like I mentioned in the talk, we would need to find a pair of job positions with similar qualification requirements, but there is an imbalance by race or age or other sensitive attributes that we're interested in. In response to your question about the limitation, one of the main limitations that we talk about in the paper comes from using location as a proxy to calculate gender because, for example, people might move between different locations or the data might be outdated. So there's some error that comes from that, and there's another limitation is just the cost involved in running this ad.

We run our experiments on multiple audiences and also different job categories seeking confidence in our results and the cost can easily add up if we want to gain more confidence in our results. So yes, there are those caveats. But yes, our methods can be extended.

DEVIN WILLIS: Thank you. Again, your paper covers one scenario where fairness can conflict with privacy. Specifically, good fairness as you were discussing earlier on equalized odds. Are there other ways that you think privacy and fairness can conflict in machine learning algorithms that you haven't already discussed?

HONGYAN CHANG: Yes, thanks for the question. I think fairness and policy could conflict with each other in other scenarios. One scenario that draws a lot of attention is when we care about the privacy with respect to all those sensitive attributes such as race, gender.

For example, JDPR restricts the racial data collection from customers. So it protects the privacy with respect to the sensitive attributes, but actually it raises a big issue for training a fair model, because researchers have found that fairness can only be achieved through awareness. So for example, if you want to build gender fair classifier and you exclude the gender attributes, it does not provide-- it does not guarantee to provide fairness and [INAUDIBLE] the accuracy of the model.

So it is because there are many other features that are well correlated with your gender. For instance, we can't make a good guess about individual's gender by knowing their favorite songs and the favorite colors. So in this case, the gender blind model may discriminate against males or females by discriminating against the people who like a particular song. So this allowing the collection of those sensitive attributes for protecting privacy indeed raises a problem for fairness.

And another important scenario is when privacy is achieved by adding a noise to individual's data. So the model can always see the noisy version of the individual's data. Actually in this case, some researchers also found that the resource allocation decisions made all this noisy data can disproportionately affect some subgroups. So overall, there are a lot of ways privacy they can conflict with the fairness in machine learning task. So basically, it tells us that when we ask for privacy and fairness, we really need to think about both of them at the same time because they can really affect each other.

DEVIN WILLIS: Thank you for that. Martin, throwing it to you, your paper focuses on the privacy aspects of data, specifically stakeholders and the deployment of machine learning model explanation tools. Are there other stakeholders including those for example involved in the design or development of machine learning or model explanation tools that you think might see their privacy eroded?

MARTIN STROBEL: Yeah, thanks for the question. So I hinted at this in the talk a little bit. So I think the biggest hurdle for explanations is not so much the privacy of the data, it's the privacy of the model and the fact that whoever creates the model wants to keep it private.

So there's already research out there that demonstrates if you just see enough predictions of a model, you can reproduce that [INAUDIBLE]. And if you explain these predictions, it just becomes easier to reproduce the model. So I think the companies might have an interest in not releasing explanations because it makes it easier to kind of extract the models and then just copy the models. So this is kind of the one big stakeholder that might see their privacy affected.

DEVIN WILLIS: I kind of want to go back to sort of like you, Hongyan were saying and also Martin sort of about the collection of data and maybe the collection of sensitive attributes versus when you might have data with added noise. Then I know there's been some technology experts who advocate for the need to have more collection of data. And so you need more information on protected characteristics of underrepresented communities and things like that while others might suggest that it's possible to mitigate algorithmic bias without collecting demographic or proxy data such as using simulated data or data where you've added in noise. In your view which approach if any, and this is really are question for all of our panelists, would you recommend, and do you think that there are any privacy risk of the approach that you would recommend?

MARTIN STROBEL: Thanks for the question. I think this is a very interesting question. So first of all, I would like to mention that it's hard to trade-off between privacy and fairness. I mentioned before, [INAUDIBLE] infinite number of samples. So it is hard to trade-off between privacy and fairness even with a limited number of samples.

But in practice, what we often observe is that if we have more data from the distribution, then the privacy risk will be reduced. In fact, in our paper, we analyzed the effects of small data collection on the trade-off between privacy and fairness. Actually, we show that while there are more data from the other representative group, the data set is more balanced. But in this case, the standard model is also less biased.

As a consequence, the policy cost of achieving fairness is also reduced. So in other words, more data collection from the other representative group can help to reduce the cost of achieving fairness. And another thing you talk about is so simulated data.

I think it's a very interesting question because to use this kind of the techniques, we must make sure that this simulated data is privacy preserving. Because [INAUDIBLE] stimulated data based on some private data set. So this simulated data may contain sensitive information about the original data set. So if we want to use this kind of techniques, we must make sure that the individual information won't be mixed through this simulated data.

And another quality that we want to make sure that any fairness guarantee a model provides on this simulated data should also hold approximatively on the original data set of all the distribution we care about. So simulated data satisfy these two requirements would be very interesting thing we should look at. In my opinion, this is the research direction we should definitely look at.

DEVIN WILLIS: I'll be looking forward to that research [INAUDIBLE]. Martin or Basi, do you have any responses?

BASILEAL IMANA: I want to add something to the Hongyan's wish list for the simulated data. So like explanations often only work in the context, so they need access to the data to kind of have the context and then explain it in the context, in different context. I definitely expect you might need different explanations.

And now if you only have like simulated data, your explanation I mean to be privacy preserving should be on the simulated data, and this is a slightly different context than the original data. So you need to kind of ensure that the explanations you gain, the simulated data is actually useful in the real world that you then base on that real data. So I want to add this to the list of Hongyan's criteria, it should not just be privacy preserving and fair, it should also be explainable ideally.

DEVIN WILLIS: And now we only have a few minutes left, and I don't know if you wanted to add anything on that, Basi. I would really be interested in hearing from all of our panelists in our last few minutes if you had any thoughts on how you think policy makers or law enforcement agencies like the FTC can help mitigate algorithmic bias in ad delivery algorithms used by platforms or any other machine learning tools while protecting consumers' privacy interests. So I don't know who would like to begin. How about you, Basi.

BASILEAL IMANA: I can start, yeah. So like I mentioned in the talk the technical evidence that we show shows that the role of platforms play in ad delivery is significant in that regulations and policies should take that into account. I think a good first step would be for technical legal and policy experts to be in the same room and look at the technical evidence from both our work and other prior audit findings to see whether those technical audits or the findings are enough to inform or enact new policies. And if not, if that's not the case, what are the additional technical contributions would be useful to inform future policies thus far our recommendation.

DEVIN WILLIS: Anyone else in the last few seconds? OK, we have got a minute.

MARTIN STROBEL: OK. Yeah, so the worst case I see would be if people look at our work and say, oh, this transparency has privacy implications. So we shut down and have less transparency. So I actually think we should have more work like Basi's.

And I think the FTC can help because like it can regulate who has access to this transparency tools and like if auditors have access and trusted auditors have access, think there's little privacy risk. The privacy risk comes from everybody having access. So if you kind of can ensure that the right people get access and more people like Basi have access, that would be good.

DEVIN WILLIS: OK. Thank you. I mean, this has been a very interesting discussion. I want to thank again all of our panelists for their amazing presentations, I mean this awesome discussion. We really hope everyone will stick around as next we're fortunate to have another presentation on auditing machine learning algorithms for bias. So thank you again to all of our panelists, and I appreciate everyone for sticking around.

BASILEAL IMANA: Thank you.

LERONE BANKS: Good morning. My name is Lerone Banks, and I'm a computer scientist at the FTC. It is my pleasure to introduce Ziad Obermeyer from UC, Berkeley.

Ziad will be presenting work his team has done to identify algorithmic bias in health care and practical steps that organizations can take to identify bias in their own applications. Please send your questions via Twitter or email at privacycon@ftc.gov, and we'll get to them after the talk. With that, please give your attention to Ziad Obermeyer.

ZIAD OBERMEYER: Thank you so much, Lerone. So I'm going to talk a little bit about work with my co-authors and the rest of my team. We've been doing to try to diagnose and fix algorithmic bias over the past couple of years.

I'm going to start by walking through a case study very quickly and then transition into some of the practical steps that I think we've learned can be really, really effective for this goal of fixing algorithmic bias. I'm going to try to wrap that up in about 15 minutes, mostly because I really like chatting with Lerone and I always learn a lot from him about the links between algorithmic bias regulation and privacy regulation. So let's start by working through a really quick case study on what algorithmic bias looks like. I think for me this is an example I learned a lot from.

This comes from the paper we published two years ago in science, and it works through a case study in health systems that are trying to target extra help to patients who need it. So all throughout our health system, there are these pockets of complex chronically ill patients. And those patients are having a very bad experience with their care, they're experiencing a lot of exacerbations of chronic conditions, and they're also generating high costs for our health care system.

And so throughout the health care world, health systems have invested in what's called high risk care management. And I think of that as just kind of an extra help program for these very sick patients. So this is a resource that's scarce and that health systems have to distribute to people who need it most.

That's like home visits and primary care slots and a lot of extra help that itself costs money. So we're trying to find people to help so that we can prevent their health care problems so that we can save the health system money. But doing that is actually a scarce resource on its own.

And so we really need to target that resource to the people who need it most, and that's where algorithms come in. So algorithms have gotten to a huge, huge scale in health that at least surprised me when I first learned about it. We studied a particular piece of software that itself was used to screen about 70 million people a year at health systems throughout the US. If you look at the family of algorithms that work just like the one we studied, those are being used for like 150 to 200 million people a year, so the majority of the US population. And so when we think about the scale that algorithms have reached in society at large, I think health is one of those places where they've just started impacting lives at very, very large scale.

So as I mentioned, all these algorithms are trying to find people who are going to get sick. And the way they do that is they predict. So algorithms are really good at looking into the future. And just like algorithms can figure out what product you're going to buy, what movie you're going to like, these algorithms look ahead and figure out how much someone is going to cost the health system by showing up in the ER and getting hospitalized.

