FTC PrivacyCon 2024 – March 6th 2024

Jamie Hine:

Good morning. Welcome to our eighth annual PrivacyCon. My name is Jamie Hine. I'm an Attorney in the Division of Privacy and Identity Protection. My co-organizer for this event is Dr. Lerone Banks, a Technologist from the Division of Privacy and Identity Protection. Before we start today's program, a few details. Linked on ftc.gov, on the event page, we have the agenda for today's event. We also have a list of biographies and we have updated versions of the papers for you to download. Today we have seven panels, as well as remarks from the FTC chair and fellow commissioners. Following PrivacyCon, we'll make today's event available in a webcast and a transcript, and that should appear about seven to 10 days after the event. As with any virtual event, we ask in advance for your patience. Technology happens, we have a team today to help out with any technological issues, but again, we'll have the webcast after the event, if you're unable to connect or you miss something.

We're live tweeting and commenting today's event. You can follow us at @FTC or you can use the hashtag, #PrivacyCon24. We encourage you to participate and also to follow along. In addition, we welcome any questions. If you have any questions for any of the panelists, you can send those to privacycon@ ftc.gov. We'll be monitoring that inbox all day and we'll pass along any questions to the moderators to allow everyone to be a part of today's discussion. Next, I just want to say thank you to everybody who's participating today. We have 20 presentations and those presentations represent a number of additional researchers who have worked as part of teams to put together the great research that we have today. So we want to thank everybody who's been a part of the work that you'll learn about and engage with today. We hope that everybody has a great day today. It is now my honor to turn the floor over to Federal Trade Commission Chair, Lina Khan. Thank you.

Lina Khan:

Thanks so much, Jamie, and thanks so much to the whole Division of Privacy and Identity Protection for putting together today's event, the eighth annual PrivacyCon. It's just such an honor to be here with you all, and really just want to start by thanking all of our panelists. The goal of this conference has always been to be bringing scholars and researchers and experts into policy conversation with the FTC on issues around privacy and technology and your expertise is really just critical to help us spot challenges before they become crises. So really, I'm just very grateful to everyone who's taking the time to share their expertise with us today. As you all know better than most, data abuses are not abstract. As firms deploy digital tools and automated decision-making technologies across sectors, how people's data is collected, used and retained can have enormous consequences for people's lives.

For example, we've seen how a woman with 16 years of punctual rental payments was denied housing due to a tenant screening algorithm or how a man was falsely accused of shoplifting and ended up spending 10 days in jail and spending thousands of dollars to defend himself, all because the police had relied on an error-ridden facial recognition software. We've seen how the algorithms that decide how much gig platform drivers get paid are now so unpredictable and opaque that drivers equate the experience to gambling in a casino. We've also seen how employer surveillance of workers' every have move, be it their trips to the restroom, their precise keyboard strokes or their participation in efforts to organize their workplace can chill workers from exercising their basic rights. And we've seen how parents across the country, too often have to either surrender to widespread data harvesting of their kids in order to let their kids learn and play online.

I could go on and I'm sure many of the folks here today could too, as the proliferation of artificial intelligence and algorithmic decision making further incentivizes businesses to vacuum up people's personal data, establishing clear rules of the road on data and privacy are more essential than ever. I see this as a moment of opportunity, but also as a moment of tremendous risk. Now is the time to ensure that we don't promote race to the bottom business models that automate discrimination, turbocharge fraud, or further entrench [inaudible 00:04:49] surveillance. As companies move at breakneck speed to deploy and monetize AI, we have a chance to make clear that firms cannot use claims of technological innovation as cover for lawbreaking. That's why it's so important that policymakers in government stay on the cutting edge as this technology develops. Working with and learning from technologists is a key part of how we do this.

And I'm so grateful to the FTCs office of technology who's been embedding with our teams across the board and have gathered enormous expertise in areas like privacy, engineering, ai, user experience researchers, data scientists. Really a phenomenal set of experts who are helping make sure that we can stay laser focused. As this technology develops so quickly. Together our teams are helping ensure that our enforcement approach is flexible and nimble enough to keep up with this moment of rapid technological change. And let me just say, I've been blown away by the work of our privacy attorneys and of the whole division of privacy and identity protection. Their impact in this past year alone, when they've brought groundbreaking cases involving geolocation data, health data, browsing data, facial recognition software, is really difficult to overstate. So my gratitude to the full team for ensuring that the FTC can use all of its tools to best protect Americans.

I'd like to just really briefly share a few high-level principles that are driving our efforts in this area. So first is that our enforcement actions are designed to account for how business incentives are driving unlawful conduct. We've seen AI model training and machine learning are emerging as features that could further incentivize firms to be collecting a lot of user data. And in our case, against Amazon's Alexa for example, we made clear that business incentives to train AI models cannot override existing privacy obligations, and firms cannot say that they have to hold on to data forever, just in case it might be useful for some future AI need. Our remedies are also focused on the various ways that firms can use lawbreaking to enrich themselves, which is why we demand not just the deletion of illegally obtained data, but also the deletion of all models and algorithms that were informed by that data.

Second, our enforcement actions are making clear that selling certain types of sensitive data is presumptively off limits. That's especially true when data can reveal intimate details about people's lives, including where they live, which doctors they visit and the websites they browse. In a series of groundbreaking cases, the FTC has made clear that there is a presumption against selling people's sensitive data and selling it or sharing that data without getting people's permission can be an unfair practice and violate the FTC Act. And third, we are looking upstream to establish liability and pinpoint the actors that are driving or enabling unlawful conduct on a massive scale. As with our actions against the data brokers, the traffic insensitive data, we are looking past the consumer-focused applications and zeroing in on the backend infrastructure that is facilitating the commercial surveillance ecosystem.

In addition to data brokers, we've also investigated and brought an action against an ad platform for its role in facilitating the mass collection of location data for children. Our work will continue to examine these key intermediaries and middlemen. Informed by these core principles, the FTCs law enforcement efforts today reflect an agency that is nimble and flexible to a constantly evolving technology landscape. And at every turn we will continue to rely on input from the academic community, including from the brilliant panelists we have here today. So thanks so much again to staff across the agency, particularly division of privacy identity protection, the folks across the Bureau of Consumer Protection, the Bureau of Economics and the Office of Policy Planning, for putting together such a timely important event. And thanks so much to all of our stellar panelists for taking time to join with us today. And with that, we will turn to our first panel of the day.

Eric Spurlino:

Thank you, Chair Khan for those opening remarks. My name is Eric Spurlino. I'm an Economist at the Federal Trade Commission Bureau of Economics. I am moderating this session along with Tia Hutchinson who's a Technologist at the Division of Privacy and Identity Protection. Our panel has three economists, we have three speakers. We have Timo Müller-Tribbensee from Goethe University. We have Sebastian Benthall from New York University School of Law. And finally we have Bernd Skiera from Goethe University Frankfurt. Timo, if you want to start.

Timo Müller-Tribbensee:

Yes, thank you for your kind introduction. Happy to present our work today. This is joint work with Klaus Miller from HEC Paris and Bernd Skiera from Goethe University. And the topic of our research paper is, Paying for Privacy, we are the means of Pay-or-Tracking Walls. Next slide please. So what we recently see in Europe is the rise of pay-or-tracking walls. So recently Meta with Facebook and Instagram introduced a pay-or-tracking wall. And what a pay-or-tracking wall does basically is, it offers the users two options to access content. The first option is if the user pays, and with this pay option, users avoid tracking and the collection of their data. And in the case of Meta, they also don't receive any advertising anymore. And the second option to use the content or access the websites is to consent to tracking, so the users will be tracked and typically they receive behaviorally targeted ads. If users do not choose one of these two options, they only have the choice of leaving and can't access the content.

So in general, this move to pay-or-tracking walls, the core element is basically that users now have to pay for privacy in Europe. Next slide please. Of course, this move to pay-or-tracking walls raises questions for multiple stakeholders. First of all, there are users who fear that privacy might become a luxury if they always need to pay. Secondly, also from publisher's perspective, it is yet unclear whether this move actually helps them economically or whether so many users leave that they make losses. And of course, from regulatory side, there's a debate in Europe going on about the compliance and especially about those prices. And the problem is, we don't know many things yet about those pay-or-tracking walls. For example, we don't know how widespread they are, we don't know whether those implementations differ, how they are priced, how users react, and what are actually the economic consequences for publishers. And we want to contribute so that those stakeholders can evaluate the appropriateness of pay-or-tracking walls.

And we do this by a series of empirical studies. Next slide please. The first thing that we investigated is how popular our payer tracking was actually. And we found out that the first pay-or- tracking wall was established in Austria already in 2018 by a news website. And right now, as of November 2022, there were already top publishers in Austria, France, Germany, and Italy using those pay-or-tracking walls and making users pay for privacy. Most of these publishers cater news, are news websites, but there are also many that cater other contents such as lifestyle magazines or cooking or business or computer magazines. Next slide please. The various implementations differ, especially in their pricing. So if users choose the pay option to pay for privacy, we find examples of prices as low as 40 cents, but there are also pay options that charge more than 10 euros per month. So there's a variation in the implementations regarding pricing. And further, there's also a big variation due to several bundling strategies.

So there are publishers who basically bundle the pay option, so the privacy preserving pay option with advertising free access, but there are also ones that don't. And then there are further publishers who also bundle their pay option with premium or additional content such as certain articles or sections. So in summary, a large variation actually, so pay-or-tracking walls are not the same everywhere. Please, next slide. Next we were interested in, so how do they actually price the pay option? What motivates publishers and websites in their pricing? And one idea that is out there is that this price of the pay option should be motivated by the alternative option, the tracking option. So the alternative revenue event tracking is possible, and we have access to a big ad price data set that allows us to retrieve ad revenues, and we compared the pay options prices with the foregone ad revenues. And what we found is that the average price for a pay option, for paying for privacy if they bundle it with a free access is around 3.24 euro.

But the foregone ad revenue on average is only 24 cents. So what we can conclude here is in terms of the price motivation, publishers do not really motivate their prices by the alternative revenue when tracking is possible, and instead those prices exceed the alternative ad revenue. Next slide please. Next, maybe how do users actually react? At the beginning I already said, if publishers introduce such a pay-or-tracking wall and those users that do not want to pay for privacy, but value their privacy, might leave. And so we conducted an analysis based on the online traffic data of multiple publishers who introduced such a wall. And what we found is that after introducing such a pay-or-tracking wall, there's actually on aggregate, no decline in online traffic. This result was confirmed by several robustness checks, but we can conclude on aggregate at least there's no decline in online traffic and users do not seem to leave.

Next slide please. So the next question is, do those users now pay or choose the tracking option? And we have had access to a data set of one top publisher in Germany, and we calculated the share for several traffic metrics using this data. And I want to focus here maybe on one metric, the number of unique users. If you look at the whole traffic of this publisher, what we see is the 99% of users choose the tracking option, whereas only less than 1% of all those users choose the pay option and actually choose to pay for privacy. I saw that this share increases over time, but the major result here is only a few users actually decide to pay. Next slide please. So that brings us to the question, what are now the economic consequences of a pay or tracking wall for publishers? And if a publisher is using a pay-or-tracking wall, the direct alternative strategy would be a cookie constant banner. A cookie constant banner with a costless option to refuse tracking. And a similar, like a pay-or-tracking wall. If you constant to tracking, you could use tracking.

And there are two potential disadvantage of a payer tracking wall compared to using a cookie concept banner. The first one is publisher could make losses due to a too low price. And the second one is a publisher could make losses because of lower demand. And for the first thing for the price, we showed in the one empirical study that the prices of those pay options are actually higher than the ad revenue that publishers can generate with the tracking option and also with the costless refuse option. So actually they don't seem to make losses based on the pricing. For the potential losses based on a lower content demand because users decide to leave and not visit this publisher anymore, our other study showed that there's actually no decline in online traffic, so publishers also don't make losses due to demand. We used all this information to exemplarily calculate how much revenue publishers actually gained with a pay-or-tracking wall compared to a cookie consent banner, and we came up with the number of an increase in revenue of 16%.

Next slide please. So let me quickly summarize the major results of our research paper. So what we found out is that many top European publishers already use pay-or-tracking walls. Their implementations differ largely due to diverse bundling strategies such as bundling it with advertising free access or bundling it with premium content. In terms of pricing or the pricing motivation, the prices of the pay options typically exceed the foregone ad revenue that publishers otherwise would generate with the tracking option, and users mostly consent to tracking. So there's no decline in online traffic, but just a few pay, but the majority really chooses to consent to tracking. And using these results, we found that publishers can increase their profits due to a revenue increase of 16%. So let me quickly share some further thoughts on implications about these results. Next slide please. So just to name a few for these multiple stakeholders.

First of all, for publishers, I mean we've seen that pay-or-tracking walls are economically profitable. So at least from an economic view, they should adopt them. However, only a few users pay. So it might be worth to experiment with different prices and a different design in order to attract more users to pay for privacy and pay for this option. From the user perspective, I mean, we've seen there's an increasing price for privacy. So publishers increasingly adopt payoff tracking walls. So this fear that privacy could actually become a luxury is there. And we've also seen that only a few users pay. So that indicates or suggests that at least on the current prices, there's a low willingness to pay for privacy. This is also relevant for regulators in general. I mean, there's a large variation in those implementations and some publishers bundle their pay-or-tracking walls, some don't.

And to ensure a fair competition and an equal level playing field here, it's probably worth to implement guidelines and specify the rules, especially also for the pricing. But we've also seen this, I mean, there's a frequent bundling of the pay option of the no tracking option with other offers, such as advertising free access or additional content. But if we think about unbundling, so offering privacy just as a separate thing, then this may even allow for lower prices for no tracking. So those are some of our implications. Happy to answer questions and thank you for your interest in our research.

Eric Spurlino:

Great. Sebastian, you can start now.

Sebastian Benthall:

Thank you. Thank you very much. This is joint work with Ido Sivan-Sevilla at University of Maryland, and we're talking about Adaptively Regulating Privacy as Contextual Integrity. Next slide please. Contextual integrity, it's a whole theory of privacy and data protection and deserve some comment. It defines privacy as appropriate information flow. It's developed by Helen Nissenbaum who started out as a Philosopher and is now a Technology Theorist. It's not privacy as secrecy about personal information, not privacy as control of personal information. It's not a fair information practices procedural definition. It's really focusing on what positive social goods or purposes result from information flowing appropriately in society, and it's best expressed by sectoral laws in the US like HIPAA and GLBA. Next slide please. So for example, we expect in a medical context, for our medical records to go to doctors under conditions of confidentiality, and that's very well established as an appropriate flow of information.

But we might ask, are queries to a search engine about medical topics appropriately flowing to marketers, even though we may have been given notice and given consent? And many contextual integrity theorists would say, "No, this is not promoting the goals of health and there's a reason to go beyond notice and consent here." Next slide please. So a lot of the contextual integrity approach is trying to imagine an alternative to the notice and consent paradigm, especially for consumer privacy online that's drawing more on the sectoral regulations. But it's a bit broader than that in that it's about three major challenges. One, operationalizing what privacy is good for, rather than focusing on individual harms, thinking about what collective or aggregate goods result from good regulation and privacy. It requires dealing with what are the actual information flows at stake. That includes inferences being done with data and addressing the opacity of information flows. We don't know what's happening in say, a B2B relationship, and these information flows are changing all the time.

And we know that we're not effectively regulating this complexity with the current strategies and regulatory instruments. Next slide please. So scholars have been proposing adaptive regulation which imagines a more empowered, actively learning regulatory agency, rather than a set and forget approach. It's really proactively engaging in these learning cycles. First learning cycles, analyzing new risks and information flows, really building a model of what information should be flowing. Learning cycle two is a real time monitoring. You can think of this as real time econometrics based on that model to see if information is flowing properly. And the third learning cycle, the slowest is to essentially validate the model that it actually is predicting the kinds of social benefits from privacy that we think it's supposed to be doing. Next slide please.