They make that forecast and then they figure out, OK, this person looks like she's going to cost a lot of money with all these ER visits and health care that she's going to consume. Let's target her with extra help. So given how widely used these algorithms are, we were really interested in this important question of whether or not they were racially biased.

Now, to study racial bias in an algorithm, you really need to define exactly what you mean by bias. And so here's how we did it. We articulated a principle of what an unbiased algorithm would look like. So the way these algorithms work at all these health systems, the one we studied, the ones we've worked with since then is that it trolls through a population of patients.

So if you're working at a hospital or an insurer, you've got a population of patients you're responsible for. And once or twice a year, that algorithm is going to generate a score, and then that score is going to determine if you get prioritized for extra help or whether you get screened out. So people with the same score are going to be treated the same way. And as a result, what we thought was those people should have the same needs in terms of their need for extra help, and the color of their skin definitely shouldn't matter.

But that's not what we found. I'm going to show you a graph and just try to walk through the axes very carefully, because this turns out to be a very general test for bias in an algorithm. So on the x-axis of this graph on that horizontal axis we've ranked patients from low to high risk. And so the top 2% or 3% all the way on the purple side of that graph and that the last little bit on the right, those are the people who are going to get fast track into this extra health program.

On the y-axis, I'm showing you-- so the x-axis is what the algorithm thinks is going to happen in terms of your health. On the y-axis, I'm showing you what actually happened in terms of these patients health. For anyone on the x-axis, this is what goes on to happen.

This is how many chronic conditions do you have that flare up over of course of that next year. And you can see that there are two lines there. The top line, the purple line is Black patients and the yellow line is White patients.

And what you can see is that no matter where we are on that graph, the Black patients line is above the white patients line, which means that they go on to have worse health. So at that same algorithm score, no matter where you look, Black patients go on to have worse health than White patients even though they're being treated the same and they have the same priority for getting extra help. So how much bias are we talking about here? It's a little hard to tell from the graphs. So let me give you some numbers.

When we looked at that program, that extra help program that people get prioritized in by the algorithm, it's 18% black today. Now, you could look at that and think, OK, well what's the population rate of Black patients that this high priority group is drawn from? And that's actually only 12% Black.

So at first glance, you might look at that and say, oh, Black patients are 50% overrepresented in that group. The algorithm can't be biased. It's overrepresenting Black people in this group. When we did a simulation though to figure out what that group should have looked like, what proportion of black patients should have been in that high priority group, it actually should have been 47% black. So it's an enormous amount of bias that reduced the fraction of black patients in that program from 47% to 18%.

So on the next slide, I'm going to show you another graph. And this graph shows you an important aspect of why the algorithm was going around. So we wanted to understand how this happened and how that bias got into the algorithm.

And one clue to that is where the algorithm was going right? So on this graph again on the x-axis, the horizontal axis, people are ranked by their algorithm score. But now on the y-axis, instead of showing you what happened to their health, I'm showing you what happened to their costs.

And you can see that those two lines are right on top of each other. So when the algorithm predicts a certain score, those people go on to have the same costs even though on the last graph I showed you, they didn't have the same health. So the algorithm is predicting total health care costs very accurately and without much bias between Black and White patients. So let me summarize this on the next slide.

The algorithm is biased for predicting health, but it's unbiased for predicting cost. And that's because Black and White patients don't have the same relationship between health and costs. White patients have better access to the health system.

So when they get sick, they're more likely to go see a doctor. They're not going to stay at home. They're going to see a doctor, they're going to generate more costs even though they have the same needs. When you have a squeezing sensation in your chest and you're sitting on your couch, you're more likely to call the ambulance, get tested for a heart attack if you're White than if you're Black.

In addition, the health care system just treats Black patients differently. And there's lots of evidence of systemic racism and how doctors actually recommend tests and treatments for Black patients. So the result of all of this is that conditional on someone's health. Two people with the same health are going to have different costs if they're Black versus if they're white. And that means that predicting cost accurately means bias when you're predicting health. So let me try to distill some lessons from that case study before stepping back and teasing out the implications.

A really important part of what we did is to articulate that ideal target for the algorithm. What should the algorithm be doing? In this case, the algorithm is being used to decide who gets what in terms of extra help for someone's health. And so those people at a given algorithm score, the algorithm should be predicting health and those people should have the same health needs.

That's how you hold an algorithm accountable by articulating what the target is that it's supposed to be predicting, in this case health, and comparing it in this case to what it's actually predicting, which is cost. That difference, even though it's subtle, is the source of a lot of algorithmic bias that we found in our work. And at the same time, when we detect that bias, we have a roadmap for fixing it.

So once we articulated this problem and we realized that the algorithm was predicting the wrong variable, we were able to work with the company that made that algorithm to retrain it to predict health rather than cost. And that really helped and it reduced the bias in that algorithm by one measure by 84%. That lesson that we have to articulate the ideal target hold the algorithm accountable for that and then make sure the algorithm is doing what it's supposed to do is a theme that's come up again and again over the past two years as we've worked with a lot of health care systems, insurers, tech companies, and regulators at the state and federal level.

And on the next slide, I just want to give you a sense of some of that work, which is that we found that this same bias, this discrepancy between what an algorithm is ideally supposed to be doing and what it's actually doing is very, very widespread throughout the health system. So the top row shows you the algorithm that I just talked about-- health care needs versus total costs. When we look at a number of other things, for example, on the second to last row, a lot of health care systems are using algorithms to figure out, oh, I booked a patient for an appointment with her doctor. Is she going to show up for that appointment?

And if that algorithm predicts you're not going to show up, it's going to take that slot away from and reassign it to some other patient. But of course, people can not show up to their doctor for a couple of reasons. One is that they decided they didn't need care. After all, their runny nose went away, their knee pain got better.

But some people don't show up because they face barriers to accessing care. And for that second group of patients who are more likely to be Black and more likely to be poor, the last thing you want to do is reassign that slot to another patient. You want to reach out to that patient and help them not rebook their slot to someone who needs it less.

And so these kinds of biases are active throughout the health care system. If you think about it, they're also very active in a lot of other industries. So in criminal justice, a lot of algorithms are trying to predict someone's innate tendency to commit a crime.

But we don't see their innate tendency to commit a crime. We see whether or not they get arrested, whether or not they get convicted. And those two are not the same, especially if you look at those two through the lens of race.

In finance, we're often interested in predicting creditworthiness. But what is credit worthiness? Instead we often predict income, and that is also not the same, especially when seen through the lens of gender, race, or socioeconomics.

So on the next slide, I try to take some lessons away for detecting bias. And when we're thinking about how to regulate this, whether we're a regulator or whether we're companies or actors that are being regulated, it's very different to work with algorithms than a lot of other things. So when we're regulating a drug, we understand that a drug should do more good than harm. And even though we can disagree about how much good or how much harm, that is the standard that we hold drugs to.

When we're regulating a toaster, rather appliances, the standard is it shouldn't catch on fire. But what about algorithms? How do we-- what vocabulary do we use for regulating them?

And what I'd submit to you is that the goalposts, the target that we want to hold algorithms accountable to is that ideal target the algorithm should be predicting. Is the algorithm doing what it's supposed to do, and is it doing equally well for Black and White patients? That's the intuition behind a lot of our work, and it's the intuition that yields a very crisp test for statistical bias which is does the algorithms' ability to predict that ideal target differ for Black or White patients? And if it's useful, we can get into this in the discussion, but that has a clear parallel to civil rights law and the use of proxy variables, which can be discriminatory.

So I want to close by taking some concrete lessons about what your organization can do to mitigate algorithmic bias. As I mentioned, over the past couple of years, we've been working with a lot of organizations in health, but increasingly outside of health and finance and other sectors as well. And here are some four steps that we found can be taken within organizations that can really help when dealing with algorithmic bias.

The first step is to designate someone in the organization that is responsible for oversight of algorithms, and importantly that person needs to be at a high level. Very often, decisions about algorithms are pushed down to technical staff and organizations who aren't empowered to make these high level strategic decisions or to engage in oversight activities about how algorithms are being used. So much like in other parts of regulation and law, we need someone at a high level at an organization who's ultimately responsible for oversight, and that person needs to be advised by a diverse group of people both inside and outside the organization who were impacted by algorithms and who are empowered to raise issues and ask questions.

Number 2, one thing we found is that most organizations actually don't know what algorithms are being used inside of their own organizations. And I think that's really surprising, because let's say you are an executive who hypothetically didn't care at all about racial bias. You would still want to know given the enormous strategic importance at algorithms that are being used in your organization, what's going on, what are they doing, what are they supposed to be doing. That inventory needs to be maintained and updated regularly so that anyone can start asking questions about what algorithms are doing and how they're performing for that task.

Number 3, algorithm performance needs to be documented. It's really surprising how often we have found that when we ask about an algorithm whether it's in the course of helping a hospital or health system do better or in the course of a civil investigation organizations and the staff that are at those organizations have no idea where the algorithm came from, what it does, how it's performing. They often say, oh yeah, Bob made that and Bob left a couple of years ago, but we're still using that algorithm. And that is a really dangerous situation to be in both again for strategic purposes and for bias purposes.

Finally, when algorithms are found to be biased, they need to be either fixed or deleted. So let me just try to articulate some use cases for these lessons. I think if you're a strategic leader at a high level in an organization, you need to know what algorithms are operating at scale in your organization, and you need to think strategically about how those algorithms are being used, where they can go wrong, and what oversight mechanisms you can put in place.

If you're a part of a technical team, you need to be able to recognize and avoid the subtle technical questions that we found can lead to bias. And if you're buying algorithms, you need to be an educated consumer of those algorithms you buy. If you're making policy or regulating algorithms, you need to have clear standards for what algorithmic bias looks like both so that you can conduct investigations and also so that you can provide guidance to industry on how to stay on the right side of the law.

The last thing I'll mention is that we tried to distill all of these practical lessons into what we're calling an algorithmic bias playbook. So the link to this playbook should be in the accompanying material on the PrivacyCon website. It's free to download and use, and you'll find a lot more detail and some summary steps about how to apply these lessons.

We're also working directly with organizations to help them implement some of these principles when they don't have the internal capacity. So I'd urge you to reach out to us if you're interested in being part of this work. And thank you very much.

LERONE BANKS: So thank you very much, Ziad, for that presentation. We've talked about this quite a bit. So I have lots of questions. But the challenge is just really picking where to start. So let me start with some of the data that you presented or question about the data that you presented.

You showed basically what was in estimate to some degree of the amount of bias in the algorithm that you were evaluating. One question I have is, are there thresholds for acceptable levels of bias? And I'm thinking about that in terms of you could detect bias, and an organization that can detect bias in an algorithm and then immediately be faced with a decision about whether or not to continue using that algorithm and potentially or discontinue using the algorithm and then potentially risk getting any of a benefits from its use, or to try to use it in some limited capacity. And so the question I have is, how do organizations sort of make that determination when bias is detected?

ZIAD OBERMEYER: I think that there are two answers to this question. I think there's a simple answer, which is that I think that in many legal settings, the standard that lots of other things, not algorithms, but any kind of potentially discriminatory policy is held to a very basic statistical test of significance. And so in our setting, what that would be is let's take a group of people who have the same algorithm score and let's separate them out into the Black subpopulation and the White subpopulation, and let's just statistically compare what ends up happening to those people. In our sense, that's on the ideal target and test whether those two groups are statistically different.

So I think that would be a standard and a lot of other settings has been held to other instances of discrimination. And so that's kind of statistical answer. Now, I think there's a deeper question that you're asking which is that there's often-- so let me just distinguish between two settings and health, we're often dealing with algorithms that are fundamentally helpful.

So for example, in this setting that we talked about, we're trying to find people who need extra help and target resources to the people who need extra help. In that setting, it's a real problem, both in terms of bias, but in terms of just what the algorithm is supposed to be doing if that extra help isn't going to the right people. What we found is that it wasn't going to the right people and the people that were missing out were more likely to be Black.

And so there's both a great business case for fixing the algorithm and a great case for anyone who's interested in promoting racial equity. And I think that's often the case in health when we're allocating a scarce resource. We want that resource to go to the people who need it, and those people are often more disadvantaged.

In a lot of other settings, for example, in finance, it's a little bit different. So because of structural barriers and historical discrimination, the people who need for example, credit are often less likely to repay loans. And so in that setting, I think there's a legal standard around business necessity, which is to stay afloat creditors can't be giving loans to people who are not going to repay their loan.

That's part of the business necessity of being a creditor is finding people who are going to pay back loans and pricing the credit accordingly. So I think that that maybe the other part of the legal standard is that we do have laws that provide for this business necessity purpose for algorithms, and I think that's the category that I would put that in.

So is it a quote unquote "acceptable" level of bias? Well, it's never an acceptable level of bias, but under the law, there is provision for business necessity. For people who are less likely to pay back a loan to be charged a higher rate to accurately represent that risk of not paying back a loan. And I think that's the standard that at least in our work with some regulators at the state level is being applied to algorithms as well.

LERONE BANKS: As moving towards, some of your work directly with organizations, can you tell-- sorry about that. Can you talk a little bit about the costs associated with applying the playbook? So you give a lot of practical tips about that organizations can actually use, but in your experience, can you discuss the costs that are incurred? So like where there's significant cost in hiring staff in order to audit algorithms, are there significant additional cost incurred by the organization?

ZIAD OBERMEYER: So the way we tried to structure the playbook was really grounded in the work that we've been doing with organizations over the past couple of years, and that work was using existing resources. So using the technical teams that are already deployed within an organization and applying some of these principles from the playbook, we were able to conduct an inventory of algorithms, identify potentially problematic ones, and adopt them without hiring new staff, without devoting a lot of additional resources to it. Now, how intensive is it of existing resources?

Well, you do need to allocate time to actually building up that inventory, and that sometimes requires putting together a lot of information from different business units in the organization. You also need to be very thoughtful about articulating OK, what is this algorithm supposed to be doing? What is it actually doing? Does that put this algorithm at risk of generating bias?

Now, is that a substantial cost? I think it requires part of an FTE for a few months to kind of do this realistically, and that's what we've seen in organizations. But I would say that it's almost an illusion to think that there's no cost to letting the status quo be the status quo.

What we've seen over and over again in a lot of organizations is that there are fundamentally flawed algorithms that are affecting thousands, or in some cases millions, of customers and patients on that scale. And so there's a huge cost, both in terms of regulatory risk and just in terms of algorithms not fit for business purpose in not doing these things. And so I think that's why I think it's an allusion to try to save costs by not doing these things. These things are eventually going to come to light whether it's in bad business decisions or in terms of regulatory exposure. And so I think this is a pretty good investment of a small amount of resources considered on an organizational basis.

LERONE BANKS: That really resonates with me because from a security perspective, we sort of make the same argument that the upfront cost that you pay almost always tend to be better than the cost you'd have to pay on the back end after an incident occurs. So that makes total sense.

ZIAD OBERMEYER: I remember you also telling me that that argument sometimes isn't that persuasive to people for security.

LERONE BANKS: It hasn't been, but it's changing slowly. Maybe some FTC fines have helped with that. Do you have recommendations for information that organizations can provide to show that their algorithms have been subjected to a reasonable process or have been audited in some meaningful way? I guess, conversely questions that consumers can ask to try to understand that the algorithms that they're being subject to actually have been vetted at least in some way to try to identify bias.

ZIAD OBERMEYER: Yeah, it's a really great question because I think that unlike in a lot of other industries like in finance, there's enormous documentation requirements that are imposed on companies by the regulatory system. And I think in these algorithmic settings, there's no corresponding need to document what the algorithm is doing or how it's performing or that it's unbiased. So let me just kind of give you two thoughts based on our work that went into the playbook.

Number 1 is that the inventory and the documentation of performance of an algorithm actually doesn't need to be public. It should be maintained internally and it should be kept on file so that if anyone asks questions, whether that's someone internally, whether it's a regulator, that information is available, and a company can very easily show that the algorithms that it's using are both doing what they're supposed to be doing and not introducing bias into the decisions. On the other hand, I think that all of the work that we've done doesn't require opening up the black box of the algorithm.

And so in order to do the work that we published in our original paper a couple of years ago or any of us work that we've done in the playbook, what we need are the algorithm scores and some readout of what the ideal target would be. So in our setting, this was, how did the patient do in terms of their health? Putting those data together is actually something that doesn't need to compromise trade secrets.

It can be done by an external auditor very easily with the right data. And so I think that those kinds of audits are appealing because they don't require us to do a lot of complex work on the inside of the algorithm or open up the box or to the previous sessions point, there are lots of transparency methods for illustrating exactly what the algorithm is doing. Our method actually doesn't require that we just need the score and then the ultimate judge of whether the algorithm is doing what it's supposed to do in the form of that ideal target.

LERONE BANKS: And I think that's really a great point that-- the point about not needing to open up the black box, because I think in certain other context, some organizations may be reluctant to share auditing information because they feel like that puts their intellectual property at risk. So it's nice to know that companies can go through this process and share information about the lack of bias in the algorithms without revealing trade secrets. I think related to that point and also mentioning something from the previous panel, I'm wondering the degree to which that's true in the phase of using proxy variables. So the previous panel talked about a proxy of variables for gender and for race. And can you still get that same guarantee of being able to thoroughly analyze an algorithm and withhold sort of the proprietary information in the face of proxy variables or with the use of proxy variables?

ZIAD OBERMEYER: In here, let me just make sure I understand these are the proxy variables that you need when you don't have access to someone's self-reported race. The ones that can be imputed using like the Consumer Financial Protection Bureau method or things like that, are those the proxies you mean?

LERONE BANKS: Or proxies that are defined internally by the organization itself. If there are different contexts, then maybe you can talk about sort of where the difference lies.

ZIAD OBERMEYER: Yeah, so in a lot of our work, for example, it's very common for health insurers actually not to have data on the race of the people that they're insuring. And so in those settings, I think there are two sets of solutions, and then I think this holds whether they're external variables like race or internally defined proxies. One solution is that you can often actually get those data if they're important.

So for example one health insurer we're working with is requesting self-reported race information on their insured population from the health care systems that they're reimbursing for care. So the health systems have those data because they can ask the patient directly. And if the insurer wants it, they can request that and emerge that into their data.

There are also a lot of places where you can purchase those data and merge them in. So just like you can purchase someone's credit score and merge that into your data set at some cost, you can do the same for race from a variety of external sources. And so I think those two options are actually both somewhat underrated.

I think historically, we haven't prioritized getting these information, and it's almost like sometimes companies don't want to know because they're under the impression that oh, if I don't know about disparities, I can't be held accountable for them. And I think from my involvement in some civil investigations that I unfortunately can't talk about, I can assure you those companies that is not the case. And I don't think that's the case at the Federal level either.

So I think those two are really important to flag. On the other hand, the Consumer Financial Protection Bureau also has endorsed a method of imputing someone's race, for example, based on a combination of zip code and other demographic information that you do have. And so I think that's a reasonable alternative in cases where you don't have the real variable and you have to rely on proxies.

LERONE BANKS: And I think that's great, and I really want to just sort of confirm my understanding of what you're saying, which is that the use of proxies shouldn't necessarily be a limiting factor in an organization's ability to audit its algorithm for bias.