So a major technical contribution of the paper, which is a bit too detailed to go into depth here, is Regulatory CI. A complaint about contextual integrity as a theory is that it's not formalized in a way that computer scientists and economists can work with. We are providing a operationalization of it that builds on causal Bayesian networks. It's a form of structural modeling that allows us to capture things like, "Why is it that an advertiser that observes social media data is able to predict the response to certain kinds of advertisements?" Well, it's because there's an underlying common cause to both of the social media data and the advertising response, which is some personality variable. And this is well understood in the statistical theory and computer science of modeling. We're trying to bring this into an econometrics framework that can be used by regulators. Next slide please.

So for example, using Cambridge Analytica as a well understood case of a privacy violation. We might imagine that the context requires determining public representatives in a legitimate way, and that means preserving voter autonomy. And we can operationalize voter autonomy as a certain independence of their judgment from the moves made by advertisers. So here, high level, the green is the social media users, the yellow is some advertisers. And if the social media users decisions here, the rectangles are independent structurally from the decisions made by the advertisers or the yellow boxes, then there's a certain amount of autonomy that is cashed out in the structural features of this model. But if you start allowing information flows here, it'll actually elicit information flows because it was based on a violation of a Facebook terms of service. For example, from M to the decision making of the advertiser, you allow for a structural dependency. That means that the advertiser starts responding strategically to the moves made by the social media user. That's a very simplified example, but it's paint the picture of what we're accomplishing here. Next slide please. So with a model in place, and we imagine that being developed by a multi-stakeholder process with a lot of inputs from civil society and industry and regulators, that enables us to do realtime monitoring of compliance with the norms of the regulations. So journalists, public interest technologists, et cetera, are already doing the collection evaluation of data, the presentation we just saw as an example of good research in this area. There's also real-time possible to real-time monitoring of privacy policies published in the wild, browser tracking information, et cetera. Acknowledging all that great work that exists, we're imagining a way to bring that together into a systemic, intrusion detection system almost, but for privacy violations based around this metric modeling. Next slide please. So for example, if the presumption is that some observable variables in society, such as the returns on advertising in a political context are supposed to be independent of something else, such as people's decision to use social media. A detection of a correlation there would suggest a structural link. And that might be the sort of thing that triggers further investigation, asking for information from companies, et cetera, et cetera. Next slide.

And then the third learning cycle is a test of whether this whole system is working. And we know that notice and consent did not help the victims in the case of the Cambridge Analytica scandal, preserve their privacy. They couldn't know what manipulation or consent might allow and their consent was violated anyway. So some validation of the overall model is necessary. And that's this third learning cycle. Next slide, please. Obviously, this is a very ambitious proposal. There's a question of who would be ready to adopt it. We've spoken to some European data protection authorities. They find this is an exciting prospect, but there's of course the low level of transparency from technology companies about what information they're actually sharing. It is possible to just regulate them and say, you have to report this to the government Digital Services Act. Digital Markets Act are steps in the right direction here, but it's an unsolved problem in the US.

There's the question of whether regulators want to embrace this kind of economic framework and do the work of building in the technical sensors needed to do the real time monitoring. And then there's the question of, "Can we accomplish a paradigm shift towards contextual integrity versus a status quo of the culture and the law?" It's possible that the Californian CCPA might be empowered to do this sort of thing in the States, but of course there's many, many open questions there. Next slide. Thank you very much, honored to be presenting this work here and happy to answer any questions.

Eric Spurlino:

Great, thank you. Bernd, floor is yours now.

Bernd Skiera:

Yeah, thank you so much for having me here and letting me present our project on Apple's app tracking and transparency in short ATT. Next slide. So let me briefly outline what ATT is all about. What you need to know, on all Apple devices you have an identifier, a unique identifier for advertiser in short, IDFA. And what that identifier allows is it enables tracking, meaning collecting and sharing information about a user over time. And this could be quite useful for firms because they can now profile and target users. And targeting users can usually mean two things, targeting users with content and of course also targeting users with advertising. And for advertising you have to share the idea of a respectively the information and that of course raises privacy concerns. So what did Apple do with the app tracking and transparency? Well, with iOS 14.5, they introduced this ATT prompt from which you can see here in the middle on the right-hand side, which is essentially a banner. And

Bernd Skiera:

In the banner asks whether the user allowed the app to track the activity or the user's activity across other companies, apps and websites. And the user has two alternatives to pick from; is to say no, ask app not to track or say yes, allow the tracking. So that was introduced as I mentioned with the iOS 14.5.

And essentially what Apple did is they moved from an opt-out approach, which was feasible before, to an opt-in approach. So before, right now, I mean the banner shows up and the user has to make an explicit decision. Before it was a bit more implicit or as I mentioned, opt out because what the user could do is the user could opt out on the device level. Next slide.

So what do I want to present here today? I want to present here today the impact of the ATT on two things. Or a bit more broadly formulated, I want to present what happens if you're using an opt-in versus an opt-out approach in multiple countries. And the two dependent variables are essentially the share of trackable users, which is the share of users for which the IDFA is available. And then I want to look at the economic consequences, more precisely at the likely impact on the advertising revenues. So that will be the main part of the presentation in the paper and I here will briefly present it in one slide at the end. I want to also look at the differences of implementing ATT in several countries. I want to look at the differences in terms of the share of trackable users. Next slide.

So this slide presents how the situation with ATT meaning with iOS 14.5 and highers differs from the situation without. The with situation is presented here on the right-hand side, the without on the left-hand side. So let me start with the right-hand side. So what EPI says is that you can only track users if those are adult users. So if you are not an adult user, there's no way that the app will get the IDFA. And then what you have, is that you have what we call implicit consent. That means what you can always do is you can opt out on the device level. But this is what we call implicit consent because the user has to be active in a sense of, has to go to the respective settings and then opt out. So if the user opted out on the device level, there's no way that the app can present the ATT prompt. But if the user didn't do so, then of course what the app can do, they can ask for the tracking requests, meaning they can present the ATT prompt.

So if the app does that, then of course the user has to make a decision whether to say yes or no. If the user says yes to tracking, then it means that we end up with trackable traffic because the user is an adult. The user didn't opt out on the device level, the app asked, and the user said yes, you can track me. And if one of those answers is no, we end up with un trackable traffic. And then of course, looking at the ratio between the trackable in the un trackable traffic, that gives us the tracking rate here defined as the share of users of all users that are trackable. So that's the right-hand side with ATT.

The left-hand side without ATT is fairly similar for the first two stages. You can only be tracked if you're an adult user. You can always opt out on the device level. What differs compared to without ATT? Without ATT, there was no opportunity to explicitly ask for consent. So essentially the blue in the right field, that would be 100%. And then so before ATT, it was feasible if you're an adult user and if you do not opt out on the device level, you have been part of the trackable traffic. And that of course also allows us to calculate the tracking rate, meaning the share of trackable traffic. Click please.

And that allows us to derive a couple of metrics here, the adult rate, the implicit concentrate, the ask rate, the explicit concentrate, and then of course we can derive those metrics and we can also look at differences. Next slide please.

So let me tell you which data we used. What you need to know is that most of the advertising is sold in auctions. So essentially what we have, we have a publisher on the one hand side wanted to sell ad space to an advertiser. And usually they do that by running real time bidding auctions. And usually the publisher doesn't do that on his own. They usually ask what's called an SSP, a supply side platform to help them selling the ad space or the ad inventory, meaning they run the auctions, the SSPs.

And then on the other side for the advertisers, usually they do not bid themselves. Usually they have also a helper, what is called a demand side platform. They bid on their behalf. So if the publisher runs an auction, essentially the publisher sends a bid request to the advertiser respectively, the demand side platform. And then the demand side platform usually bids on behalf the advertiser. And then if the demand side platform wins the auction, they can serve the bid. If they lose, they cannot serve the bid.

So what is the information that we use? Well from the bid request we get information about whether the IDFA is available or not. So that essentially allows us to derive the tracking share or the tracking rate, meaning the share of trackable ad impressions or ad impressions are served to trackable users. That's the one type of data set we are going to use. And then in case that the DSP is bidding, they receive information about what was the price at which the auction ended. So conditionally upon bidding the DSP also receives price data and essentially use that price data to derive the differences between auctions for the prices of auctions for trackable users, prices of auctions for non trackable users, of course controlling for a couple of coverts that allows us to derive the differences in prices. Next slide please.

So here you can see how quickly the ATT was adapted. You see the curves on the left-hand side on the bottom for all countries. And then you see it also separated for 19 different countries. The first red bar is April 21 or 26, sorry, the date when ATT was introduced. The second red bar indicates October 1st. And what you can see that most of the adoption occurred between April 26th and October 1st. So we had an extremely quick adoption across all the countries. And what we're going to subsequently do is that we look at the period before the introduction of the ATT. That's essentially April until April 26. And then we compare with the period after October 1st until March 31st. So approximately a year later. Next slide.

To illustrate our calculations, what I'll do in the following, I just look at two points in time for the United States. I look at April 1st, 2021. So that's a few weeks before the introduction of the ATT. And I look at about one year after the introduction of the ATT, that's March 31st, 2022. Let me start with the right-hand side. What you can see, our data shows an adult rate of 93% and we have an implicit concentrate on the device level of 54%, which means that obviously 46% of all users opted out on the device level. Then we have an ask rate from the apps of 93%, meaning 7% of the apps never asked the user for their consent. So of course they cannot track, but 93% did and then conditionally upon being asked the explicit consent rate is 36%. And if we add that up, we end up with a tracking rate of 17%.

So that's after ATT. If you look on the left-hand side to the respective numbers, we have the same adult rate. We have an implicit concentrate of 79% indicating that 21% of users opt out on the device level and that leads to a tracking rate of 74%. So then, click please, what you can see the 57 percentage point differences. That's the difference between the 74% tracking rate before and the 17% tracking rate after. Click please.

And then what we do is we evaluate those differences in the tracking rate by the price differences between the trackable and the non trackable traffic. So 0.51 means that for the trackable traffic I get a 51% higher price than for the non trackable traffic. If you consider that information, click please, what we can see is that after ATT on the right-hand side, we have a price index indicating the advertising revenue of 1.09 and that's obviously lower than the price index before ATT, which is a 1.38. And if we look at the percentage difference, that's a minus 21%. So that's indicated here.

So this is just two points in time and now you could say, okay, we don't have a control group, but the track of the traffic for Google, for Android phones, that more or less remains the same across the whole observation period. Next slide please. So before I just used two points in time. Now I use the full data set that is available so we could run a regression. What is marked here in red on the right-hand side is the result for the United States. And you can see across all the 19 countries that we observed, the United States has a rather high tracking, sorry, has a rather high drop in the tracking rates. If you move to the next slide, what you can see here is the impact of the IDFA, meaning trackable traffic on the ad price.

What you can see for the US we had the 51%, which I just described. That's a bit higher than the average price increase, which is a 45%, but what you can see is a huge variety across different countries. And if you move on one additional slide, next please, what you can see here is the impact on the advertising revenue. So for the US we have a decrease in the advertising revenue of a 20.55%. And you might recall that meta complained about an advertising revenue loss of about 10 billion US dollar per year. If you use our data, we come very close to that estimate. Move on the next slide.

So this is just very briefly from study two. What we can see is the tracking rates after ATT team controlling for differences across the app. And what you can see is that the United States has fairly low tracking rates compared to other countries, but also for the countries that allow for where you can track more, the tracking rate is far below 50%. So most of the mobile air traffic on Apple devices is nowadays un trackable. So next slide.

So let me come to my conclusion. First of all, I want to outline, we have a lot of heterogeneity in the impact of tracking rates across the different countries. I want to also highlight that the trackable traffic is much more valuable for publishers, respectively advertisers than the un trackable traffic. What I didn't talk much about is whether there are spillover effects from Apple users or Android users in the sense of that Android users become more privacy concerns yielding to less regular traffic.

That's indeed not the case. So that allowed us also in our setting to avoid a different setting. What I would like to emphasize is that also Apple managed to have a rather quick rollout of the privacy policy and I would like to highlight that the design of the privacy policy matter quite a lot. So there's a huge difference between an opt-out and an opt-in approach. And from my point of view that all the races, some concerns whether privacy laws such as GDPRR should remain rather wake regarding the precise implementation of how to ask for consent.

Well, with that, I would like to thank you for your attention and I think leave the floor for our discussion.

Tia:

Thank you [inaudible 00:47:21] Bernd, Sebastian, and Timo for our presentations. We will now move into the Q and A session. So this first round of questions, we would like to go beyond the research and look at upcoming trends across the market. So my first question is to Timo. In your research looking at the payer tracking walls, do you believe that this implementation of consent will continue to spread and will there be widespread adoption?

Timo Müller-Tribbensee:

Thank you for a question, Tia. We generally expect that payer tracking walls will spread. We already see recent examples. I think a couple of weeks payer tracking walls started to expand, for example in Spain, across Spanish publishers. And I think always if a state or a region has an opt-in requirement, then this usually leads to lower constant rates for tracking compared to an opt-out situation. And of course, a payer tracking wall is basically a means to enhance consent rates and publishers are better off with higher consent rates can earn more revenue. So that's why we expect an increasing adoption of payer tracking walls, especially if there's an opt-in regime like the GDPR.

Tia:

Thank you Timo. Our next question will be for Sebastian. So looking at contextual integrity, it grounds itself from a privacy perspective in social spheres and contexts. You've discussed this in your paper looking at healthcare and the financial sectors as examples, but we know that the consumer internet spans many social contexts and it thrives commercially on the reuse of consumer data. So from an internet perspective, how can that internet activity be segmented into context so that CI can be applied and appropriately regulated?

Sebastian Benthall:

Great, thanks. There's several different layers of the stack where that kind of segmentation can be implemented. Mozilla's containerized browsing setup would be one kind of technological solution to actually build into the user agent a sense of different contexts. But there's also a question of communicating that to users. So is there some kind of visual identity that could be provided to users so that they know that they're in one context versus another? I think that the key idea is to rather than have privacy be a matter of individualistic control over a lot of legalese, which is not well negotiated and not well understood, to somehow organize data flows, information flows into well understood sort of sets of expectations that are used across several different applications or websites, et cetera.

Tia:

All right, thank you Sebastian. And moving to Bernd, our question for you is, given that Apple's implementation of ATT has resulted in a significant loss of advertising revenue for publishers, do you think that this loss of revenue will motivate publishers to move access to their content and resource to more of a paywall?

Bernd Skiera:

Yes, I'm pretty sure that will happen. I mean, first of all, most publishers don't make a lot of profits, so now they have less revenue, so they have to do something. And the one way would be to reduce the costs and that could happen by reducing the quality of content. We have seen that to some extent when we look at the effect of ad blockers. And of course the alternative is to move content behind the payroll so users would have to pay all by sudden for the content. And that of course can also have negative effects for society. If you think about what would be the users who can't afford to pay, they might rely on different content, maybe on fake content, and that could have quite a number of unexpected consequences also for the society.

Eric Spurlino:

Great, thank you Bernd. So we want to move to maybe a second batch of questions. These maybe have to do with, obviously since you're presenting at the Federal Trade Commission, we're obviously interested in how we can take the ideas from your papers and apply them to regulatory ideas both with rulemaking and enforcement. So my first question would be for Timo. So I know in your paper you note how the balance between charging for privacy, charging for data sharing is a bit controversial with respect to GDPR and things like this right now. So we're wondering just how can regulatory agencies such as the FTC or other national protection authorities guide entities in striking a balance between maintaining profits while also providing fair options and pricing to consumers?

Timo Müller-Tribbensee:

Thank you for a question, Eric. I mean, that's really a tough question to answer, but I think what's really important also from regulatory perspective is really to gather the knowledge about potential outcomes. For example, through the means of research. So conduct experiments, gather knowledge even before putting out a guideline or putting out policies.

For example, in the case of payer tracking walls, which I presented about, regulators could also try to start experimenting with different prices together with publishers and then they can use this knowledge to come up with better guidelines that help to find a balance for users and also companies. But what I also think is, I mean really through the means of guidelines, they're really important because for example, in the case of payer tracking walls, I guess there's a high uncertainty for companies and there's of course also high uncertainty about this model for users.