ZIAD OBERMEYER: Yeah, I think that's correct because it's certainly those proxies even if they're imperfect, are certainly going to give you a readout. They're going to be correlated to the real variables of interest with all the caveats, but they're not exactly right. And I think, from an optics point of view, one thing I found is that regulators genuinely just want biased algorithms not to be used. At least the ones that we've been working with, there hasn't been like a punitive or unreasonable standard. And so if you're a company and you're making good faith efforts to understand the amount of bias in your algorithms and reduce them, I think that goes a long way, and those proxies can certainly help with that.

LERONE BANKS: All right. Well, Ziad, we can go on about this for hours, which we've nearly done in a few of our previous conversations. I really want to thank you again for taking the time to present to the PrivacyCon Community. And let's see. I think next up we have a short break, and we'll reconvene with the next panel at 10:55. Ziad, thank you again very much. It's been great talking to you.

ZIAD OBERMEYER: Thank you so much for having me.

DANIELLE ESTRADA: Hello and welcome to panel 2 of PrivacyCon 2021. My name is Danielle Estrada, and I'm an attorney in the Division of Privacy and Information Protection at the Federal Trade Commission. I'd like to welcome you to this panel entitled Privacy Considerations and understanding.

We look at issues like how do we ensure that users find and understand privacy notices, how do we understand and measure their privacy choices, and when they are affected by data breaches, what can we learn from their responses. I'm joined today by a group of distinguished scholars who will be presenting their research addressing these different ways to measure and understand user awareness of privacy policies, data breaches, as well as different approaches to improve user decision making and increase awareness. You will hear from Nico Ebert of Zurich University of Applied Sciences presenting the paper Bolder is Better, Raising User Awareness Through Salient and Concise Privacy Notices, Siddhant Arora of Carnegie Mellon University presenting Finding a Choice in a Haystack, Automatic Extraction of Opt-Out Statements from Privacy Policy Text, Cameron Kormylo of Virginia Tech presenting his paper Reconsidering Privacy Choices, the Impact of Defaults Reversibility and Repetition, and finally Peter Mayer of Karlsruhe Institute of Technology presenting Now I'm a Bit Angry, Individuals' Awareness Perception and Responses to Data Breaches that Affected Them.

As before, if you have questions for any of the presenters, please remember to submit them via email to privacycon@ftc.gov or via Twitter a hashtag PrivacyCon21. I will be asking each of the presenters questions after they present their papers and then open up discussion amongst the group at the end if we have time. Finally, I encourage you after the presentation to go to the PrivacyCon 2021 page at ftc.gov to access their full papers. With that, I would like to turn it over to Nico Ebert to present his research.

NICO EBERT: Thank you very much, Danielle. My name is Nico Ebert from Zurich University of Applied Sciences or ZHAW. I'll be talking about privacy notices, probably one of the most boring topics in the whole world. But I'm going to hope to show you that it doesn't have to be this boring at all. This is work we have conducted together with Kurt Ackerman and Bjorn Scheppler, and our paper is titled Bolder is Better, Raising User Awareness Through Salient and Concise Privacy Notices. So next slide, please.

And the question is like is it possible to raise privacy awareness with short privacy statements? We have all seen these kind of like short notices we might not like actively have looked at them. But companies have started to use them like, for example, Apple and Apple Pay has these small notices in their apps.

Recently, WhatsApp has used short text hints in their app when they change their general privacy terms and conditions, or let's say they try to change their privacy terms and conditions. And the question is do these work in any way? Do customers perceive these kind of short notices in any way, or do people just ignore them like they ignore traditional long legal privacy policy statements that are legally required?

So is this more effective than what we had before? That's a question. And in order to answer this question, we did an online experiment.

To the next slide, please. Located in Germany, we created a fictitious fitness tracking application, which looked pretty real and we asked participants in an experiment to give us feedback to this fitness tracking app. What the participants didn't know at the time was that we put in privacy notices like very short privacy notices in different ways.

To the right, you see what was inside these notices, which text was inside these notices, and they were embedded in this app. We had about 2,000, more than 2,000 participants that used our app with these notices deployed in the app. So we changed different things with regards to these privacy notice.

And the first thing, and this is what brings us to the next slide, was a level of saliency. So we changed it in three different ways. We had these short notices with privacy information just hidden behind a link. This is still very common in practice that you have to click a link in order to get to the privacy information.

We call this policy via click. We made an exclusive presentation, meaning that every user would have to see the privacy policy. So basically everybody should have clicked through the app and would have seen the privacy policy.

And in the last design, we had users that saw privacy information just below at features. So we call this embedded. So whenever we had a specific feature of the app, we had the related privacy information next to the feature, which is very comparable to the way Apple did it with Apple Pay in my introduction. So that was one thing, that was how salient our privacy information was embedded in the app.

And the second dimension, next slide, please, was the level of risk of the information. So what we did was we had like a very privacy friendly version of our privacy policy, and we had a very aggressive privacy intrusive version of our privacy policy, the one that probably no company would ever use if they are not forced to do so, which for example, had stuff in it. This app records everything you do with your microphone, this app stores your location data for ever, this app saves your listening habits, songs you listen while running, and if they are pirated, it's directly reported. So this was very aggressive text because that was our intention or a hypothesis.

Maybe nobody will ever read those texts. So let's at least try to make them very aggressive, very privacy intrusive to see if we can actually have some kind of reaction. Next slide, please.

So in the end, we ended up with a 3 by 2 design, meaning that we had these more than 2,000 participants and we assign them to different groups. So we had one group that saw the low risk policies, privacy friendly policies, and one group of these high risk policies. And then we had the subgroups where we had policies hidden behind a link where we had this exclusive presentation where we had this embedded policy. And then as a seventh group, we had a control group where there was no policy text at all included.

As I told you, we didn't tell the people that it's about privacy policy text, but about testing the app. So we then did some destruction questions, asking them how did you like our app, would you recommend it to your friends, and everything. But then suddenly, we asked people for recall. Do you recall where the app saves your data? Do you recall if the app uses the sensor?

So people had to take a little bit of a quiz right at the end of our experiment, and that was basically the essence of experiments in order to see if the stuff was really working, because a lot of previous experiments told people to read the policies, but our assumption is that basically the behavior is very different if it's a more natural, if you're not telling people to read the stuff, but still ask them to recall the information. So in the end, I now show you the results on the next slide. We asked eight questions in total.

People had four possible answers of only which one was correct. And I told you, people could also guess. We had to account for this guessing effect, and that's what our control condition was good for where no privacy policy was included.

So this was some kind of a baseline for guessing. If people really don't remember anything, they would probably have a score of 2.5 correct answers. And on the left hand side, you see the net recall score which already accounts for this guessing effect.

So you could also say this is basically true knowledge. People really remember stuff. So minus is also already accounting for this guessing effect.

And as you can see, in click condition, people don't remember anything. And this is easy to explain because simply nobody clicked on the link out of, as far as I remember, it was like 7, 800 people. In this condition, basically nobody-- I think 16 people clicked the link.

However, in the exclusive condition when it was very bold, everybody had to see it, people start to remember stuff. So for example, in the privacy friendly condition, people remembered two-- could answer two questions correctly. And in the privacy intrusive condition, it was even close to three answers made correctly.

The embedded condition was less effective. So when the stuff-- when the privacy information was embedded below other information, the recall score declined, but still more effective than placing it behind a link. On the right hand side, you see the time that participants spent in the conditions and you see that they actually spent more time in conditions where the privacy information was presented in a more salient way, which demonstrate that people actually spend time reading the information, which explains the recall that we saw in the record score, which brings us to the very last slide.

So what did we learn as this experiments? Basically, we conclude that these concise and very short privacy notices are a very promising approach to increase user awareness in terms of recall. Saliency has a very huge effect on the awareness of the data practices that is measured by a recall performance at our experiments.

So if you make it not very salient at all, there's no effect. If you make a highly salient, you have, what we would say, huge effect giving or taking into account that basically nobody or a lot of people are probably not even interested in this information. So saliency has a big effect. Making it like bold is better than just embedding it and very much in compliance with our expectation if it's risky then people we call it better.

We have chosen a very specific context. It was also just a lab experiment. So it's not a field experiment that would have to be done in the future. Really trying it out.

But you could also say that our conclusion is that this result is very similar to what probably people in marketing research would confirm. So it is possible to basically make it relevant information like perceived by people. So it is not a natural law that privacy policy and the information that's inside is not perceived by people. So it is possible if you really wanted to do this to present them in a form that is perceivable by the people. Thank you very much.

DANIELLE ESTRADA: Thanks, Nico. I wanted to follow up on-- to start on your discussion of brevity and sort of how short policy texts can be useful to create better privacy awareness. Can you elaborate on that and how that can help users?

NICO EBERT: Yes. Maybe we can switch to the slide. Again, I'm not sure if the slide is still open. I'm pretty sure that you all have seen this WhatsApp. I mean I guess many of you are using WhatsApp.

They have exactly used the same approach. So they picked up information they consider relevant, and I'm pretty sure a lot of people were able to perceive these kind of information if presented in this form. It's basically still a big challenge what information to pick and what information you can choose.

In order to present it, you cannot simply compress your 10 page privacy policy into like 5 sentence. That's for sure. So one of the main challenges is going to be what information is relevant to the people if you want to use these formats.

And also people in marketing research have answered this question. So it requires continued research. Probably also regulators have a say what is relevant. But basically you have to discover now what's relevant in order to display this in an adequate textual form.

DANIELLE ESTRADA: And sort of following up on that, have you found in your own research how consumers decide what information is relevant in these texts?

NICO EBERT: Yes, so one thing that immediately came out of this paper is obviously information stuff that is risky or that is potentially risky is considered to be relevant. And previous research also of ours has shown that it's mostly to do with third party data sharing. That's, for example, one classical risk that seems to be relevant for what seems to be a relevant concern. That's one example of what people would probably consider as a relevant privacy information.

DANIELLE ESTRADA: Thanks. You also mentioned earlier the issue of icons and the use of icons by organizations, and that's something we definitely have seen a greater increase of in terms of using icons in connection with privacy notifications. Maybe you can talk a little how text could be combined with icons and what your researchers found there.

NICO EBERT: Yeah, so we have our own carriage research about this topic of using icons. But there is a lot of research already. They're starting to get more research on these icons.

And basically, yes, you can combine them. This is also what companies does. But I would say that can use them both ways. You can use them to warn people to get their attention, you can also use them to make a cozy atmosphere so that they probably wouldn't even read the text.

So it's like science on the street that tell you about the temple limit. You could imagine different forms of design with the same information, but with different outcomes. So this really needs investigation because I would argue that you can have any kind of effect like remember this Apple sign of these two shaking hands would be interesting to see if people-- if this already raises trust and probably nobody ever reads the information below anymore. So that's an interesting question to study. But generally, I think it's possible to combine them-- combine text also with icon, and that's an efficient combination.

DANIELLE ESTRADA: Once again, interesting extension of what you've been doing.

NICO EBERT: Yes.

DANIELLE ESTRADA: You've talked a lot about this use of short policy texts and distilling the information within longer policy notices. Do you have a view on whether traditional policy disclosure documents are needed any longer?

NICO EBERT: So my assumption is that they are just needed by law. I'm not a-- I'm not a lawyer. Like the companies I've been talking to would tell me that it's required to have one as it's required to have terms and conditions. But they are aware it's not an effective information measure.

So what you could do is still have your old long policy text required by law, but use more like salient, shorter, user friendly, user understandable ways for this part of information that should be really perceived. So I think we will end up with a combination, and that's, for example, already what Apple and also WhatsApp and Facebook did was just using them in combination with the policies that they have already.

For example, the TikTok would be nice example of having kids friendly privacy policies. They are still, they have a very kids friendly app, but their privacy policy is more loyal friendly although they made it easier already. But I think you can combine it very good with long policy texts.

DANIELLE ESTRADA: And do you have a view on how to enforce or kind of ensure that more companies or more organizations are using these short salient policy texts that you found, that your research found to be effective in reaching consumers?

NICO EBERT: Yes, that's a very good and interesting question. I mean, it's very difficult to enforce this. And there are for sure companies that have an own interest to create awareness of the privacy practices.

If it tends to be or if it's done via regulation, I guess it's getting complex. You would probably need very precise, like for example, also design recommendations on what has to be presented, because otherwise you will always find ways around design ways. Basically, you can beat salience with salience by making some other thing more salient.

That's what we also demonstrated. So if you embedded the text beneath a very nice image of a landscape, nobody will read the text anymore. So basically, if you really want to regulate this kind of topics, you would have to look at other areas of regulation. For example, in Europe, we have discussed nutrition labels that are basically very highly standardized on a pixel level. You would have to do this in order to reinforce this.

DANIELLE ESTRADA: OK. Thanks, Nico. That's all I have for now. I'd like to-- I appreciate your-- this was a very interesting presentation and I appreciate the time taken to answer my questions. I'm now going to turn it over to Siddhant Arora to make his presentation.

SIDDHANT ARORA: Hi, everyone. I'm Siddhant Arora from Carnegie Mellon University. And I am here to give a presentation on our recently published work on Extracting Opt-Out Statements from Privacy Policies. This work was presented [INAUDIBLE] 2020 conference and conducted as part of the usable private policy project. So slide, please.

Our paper concerns opt-out choices. These choices allow users to opt out of companies sending them email communications, targeting advertisements based on their behaviors, and sharing personal information with third parties. But these options are often buried deep in policy text, and many users do not know that they're even there. Our goal in this work is to help these users.

In this work, we want to understand if we can get computer these privacy policies. We have previously had some success in automatically extracting useful information from privacy policies. We ask ourselves whether similar approaches could be used to automatically extract opt-out choices from the text of privacy policies and make them more readily accessible and visible to the end users.

So we are now going to watch a short video motivating the impact of this work. Next slide, please. Yeah, can you play the video?

[VIDEO PLAYBACK]

While many websites offer users choices to opt out of some of their data collection and use practices, most of these choices are buried deep in the text of long jargon-filled privacy policies and are never seen by users. Different privacy regulations grant users the right to opt out of practices relying on the collection and use of their data. This includes the right to opt out of having one's data shared with third parties for different purposes, the right to opt out of receiving marketing emails, the right to opt out of cookies, and more.

But as it stands, most websites don't offer easy and practical access to these choices effectively depriving users of their rights. To help make opt out choices more accessible to users, a team of researchers from Carnegie Mellon University has developed a browser extension called Opt-Out Easy, which uses machine learning technology to automatically find opt-out choices for users as they browse from one website to another. Opt-Out Easy is available to both Chrome and Firefox users. By clicking on the extensions icon, users are presented with Opt-Out links found in the privacy policy of the website that they are currently visiting allowing them to, for example, opt out of analytics or limit marketing emails. Start practicing your right to privacy with Opt-Out Easy today.

[MUSIC PLAYING]

[END PLAYBACK]

SIDDHANT ARORA: The major research contributions of our work are the following. We built machine learning classifiers to automatically extract these opt-out choices from the privacy policies. We built a browser extension to suit the opt-outs for a given website. The browser extension is now publicly available and can be downloaded from the links shown on the slide. Another benefit of the automatic classification approach presented in this work is that it actually enables people to more systematically analyze opt-out demographics within and across different categories of websites.

Next slide. The privacy policies are presented on webpages, but there are often no standard location for the privacy policies. We built [INAUDIBLE] model that found the page containing a privacy policy for the given website. We were able to obtain 236 web pages containing privacy policies.

We further split up this text of the privacy policy into what we call segments based on [INAUDIBLE]. We relied on links to third party services like DAA and NAI to automatically identify these opt-outs. Before the machine learning classifiers for the remaining hyperlinks that are more difficult to identify as opt-out. To train these classifiers, we manually annotated 2,692 hyperlinks.

Next slide, please. Up until this point, we have discussed about the pipeline we built in order to collect the annotations. Although that information is useful in itself, it is of paramount importance to understand the type of an opt-out. Hence, we decided to do a fine-grained analysis of the opt-out choices.

During the data-collection process, we would annotate each hyperlink with up two data practice categories. These categories were based on the privacy regulations proposed in Europe and US like GDPR and CCPA. Some of these opt-out categories are even required by law. Next slide.

So we [INAUDIBLE] text-based classifiers where we would generate features based on the segment text, DUI of a hyperlink, and the anchor text associated with the hyperlink to automatically categorize a hyperlink as opt-out. In the example that we see in the slide, we can see how the hyperlink text go to ad settings and the surrounding text discussing managing ad preferences can help to classify the given hyperlink as opt-out. Overall, our classifiers we're able to achieve a precision of 0.93, that is 93% of the hyperlinks classified as opt-out were in fact opt-out, and the recall of 0.90. That is classifiers were able to successfully extract 90% of the opt-out hyperlinks. So next slide.

After building classifiers which are able to categorize the opt-outs into different data practices, we wanted to study the demographics of these opt-outs. Hence we performed an analysis on around 7,000 privacy policies. Here are the three questions, which we want to answer. Next animation.

Out of the websites which we analyzed, how many websites had opt-outs? Next animation. We can see that most of the policies do not have any opt-outs which is consistent with the previous findings. Next animation.

What is the average number of opt-outs per website, and how is it related to the popularity of the website? Next animation. So in the graph that we see on the slide, we see that, the mean number of opt-outs based on the Alexa rank of a website, we observe that higher ranked websites had more opt-outs in them. Next animation.

We also wanted to understand the distribution of opt-out categories. Next animation. This graph shows the distribution of various opt-out categories that we have recognized for the top 200 most popular websites.

The distribution of opt-out hyperlinks are skewed with most of the websites providing advertising opt-out hyperlinks. It was also observed that these trends were similar irrespective of the websites' popularities. Next slide.

So up until now, we have discussed ways of finding opt-outs and doing an analysis of the opt-out categories on the web. But our work will have more value when we can provide this technique as a service to the end user so that privacy takes a front seat. In our opinion, the best way we could package the service is with the help of a browser extension.

So we have built an extension called Opt-Out Easy, which would make it easier for people to find and opt out of data practice controls. This extension is publicly available with download link mentioned on the slide, and we encourage you to download it right now. So in this extension, we used an iterative design approach and came up with four important screens in the end. Next animation.

The first one shows you the opt-out practices and the kind of opt-outs for a given website. Next animation. The second screen displays the list of websites you visited and all the opt-out controls, which are associated with that particular website. It will show you the opt-out control, which you have visited in blue, and it proactively encourages you to take action and opt-out of unwanted data practices. Next animation.

We also have help page in the extension that will show users the working of the plugin. Next animation, please. We analyzed the privacy policies offline and stored the results in the database.

We only showed the results of analyzed policy to the users how Opt-Out Easy also allow users to request for websites that have not been analyzed yet. We then done our analysis to populate the result for those websites and show them later. Next slide.

So as we have seen, our technology does a pretty good job at extracting opt-out choices. But how useful and usable is the browser extension that we have developed? To answer this question, we decided to run a human subjects study.

We performed a controlled experiment with eight participants. The treatment group was explained and given access to the browser extension. So we asked users to perform five opt-out tasks on four different website.

This task was to opt out of a data practice category based on the prompt, which the test subject was given. We see that the time taken for opting out in almost all task is much more in the control group than the treatment group. Also, the success rate is higher for the treatment group over the control group. This is because the users get fed up of searching for an opt out and eventually decided to give up. Next slide.