So guidelines could help to create a more fair competition and an equal level playing field, especially maybe also in shared markets like the European Union where we have different countries and national regimes, but not always one shared opinion about privacy. I think coming up with a general guideline that's international or for example in the US across all states, that could really reduce uncertainty and lead to a fair competition, I think.

Eric Spurlino:

Thanks, Timo. That kind of transitions nicely into a question for Bernd, when you're talking about how much should we provide guidelines for these companies and things like this. So a question for Bernd would be, should privacy laws and regulations be more prescriptive in defining how safeguards such as consent approaches should be implemented? As you mentioned in your paper, this is a bit of an area where you think there might be some improvement, so I'd be interested to hear your thoughts.

Bernd Skiera:

Yeah, I think so. I mean what you want to achieve is first of all, you want to have an acceptable level of privacy, whatever acceptable means that requires to trade off between privacy and profit. But at a second time you also want to make sure you have a level playing field for competition. So you don't want to have one company using one approach that, for example, in my setting, might lead to more trackable traffic than another one.

And if we also compare what we have with the ATT, I mean with the ATT and the ATT prompt, the user has to make a decision. So I mean we force them to make a decision, but of course even before the ATT, the user could easily opt out on the device level. I mean it's not much more complicated. I mean you can look it up anywhere, but I mean what our results show that with the explicit question, the opt-in approach, we end up with a much higher share of, no, I don't want to be tracked compared to the opt-out approach where the user had to be a bit more active.

And so in terms of the level playing field, I think it's important to avoid that companies gain advantage by claiming or by implementing the privacy of users differently. So I would be very much in favor of having a rather clear guideline. Yeah, I think that that would help maybe to achieve better privacy, but at the same time also also make sure that you still have what I would consider to be a fair competition.

Eric Spurlino:

Thank you so much for your answer, Bernd. So to turn things to Seb, so one of the real world limitations that you highlighted in your paper is kind of the limited technical expertise of the regulators spanning all these different possible contexts. So in your opinion, what kind of additional training or experience or resources would help bolster regulator's ability to effectively utilize this regulatory contextual integrity framework?

Sebastian Benthall:

Thank you for the question. And I realize I have no intention of insulting the existing technologists and people working at the FTC, but I think from my perspective, what's interesting is that there's a lot of very active research now about defining what privacy is technically.

So the technical definitions of privacy in computer science have been dominated by the differential privacy community for a long time and they're starting to see the limitations of that approach and trying to find expanding into more socially meaningful or legal definitions of privacy, but trying to lay that out in computational terms. So that's an active research field.

What I don't see a lot of, and which we're trying to do it is hard, is how do we feed that progress in computational definitions of privacy back into economics, which seems to be the main vector of scientific understanding into policymaking and really trying to figure out how to get the technologists and the economists to work better together in a way that really operationalizes the law and sort of social expectations. It's really a much more synthetic thing that's needed rather than people needing to learn stuff that's sort of further out for their field.

Tia:

Thank you for your answers Sebastian. So with a few minutes left, I'm going to turn to more personal question of what is on the horizon, what is some of the upcoming research areas that each of our panelists are looking at? And I'll start with Timo.

Timo Müller-Tribbensee:

So of course, I mean the topic of payer tracking walls is just starting off. And so we also conduct further research in that direction and look how different prices might affect consent rates. And we also look in the tracking that's going on, on websites. So if you choose the pay option or if you choose the tracking option. It's not always clear that if you pay for privacy that this always means that you are not tracked at all. There might still be some trackers, maybe not that privacy infringing, but that those are directions that we currently look at. So I think it's a very broad topic and we further go into that here often. Thank you.

Tia:

Great. Bernd, same question.

Bernd Skiera:

Yeah, there are two things I'm going to do in the future. So in the project that I just presented, we came up for the US of a price difference between trackable and non trackable traffic of 51%. So in trackable traffic leads to an advertising price that's 51% higher, probably because the advertisers can better target the user. But what we only could do in this project, we could evaluate the tracking rate differences at this 51% for the US. What we couldn't do is we couldn't observe the development of the prices over time. And of course it would be interesting to see, for example, in the case of Apple, we all by sudden have fewer trackable traffic. Does this lead to a stronger increase in the price differences? So therefore we need observational data over time. So that's something we are currently trying to do. So to even better understand the price differences, that's one thing.

And secondly, what we do is we look at compliance. So for example, on the Apple App store as well as on the Google Play Store, what you need to do is as an app, you need to tell the user which type of data you're using, including, for example, the usage of the IDFA of the identifier for advertiser. And what we want to do is we want to look at what the app's claim is actually what they at the end of the day also do. So where's the claim behavior corresponds to the true or to the actual behavior? And if you would see a huge deviation, then we could conclude at least there is to some extent the non-compliance going on, which of course would have all kind of consequences. So for example, for the firms we would see there might be unfair competitions. For the platforms we would see that they might have to do something to ensure compliance is going on. And of course they would also justify a regulator's behavior towards more regulation for platforms in that area.

Tia:

Thank you. That sounds very interesting. Sebastian, finally for you.

Sebastian Benthall:

So for next steps for research, there's really just so much. I mean the proposal we presented in the paper here, leaves a lot to the imagination. A number of areas that I can imagine next steps. One is, suppose there were a way to get reports about data sharing from say technology companies, data brokers, et cetera. Suppose there were the instruments to get that to be reported to a regulator. How should that be reported? Is there a technical machine learning, sorry, machine-readable format for describing information flows between parties that could then be used and combined in order to create a kind of network diagram or model of information flows at large that is still preserving the privacy of the individuals in it. So operating at the aggregate level, there's a question.

Another question is can a good context specific model of information flows, appropriate information flows, and the good social goods that are supposed to come from privacy in that area really be satisfactorily created with a multi-stakeholder group of experts and building a structural model of that domain. And how can any of the particular variables in that model be operationalized and instrumented in real time? So it would be sort of an institution building effort.

Tia:

Great, thank you. And with that, we would like to thank our panelists for their insightful discussion around how entities are gathering consent and how they're monetizing it and the effect that these trends will have on our collective browsing experience. We will now move to our second panel of the day, which is entitled Consumer Attitudes and Behaviors.

Robin "Bobbi" Rosen Spector:

Good morning and welcome to panel two of privacy con. My name is Bobbie Spector and I'm an attorney in DPIP at the FTC and along with my colleague Bhavna Changrani. We will be moderating this panel and today we will hear from Byron Lowens from University of Michigan, Monika Leszczyńska from Columbia Law School, and Klaus Miller from HEC Paris. And Byron, you can go ahead and present first please.

Byron:

Thank you. I'm happy

Byron M. Lowens:

I'm happy to present my joint work among collaborators and I, namely our investigation into individual's awareness, perception, and responses to breaches that affect them. Next slide please. Data breaches have become increasingly common, with a notable upward trend observed over recent years. Since 2016, the United States alone has experienced over a thousand breaches annually, resulting in exposure of more than a billion records. Now this staggering statistics highlight the escalating challenge of safeguarding personal data in the digital age. Next slide please. Now, previous studies have often focused on participants general experiences with data breaches or their hypothetical reactions to such incidents.

In contrast, our study consisted of presenting participants with specific real world breaches that directly impacted their personal information. By following up six months later, we were able to gain insights into the actual responses to these incidents. Using this approach, it allowed us to enhance the ecological validity of our survey as participants could directly relate to the breaches in question. Additionally, it helped reduce recall bias. Many participants were unaware of the breaches until our study and this allowed us to capture their immediate reactions and actions provided a more accurate picture of post breach behavior. Next slide please.

Now to implement our approach, we developed a custom survey platform integrated with the Have I Been Pwned API. Now this web service allows users to check if their email addresses have been compromised in any known data breaches. Leveraging this capability, we were able to query participants email addresses to identify specific breaches that affected them. This strategy allowed us to gain insights through personalized questions about the data breaches directly linked to their email addresses. Next slide please. Now our study was structured into two separate phases, so an initial survey involving 413 participants from which we collected 719 detailed responses about specific data breaches. Now this covered 189 different breaches across 66 types of exposed data such as passwords and physical addresses. Findings revealed that 73% of participants had at least one breach associated with their email, averaging 5.4 breaches per person. Notably, participants were unaware of 74% of these breaches, only recognizing 18% of them. So this results highlights a significant gap in awareness about the extent of personal data exposure. Next slide please.

Now our aim in this study was to understand the behavioral intentions of participants upon learning about a specific breach affecting them. Among the 10 potential actions we explored, changing passwords emerged as the most common response. So preference suggests that exposure of passwords prompts greater concern compared to other types of data. Now conversely, actions such as filing complaints or pursuing legal avenues were less frequently considered. So this may be attributed to the nature of the exposed data, as most breaches involve nonsensitive information, unlike scenarios involving credit card details of government IDs which may elicit stronger reactions. Next slide please.

Now in our follow-up survey conducted six months later with 108 participants, we revisited the actions that participants actually took in response to being informed about the breaches. Now this method allowed us to compare their initial intentions as stated in the main survey with exact concrete steps that they took. Next slide please. Okay, now we're going to look at how intention actually translated into action. So for this, we have these graphs here. So if you look on the left side of each individual graph, you can see in green people who said that they would want to take action. And then in the blue there are people who were unsure and in the tan-ish color are people who indicated that they would not take the action, and the actual action is below each graph. Now, on the right side of the graph with the dark green and the dark red, you can see whether people actually took that action or went against it.

So as you can see from this graph, people are actually doing some stuff, but it's really low effort stuff like reviewing a credit report, checking their accounts. Whereas things that would actually protect them, like enabling two factor authentication or deleting the account or taking legal action, most people aren't doing this, even the ones that said they would do it. And this is very concerning, right? People are not taking proactive actions that we would hope they would be taking even if they say they would do it. So we try to understand why is this the case? Why don't people take these actions? Next slide please. So in our qualitative data, some themes emerged and uncovered some key motivation and hindrances that affected people's responses to breaches. So one of the primary hindrances is the difficulty in taking action, rooted in this unawareness of an account's existence or this unfamiliarity with this breach site.

So this challenge is exasperated by the intricate data and collection sharing web where personal data circulates invisibly among data brokers and advertising networks, often without user's consent. So a notable example from our study is the breach from verifications.io. This highlighted how businesses often share consumer information with third parties such as data brokers and advertising networks even without individual's knowledge. Now this breach underscored the pressing need for greater transparency in data practice, stronger privacy protections to empower individuals in managing their digital presence. Now echoing our initial survey, follow-up responses revealed a pervasive sense of resignation towards data breaches leading to a preference for inaction. Participants expressed feelings of resignation regarding future breaches and a perceived lack of power in mitigating potential damages caused by breaches. Now this sentiment is critical for intervention, emphasizing the importance of fostering a more proactive, empowered stance towards personal data security.

Next slide please. Now one really interesting component that we uncovered in our study is this significant intention behavior gap. So what this gap reveals is there's a disconnect between individuals plans to react to a data breach and their actual actions. So our findings show that while intentions to change passwords or sign up for monitoring services are often followed through, like I mentioned before, those actions that require a little bit more effort like legal actions or account deletions, we see this very wide gap. So remarkably only a fraction of participants acted on their intentions to review credit reports or financial statements illustrating that intentions don't reliably predict actions post breach. So this challenges us to find ways to sort of bridge this gap. So leveraging insights from security research, strategies like reminders, implementation nudges, these methods could help convert intentions into actions. So our study not only sheds lights on the complexity of human responses to digital threats, but it also offers a pathway for encouraging proactive behavior in the face of data breaches.

Next slide please. Now our study highlights essential policy implications for enhancing consumer responses to breaches. Now first, this is going to be critical for companies to improve how they notify consumers about breaches and steps for remediation. Clear, effective communication is a key to empowering individuals. Furthermore, we stress the importance of providing stronger, user-friendly protections. Simplifying security measures can significantly reduce the intention behavior gap observed in our research. So we recommend companies adopt proactive tools that offer a robust defense mechanism, early warnings of security risk, email alias generators, and password managers. These tools can only help mitigate the impact of breaches but also potentially protect them. In essence, advancing these policies will significantly bolster consumer protection and fortify our digital infrastructure against future breaches. Now this collective effort of policymakers and companies is vital in driving these improvements. Last, next slide.

So to summarize, let's dive into some key takeaways that are crucial for shaping future policies and research. First, as we mentioned before, bridge the intention behavior gap. As I noted before, our research reveals a significant gap between what people intend versus what they actually do after a breach. Specific actions like reviewing credit reports or financial statements, we see an increased follow through, which was rare among the other behaviors we studied. And this behavior varied by action, highlighting a need for targeted interventions, as I mentioned before, like using reminders, commitment nudges. These interventions have shown promising, encouraging individuals to follow through with their intentions. So it's going to be essential that we continue to explore this gap, employing both self-report and real-world data to find effective strategies that help people translate these intentions into actions. We also need to consider the effort, budget, and situations of certain individuals.

The effort required for consumers to protect themselves and how their specific situations impact their ability to act is really important for us to acknowledge. So we need to be able to design interventions that reduce effort and cater to individual circumstances. And this can significantly enhance response effectiveness. We also need to address misconceptions. This is key for encouraging proactive responses to data breaches. Many people misunderstand breach notifications and security measures thinking that they're costly when they're actually free, many of them. Also misplaced blame on personal habits instead of organizational security lapses also can misdirect focus. Educating consumers on these points, emphasizing collective actions for redress, and using teachable moments can shift perceptions and promote informed actions all within a framework that avoids this unnecessary fear. Now, developing tools to assist consumers in responses to breaches is crucial. While not all breaches can be prevented, actions like password changes show high adoption and it can mitigate their impact.

However, our study reveals a need for more accessible and effective proactive measures. Simplifying these actions aligns with the protection motivation theory by reducing response costs and encouraging proactive behavior. A significant insight is that heavy email use increases breach reach, yet advising people to use email less isn't really practical. Instead, we need innovative solutions offering a promise of direction. Integrating such tools into widely used platforms akin to signing in with Apple or Firefox Relay can streamline security practices. So future research should explore the effectiveness and user management of these tools, aiming for a broad ecosystem applicability. And lastly, calls for stricter legal requirements on breach notifications are crucial and this highlights this wide gap in consumer awareness.

Enhanced notification laws should mandate comprehensive outreach for all breaches, using multiple communication channels to ensure informed consumer responses. Beyond notifications, businesses should actively support consumers with effective tools for data protections like password managers rather than relying on traditional but less effective credit monitoring services. Regulatory incentives such as fines are key to encouraging businesses to adopt proactive measures and aligning consumer protection with business interests and trust maintenance and legal compliance. Thank you all for attending my talk. And if you would like to learn more about our paper, feel free to click the QR code at the top right. Thank you.

Robin "Bobbi" Rosen Spector:

Thank you so much Byron. And now we'll hear from Monika.

Monika Leszczyńska:

All right. Hi everyone. Thank you so much for having me here and my name is Monika Leszczyńska. I'm an academic fellow at Columbia Law School and assistant professor at the Maastricht University Faculty of Law. In the project that I'm presenting today, I focus on so-called dark patterns. Next slide, please. And you probably encountered many of these dark patterns in your everyday use of mobile applications and web-based services. These are all these different online marketing strategies where companies are exploiting people's cognitive limitations and biases to influence their consumption choices. One example of such a strategy, it's called nagging, is when users are set these constant notifications urging them to, for instance, share that geolocation. Another example that you probably encountered when for instance, booking your flight tickets where you're offered with all these different travel insurance options. And the option that allows you to reject these offers is buried somewhere at the bottom of the website.

This strategy is called aesthetic design modification. And then the final example that I want to provide you with is called roach motel technique, after this very famous commercial of roach baits and this very famous slogan saying that roaches can check in but they never check out. And it very nicely describes how this strategy is actually working, when it's very easy for consumers to subscribe to services, but then it's extremely difficult to cancel them. What all these strategies have in common is that although they do not involve deception, they pose specific risks to consumers. So first of all, they may negatively influence consumer welfare when they make decisions that are not in their best interest. They may also influence consumer autonomy and undermine consumer trust in the overall online economy. Despite all these risks, it's still unclear how exactly the law should react to those strategies, especially if they do not involve deception.