So here are some of the discussion points from our user study. Users are often unaware of the available opt-out choices and sometimes lack the necessities knowledge needed for them to opt-out successfully. The opt-out hyperlinks are often broken and take too much time to respond, which makes the user give up and quit out of the opting process. Due to all these reasons, we believe that privacy laws should put pressure to ensure that services are always available in the form of standardized APIs.

Next slide. So the final takeaway from this presentation are that we have developed techniques that are capable of identifying opt-out texts from privacy policies. We have presented a browser extension, which is available in both Google Chrome and Mozilla Firefox. We encourage you to download the browser extension right now and take privacy in your own hands. Thanks for your attention.

DANIELLE ESTRADA: Thanks, Siddhant. That was a very interesting presentation and very interesting tool you've put together here for us. I want to start by asking you the point you raised a couple of slides ago, which is from a regulatory standpoint as we hear at the FTC, our regulatory agency, what does your research suggest about future regulations for opt-out choices for users?

SIDDHANT ARORA: Yeah, so that's a really good question. So our research makes three suggestions about future regulation of opt-out choices. The first suggestion is that the privacy laws should put pressure to ensure that these services are available in the form of standardized APIs.

Like I talked about earlier in our user study, we observed that not every website offers the same number of opt-outs, and these opt-out hyperlinks are often broken and took too much time to respond, [INAUDIBLE] multiple levels open directions which would finally show that the service is temporarily unavailable. So because all of this waiting period, the users eventually just give up and quit out of this opting out process. So you do all these reasons, despite our classifiers having very high precision and recall, it is often difficult for the end users to opt-out.

So we believe we should not have to rely on machine learning, but these opt-out links should be readily discoverable in the form of standardized APIs. Also once we have these APIs, user would no longer need to do this for website to website, but can always choose to opt out by setting up preferences in a plugin like the Opt-Out Easy browse extension that we have just made.

Our second consideration was about opt-out settings. So our lab has conducted prior research on people's preferences to opt-out practices through qualitative and quantitative surveys. And what was observed was that settings which allow these intrusive practices by default were more burdensome to end users than the settings which are contextualized based on website categories. So that's another interesting direction. And third suggestion was that there needs to be focus on nudging users towards making beneficial choices pertaining to privacy decision making, and our lab has done a lot of research focused on that.

DANIELLE ESTRADA: Thank you. I want to turn now to the tool you've created and find out whether and how you plan to continue developing it and improving the performance of your system in finding and categorizing opt-outs? I know you've tested it to some extent, but it's still a new tool. So how do you plan on continuing to develop it?

SIDDHANT ARORA: That's a really nice question. So our classifiers are currently trained on a corpus of 2,700 hyperlinks. So we believe that increasing this corpus size by manually annotating more opt-out will likely improve the performance of our classifiers.

And also we plan to do future work on additional feature engineering to improve deep performance of our system. Another direction that we can express that currently we use a classifier for data mining whether web page contains a privacy policy and what is the location of privacy policy for a given web page. So improving the performance of this classifier can improve the performance of our end system in extracting the opt-out choices. Also currently, our methodology is limited to extracting opt-out links that use anchor tags.

On manual inspection, we observe that opt-outs can also occur as like non-anchor tags with JavaScript event handers that would automatically redirect the user. So we plan to extend our methodology towards capturing such opt-outs links as well. Another thing that we are currently exploring is that we are going to try following given hyperlink and downloading the page that the hyperlink leads to. We believe that this could also help in detecting if the given hyperlink is an opt-out or not.

DANIELLE ESTRADA: Thanks. It sounds like a lot of interesting avenues to explore there. What kind of analysis is facilitated by your research on automatically identifying and categorizing opt-out? So like what other analysis is born out of your research.

SIDDHANT ARORA: Yeah, so that's a very interesting question. And one of the benefits that we think of this automatic classification approach is that it would actually enable people and regulators to more systematically analyze the opt-out demographics within and across different website categories based on different website popularity's, websites sectors, and so on. We hope that moving forward, this type of systematic analysis will be used to inform public policy debates.

We also believe that our work has a lot of potential in being used for compliance. In particular, with like the introduction of the California Consumer Privacy Act, which requires an opt-out on the sale of one's data, it would be interesting to see if we can extend the approach presented here and do a systematic analysis looking at the present of opt-out hyperlinks focused on this requirement. That is like what percentage of websites are in compliance with this rule? How does this compliance with website popularity and website sectors and so on?

We are also currently are looking into a more extensive study at how sectoral regulations can affect the presence of opt-outs, like US financial organizations are required by [INAUDIBLE] to have these opt-out notices. So that's another direction. And future work might also examine the jurisdiction under which different sites operate and to what extent do these jurisdiction affect the number and type of opt-outs. For example, we are currently looking at the US and German policy for the same website and trying to analyze how do they differ in the number and type of opt-outs and how can this be attributed to the location-specific privacy regulations like [INAUDIBLE] and so on.

DANIELLE ESTRADA: Great. Finally, I just wanted to ask you if you could remind us is your [INAUDIBLE] Opt-Out Easy is available to the public and can I use it now?

SIDDHANT ARORA: Yeah, you can. So the Opt-Out Easy is publicly available as a browser extension which is available in both Google Chrome and Mozilla Firefox, and we strongly encourage you to download the browser extension right now and take privacy in your own hands. Thank you.

DANIELLE ESTRADA: Thank you, Siddhant. That was a really-- that's just a really interesting tool that you've created for people to explore. I'm now going to turn to Cameron Kormylo to present his paper. Cameron?

CAMERON KORMYLO: Hi. Thank you, Danielle. Go to the next slide, please. So as Danielle said, my name is Cameron Kormylo. I'm a third year student at Virginia Tech.

And my co-author on this paper and my advisor is Dr. Idris Adjerid, also Virginia Tech. He studies and has inspired my interest in Economics of Privacy. Next slide, please.

So the problem that we are addressing in our paper kind of arose out of this sort of frightening reality, that is the current state of online consent. So as you can see, I chose my background today to be the Pont des Art bridge in Paris or more commonly known as the love lock bridge. And I felt that this was a pretty good visualization for the state of today's privacy landscape.

So the bridge itself, you can think of as representing one's own personal privacy, and each lock is another decision that needs to be made. Do I turn the key or do I throw it into the sand? Do I consent to some online data practice or not?

And as I'm sure, many of you also in 2014 guardrails from the bridge began to collapse under the weight of the locks and deteriorating the safety and structure of the bridge itself. And as you may expect, consent rates for these decisions are astronomically high and the tools that industry has or regulation has used to kind of prevent this have been largely ineffective. So for example, the ad choices program, which gives users the ability to opt out of behaviorally targeted ads was only using 0.23% of all American ad impressions, and this kind of phenomenon is similarly substantiated by academic research.

And there have been past papers that have seen almost universal acceptance to privacy policies even when they include the naming rights for their first child, access to the airspace above their homes for drone traffic, and sharing all of their data with the NSA. So as you can see, this is kind of a very strong and frightening phenomenon. And the causes of this have been discussed and disputed over the last decade or two.

So some industry professionals and academics cite consumer indifference to privacy concerns maybe people just aren't that concerned about their privacy, or comparatively they have very high valuations for the online services and those valuations can overpower any concern they do have for privacy. However, a significant portion of the research has kind of converged around the idea that most consumers do not actively evaluate the costs and benefits of consenting to these online data practices. And this is especially true for the most important privacy decisions that are often implicit and are difficult to reverse were being covered up by the complexity of the choice presentation. Can we go to the next slide?

So regulators and policy makers have largely taken notice of these concerns. And this is reflected in the enactment of both broad and all encompassing regulatory changes like GDPR and CPRA as well as the more specific federal regulation that is largely championed by our hosts today. And with each new regulation that's passed, we have new opportunities for research.

However, a lot of the current research focuses on the broad policy effects. And it's nice to know that GDPR as a whole has a positive or negative effect. But this only gives lawmakers a small sliver of the story where in reality these primary effects seem to be further broken down and differentiated.

We need to consider specific tenets of the regulation. What parts of GDPR, for example, increased rates of consent? What parts decreased it? And this more specific consideration can then further inform future regulation and allow for a much more detailed formulation of policy. Can we go to the next slide?

So our work specifically isolates three tenants of GDPR and aims to understand their individual effects as well as any interactive effects that they may have within each other. So first we look largely at the change in consent structure required by GDPR article 9. So this requires that consent elicitation is explicit. It bans the use of implicit consent that utilizes a default opt-in structure.

So an opt-in default in this context would implicitly allow for consent while requiring the consumer to make an active change if they decided that they did not want to consent. And in academic literature, the consideration of this type of change in choice presentation is called choice architecture or how the design of a choice can differ when presented to consumers and how these differences impact the subsequent decision making. So default choices, as discussed here, are a very popular tool of choice architecture, and they can take advantage of consumer decision biases and encourage some sort of particular outcome.

So second we consider reversible consent required by article 7 GDPR, and this drastically refigures the structure of consent in such a way that consumers now know that the choice they're making is not permanent. It can be revisited at a later date. And while this is meant to give individuals more control over their privacy, this could also lead consumers to viewing the choices maybe less serious or less pressing, and this could even encourage them to be more lax with their decision.

And finally, we look at a largely implicit change from GDPR that has resulted in consent elicitation being highly repetitive. And we can see this represented by the countless locks here on the love lock bridge. Previously, privacy decisions were explicitly made only occasionally when signing up for new service or creating a social media platform account, for example.

However, as we have all seen, almost every interaction with the website is now accompanied by a cookie banner asking consumers to continuously make content decisions. And this repetition has the potential to further influence consumer choice. It may lead to a sense of fatigue where they give in and consent all the time or it could do the opposite in which they adjust their belief system slightly each time and eventually learn to make a more informed decision.

Please go to the next slide. So given that consideration, we can summarize our research goals as follows. So first, we evaluate the effect of changing choice architecture or specifically, the default consent choices on the outcome of consumer privacy decisions. And then second, we explore how reversibility and repeated exposure impact decision making across these differing choice architectures.