What I do in my study is I look at people's perception and I leverage these insights to inform the answer to this question as to how the law should react to those strategies. More specifically, I examine, I test what are the factors influencing people's perception of online marketing strategies and whether these factors correspond to the elements that the law would look into when evaluating whether those strategies are in compliance with consumer protection law and specifically with unfair trade practices laws that are currently the most available to actually challenge those strategies. Now what I did, initially I identified those key elements under both state and federal unfair trade practices laws that would allow us to challenge those practices. And according to my 50 state survey, I analyzed the state laws and identified 17 states that rely on morality criteria to assess whether a practice is unfair.

This would mean that in order for an online marketing strategy for dark pattern to consider unfair, it would need to violate moral norms. Now, according to the federal law and according to state laws in eight states, three elements would need to be present to deem a practice unfair. First of all, this practice needs to be leading or be likely to lead to consumer injury, to substantial consumer injury predominantly understood in monetary terms. This injury cannot be reasonably avoidable by consumers, meaning that consumers needs to have freedom of decision making, and it cannot be outweighed by countervailing benefits. Now given those elements, what I did in my study, next slide please, is that I wanted to find out which practices people perceive as less morally acceptable to inform the application of unfair trade practices in those states that rely on morality criteria. And I also wanted to find out which practices are perceived as more threatening to consumer freedom of decision-making to inform the application of unfair trade practices laws under the federal law and in eight states.

I also wanted to find out whether moral acceptability is affected by the presence and type of harm, to find out whether the moral criterions actually overlap with this free element test under the federal law and in eight states. Next slide please. So to address these questions, I run an experimental vignette study in which I presented participants with a hypothetical scenario describing a mobile application. Now I don't want you to read this scenario, it's pretty lengthy. I want you to only see how it looks like for participants. So in this scenario, participants read that there is this dating app that they can use for seven day free trial period, and then they can extend it for another 30 days. And I use this moment of the extension of the app to introduce my experimental manipulation. Next slide please. So let me go to the details of the experimental design.

This study was run with over a thousand of participants recruited on Prolific. I made sure to have a US-based representative sample with respect to gender, race, and age. I implemented a three by four between-subject design. It means that each participant saw only one scenario in this study, and I manipulated two factors, the harm factor and the tactic implemented to influence this choice about the extension of the app. When it comes to the harm factor in one set of scenarios, the app was extended for another 30 days and for free.

So nothing else except maybe for taking some space on people's mobile phones happened to the users of this app. Now, in another set of scenarios, the app was extended for another 30 days, but now the users were informed that all personal data that they shared with the application will be now also shared with third parties. In another set of scenarios, users read a notification informing them that the app will be now extended for another 30 days, but it will now cost $9.99 per month. So these were the monetary harm scenarios. Now when it comes to the tactic factor, next slide please.

I implemented four different tactics to influence user's choices. So in the baseline it was like a neutral notification. Users were able to either cancel the application or extend it. In the graphics scenarios, they read exactly the same notification, but now the button that allowed them to extend the app was highlighted in green and the one allowing for a cancellation of the app was barely visible in gray. In nagging scenarios, the notification looked exactly the same as in the baseline. It's just that participants read that users will receive this notification every day. And in the roach motel scenarios, users saw this notification informing them that the app will be now extended for another 30 days by default. And to cancel it, they need to send an email to customer services. Next slide please. I implemented two measures to measure people's perceptions of those strategies and they were divided into stages separated from each other by a couple of days to make sure that people's responses in the first stage do not influence their responses in the second stage.

But they saw exactly the same scenario in both stages. In the first stage, I asked them a couple of questions designed to measure the perceived threats to freedom of decision making and also a couple of questions designed to measure the understanding of the tactic. In the second stage, I had this one question where I asked participants to respond whether they see this strategy as morally acceptable or unacceptable, and I indicated that this means that they personally think that it's a correct or right practice to be used by this application. I also had a couple of questions designed to measure people's demographics. Next slide please. So let me go to the results on the left-hand side here, on this graph you can see the average responses to the threat to freedom of decision-making questions. On the hand side, you see the average responses to the question about moral acceptability, depending on the presence and type of harm.

The red bar represents responses in no harm scenarios, the blue bar in privacy scenarios, and the green bar in monetary harm scenarios. Now what you can see here is that participants perceive the strategies that were targeted decisions that may potentially result in privacy harms as more threatening to that freedom of decision making and less morally acceptable than those strategies that were targeting a decision that was unlikely to result in any harm. Next slide please. Now here on this figure you again see on the left-hand side the results for the threats to freedom of decision making, on the right-hand side, the more of acceptability question. Now here you can see results depending on the strategy implemented to influence people's choices.

What I observed here is that people perceive those graphic strategies when the button allowing for extension of the app is highlighted in green and also the roach motel technique as more threatening to their freedom of decision making, but only the roach motel technique is perceived as less morally acceptable than the baseline treatment. Next slide please. Now here on this graph you see results for all the 12 treatments, but I want you to focus on only on these extreme points. Here you see results comparing the baseline scenarios to the roach motel technique.

What you can see here is that regardless of the presence and type of harm, people perceive the roach motel strategy as more threatening to their freedom of decision-making and less morally acceptable than the baseline treatments. Next slide please. So what are the implications of these results? They will, of course depend on normative framework, right? What do we want consumer protection law and unfair trade practices laws to achieve? But if we want those laws to take into account moral perceptions of consumers, then online practices that lead to privacy harms should also be scrutinized as potentially unfair because these are the practices that people perceive as less morally acceptable than the practices that are unlikely to lead to any harm.

Now, if we want the law to not only reflect or take into consideration people's moral perceptions, but also to protect people's sovereignty, they are sovereign consumer choices, understood as consumers having this freedom of decision making, then both the roach motel technique but also aesthetic design modification should be considered as potentially unfair. These are the techniques that were perceived by participants as more threatening to their freedom of decision making. Then finally, in my 50 state survey, I also observed that in many states, also in those states that rely on morality criterion to assess whether a practice is unfair, for consumers to be able to bring a private action against those strategies, they still to show economic harm resulting from those strategies.

So because consumers also perceived as less morally acceptable, those strategies that result in privacy harms where showing of economic harm might be very challenging, I suggest based on my results that consumers should be given a private right of action also when a practice targets for decisions concerning privacy. This will not only incentivize companies to avoid using such practices but also serve as an expressive function to signal to consumers that this is a value, this is a good that should be protected by law. And thank you very much. Next slide please. And I'm looking forward to questions and comments.

Robin "Bobbi" Rosen Spector:

Thank you so much, Monika. And now we will move on to Klaus.

Klaus M. Miller:

Good morning, and thank you very much for giving me the opportunity to present our recent research using the dual privacy framework to understand consumers perceived privacy violations under different firm practices in online advertising. I'm Klaus Miller from HEC Paris, and this is joint work together with Kinshuk Jerath from Columbia Business School. This is actually the first time presenting this research, so I'm particularly interested in feedback throughout the discussion afterwards or after the event, and I'm happy to be here. So thanks for inviting us. Next slide, please. The online advertising industry has been innovating and developing privacy enhancing technologies to address consumer privacy concerns on the one hand side and to adapt to stricter global privacy laws. So we have the GDPR, for example, in Europe, we have the CCPA in the US, we have Chinese privacy laws, we have privacy laws in India and everywhere.

The goal of privacy enhancing technologies is to better protect consumer privacy, but at the same time maintain economic efficiency from, in this case, tracking and targeting consumers. An example of privacy enhancing technologies in online advertising is the, so-called Google Privacy Sandbox. The Sandbox Initiative consists of several sub initiatives, and the main idea of the Google Privacy Sandbox is to better protect consumer privacy from a technical perspective as the data stays on the consumer's machine. So potentially giving the consumer more control over their personal data. But the firm in this case, Google, is still using the data and targets consumers with the digital advertising at the individual level under the protected audience initiative or at the group level under the topics initiative. And this is basically the background for our study. So we are interested in whether the technical solutions that better preserve consumer privacy also lead to better perceptions of being better protected by these privacy enhancing technologies, or whether consumers still perceive their privacy to being violated or not under these privacy enhancing technologies.

So we use theory, the dual privacy framework, which postulates that consumers have intrinsic and instrumental preferences for privacy to better understand if consumers perceive their privacy to be violated under these privacy enhancing technologies. Next slide please. So for this research, we're considering different practices in online advertising. This ranges from the current industry standard, which is all the way at the bottom, scenario F, behavioral targeting, to basically a world where there would be no ads or no tracking. Untargeted ads, contextual targeting, or various proposals from the Google Privacy Sandbox, which entails group level targeting under the topics initiative and individual level targeting under the protected audience initiative. And these practices in online advertising, they differ with their degree and mode of tracking and targeting. So for behavioral targeting, tracks the user at the individual level and the data leaves the machine. The Privacy Sandbox proposals essentially also track the user at the individual level, but the data remains on the consumer's machine.

So potentially providing better privacy protection. Contextual targeting also tracks the user, but only on the focal website. It doesn't use past browsing data. And of course under untargeted ads and no ads, no tracking, there would be no tracking. And these different practices also differ with regard to targeting. So behavioral targeting would target users at the individual level using past browsing data. Similar to the protected audience initiative. Under group level targeting, you would gain a little bit more privacy because you're able to hide within a group of, for example, people that are also interested in baseball. So you're tracked at the individual level, but your targeting would be in the group. Contextual targeting only targets you based on the context, and of course there would be no targeting under the untargeted ads condition and the no tracking condition. So these are the conditions we are looking at.

And now we are using the dual privacy framework on the next slide to derive expectations on perceived privacy violations under these different practices in online advertising. The dual privacy framework has been proposed by Becker in 1980, and basically it says that privacy preferences consist of two components. There's first of all, an intrinsic component which refers to consumers taste for privacy, basically arises from a consumer's desire to control one's personal information. So information is either private or it's known by the firm, the online advertiser in our context. And then there is the second component, which is the instrumental component. And this refers to the economic consequences of revealing personal information to the firm. It involves the trade-off of the costs and benefits of sharing personal data, and it arises from the firm's usage of a consumer's data. Next slide please. So what we do now is basically we use the dual privacy theory to derive expectations with regard to violations of this intrinsic dimension and the instrumental dimension of consumer's privacy.

So for example, with regard to behavioral targeting, we posted that basically the behavioral targeting practice tracks the people at the individual level and the data leaves the machine. So this highly violates relatively speaking, the intrinsic dimension of consumer's privacy preferences. Whereas the privacy proposals and contextual targeting keep the data on the machine or only use the data of the focal website. So the perceived intrinsic disutility will be lower or it'll be zero when there is no tracking at all under the untargeted or no tracking condition. Then with regard to the instrumental perspective, we again say that instrumental disutility, the usage of the data from the firm is high under behavioral targeting and under the individual level targeting protected audience initiative, because consumers are still targeted at the individual level. It will be medium under the group level targeting topics initiative because we are able to hide in a larger group of other people with the same interests and it will be low under contextual targeting.

Again, because we're only using contextual information on the focal website that we are on, instead of using the entire browsing history of a user. And then under the untargeted ads conditions, there is no cost of being tracked and there could be potentially a benefit of seeing ads. And this depends of course on the perception of the consumer, whether he or she sees the benefit in ads, the perceived disutility will be negative zero or positive, and there is no instrumental disutility under the no ads, no tracking condition. So these are the conditions we will look at in our experiment, next slide, please, to elicit consumer perceived privacy violations.

So what we do is we run several studies, I will here only present our first main study, which is an online experiment in the United States to about 1,700 consumers. We use a between-subject design, so everybody receives only one of the seven treatment groups. We had two for contextual targeting describing how advertising could work in the future. We also ran three follow-up studies, two in the US and one in Europe with statistical identical results. And our main measure essentially is our perceived privacy violation measure, which measures whether a consumer perceives his or her privacy to be violated by the description of the respective practice. Next slide, please. So here is an example of how one of the conditions looked like for group level targeting. You can also see that

Klaus M. Miller:

... that in the paper which has been posted on the FTC's website. Due to time constraints, I won't go into any detail, so we can move on to the next slide and have a look at our main results.

So, here, you can see the perceived privacy violation per experimental group. You can see behavioral targeting at the top, which is the current industry standard. Maybe as expected, we can see the highest perceived privacy violation by consumers, followed by the two privacy proposals, which entail individual-level targeting and group-level targeting, but keeping your data safer on your individual machine.

So keeping your data safer on your device seems to help in terms of consumer perceptions, but it doesn't make any difference whether the firm is targeting the consumer at the individual or group level in the perceived privacy perceptions. What helps is contextual targeting. So not tracking consumers across different websites seems to considerably reduce perceived privacy violations and maybe also as expected untargeted advertising and no ads, no tracking have the lowest perceived privacy violation.

Interestingly, consumers seem to be indifferent to seeing ads or not seeing ads if they are untargeted and there is no tracking. So it's not necessarily the ads that bother the consumers based on our results, but it's the tracking and targeting that seems to elicit these perceived privacy violations.

Next slide, please. So to explore the face validity of our results, we also ask respondents why they provided a specific score for perceived privacy violations under the different conditions. We used these statements for textual analysis, so we conducted a LDA topic analysis, and we identified two distinct topics in line with our theory.

We find one topic which is intrinsic privacy violation or intrinsic disutility and another topic which refers to instrumental disutility. And you can also see the top keywords that are associated with those two dimensions. For intrinsic disutility, the keywords refer to data, feeling protected, data leaving the device, also in line with the theory of keeping data private or releasing the data and, therefore, experiencing those privacy violations. And the instrumental utility talks about being tracked, targeted, and using the data.

Next, slide please. So, to summarize our results, what we did is we asked consumers about their perceived privacy violations on the different practices and new proposals that differ in their degree and mode of tracking and targeting. We saw that PETs such as the Google Privacy Sandbox lower perceived privacy violations relative to the industry standard of behavioral advertising, but the decrease is actually quite small.

Perceived privacy violation is reduced because data stays on the consumer's machine, but there's no difference in perceived violations between group and individual-level targeting. What really seems to drive down perceived privacy violation is contextual targeting, so not tracking across websites. And without tracking, consumers are indifferent between seeing untargeted ads and no ads.

Next slide, please, which leads me to the last slide and our conclusions and implications. So we have arrived at a better understanding of perceived privacy violations for different practices in online advertising and privacy proposals with hopefully important implications for the online advertising industry and policymakers and consumers.

We saw that PETs that technically improve privacy don't necessarily lead to better perceptions of privacy, so we need to work on this, and we need initiatives that basically enhance technical privacy and perceived privacy. And we could also see that the dual privacy framework seems to be useful to form privacy [inaudible 01:42:53] here in the context of online advertising, which may also be suggestive that this framework may be useful in other contexts, for example, when it comes to deriving consumer expectations, for example, with regard to medical data or genetic data or, for example, census data, which are also sensitive areas.

So we hope there will be more research using this framework. And of course, we'll also continue to work on this paper and other papers related around this topic of consumer privacy.

So thank you very much for your attention and, again, giving me the opportunity to present our work here. And I'm looking forward to the discussion and your questions and comments to this paper. Thank you.

Robin “Bobbi” Rosen Spector:

Thank you, all, for those presentations. We will now lead to some questions. We will proceed in the order papers were presented.

First I'll ask questions to Byron about his paper. First, I wanted to ask was that your study found misconceptions about the cause of breaches. Consumers maybe blame their own email habits. I'm wondering if you think that FTC could help in this area with consumer education to correct some of these misconceptions? You are on mute.

Byron M. Lowens:

Sorry about that. I hear I'm on mute a lot when I'm first to get a question. But yeah, there's a lot of things the FTC could do. I think the first thing would be to address these misconceptions through education, right?

So, for example, the FTC could launch targeted educational initiatives or maybe digital literacy campaigns. And these educational initiatives and campaigns could clarify that breaches often result from systematic security lapses rather than individual email habits.

These campaigns could also include easy-to-understand guides on how these types of breaches occur, the importance of corporate responsibility and safeguarding data and practical and realistic steps consumers can take to protect themselves.

I think also leveraging social media platform. Everybody uses social media platforms, so I think leveraging those platforms and interact with the online platforms for these educational efforts can maximize reach and engagement, and this could ensure that consumers are well-informed and empowered to protect their digital identities more effectively.

Robin “Bobbi” Rosen Spector:

Thank you. In terms of the solutions, you had suggested that companies provide options for consumers, such as email alias generators or password managers. I'm just wondering, password managers actually require a fair amount of effort on behalf of the consumer to input all the usernames and passwords.

And considering that your study found that... That requires a lot less time than, let's say, enabling two-factor authentication or putting in a credit card freeze, but you only had 18% follow through on the two-factor authentication and only 4% on the credit freeze. So why would we be pushing these companies to offer these if we think consumer adoption is going to be so low, and how could we potentially increase consumer adoption?

Byron M. Lowens:

Well, my background is in human-centered computing, so where we focus on making things easy to use. So I think to encourage the adoption of these tools, we need to really think about and focus on making that setup process as straightforward as possible, right?

I think working to integrate these services also directly into platforms and tools consumers already use every day could ensure that enhancing their online security is a seamless part of their digital routine, something simple as logging into the computer, right?

But I think by clearly communicating these long-term benefits of using these tools also, and not in terms of just security but also convenience. I think we should make security a priority but also make that convenience a priority as well, especially from some consumers. I work with at-risk marginalized population, so we know there's a technology gap there, and so we need to make these technologies really easy to use.

I also think we should aim to highlight the minimal effort required for significant protection gains and that'll be required to conduct studies and things like that. But also, offering incentives, that would also be great. This could make this transition more appealing for users and provide tangible rewards for taking steps towards better security.

And as I've already mentioned in the study and I reiterated, it's important to educate consumers on the critical need for such measures and protecting themselves. I think if they better understand the risk of not using these tools, it's going to be a key to appreciating the consumer value. So I think through these efforts, I believe we can simplify the adoption process by just making it an easier choice for consumers to enhance their online safety.

Robin “Bobbi” Rosen Spector:

Yeah. And related to that, with respect to the notices of breaches, you talked about notice fatigue, which we all know is a problem. Do you have ideas for how businesses can better inform consumers about these breaches and reduce without contributing to notice fatigue?

Byron M. Lowens:

I think just trying to find the right balance, right? I think this could be done through maybe a targeted and a layer approach to notifications. So the first thing could potentially be to tailor communications based on the severity of the breach and the risk to the consumer and make sure those messages are relevant and actionable, so also using different communication channels thoughtfully and also reserving more direct communication.

Right now, current breach notices are difficult to read, they're difficult to understand. So a lot more needs to be done not just to require companies to send out breach notices but also make sure that people can make use of them, so clearly highlighting what actions are available and which ones are of higher priority.

I mean, I think that's the issue that we saw in our study. These people are doing this low-effort stuff, but the low-effort stuff isn't really the things that's protecting them. And I think educating them on more of the high-effort stuff. And also, a lot of consumers are confused about the meaning and effect of something simple as a credit freeze or a fraud alert and which ones are going to provide them better protection.

So I think clear communication that helps prioritize what actions to take could help a lot here. In this study, we also found that people struggle to make those changes for various reasons, right, usability issues when changing passwords. So these are some of the things that could potentially be addressed.

Robin “Bobbi” Rosen Spector:

Thank you. That's very helpful. Monika, I'd like to ask you, your article states that consumers found some of the design features threatening from a freedom of choice and less acceptable from a moral perspective and positive that those considerations should be part of unfairness analysis.

I'm just wondering, if we consider harm under unfairness from a much broader lens than financial, do you think that we still need to discuss morality in this context?

Monika Leszczyńska:

Yes, so morality is indeed important, nevertheless, for two reasons. One is that it is a criterion that the court would look into under state consumer protection laws when evaluating whether a practice is unfair. So studying whether and what factors are influencing people's moral perception is also important under the state law.

And I think that it's really important to study not only in general whether people are perceiving some strategies more and less morally acceptable but also be more specific here and to look into the specific factors that are influencing those moral perceptions. So, for instance, this is exactly the key finding of this paper is that consumers are perceiving those tactics that are targeting the decisions related to their privacy and personal data as less morally acceptable.

So it shows that those courts that will look into morality criterion can also look into those strategies that are targeting decisions related to personal data, so not only requiring economic harm in those situations. So that's why I think that looking into moral perception from the perspective of state law is also important.

Now, the second reason why it's important is that historically, moral norms, moral criterion was also important under the federal law. Now, under a heavy criticism, we moved away from this criterion because it was considered to be too vague, right, and we moved to a criterion of monetary harm. That was one of the elements for the unfairness standard.

Now it seems that although it has been changing, the federal courts are still kind of reluctant to consider non-monetary harms at least under the Article III standing. Now, the question is still open how they will react to cases involving unfair trade practices laws under the federal law, whether they will be willing to consider those practices as also unfair when they involve non-monetary harms.

And my study shows that, at least according to consumers, this is something that they don't find acceptable, the practices that are targeting the decisions related to their privacy, so not necessarily resulting in monetary harms.

Robin “Bobbi” Rosen Spector:

Thank you. You also recommended that consumers have a private right of action to challenge design features. And how do you think a private right of action can help consumers effectively challenge our patterns and design features and actually result in industry changes and practices?

Monika Leszczyńska:

Yeah, so I see it as one of the different tools that can be implemented to incentivize companies to stop using those practices. I don't see it as the only way. Obviously, consumers, they do also have restricted resources. Class action has also restricted possibility. So I don't see it as the only way to incentivize companies, but I see it as next to public enforcement.

But we also know that public enforcers have also restricted resources. That's why I think it's so important to join actions and to enable consumers to also bring private right of action, but not as the only way to challenge those practices but as a complementary way of challenging those practices to the public enforcement.

Robin “Bobbi” Rosen Spector:

Thank you. And just one quick last question. Are there any other dark patterns or features that you've seen in your work that might benefit from future research beyond the three that you mentioned?

Monika Leszczyńska:

Yes, yes. So there are many of those patterns, and researchers are currently working on categorizing them, identifying different features that it will be possible to classify them.

Just to give you some examples, for instance, companies are using something that is called confirmshaming, so basically blaming us for making specific choices for resigning from their services or not agreeing with their free trials. This is one of the practices that do not involve deception but can be still problematic because it relies on triggering specific emotions in consumers to make them choose and make specific decisions, specific consumption decisions.

Another example is social proof tactic, where companies are referring to how many people are using those services, how many people are using those mobile applications. And this way, they are trying to influence our choices. As you can see, they kind of feel problematic, but it's unclear whether the laws could do something about those practices.

Is it something that we can say, "Okay, it's fine for companies to use that. It's one of their tactics, similar to advertising, to influence consumption choices"? So there's plenty of those tactics, and I think it would be really important to study them further and study also people's perception of those tactics and try to identify, what are the features? What is, for instance, so special about the roach motel tactic that makes people think that this is a practice that is less acceptable? And what is so special about other practices that make them more acceptable by consumers?

So I think that this would be a really interesting future study to really try to identify specific features of those practices that trigger those reactions.

Robin “Bobbi” Rosen Spector:

Great, thank you. And I'll turn it over to [inaudible 01:56:31].

Bhavna Changrani:

Thanks. Klaus, you're up next. I have a couple of follow-up questions for you and your research and your study. Can you elaborate on some of the privacy enhancing technologies and the ones that you think might best offer protections for consumers?

Klaus M. Miller:

Sure. Thanks so much for the question. So, first of all, I think we have to acknowledge that we study consumer perceptions in our study. So we don't study whether these technologies actually protect consumers better. And based on our perceptual measures, we actually find that behavioral targeting has the highest perceived privacy violation. And proposals for better preserving consumer privacy, under the Google Privacy Sandbox, for example, they are able to reduce perceived privacy violations to some extent, but the decrease is actually relatively small.

Where we see a larger decrease is under contextual targeting, which has also been proposed by [inaudible 01:57:36] as a potential replace because we are only relying on first-party data on the respective website that a user is browsing. So if you are on a specific news website, for example, the technology would only be using the data from this focal website, but it would not be using all other data from your past browsing history. So, in that sense, we find it leads to lower perceived privacy violations, and objectively, it will also be a way to better protect consumer privacy.

Bhavna Changrani:

Thank you. Towards the end of your presentation, you noted that initiatives are needed to enhance perceived and technical privacy. What do you envision here, and who do you think is best suited to implement such initiatives?

Klaus M. Miller:

Yeah, I think it should be, first of all, the responsibility of the online advertising industry and maybe second the government. I think what we see in the marketplace is basically these solutions are very engineering-driven. So we are looking for technical solutions to improve consumer privacy, but the consumers don't seem to basically share that perception of that technical improvement. And I think to also give consumers the feeling that they're better privacy protected, we would need educational initiatives to basically communicate what online advertising is doing, how it works, what the firms are, how they are using the data, and how these new proposals are actually protecting consumer privacy better.

So, in that sense, I think we should work towards improving technical privacy, but we should also work towards improving perceived privacy. And ideally, of course, those two would go hand in hand and not only work on one or the other.

Bhavna Changrani:

Got it. And then, I don’t want to put words in your mouth but it sounds like what you're envisioning or proposing is some combination of partnership with government and maybe industry self-regulatory organizations or even businesses to achieve this.

Klaus M. Miller:

Yeah, exactly. We conducted the first study last year in February, and we conducted the last wave after there was a large public announcement of the Google Privacy Sandbox, which we use as an example for those privacy-enhancing technologies.

And there was a lot of buzz talking about and explaining these solutions, but to be honest, I think a lot of this information didn't reach the end consumer. And to be more honest, sometimes it's even difficult for experts to really understand what's going on. So I think this clearly shows that we need to work on this.

And what we see in our data is that the perceived privacy violations did not change, although we had a difference in the broader information environment between those different waves that we conducted. But essentially, our results regarding the perceived privacy violation stayed the same, so they didn't change much. And I think there's a clear signal that we need to do more on the consumer side, focus less on the technical aspects. The consumer is maybe also less worried about the process of online advertising, like the data staying on their machine, but more worried about the actual outcomes.

So if I still see ads or I still have the feeling that these ads are targeted, maybe I'm still worried about my privacy, although my data stays on my mobile phone, for example, and is technically better protected.

So we are basically calling for this consumer perspective to online privacy and also considering consumer measures. And we are kind of proposing one additional measure to get an idea of this perceived privacy violation, which gives maybe some indication on the potential risks of these technologies.

Bhavna Changrani:

Thank you so much. Thank you to our entire panel. Thank you to all three of you for undertaking such interesting and insightful research and presenting it to us and to everyone that's tuned in to listen to us.

We'll be taking our first morning break right after this panel, but please return right at 11:15. Commissioner Slaughter will be starting the session after the break with her remarks. And thank you again for joining us. And we'll leave our three panelists here, and if you have any other follow-up or questions, you can reach out to all three of them. Their information is on our PrivacyCon page. Thank you.

Jamie Hine:

Welcome back from the morning break everyone. We hope you're enjoying the presentation so far. Just want to remind everyone, please online follow along at @FTC. You can also use the hashtag #privacycon24 and any questions you may have for any of the panelists, please send them to privacycon@ftc.gov. It's now my honor to introduce Commissioner Rebecca Kelly Slaughter for some brief remarks.

Rebecca Kelly Slaughter:

Thank you, Jamie. Good morning everyone. As Jamie said, I'm Commissioner Rebecca Kelly Slaughter and I'd like to echo the Chair's thanks to the FTC staff for all their work in putting on this important event. Thanks as well to all of our panelists whose insights help us contextualize and advance our work. I last had the chance to speak at PRIVACYCON in 2021. At the time, I thought virtual conferences might be a passing necessity, but I'm glad our comfort with these tools allow us to hold these convenings and have conversations with distinguished experts from all across the country and the world. I'm excited to see how much privacy scholarship, the industry and our work has evolved since then. In 2021, I hoped we'd continue to take a broader view of the field and examine questions of data use and abuses beyond just the narrow framework of who has access to our private information.

I challenged our participants to help us move away from the outdated notice and consent privacy framework to work to help the commission more closely examine the harms of indiscriminate data collection itself and for us to underline the principles of data minimization in our orders. I'm really excited that we are on the right track since then. This consumer-focused view of our law enforcement mission is well reflected in today's agenda, and I look forward to hearing about the work from all the panels and especially about developments in user interface design, about the intersection of health data and race information, and about advances in machine learning and combating deepfake abuses.

Beyond today's agenda, I'm proud to see that addressing data collection and abuse has been integral to the work of our division of privacy and identity protection over the past few years. I believe that the courts have taken notice too. I'd like to take a few minutes to reflect on some of that work and talk about the watershed holding that our complaint against the data broker Kochava is sufficient to proceed to trial. Beginning with Drizly at the end of 2022, we began to require strict data minimization, collection and retention limitations in many of our orders.

We've now implemented these provisions in orders against ed tech providers like Edmodo and against prison phone companies like Global Tel Link. Putting force behind this principle keeps consumers safe and our data secure. What isn't collected or retained can't be hacked, stolen or otherwise used to harm consumers. We've worked to ensure that sensitive biometric information isn't used as part of a facial recognition system to unfairly target and discriminate against consumers in our Rite Aid order, and we've prevented the sharing of sensitive health information in cases such as GoodRx, Premom and BetterHelp.

The commission has also been active in protecting consumers from data brokers seeking to exploit our precise geolocation information. We resolved allegations against X-Mode, which we allege collected and sold the precise geolocation data of consumers to hundreds of clients, including government contractors. And that company is now subject to a ban on the sale of that sensitive location information. In another case, InMarket, the company used location data to sort people into audience segments that they then sold to advertisers without notifying consumers. The company must now delete or destroy any location information it collected as well as products produced from that data unless it obtains consumer consent or insures that the information has been rendered nonsensitive.

Just a few days ago, our staff published business guidance on our crackdown against mass data collection citing these recent actions. The bottom line of this guidance is clear. Unfettered data collection can easily put you in violation of the FTC Act. When we haven't been able to reach a resolution that protects consumers with a company, we've done the right thing and challenged that behavior in court.

I hope that today's listeners and participants pay close attention to Kochava. We allege that the company sharing of geolocation information violates people's privacy and exposes them to the possibility they'll be victims of stalking, discrimination, violence and other secondary harms. Kochava moved to dismiss our original complaint, which alleged that the company's disclosure and linking of geolocation coordinates to mobile devices invades consumer's privacy by creating a risk that consumers will be targeted based on their visits to sensitive locations such as abortion clinics or places of worship.

The court held that our theories of harm were plausible, but required us to file an amended complaint with more information to assess the viability of our legal claims. We did that and the court ruled and I strongly encourage everyone to read that opinion, resolving Kochava's second motion to dismiss. It is an avalanche of positive developments in privacy and consumer protection law.

After receiving the more detailed information from the commission, the court affirmed that Kochava's practices that exposed consumers to significant risks of secondary harms including stigma, discrimination, physical violence, and emotional distress are a plausible violation of section five of the FTC Act and it affirmed that the unfettered collection of geolocation information itself could represent an injury to consumer's privacy in violation of the law.

The court also accepted our contention that Kochava's acts or practices are likely to cause substantial injury to consumers when we presented evidence that harm has already come to others through the disclosure of their geolocation information from other companies, whether or not we can demonstrate that Kochava's actions have caused that harm. I'd like to celebrate this victory as a vindication of our staff's incredibly hard work building a record that the collection, disclosure and use of sensitive information can represent violations of the FTC.

Equally important, the court's affirmation that we can act to stop prospective injuries, allows us to step in to protect consumers before they're irreparably harmed. Our case against Kochava will proceed. I'm proud of the team for all the work they've put in to protect consumers from data brokers and I'm looking forward to watching the case closely. As you can see, the FTC's dedicated staff have made serious progress advancing the law to protect people's privacy and security. We've still got much to do to protect children's information, to continue to address data collection by social media companies and to create more fairness in digital markets. Thanks again to everyone from the FTC that made today's event possible, especially our agency speakers and moderators and the support staff. I look forward to your presentations and to reconvening next year to reflect on what I expect will be more exciting progress. And with that we'll move to panel three.