Please go to the next slide. So to study this, we conducted an online experiment that asks participants to make a real privacy decision where they had to decide whether or not to forgo anonymity in the face of a sensitive disclosure. We had roughly 1,500 participants in the study and it was structured as a 2 factor 3 by 3 experiment where 2 factors were choice architecture and the reversibility of the choice.

And then participants took the experiment 3 times with differences only in the context of the disclosures resulting in a panel data structure that allowed for us to consider the effect of these repeated privacy choices. As you can see on the table here, participants were either in a universal opt in structure that default of them into consenting, an active choice structure where they had to explicitly choose to consent or not to consent or a more protective opt-out structure where they were defaulted into not consenting and had to actively change their decision in order to consent. And additionally, they were given either no information as to the reversibility of the choice or they were explicitly told that the choice was either reversible or irreversible. So we'll go to the next slide.

So this is kind of the procedure of our experiment. Participants were told that they were taking part in a number of surveys that had to do with sensitive information such as criminal activity, sexual history, and romantic involvement. So they started by creating a research profile.

In this kind of mimic what a consumer may create when signing up for a social media site or some other online service, it asks for demographic information, such as gender, race geographic location even in the form of zip codes, and then participants were directed to the main treatment. And this is where they were asked whether or not they would like to in essence sign in to the research profile, which would link their subsequent disclosures back to them. So this image here shows a picture of this decision where the choice is reversible.

You can see in that bolded statement we tell them that they can change their decision at any time and it's presented as an active choice where they have to either click, sign in to My Research Profile, or click Continue As Guest with no option defaulted. So after making this decision, they were directed to a survey that had the sensitive disclosure questions that I had mentioned. And if they had chosen to log in, their research ID was listed in the top corner of the page making it very salient that their answers were being linked back to them. And then after the survey, there was a time buffer in the form of a context specific video that they were asked to watch before directing them to the next of the three surveys where the consent decision would be presented again.

Please go to the next page. So to get into our results, so our first research goal was to identify the effects of changing choice architecture. So expectedly, there is a very significant effect of this treatment. So those in the control group, which was the universal opt-in, chose to log in on average 92% of the time.

So it's very significant. An then those in the active choice condition, which is largely the structure that's encouraged by GDPR, participants logged in around 11.5% less around 80% of the time. So this definitely had an effect, not too drastic of an effect, but still very significant. And then last impressively, the protective opt-out manipulation produced a 41% decrease in logging rates with those participants logging in only about half of the time. So you can kind of see the different tiers of protectiveness of these different choice architecture structures, something that we can use going forward to further kind of find a proper balance for privacy regulation.

Can we go to the next slide, please? So perhaps, our most interesting results came from the consideration of reversible consent. So we found that when paired with a protected opt-out, both reversibility and irreversibility have strong negative effects on logging in. So this was incredibly surprising to us given that seemingly opposite constructs-- reversibility and irreversibility had the same directional effects.

So what this tells us is that despite our initial thought that reversibility may make individuals more lax and lead to higher rates of consent, giving the user information on reversibility at all is a signal as to the seriousness of the decision. So this in essence kind of scares the users out of consenting. And similarly, this effect is only found when paired with an opt-out default.

So this tells us that users also recognize that when presented with a privacy protective choice architecture, it's likely due to some sensitivity regarding the decision and the choice should be made with care. So these two effects interact to very strongly influence the log in decision.

Go to the next slide, please. So finally, we look at the effects of repetition across the three iterations of the study. So we can see that the effective repetition interestingly is also dependent on reversibility. So without any information on reversibility, those in the condition that receive no statement, the effect of the opt-in default frighteningly get stronger over time.

So you can see here the coefficients for the constant variable. You can think about that as the average of those that logged in the universal opt-in. So you can see that study one saw 92% of participants logging in, and by the final exposure, they were up to almost 96% of participants logging in. This is kind of the worst case scenario in which a choice architecture that really takes a hold of people's cognitive biases not only has a very significant effect up front, but continues to get stronger over time.

And however, when they were given an opt-in structure and are given information on reversibility, the effect of the default is still strong without a doubt, but it remains constant over time. So the last three columns here you can see that it similarly starts around 92% and that remains constant across the iterations of the study. So this tells us that reversibility and irreversibility counteract the growth of default effects seen otherwise.

And we see a similar pattern for protective opt-out defaults where their effects grow stronger over time in absence of reversibility. But when reversibility is introduced, the effects are kind of all pushed to the forefront where there's kind of a stronger initial impact of opt-outs, but that remains constant over time. So if we could go to the next slide.

So basically, we have to ask what we can take from these results. And largely, what we conclude is that there's a delicate balance between protectiveness and economic benefit. So individually, each change that we enacted had the desired effect of consumers choosing options that may better reflect their privacy concerns. So for example on active choice structure, decreased logins allowed for individuals to explicitly choose which option that they felt most comfortable with.

And additionally informing a user as to the reversibility of the choice can counteract the growth of default effects over time, which is very desirable. However, interactive effects have the ability to produce very large swings in consumer outcomes. So for example, like we said when the effect of reversibility was paired with a protective opt-out that drove much further down than we had originally anticipated.

So these findings provide very specific insight to policymakers. So we don't give answers relating to the broad effects of privacy policy. But what we do is we isolate specific changes and provide a better understanding as to their effects as well as how they interact with other changes. This allows for a much richer conversation around future regulation and is central in striking that important balance between privacy and economic benefit. So with that, I thank you all for listening.

DANIELLE ESTRADA: Thanks, Cameron. I want to start by asking you your study considers the initial choice to consent or not consent to tracking, and maybe you could talk more about what downstream effects this may have on subsequent behavior such as disclosure.

CAMERON KORMYLO: Yeah, absolutely. Thank you for that question. So as I said, the structure of this experiment was first the consent decision and then subsequent disclosures. And we did see some slight increases in disclosure differing by choice architecture.

So for example, those in the protected opt-out default conditions disclosed slightly more than those in the universal opt-in condition. However, these effects were very, very small and were not statistically significant. So this tells us that largely consumers don't intertwine the consent and disclosure decisions, which could be concerning.

We would expect that maybe more laxed privacy settings would result in more trepidation around disclosing sensitive information, but we just don't have the evidence to support that. Now, with that being said, we didn't see effects really in the choice architecture, but we do see some interesting effects relating to reversibility. So participants that were in a condition where the decision was explicitly reversible did disclose almost 20% more than those that were given no information about reversibility. So this could tell us that changes like article 7 of GDPR that require reversibility, they could lower overall consent rates as we saw, but that may have some impact in increasing disclosure downstream. So that's definitely something that needs to be considered when crafting these policies.

DANIELLE ESTRADA: I also was hoping you might be able to talk about the future of your research stream and how you see privacy policy evolving.

CAMERON KORMYLO: Yeah, absolutely. So that's a very important and also very broad question. So first and foremost, I truly do hope that more research is produced that focuses on just a few aspects of privacy regulation in the potential for new and exciting insights to arise and the huge amount of changes that we've seen in the last decade is very exciting.

But in terms of the future of privacy policy, generally I do think that one change we need to see is a refocusing to the individual. So we've made some very important strides in ensuring that companies are behaving responsibly and that consumers have the tools to make responsible decisions. But at the end of the day, the individual remains the decision maker.

So our society very rightly so has addressed so many important issues in the last few years through social change and outreach, and I think the privacy could benefit from being one of those next changes. I'd really love to see the privacy concerns discussed in the panel today, and in other panels kind of become general knowledge among the population. And without that policy, well, incredibly important can only get us so far.

DANIELLE ESTRADA: And also I wanted to-- and that kind of ties to my next question, which is there's been an abundant stream of research surrounding the effects of GDPR and other regulation. But maybe you can talk about how you see your research and where it specifically contributes to these discussions in terms of decisions about privacy policies and research of them.

CAMERON KORMYLO: Absolutely. I think that largely a lot of the research falls into two pools. So either you're a free market advocate and you trust consumers to make their own decisions or you think we need more regulation to protect consumer's.

And both pools have produced groundbreaking findings. They've informed the discussion around data privacy considerably. But what we tried to do differently was kind of remove any pre-existing idea of how privacy should be handled.

So we look specifically at what has changed and how it impacts consumer decision making, and our results kind of reflect this approach. We show that it doesn't have to be one side or the other. There can be a balance between the market and regulation.

And importantly, policy also doesn't have to be one size fits all. So the FTC has shown us that sector and medium-specific regulation can work. For example, maybe health care privacy needs to lean more on the protective side and encourage lower rates of consent through these different architectures. But there may be other realms browsing data online that might benefit from letting the market take more of a role. This is the conversation that I'm hoping to start and hopefully that future research can continue.

DANIELLE ESTRADA: Thanks, Cameron. I share your hope that we can continue this these conversations with further research because these are important issues. And I think as Nico started us out saying, there may be issues that people view as having been researched extensively in the past, there's a lot there's a lot left to explore as all of you have been addressing in your papers.

And I'm excited to see where these conversations go. So thank you for that. Last and certainly not least, I want to turn it over to Peter Mayer who will be presenting his paper. Peter?

PETER MEYER: Yeah, thank you, Danielle. So my name is Peter Meyer, and I will be presenting the work of my colleagues [INAUDIBLE] and myself, and that is our investigation into Individuals' Awareness, Perception and Response to Data Breaches that actually affect them. Next slide, please.

Now, most people here will be familiar with the term data breaches, but let me quickly define what that meant for us in our research. So for us, data breach was an event in which private, sensitive, or confidential personal information is leaked to unauthorized third parties. Such data breaches can cause tangible harm when the exposed data is misused for identity theft or account hijacking. And even if these events have not occurred yet, individuals may experience emotional harm as they feel vulnerable or anxious about exposure of this data and misuse in the future.