David Walko:

Good morning. Welcome to panel three, Privacy Technologies and Design Analysis. I'm David Walko, one of your moderators. I'll be joined by my colleague Ayesha Rasheed. Today we have three panelists, Jane Im, Patrick Parham and Sebastian Zimmeck. We'll first start with Jane. Jane, please proceed.

Jane Im:

Hi everyone. Today I'll be presenting our work on improving findability and actionability of privacy controls for online behavioral advertising. My name is Jane and this is work done with collaborators, Ray, Weikun, Nick, Hana, Lorrie, Nikola, and Florian. Next please. So ad settings are the most basic way for users to have a say over their data, but research has shown that ad settings are hard to discover and find on platforms. Some people in the audience might think, "Well, isn't that obvious due to tech companies' business models?" But the key angle to think here about is regulation. Many regulations across countries say that companies should provide ad controls to users, but the problem is they do not concretely specify how the control should be designed. Next.

So our work is motivated by the questions, "How can we design ad settings so that they're more findable and once designed and deployed, how do they impact users' behavior and sentiment towards the settings and the platform?" Next.

Our study had three steps. The first one was the formative study where we conducted interviews with 20 participants to explore and finalize designs. In particular, we explored two design variables, which I'll explain them soon. So after finalizing the designs, we built a Chrome extension to augment our designs on Facebook. I'm guessing that everyone in the audience is familiar with what a Chrome extension is, but it's basically a piece of software that injects code into Chrome that can change how the website looks or functions. So we built a Chrome extension to change how Facebook's ad controls look. Then we conducted an experiment on Facebook with 110 participants who installed our extension to experience one of our conditions so that we can measure the design's impact on user's behavior and sentiment in a realistic setting.

We did not tell participants about what the new designs are. So in short, participants were able to experience our designs similarly to how Facebook silently rolls out their new designs. Next. We explored the design space of ad controls based on two dimensions, which are entry points and level of actionability. Entry point here means the initial interface a user would click on to find a path that leads to the correct ad setting that they were looking for. In particular, we were interested in the entry points in the main feed where users spend most of their time, such as top of the feed, left menu bar within ads, et cetera. So we explore these entry points in the interviews to finalize what to consider for our main experiment. Another design variable we were interested in is the level of actionability, which means how actionable are the provided options in the ad control interface. Let me explain what we mean by this. Next.

On the left, you'll see a screenshot of Facebook's dropdown menu that surfaces a link to the general settings page that surfaces another link to another page where many of Facebook's ad settings are located at. Platforms typically provide ad controls with low actionability because they provide links to the general ad settings page where the user has to spend time searching the page to find the right setting that they want to use.

In contrast, the right screenshot here shows one of the final designs that we use in our experiment, which I'll discuss in depth later in the talk. It has high actionability because the interface directly surfaces links to specific functionalities. For example, if you take a look at the top option, the interface surfaces links like stop using data from partners to personalize ads. When a user clicks this link, a specific popup that the user was looking for opens up so that they can start taking actions right away.

Through the interviews, we explored the design space and finalized the experiment to have five conditions, so four treatment conditions that vary based on the location of entry points and level of actionability and one control condition, which is Facebook without any design changes. So I'm going to now show the designs to concretely describe what I mean, but I recommend checking the paper to understand how we made the final design decisions. Next.

So the first two designs of our four treatment conditions are what we call as ad menus, one with high actionability and one with low actionability. So for both you can see that the location of entry point is within ads. We made the ad button more noticeable based on user studies and in the dropdown menu we kept Facebook's original options under the section for this ad. The right interface is the ad menu with low actionability because using the Chrome extension we surface new links to Facebook's ad settings page and account settings page under the section for all ads. Compared to this, the left interface has high actionability. It surface links to specific ads and functionalities that we consider the most important based on prior research. So you can see options that says that say "stop using data from partners to personalize ads," "manage ad topics," et cetera.

Next, the third and fourth designs are what we call as feed dashboards. So the location of the entry point is at the top of the feed and the interface is in the form of a dashboard. Just like the ad menu, we tested the feed dashboard with both low and high actionability conditions. Next. Okay, so now before jumping into the findings and describing what this graph means, I first want to quickly summarize what our experiment was like.

To recap, we had four treatment conditions, ad menu with low actionability, ad menu with high actionability, feed dashboard with low actionability, feed dashboard with high actionability. And we also had one control condition, which is Facebook without any design changes. So we considered the control condition to have low actionability because the Facebook's feed surfaces links to general ad settings page. So during the experiment, we asked participants to complete three tasks which involved finding an ad setting about reading scenario-based prompts, for example, in one task, which is what this graph is about, we asked the participant to review advertisers that targeted ads using lists of personal information.

The correct answer was to find a setting called "audience-based advertising." So overall, our findings show that ad controls within ads and at the top of the feed as well as high accessibility increased the findability of ad settings. So when we focus on this graph, the blue color show the ratio of participants who are able to find audience-based advertising versus the red shows the ratio of participants who are not able to find it regardless of whether they previously saw the setting or not.

I want everyone to look at the very bottom row. It's the control condition and we can see that about 23% were able to find the audience-based advertising. But in the feed dashboard with high actionability, which is the very top row, the findability rate was 64%, so there was about a 41% increase in the findability rate.

Next. Another finding that was interesting was that ad controls within ads and at the top of the feed as well as high accessibility both positively impacted user's perception of existing Facebook ad settings. So this graph here shows how complex participants found Facebook's existing ad settings interfaces to be for participants who were able to reach them during the study.

So again, the participants had to find ad settings on Facebook and this graph shows how they perceive them after they found them. Here the orange color shows the control condition and the other colors stand for the four treatment conditions and you can see that participants in the control condition perceived the ad control interfaces to be more complex. I want to emphasize that all participants eventually ended up at the same setting offered by Facebook and our conditions changed how they got there. So this demonstrates how findability and actionability have a substantial impact on how users perceive provided ad settings.

Next. The last major takeaway is that participants did significantly prefer the ad menu compared to the dashboard in terms of usability, although I will note that for both the ratings based on the Likert scale questions were high. So the median were at least four for both ad menu and dashboard across all Likert scale questions. Next. So overall, the takeaways of our study is that first our findings show that clear entry points for ad controls, they do indeed increase their usability.

In particular, I think the strength of this work is that we tested very concrete designs and the ad menu increased the findability of ad settings and users also preferred it more than the feed dashboard, which makes it a promising design companies can potentially adopt. And while the dashboard is more findable, it should be designed in a less intrusive way for users. We also believe that regulators should provide research informed requirements to companies for ad control designs potentially based on studies like this. And lastly, we caution against designs that hide important functionalities under the guise of being minimalist because I think sometimes even experts like researchers are very used to the minimalist designs that platforms often provide us. Thank you for listening.

David Walko:

Thank you, Jane. Our next presentation will be from Patrick. Please go ahead.

Patrick Parham:

Good morning, I'm Patrick Parham. I am a PhD candidate at the University of Maryland High School. Today I'll be presenting a work that my colleagues and I completed that analyzes and compares different definitions of privacy employed by companies designing privacy preserving ad tech, specifically their proposed attribution solution products. If you'd like to read the full work, it was recently published with New Median Society and will be posted following the presentation. Next slide please.

Specifically, our study seeks to analyze the meanings and technical mechanisms of privacy that leading advertising technology, ad tech companies are developing under the banner of privacy preserving ad tech by examining documents where Meta, Mozilla partnered, Google and Apple each proposed to provide advertising attribution services. This clarification of their definition of privacy is of the utmost importance because currently regulators and policymakers around the world are seeking to codify privacy and digital environments.

Meanwhile, ad tech companies are appropriating the term "privacy" and "public relations" and using their positions to encode strategic definitions of privacy into information and market infrastructures. This is a key time in ad tech primarily because the shift away from third party cookie capabilities or the cookieless future is presenting new proposals for attribution services at a suite of other ad tech products that are staking out legitimate boundaries of privacy. It is a critical moment, therefore, to clarify meanings, contradictions, influencing forces and implications of privacy preserving ad tech. We are interested in what definitions of privacy can be reconciled with attributions basic processes. To do so we ask in the following research questions, what do these companies mean when they talk about privacy? How do their solutions differ from each other in terms of privacy and how might each company's approach to privacy reflect political economic factors?

Next slide please. A little bit of background on attribution before I begin for those not familiar. Attribution is a process that documents users engagement with advertisements and connects those records with observed marketplace outcomes. An example would be tying actions such as purchasing an item online, placing an item in a cart or simply clicking or visiting a page. These actions are recorded and associated with or attributed to the ad or media placement that led to the action online the purposes for advertisers and those running their media to measure the cost of their advertising objectives. And this measurement also impacts the ability for non buy side actors such as publishers to generate revenue based on attributed actions generated for advertisers on their pages. Attribution currently uses third party cookies or mobile device identifiers to recognize individuals across sites for apps and conversion pixels to record user behaviors.

The solutions we analyze employ various technical mechanisms in their attempt to preserve privacy. Surveillance and identification are critical to attribution since it is, at its root, a claim about advertising effects. Attribution requires detailed accounting of user behavior to confirm causation and attribution is usually derived from a last click or simply the most recent advertising event gets credit for the action. Increasingly though, advertisers are using a multi-touch approach wherein credit is divided across all the events deemed to have contributed to the outcome.

Multi-touch attribution is a more surveillance process since it implies a fuller inventory of the possible advertising influences on consumption or actions online. With a move away from third party cookies to privacy preserving ad tech solutions, companies are also seeking to maintain existing capabilities while complying with new rules and norms. Therefore, it follows that the practice of attribution raises privacy concerns and raises a key dilemma, what definitions of privacy can be reconciled with attributions basic processes? This leads us to ask whether we should be satisfied that attribution and attribution solutions that claim to preserve privacy. Next slide please.

In order to answer this overarching question in our specific questions, we analyze ad tech's privacy discourse by performing a critical discourse analysis by assembling publicly available documents, detailing the purpose and functionality of attribution solutions from large ad tech companies. These included Google's attribution reporting API, Meta and Mozilla's partnered interoperable private attribution, and Apple's SCAD network and private click measurement.

To clarify, the documents we included were only materials published by these companies and their employees on company sites, company developer blogs and company GitHub pages. Overall, the corpus of documents is 18 texts in total and six for each company. We chose these attribution solutions to analyze specifically because these firms operate large platforms, set commercial terms, generate significant revenue and are positioned differently in the ad tech landscape and can command a significant amount of influence in the ad tech industry. Next slide please.

We coded the documents using an iterative process to generate an initial list of privacy meanings and then further refined and consolidated these meanings. We ultimately arrived at five categories for classifying privacy meanings across these attribution proposals and then we coded statements that articulated or implied privacy meanings or described a method or mechanism for achieving privacy.

In coding, we also noted instances in these documents that referenced trade-offs or tensions between privacy and commercial objectives that these tools are designed to achieve. We also contextualize the findings by considering each company's position in the ad tech industry. The meetings we observed include anonymity, limiting access, anti-tracking, control, contextual integrity. And the privacy mechanisms we observe included differential privacy, local or on device or on browser processing and multi-party computation, data aggregation, obfuscation, encryption. I'm not going to walk through our analysis of the individual solutions detailing where we observed these meanings and mechanisms, but you can find this in the paper or email myself and my co-authors. Instead, I'll discuss our primary takeaways addressing our three research questions. Next slide please.

We see that anonymity, limiting access and prevention of third party tracking are the most dominant privacy meanings invoked in the corpus. Overall, these companies are vague in selective and how they define privacy. Yet they're leveraging the term's positive connotations to justify self-regulatory solutions that will structure data governance relations with users, customers, and competitors going forward.

Here rather than grappling with deeper issues, the privacy preserving attribution solutions we looked at focused larger on complying with the basics of a post cookie world and disavowing the creep factor associated with third party tracking. Notably, the companies exclude from their concerns the ongoing data collection and usage conducted by first parties and they neglect the largest surveillance assemblage of complimentary ad tech suite of products in the larger industry.

Here it is implicit that privacy violations are unsanctioned actions and attribution itself is not questioned. Instead, attribution is positioned simply as a capability to be preserved to serve business needs. It's clear that the term "privacy" is being wielded here without a fully outlined and encompassing definition and these firms would be better served to engage with the theory or a framework of privacy in the design of their solutions. Next slide please.

While the solutions have a lot in common, they do have differences. Overall, we see they were written for different audiences. Meta and Mozilla emphasized general features of the solutions, while Apple details a significant amount of technical info and Google details instruction on how advertisers can incorporate the solution into their workflow and impact on advertising campaign objectives. These differences in how they were written reflect variations in their current implementation status. Apple has not been widely adopted yet. The Meta and Mozilla solution is still perspective while the Google solution is already in use with its other suite of ad tech products.

In addition, each company's market position may help to explain differences in these solutions. Apple and Google who own browsers, devices and operating systems restrict the capture of information to within their own ecosystem, legitimizing the enclosure of data within their walled gardens. This makes their products more valuable to advertisers and excludes third party competitors from accessing the walled gardens.

The partnership between Meta, a social platform, and Mozilla, a browser advances across device and cross browser solutions that would require other browser and device operators to organize around a single standard. Here Meta may be attempting to shift the advertising industry's dependence to the software or application level where it has advantages in scale and reach.

And finally, Apple's emphasis on control via consent and tracking prevention reflects its position as an operating system that can control privacy permissions and also its current minor stake in digital advertising. Apple exploits its position by making requirements of competitors like Meta and at the same time exempting its own first party data measurement from its definition of privacy. Here Apple translates its position into a possible greater market share while impairing other surveillance advertising and sanitizing its own expansion into ad tech space. Next slide please. In our final takeaway, we saw instances where these firms use language of existing theories and frameworks of privacy, such as "contextual integrity" when they promise to prevent cross context tracking of information flows and or respect user expectations. However, what constitutes a context is not defined. For example, in describing its privacy click measurement solution, Apple says it is intended to support privacy preserving measurement of clicks across websites or from apps to websites. It is not intended to be used to track users, events, devices across these contexts. This seems to imply that each site or app represents a context.

Each of these solutions makes a point to distinguish tracking and measurement also. The documents encode tracking with a negative connotation as a way for someone to follow and record individual's activities without justification or consent. But by contrast, measurement is presented as a legitimate necessity authorized by the premise that advertisers need and deserve to know how efficiently their campaigns are achieving sales or their objectives.

Basically, these firms are given license to measure user clicks across context while other players such as ad network's methods of following the causal chain of attribution bars illegitimate tracking. Here attribution implies that marketers are not just entitled to measure audience attention to confirm that their ads are distributed properly, but markers are entitled to measure the effects of advertisements by following audiences beyond sites of ad exposure and into marketplaces where those audiences become active consumers. This is a corporate imposed shift in relationships that require scrutiny going forward. For attribution to be privacy preserving in the sense of comprising-

David Walko:

Patrick?

Patrick Parham:

Yeah?

David Walko:

We're at time. So thank you.

Patrick Parham:

Okay, no worries. Thank you.

David Walko:

Thank you. Next up, we'll have Sebastian. Please go ahead.

Sebastian Zimmeck:

Thank you, David, and it's a pleasure to be here. Thank you so much for the opportunity here. So I'm going to talk about generalizable active privacy choice and GPC. So as you see on the next slide, what's our motivation here? We have a problem. We have a problem with opting out. The basic trade-off of much of the internet, in particular, the web is content against data. I get to see newspaper articles, blog posts, videos, but you get my data for it. And that is recognized for a long time and many laws now in the United States, state laws, the California Consumer Privacy Protection Act, for example, provide now rights to people to opt out to say, "I don't want my data to be sold or shared." And what we were thinking about is how can we actually implement this right on the web? How can people actually do that technically?

And so as we see on the next slide, we tackle this with what we call Global Privacy Control. So Global Privacy Control is essentially the idea that you click a button and you are opted out. And so Global Privacy Control could be a setting in your browser. It's a binary switch that you turn on or off depending whether you want to opt out from a site. And when you visit a site, then you would be opted out if you have turned on the setting.

So the idea is that you don't go to individual sites and make your choices, but rather your user agent or your operating system, for example, if GPC is implemented on mobile devices, makes these choices for you or helps you with these choices. And so as we can see on the next slide, GPC is actually mandatory in California. Some call this here, "regulation by tweet." We were fortunate that the former California attorney general supported us and said, "Yes, if you are addressing California consumers, you need to respect GPC." And that was actually also enforced against

Sebastian Zimmeck:

Against Sephora. That was the first enforcement action that said you did not respect opt-outs via GPC, and so you have to pay a fine for that. We have various other privacy laws coming online. The Colorado Privacy Act is the next. That also said GPC is a universal opt-out mechanism under the Colorado law and needs to be respected. There are other laws in the United States that have these requirements, but GPC could also apply to the GDPR, for example. The idea is that it is really agnostic as to the devices and also to the laws to which it applies. Every regulator themselves can say, what does GPC mean in my jurisdiction? Right? And so we think we will ultimately converge onto a meaning what selling and sharing means, but in principle, we don't specify what it needs. If we go to the next slide where you see substantial adoption already. And so if you look on the left side, what we have is we have adopted GPC in browsers. So for example, in Brave or in Firefox as well as in DuckDuckGo and in various browser extensions. I can recommend OptMeowt. It's done by my students. They have done a fantastic job implementing this here.

And then on the other side, we have the sites that respect GPC, so publishers, retailers, other companies. And so here you can see some of them and they have various implementations. Sometimes you see a little pop-up. We are respecting your privacy choices. I just went to AT&T, for example. Very nicely done. And then if you don't want to do this yourself, you can also use content management platforms. So all the major content management platforms have integrated GPC.

And so what I want to convey here is this is an effort that requires everybody working together. We need to work with the browser vendors, we need to work with the sites, and maybe most importantly, work with the regulators. And so that's why I'm particularly pleased that we have this opportunity here at the FTC.

As we can see on the next slide, when we want to implement GPC, we have a major problem and that problem is usability. So we heard it from Jane and Patrick talking about this in their previous talks. We need to annoy people to some extent and ask them, yeah, you need to make a choice. You need to look at this. And this is a secondary task. People are usually not coming to the web or the internet to do some privacy. They came to do something else. And so we are disrupting, we are interrupting them. And so we want to have this interruption as little as possible, make the experience as smooth as possible.

And so when we display now GPC options, how can we do this so that we can get a meaningful privacy choice, but on the other hand, keep the level of annoyance as little as possible? And so if we look at the next slide, what we will see is some of the designs that we tested for opting out using GPC. And so on the left side what you can see is what we call a banner scheme. So this would be a privacy opt-out scheme or choice scheme where when you visit a website, you are presented with a banner and you can make a choice for the individual website that you're on, but you can apply this choice to all websites that you're going to visit in the future.

And so that is the element of generalizability. You're making an active choice, but also you make it a general choice. And so that is the idea. Along some dimension, you generalize your choices so that you are not disrupted too much in what you actually came to do.

A similar approach you can see here in the middle right where we have what we call a category scheme. So we would say, for example, yeah, if a site is an advertising site or a social media site or has crypto mining, then I would like to apply my opt-out right and I would like to turn on GPC. And so these are just two examples of the different schemes that we were thinking about. And those schemes are tested in usability studies. And as we can see on the next slide, we have the two elements. Generalizability so all these schemes have some sort of generalizing your choice, your privacy opt-out choice towards a larger set of choices, and we have the activity element. Both of these are important. The activity is particularly relevant because we need to show some intent, intentional opt-out. That's relevant for the law. It cannot be hidden. It has to be an active choice of the people.

And so what we are evaluating here is done in the context of GPC, but you could also apply this to cookies or any other kind of choice mechanisms really. There are different use cases. This is just here in the context of GPC.

So let's get a little bit more into that on the next slide. We can see here one result. And essentially what this is telling you is, okay, the more you look at the top, the more red you see. So what does it mean? Well, people were less disrupted. So these are the different schemes that we tested. And so if you look at the bottom there is SB base and S0-Snooze. What this means is essentially you were asked on every website to make an opt-out choice for GPC, right?

And I was actually surprised. People are pretty tolerant. I would have thought they say, "Okay, after one day, let me out of the study, I don't want to participate anymore." But actually nobody said that. When you present people actually every site they visit with a choice, they actually do it. To some extent it is disruptive, but I would've thought we cannot even do this, which is why we had this snooze button that people at least could snooze for a few hours this and not be annoyed with banners. But as you can see, the difference is actually not that big. But the point here to make is that for these data, S7-Data or S6-Universal or S5-Learn, which you can all read up in the paper, for the time being, this is just different generalizable active privacy schemes, the level of disruption is lower. And so that is really the main point here.

If we go to the next slide, what we see is the general result here, generalizability features tend to decrease the opt-out utility slightly because you make your choices a little bit broader. They're not as fine-grained anymore, but they increase substantially your opt-out efficiency. So you are much more happy that you are not disrupted.

If we check the next slide, then the main points to save from our paper are regulators should require publishers to honor GPC signals. We think that's a good thing. Browser vendors should integrate GPC. Those that haven't... Some already have, but those that haven't would be great if you do it as well. And publishers should also come on board honoring GPC.

And I want to point out one thing. If you have a privacy preserving product, let's say a privacy preserving browser extension or browser, then you could also have no interface. You could actually turn on GPC just as is because by selecting the product, you make an active choice that you like to send GPC signals to every site that you visit, for example. Okay, if we could see the final slide, I would like to thank this wonderful team here. I had the chance to work with four of my undergrads, Eliza Kuller, Chunyue Ma, Bella Tassone, and Joe Champeau, and they did fantastic work. And this work is also supported by the NSF, the Alfred P Sloan Foundation, and Wesleyan University. Happy to answer any questions on that. Thank you.

Ayesha Rasheed:

Great. And with that we're going to head into Q&A. And to kick us off, our first question for the panelists is how might your findings be extrapolated for other market players that have, one, less sophisticated ad environments, and two, perhaps less of a public footprint than say Meta?

Patrick Parham:

I can go first. I guess when we're talking about attribution, it's just one specific process and the sequence of ad activation. And we're talking about in our paper the sanitization of the boundary between media and just consumer marketplaces. But a large portion of what's being sanitized in these privacy preserving ad tech solutions is the idea that first party is data is just acceptable to use in all cases. And when looking at different ad tech privacy preserving solutions, a lot of them are based on first party data specifically. And going forward, I think it's delegating a lot of control to individual actors from larger platforms and smaller ones that are less sophisticated and have less of a public image and maybe they're not as equipped to implement sophisticated privacy policies. And perhaps that should live at a more sophisticated and higher level in the ad activation process to enable more control versus them maintaining and leveraging first party data how they choose.

Sebastian Zimmeck:

If I could add one point to this question, I think in my experience I've been working with some companies smaller and bigger on implementing GPC. So that's kind of where I'm coming from on this question.

And in my view, what is actually really critical is that you need to think about your business model. So if you don't want to have a do not sell link, then maybe do not sell. And if the technical aspect, I think at least I can say that for GPC is fairly minimal, but what you need to do then is, yeah, if you are saying that you're not selling, then you need to make sure that either if you have ad networks that are buying data from you, that you transmit the signal so that it's propagated through the ad ecosystem or that you use privacy preserving ad networks that maybe are not based on this kind of data practices.

And so I think the critical point is really to take a step back and say, what is my business and how should it be, how should it look like in the future? And I can only offer my opinion here, but I think the future is pretty clear. I think it will be more privacy, and I think everybody who says, "I want to do it the old ways," you are standing in front of a moving train. You can stand there, but it is moving this direction.

Jane Im:

My answer I think is pretty similar. So I think the findings for our paper also actually hold for other market players that are smaller and with less of a public image. Because if you think about it, I think the findings from our study of bringing the ad settings to, for instance, the top of the feed or within ads is a pretty straightforward interface change. So I'd argue that actually smaller market players can implement this, and if they do it right, it can actually potentially give a positive impact to how you just perceive those businesses. But I think Sebastian made a really good point that probably the underlying question is the company should think harder about what their business actually is.

David Walko:

Thank you. Could you speak more about privacy-enhancing technologies and how they could potentially harm consumer engagement or happiness with the internet? The classic example we hear is that consumers do not want to click through a bunch of privacy disclosures.

Sebastian Zimmeck:

Yeah, I can start on that. And one thing that I was surprised, as I mentioned, was the tolerance of people to actually click on every website their privacy choice. And I underestimated people's tolerance of that.

Now, do I think it's a good design? No, I don't think so. But it's also not true, that's at least my finding from this study that we did, that people don't want to do anything. They expect privacy to come for free or everything kind of stays as is, but more private. I think it is clear to most people that they have to invest a little bit of work to protect their privacy. And my sense from the people in our study is that they are okay with that. And so I think what we would need to do is generally, the way I think about it is maybe not be a perfectionist about every little detail, but get the big picture right.

And so that's sort of also where we started with GPC, very simple binary switch. We did not think, okay, there are these 5,000 settings that somebody can have and there are these different jurisdictions where you can have all these subtle differences. And so I think taking a big picture view and trying to come up with a design that is sort of capturing the main point and designing that iteratively converging onto something that is repeatedly tested, user tested, and gets better over time. I think that in my view is a good approach.

Jane Im:

I'll also jump in and say that I agree. And so some findings that I couldn't really talk about in the presentation is quite similar, I think, to Sebastian's findings is. So in the presentation I talked about different levels of actual ability, so high actual ability and low actual ability. So actually before conducting the study, I actually thought that participants might not like the high actuality interfaces because it's very, it's in the opposite direction of a minimalist design. It provides a lot of options and the user has to look at the menu and choose from more options than the minimalist design that Facebook, for instance, has. But what we found out was that actually participants in both our interview studies and the main experiment, we're pretty open to those un-minimalist designs.

So I actually think we regulators and platform designers and developers need to think out of the box. And it's true that if you make the settings too intrusive, it will of course harm consumer engagement and satisfaction. But we also I think need to venture out a bit more and test out new kinds of designs that Sebastian, for instance, talked about and I talked this study.

And I'll also add that one direction that I think is personally promising, but I haven't personally explored, is surfacing more options potentially with more pages, but also designing the interface in a way to customize the surfaced options per different user groups. So you can kind of imagine when a user comes to website and they sign up, the signup process requires users to answer about their privacy preferences. So very similar to what Sebastian has talked about. And so based on the user preferences, then the ad setting interfaces would surface the settings that the users will probably be more interested in using. So yeah, I think those are my two sets.

Patrick Parham:

My last thing to add on this question would be in looking not just at attribution, but the sequence of ad activation. While attribution employs encryption and technical fixes, the other steps in ad activation in the proposals for the other product of suites are leveraging universal IDs going forward. And we've seen in our research is that universal IDs are more persistent perhaps than third party cookies. And while attribution and proposed targeting techniques and ad serving techniques promise to not track users across context, it very much enables users to still be tracked across different contexts on the websites. So I think that would interfere with consumer happiness when browsing the web going forward.

David Walko:

Thank you. Our next question, some commentary has suggested that these privacy proposals are anti-competitive in nature, or at least might be. We at the FTC, of course, have a competition and a consumer protection bureau. How do you think our two bureaus should address the potentially anti-competitive nature of these privacy proposals?

Sebastian Zimmeck:

I say something about it, but I would say generally I think the angle of antitrust usually comes in at companies if they have a walled garden. If they have a certain area where they have control, something like an app store, they are operating system or some other layer of software that they control where they can then make decisions for themselves differently than for third parties that are also present on these platforms.

And so generally speaking, if these are general mechanisms that work the same for everyone independently of whether you're the operator, whether you're the owner of that platform or whether you are a participant, I generally don't see how these could create antitrust concerns. And some of these contributions we heard about here from Jane and Patrick. I think they were specific to, for example, Facebook or very specific context. And so I think for them then it's just a decision of the company. Yeah, do we want to do this? Yes or no? Just like with other ad settings. So I think maybe there are concrete problems in this regard, but in general I think they're solvable.

Jane Im:

Yeah, so when I first heard the question, I guess I also wasn't so sure if I agree with the premise that the findings here could potentially be anti-competitive. I think that's why I was kind of taking my time to think. So yeah, I guess I'm a bit cautious in trying to answer this question. I'll just say that I think it is important future work to actually potentially flesh out these designs on smaller companies that are not Meta and Google and see how they play out for users. So yeah, I would say I think that's an important question maybe to be potentially addressed by research. But yeah, when I first heard the question, I guess I'm not so sure if I agree that the findings here could potentially all be anti-competitive.

Patrick Parham:

I don't have anything to add further to this question beyond what Sebastian and Jane have already said.

Ayesha Rasheed:

Then we're going to wrap up with our final question for this panel, and this one's for all of you as well. So if you could pick just one design change of the various ones outlined in your studies, what would you like to see most for developers to introduce or regulators to mandate?

Jane Im:

So for me, I think I would obviously want to see companies making the buttons and the ads a bit more larger and noticeable. So many companies tend to just have three dots in the ads, but I actually think making them maybe text-based or changing it to a more visible icon could actually create a big change, especially for users who tend to be old who are not that sophisticated in terms of technology. So for those users, I think a simple change like that could potentially create a huge impact.

Sebastian Zimmeck:

Yeah, I would say speaking generally, I would say it would be important to surface privacy choices to people. So keeping it very, very general. Obviously, these privacy choice mechanisms need to be usable, but just surfacing them, showing them to people, not hiding them in hard-to-find places where you have to click through five different sites to actually get to, but rather making them available and accessible. I think that is my general hope that we are moving in this direction.

Patrick Parham:

A simple response. But it would be nice to see developers and regulators introduce their mandate, just a way for researchers and those evaluating these technologies to just follow flows of information and to more easily assess if they conform to the appropriateness of just privacy and just user expectations in general. Because the way these are presented now and these attribution solutions, it's very difficult to assess fully and get a transparent picture of what is actually going on in the full picture.

Ayesha Rasheed:

Thank you. A big thank you to all of our panelists today. We'll now move to lunch and to let everyone on the webcast know, we'll resume at 1255 with remarks from Commissioner Bedoya. Thank you.

Jamie:

Good afternoon and welcome back everybody. We hope you had a good lunch break. If you weren't able to join us this morning, a reminder that seven to 10 days after the event, we'll have transcripts and videos available on the event page at FTC.gov. For those of you who are following along or who would like to participate with social media, we're doing live tweeting commentary. You can follow along at, @FTC or at #PrivacyCon24. Also, if you have any questions for any of the panelists, you can send those to PrivacyCon@FTC.gov. We're monitoring that inbox and we'll pass those questions along to the moderators.

Now, without further ado, it is my honor to introduce Commissioner Alvaro Bedoya for some afternoon remarks. Thank you.

Alvaro Bedoya:

Thank you, Jamie, and welcome back everyone, to PrivacyCon. Before I start, I want to thank everyone who made this event possible, particularly folks in our Division of Privacy and Identity Protection. For a long time, I thought that this event was really terrific and now that I'm a part of the commission, I think it's a really great way for commission and commission staff to stay ahead of the latest developments and to help spread the word about that research to the broader public.

Our Division of Privacy and Identity Protection, or DPIP, is bringing so many high impact cases with such frequency that sometimes I think it can be hard to see the broader picture of the enforcement agenda that I think is taking shape.

And so, in the time I have with you all today, I want to flag a few high-level trends in our privacy enforcement program that are really important to me personally. Before I do that, I want to clarify, I'm speaking only for myself. These are my opinions and views on these enforcement actions and I'm also speaking informally. And smoo, for the details of all of these cases that I will discuss, please refer to the complaints and settlements in each of these matters.

So with that disclaimer, I will dive in. The first trend I want to underscore is that in settlement negotiations following serious law violations, this commission will not hesitate to seek outright bans on specific commercial practices. A lot of attention has rightly been given to the fact that the commission has required algorithmic model deletion where those models have been built using illegally obtained data or through illegal practices.