Next please. When we look at the number of data breaches and exposed data records over time, we see that data breaches are on the rise. For the United States, we see that there were more than 1,000 breaches each year since 2016 leading to more than a billion exposed records overall.

Yet, despite this large number of breaches, recent research also suggests that affected consumers only rarely take action [INAUDIBLE]. So we wanted to have a closer look with the methodology that was different from what had been done before. Next slide, please.

So prior work primarily asked participants about past experience of breaches in general or asked them to describe intended reactions in hypothetical scenarios. For example, participants should imagine being affected by a specific breach. In our work, we [INAUDIBLE] participants with real world breaches that are known to have exposed the personal information. Therefore, our survey has a greatly increased ecological validity when compared to this prior work since participants were more likely to relate to these features and our work also mitigates recall bias as many participants first learned about the features in our study and provided immediate responses. Next slide, please.

So to achieve this new methodology, we built our own survey platform that's [INAUDIBLE] is a web service that collects data from data breaches and allows visitors to enter the email address on the website and see a list of known data breaches tied to that email address. Overall, 413 participants were recruited off the prolific panel [INAUDIBLE] study and went through the survey in three stages.

In the first stage, we asked participants to provide the most commonly used email address for [INAUDIBLE] API, followed by questions about several properties of email address such as its frequency and purpose of use. In case the participant's email was not tied to any breaches, the participants were given the opportunity to enter another email address, which they believed to be more likely to be involved in breaches. In the second phase, all participants that were affected by at least one breach represented up to three specific breaches from the full set returned by [INAUDIBLE]. For each breach then, we collected data relating to our participants awareness of the individual breach before our study, their perception of causes and impacts of being impacted, the emotional reactions and if they've done or intend to do anything in response.

And in the end we collected the participants demographics and showed them the complete list of known breaches that included the email address to ensure that they're aware of all the risks. And additionally, we provided resources to help participants in taking action and dealing with the potential aftermath of the breaches we showed them. Next slide, please.

So using the data from the survey, we aim to answer five research questions namely, the factors that influence the likelihood of an e-mail addresses' exposure to data breaches, the participants' perception of causes and impacts affected by data breaches, the awareness of the data breaches, the emotional reactions, and the behavioral responses to the data breaches. Next please. So in this talk, we will only-- I will only present results regarding four of these research questions namely question 1, 3, 4, 5. So what did we find? Well, next slide, please.

For the first research question, we investigated the factors that influence e-mail addresses likelihood of exposure, and we found that many participants were affected. Specifically 73% of participants appeared in one or more breaches with an average of 5.4 breaches per participant. Therefore, it becomes immediately apparent that most consumers seem to be affected by data breaches.

Using [INAUDIBLE] some regression, we found that the number of breaches associated with an email address increased by 8% per year of use. While 8% might sound like a rather small number, the figure on the right shows how this effect actually builds up over time. Next slide, please. Regarding participants' awareness of data breaches, we found that participants were unaware of the majority, namely 74% of the 792 breaches they saw during the survey, and they were aware of only 80% of them. Next slide, please.

In research question 4, we found that participants like responses show a low concern for the breaches overall as the median was only somewhat concerned. This sentiment was also reflected in the qualitative data we collected as illustrated with a code on the right here. Now, these two aspects, the low variance and the long concern are actually quite critical. Next slide.

Because in the investigation pretending to research question 5, we found that both awareness and concern are key predictors of consumers taking action in response to a breach. So to sum this up, we found that most consumers seem to be affected by data breaches, but are largely unaware and unconcerned of breaches that effect them. [INAUDIBLE] leads to decreased action taken by consumers after a breach. So what are the implications? Next slide, please.

About 74% of the breaches were unknown to our participants indicates that current ways of notifying consumers about breaches may not be effective. Therefore, we argue that an important aspect of addressing this issue is the need for regulators to push for stricter requirements for breach organizations regarding when and how to notify the customers. Next please.

Ideally, the notification can be delivered in multiple channels such as a written letter and an email or when a customer actually calls in to a company, this can also be an opportunity to make that customer aware and inform about mitigating actions. Using all these channels allows increasing the chance of reaching the affected individual. But the notification must also be understandable and usable for everyone.

For example, this could be made better by including easy to enact mitigation actions. The important bottom line here is that requiring breach notification is not sufficient to reach consumers. It also matters how the information is provided [INAUDIBLE] to make sure people really pay attention, understand the risks, and are motivated to take protective action. So the question is what to do about this?

Next slide, please. And here, we argue that notification alone is not enough and companies should be required to stay involved in helping affected individuals recover from the purchase. Rather than providing free credit or identity services, which has limited preventive protections, regulators should encourage companies to offer data protection tools. One example here is tools that allow creating unique e-mail aliases during configuration. Next please.

For example, sign in with Apple allows users to provide an email address, but they can also choose to hide that, which means Apple will create another e-mail address for the sign in and forward the incoming correspondence to the users with e-mail address. Assigning with Apple and similar tools see widespread deployment. More research is needed to understand motivators and barriers behind adoption of such tools.

But offering these tools in a well integrated way would enable users to protect the data with basically no additional friction in the process. And additionally, having these unique email addresses which allows users to identify which services have leaked or sold the data in case they appear in scam, spam, or fishing mails. Next please.

Further more, [INAUDIBLE] notification methods. For example, integrated into positive management such as Firefox lock wise, which you can see on the right here. Let users learn about the breaches and take the available action in the moment as they visit the breach site or start out their credentials in this [INAUDIBLE]. And so both of these technologies I just mentioned can more fundamentally help consumers manage their online presence and stay secure by offering benefits beyond the context of data breaches, but in particular their. Next slide, please.

And this brings me to the end of my talk. This research was done by my colleagues [INAUDIBLE] and myself. And if you want to check out the full paper, you can find a link and the QR code on the slide. And thank you very much.

DANIELLE ESTRADA: Thank you, Peter. I wanted to-- this is fascinating and it's really interesting to hear your views and your research on consumers' awareness or lack thereof of their data of the many breaches that at least the consumers in your study were affected by, and then the rules that everyone can play to help them and to mitigate the effects. I wanted to start by asking you about what consumers can do to protect themselves and to mitigate the effects of data breaches, and also what consumers can do in response when they do find themselves to be affected by a breach?

PETER MEYER: So the most effective ways to mitigate effect is proactive measures. So being proactive when creating accounts, making a conscious decision about whether I need the account and which date I actually have to provide to create this account. And to protect the data, that is actually needed to create an account.

We have see that there are proactive measures such as the email aliases, and there are several technologies available. Integrated options like sign in with Apple work if you actually have an Apple device. But there are other players in this market. Mozilla has a similar service, and there are others. So using these proactive measures is one of the best choices when you actually have to provide data to protect.

And then as responses to a breach, the most important thing is to first see which data is actually affected because the response depends on which data has actually leaked. For example, if the passwords or if a password is leaked, you should change that password as soon as possible, as soon as you become aware of that breach. But also if that passport was used on different websites, you should change it there as well, because there are attacks that just reuse what has been leaked on other websites. And so it's important to not view this one service done in isolation, but see where this might cause other problems. And this might actually be a good chance if you're creating a new password anyway to adopt a different strategy to manage your online presence, for example, password manager, for example, with these built-in notification options that then helps you to stay on top of things even more.

DANIELLE ESTRADA: Thank you. Those are all really helpful suggestions. I wanted to then turn to the actions that organizations can take to mitigate the effects of breaches. I mean there's so many different layers in this ecosphere of different entities that consumers will interact with. But specifically, what did your research show in terms of breached organizations? What sort of effects can-- or what actions can they take to best mitigate the effects of the breaches that they've encountered for consumers?

PETER MEYER: I think the biggest factor that we identified here that is relevant to this question is the lack of awareness that we saw in our participant sample. And this indicates that companies really need to be more active when notifying their consumers because before I can act, take protective actions, I need to be aware that something has happened.

And so companies should notify their customers as soon as the company becomes aware of the breach. And well, to become aware, they should have a monitoring of their systems to actually see if something gets lost. And then if actually something happens, then they should use every channel they have at their disposal.

The traditional ways have proven to be not too effective. And so on interacting with customers on that particular breach, like if they called in to order something and you know that they have not changed their password yet, this might be a perfect opportunity to call them on it and to really go out there. It might not be really for the company a desirable thing to have a big banner on their front page that says we've been affected by data breach, but it would definitely help make people aware.

And so companies should be more open to take creative approaches to notifying people. And now for organizations in terms of developers, they might also want to integrate these technologies into their tools. For example, allow sign in with Apple or similar technologies support this in the app so that consumers can actually choose this technology. And I think we would like organizations to be more proactive and really notify about all the breaches, not just high risk ones, because it has actually been shown and it's not our research, but related research, that companies that take responsibility and then help people in this situation actually face less severe consequences in terms of lawsuits, for example.

DANIELLE ESTRADA: Thank you. And then finally, I wanted to ask you we've now talked to consumers and organizations, but from a regulatory standpoint, what does your research suggest about future regulations for data breaches, and what you found to be effective or found not to be effective?

PETER MEYER: So I think the most important thing is that organizations need to be nudged if not mandated to be more proactive with the notifications. And not just with high risk ones, but also with any breach that occurs because it has been shown that in court it's often unclear how high risk a breach actually is. And so it makes sense to just notify customers whenever there is a breach. And overall, gets organizations to take more responsibility there and to take creative approaches to help raise awareness about the breaches that occur.

DANIELLE ESTRADA: Thank you, Peter. And I wanted to thank all of our panelists-- Peter Cameron, Nico, and Siddhant today for presenting this research. I think it is all very interesting and novel research on this issue of privacy notices and data breaches, which has been addressed before. These are new and different and innovative ways to explore them, and hopefully will be something that you and others build on in the future.