I want to further point out that in at least three instances in the last several months, the commission has flat out banned companies from engaging in specific conduct as part of proposed settlements. In December, we proposed a settlement banning Rite Aid from engaging in face surveillance for a period of five years. In January, we proposed a settlement permanently banning InMarket from selling or licensing any precise location data, and also banning them from selling or licensing or sharing any product or service that categorizes or targets consumers based on sensitive location data. And just last month we proposed the settlement permanently banning Avast from selling browsing data for advertising purposes. If you break the law and violate people's privacy, our staff will seek the strongest protections for consumers to stop that.

Second, I want to highlight that our investigations and enforcement will reflect the reality that sensitive data usually doesn't require further identifiers in order to be sensitive and require protection. That's a really long way of saying something that our technology blog said much more simply, and it said browsing and location data are sensitive, full stop. What all that means is that if I have a collection of just one specific user's browsing data or just one user's location history and nothing else, that information can in and of itself be highly revealing and can be used oftentimes to re-identify that user often with frightening ease.

We allege that EXMO, for example, was able to use location data unassociated with any other traditional identifier to figure out who visited a cardiologist, an endocrinologist, or a gastroenterologist, and then visited a pharmacy, or to figure out who visited, "size inclusive clothing stores," or military bases." We allege that the company InMarket was able to use sensitive location data unassociated with any other traditional identifier to identify low-income, millennials, single moms, single dads, and parents of preschoolers. We allege that Avast disclosed browser data again, unassociated with any other traditional identifier to disclose visits to a French dating site, to a Spanish language children's video on YouTube and to specific job searches near NSA headquarters near me right now in Fort Meade, Maryland. It's easy to see how this data can harm people and FTC and its staff will not wait for it to be associated with a traditional identifier before it moves to protect that data.

Third, the FTC will be skeptical of claims that companies need to retain sensitive data for long periods of time. Right now, there is a race to acquire and keep as much data as long as possible. And that's because data generates value, whether it's used for advertising or to train a machine learning model. The value that a company claims from obtaining or retaining consumer's data cannot justify breaking the law in order to get or keep that data. In our settlement with Alexa, we rejected the company's claims that it needed to keep recordings of children's voices indefinitely, in part in order to train their voice recognition algorithm because our kids' privacy law prohibits indefinite retention of kids' data. In the InMarket case, the company in that case, InMarket, retained precise records of consumers' daily movements for five years, much, much longer we alleged than it needed, to help consumers earn shopping points or make shopping lists. The ostensible reasons given to consumers for that collection of geolocation data. We alleged in our complaint that that excessively long retention significantly increased the risk of consumer injury.

Fourth, I myself will press companies to set privacy settings for children and teens, at their maximum. A lot of the allegations around teen mental health online focus on manipulative design practices that trick kids and teens into staying online longer than they want to, or content recommendation systems that allegedly show teens pro- anorexia or eating disorder content. Those allegations absolutely merit careful scrutiny. But we can't forget that abuse and harassment online also contribute to teen mental health issues online. In our Fortnite case from December 2022, we alleged that the default privacy settings for that video game let adult strangers engage in what I think are horrific cases of harassment against kids and teens using that platform. And our settlement required the company to set kids and teens privacy settings at their maximum as a result of that.

Fifth, and lastly, we will closely scrutinize algorithmic decision-making systems when they substantially injure people. In our December complaint and settlement with the company Rite Aid, we made clear that our authority against unfair and deceptive trade practices absolutely cover situations where biased algorithms injure people. In the case of Rite Aid, we alleged that a faulty algorithm resulted in shoppers across the country being stopped, searched, detained, accused of crimes they did not commit, and in many instances, had the police called on them for crimes they did not commit, as a result of that faulty algorithm. One instance this involved an eleven-year-old girl who'd done absolutely nothing wrong.

I'll wrap there, but I'll just say that while our privacy work often receives a fraction of the attention of our competition work, in my mind it is absolutely no less consequential and I am profoundly proud to serve alongside the staff and leadership of our Division of Privacy and Identity Protection, our Office of Technology and our Bureau of Consumer Protection. Thank you for your time. With that, I will pass the baton over to our fourth panel, which I believe is chaired by Elisa Jillson. Over to you, Elisa.

Elisa Jillson:

Thank you, Commissioner Bedoya. And good afternoon and welcome to the fourth panel of today's PrivacyCon on the Consumer Protection Implications of Health Technology. My name is Elisa Jillson. I'll be co-moderating the panel with my colleague Crystal Grant. I'd like to welcome our distinguished panelists, Hiba Laabadli, Dr. Ari B. Friedman, and Dr. Jesutofunmi Omiye. Our panel today will address a range of health technology issues, including reproductive privacy, hospital website privacy policies and large language models that may propagate race-based medicine. Our first panelist is Hiba Laabadli, speaking on her research, *I Deleted It After the Overturn of Roe v. Wade: Understanding Women's Privacy Concerns Toward Period-Tracking Apps in the Post Roe v. Wade Era.* I'll turn it over now to Hiba.

Hiba Laabadli:

Good afternoon everyone. So I'm going to share how our study examined how the overturn affected women's privacy attitudes towards period tracking apps. Before I start, I'd like to mention that our paper was accepted for publication at the ACM conference on human factors and computing systems. And if you want more details, feel free to shoot me an email. My email address would be on the last slide of this presentation. Next.

So let's start with a brief overview. In 1973, the US Supreme Court established constitutional protections for abortion rights in the us. However, as you all probably know, this landmark decision was overturned in 2022, granting individual states the authority to legislate abortion laws. Now, this in turn has led to the complete ban of abortion in 14 different states in the US. Next slide please.

After the overturn, as you can see from these articles, discussions around the safety of using period tracking apps noticeably increased, as prior research has found that these apps engage in unsafe and unfair privacy practices, including misleading privacy messages, inadequate de-identification measures, limited user control, pervasive tracking of the user's online behavior and retention of data for years after the consumer stops using the app.

Now, while these findings are alarming on their own, they mostly reflect the European context and focus on the practices of the apps themselves. In other words, little work has been done to understand users' perceptions of privacy towards these apps, particularly in the US and especially after the overturn. We believe that such understanding is crucial to inform the design effective strategies for policymakers and app developers for enhancing women's reproductive privacy. Next slide please.

This gap is where our research comes in. We were mainly interested in understanding first the factors that influence women's privacy practices towards period tracking apps. Second, their knowledge of Roe v. Wade, and its impact on their privacy practices and concerns. And lastly, if they had any expectations for privacy enhancing features for period tracking app stakeholders. Next slide please.

To answer our questions, we conducted a vignette study with 183 participants from the prolific platform. We presented each participant with four randomly selected scenarios were we systematically varied for different factors. So data storage, meaning where the data is stored, the type of data collected, with whom the data is being shared, and how much control the user has over their data. For example, in terms of user control, our scenarios varied from having no control at all to being able to delete your data or choose with whom your data is shared with. We also presented our participants with open-ended and multiple choice questions, including demographic information that acknowledged others.

To analyze our results, we used both qualitative and quantitative analysis. For quantitative analysis, we statistically modeled participants' concern towards the data practices presented in the vignette using the cumulative link mixed model. We wanted to identify the relative importance of each privacy factor in influencing user's concern. For qualitative analysis, we opted for content analysis. Next slide please.

So, this is an example of the scenarios we presented. The factors I mentioned in the previous slide are highlighted in yellow here. So, for instance, we have menstrual cycle data and physical health data for the type of data collected, the cloud for the data is stored, the app company and healthcare providers for with whom the data shared, and no option to control your data for user control. I'm going to read this scenario to give you an idea of how it looked like for our participants. So, imagine you're looking for a period tracking app to install on your phone to keep track of your menstrual cycle. You see a period tracking app with the following data practices. In addition to your menstrual cycle data, this app will collect your physical health data. This app will store your data on the cloud. And in addition to the app, your data shared with your healthcare providers. You have no option to control your data. After presenting our participants with a scenario like this, we ask them to rate their level of concerns toward the practices described.

Next slide please. Oh, I think we're missing a slide. Could you go back to the previous slide? Well, we can just go back to the next slide, sorry.

All right, so a noteworthy finding of our results was the underestimation of risk in digital communication of period tracking information. So, we asked participants to rate their privacy concerns toward various privacy practices. And as you can see from the figure on the right, our participants were most concerned about posting online about period tracking and fertility-related topics. Interestingly, they were significantly less concerned about searching online about period tracking fertility-related topics. This is particularly concerning, especially because law enforcement currently heavily relies on search history to target and criminalize abortion seekers, as you can see from the article on the left. Next slide, please.

Oh, now this table summarizes the results of our cumulative link model. The model, as I mentioned previously reconstructed was to measure various scenarios factors impact on participants' level of concern. Essentially, a positive coefficient indicated a positive correlation between the tested level and participants' concern, or a negative coefficient indicated the opposite. We also report the p-value for each level tested. The model's main findings are that first sharing data with law enforcement is most concerning and this is indicated by the fact that it resulted in the highest coefficient magnitude out of all the levels tested. As you can see in the table, it's around 4.5. So as compared with the baseline, which was no data sharing at all, sharing data with law enforcement or third parties significantly increased user concern. Some participants even mentioned being concerned about insurance companies using their period tracking data against them.

Our second finding is that user control is the second most important factors in explaining participants' concern. As you can see by the negative coefficient for the lead data and outcomes from sharing data levels, increasing user control significantly decreases concern. Third, collecting location, intimacy, mental health data also increases user concern. Further qualitative analysis showed that our participants regarded these as highly personal and unnecessary for period tracking.

Finally, we also tested our model against various demographical factors including political party. Interestingly, we found that Republican women tend to express higher levels of concern compared with their Democratic counterparts. We believe that this has to do with where our participants live. As half of our participants who identified themselves as Republican were living in states where abortion was banned. Therefore, such a higher level of concern towards period tracking and privacy practices might be attributed to the legal landscape of their states.

Next slide please.

So, the lack of awareness I mentioned earlier extends to the privacy practices of our participants' apps. As you can see from the figure on the left, 40% of our participants reported not at all informed of the privacy practices of their period tracking app. And while 60%, as you can see from the figure on the right indicated that the overturn has had some sort of impact on their privacy concerns, their practices in reality, only 8% actually attributed their concern to *Roe V. Wade* in this scenario, that is before we explicitly mentioned it to them. This speaks to the general unawareness of the potential implication of *Roe v. Wade* on privacy risks. Next slide please.

We also asked participants what privacy features they would like to have in period tracking apps and there was a strong call for usable controls. Our participants expressed the desire for more intuitive ways to manage their data, like straightforward options to believe their information. They also seek end-to-end encryption and granularity. Participants also emphasized the importance of transparency. They want clear privacy policies and more explicit information on how their data is shared, used and the associated risks and benefits. Lastly, our participants voiced the need for better privacy laws as reproductive data protections are poorly defined under several major legal frameworks in the U.S. Next slide please.

So, to summarize, our findings suggest that despite showing significant privacy concerns, participants generally lack the awareness and information about period tracking apps data practices in the post *Roe V. Wade* era. Therefore, we present the following recommendations. We believe it is imperative to define necessary data practices so that period tracking apps limit their data collection to what is essential for their functionality. This means avoiding the collection of sensitive information that doesn't directly contribute to the app's core services, thereby, respecting user privacy and minimizing potential risks. While some policies have been released post the overturn, like the My Body, My Data Act in 2022 and My Health, My Data Act in 2023, which call for controlling the sharing of reproductive health data except, and I quote here, as strictly necessary. There is no definition or any information regarding what exactly could be considered as strictly necessary.

We also advocate for increasing data transparency and user control in period tracking apps. This includes clear privacy policies and mechanisms like privacy labels to enhance users and understanding of how their data is used. We also emphasize that a companies must go beyond simply offering more user control settings. They should prioritize making these settings accessible and straightforward. Our recommendations extend to policy reform, especially in defining data protections from law enforcement more clearly. Future policy should address whether and how law enforcement can access reproductive health data. Lastly, we call for increasing awareness about the privacy implication of period tracking apps through media and educational programs, especially in the context of the overturn.

Next slide please.

So that is it for our presentation. Thank you for your attention. Feel free to send me an email if you have any questions.

Crystal Grant:

Thanks so much, Hiba, for that great presentation. Our next presenter is Ari B. Friedman, who will be speaking about his paper, *A Nationally Representative Content Analysis of Hospital Website Privacy policies*. Ari?

Ari B. Friedman: Thanks, Crystal. So, I'm presenting joint work today with Matt McCoy, Angela Wu, Sam Burdyl, Yungjee Kim and others. So, we are going to tell you about hospital website privacy policies and what we found.

Next slide please.

So, we focus on health system websites. I think I need little setup here after all of the introductions today. Thanks to [inaudible 04:12:31] Chair Khan for highlighting the importance of patient with browsing data and privacy and individual welfare. It's really been remarkable to see over the four or five years we've been doing this work to see the rate at which academic, investigative journalism and governmental research on this specific topic, on website browsing and health, have been incorporated into policymaking, really to the benefit of Americans who have health conditions or who will someday, which is all of us.

So, we started studying web browsing in part because it was minimally discussed in the literature. As Hiba's work shows, it's often overlooked by patients as a source of significant privacy risk. But we believe, and I think in this audience, many believe that small clues about your health can actually be quite revealing of your health. So, we have some quotes on the right from work by David Grandy and [inaudible 04:13:25] Merchant, suggesting really anything that your searches reveal about yourself would also be a health-related search and that your purchases as well can be quite revealing, which is also a common source of online browsing data. So, you tell me what you buy, and I'll predict what your A1C is. It's a measure of diabetes, based on your age, your food and the drugs that you get. So, these little clues when combined with data on what everybody else who is somewhat like you in these ways is combined with the kind of web scale machine learning algorithms can be quite revealing of your health in a statistical sense. And then we focus on health system websites because they can be more revealing about your health, right? So, if you browse to your local hospital's webpage on where do I find a doctor who specializes in diagnosing dementia? That tells you actually quite a lot about your health conditions, about your susceptibility to financial scams and it's quite revealing of your health in and of itself. And then we also believe that healthcare entities have a higher obligation to their website browsers than general websites. I think that CNN, for instance, or any news site, has an obligation to people to minimize the amount of tracking that they're exposed to. But it does read differently when it's your local hospital that installed tracking code probably without realizing it, that is turning your web browsing information over to third parties, or when an abortion clinic has that on their website and doesn't realize that potentially that they're revealing the fact that people are seeking abortion care to third parties, who can do whatever they want with the data. Next slide please.

So, in this work, we are building on previous work where we showed that tracking is ubiquitous across many aspects of the health system. And in particular, we have previous work showing that tracking on hospital websites across the entire country, every single hospital website, tracking is ubiquitous. So, it's 98 or 99% of websites have some tracking. For this work, we wanted to see what hospitals tell their patients and website visitors about that tracking through their privacy policies.

And so, to do that we had analyze privacy policies. We had two reviewers, both of whom were law students specializing in privacy law who independently located, assessed and codified privacy policies on each website. And then we had consensus meetings weekly to resolve any disagreements of which there weren't many. And to do that we limit ourselves to a hundred US hospitals. And so we started with the American Hospital Association with canonical database of hospitals in the U.S. We focused on non-federal acute care hospitals and then we took a simple random sample. So, these results are nationally representative and then we use the tool called webXray to measure the amount of tracking was on these sites. Next slide please. So, to understand how webXray works and how tracking works. We have to talk a little bit about how web browsers work. So, when you go and visit a webpage, say it's Mylocalhospital.org's homepage, you start with your browser. Your browser initiates a data request to that web server that says, hi, show me this particular page. Next slide please.

That server then transfers back to your browser, both a webpage and code that your browser uses to display that webpage. Next slide. And then your browser, along with that code, can be included tracking code, which is code that instructs your browser to initiate a data transfer to a third party. So that could be Google Analytics, or it could be a data broker, or it could be a tool that does not perform tracking, although the majority do. That uses that data to provide a service. And so that information can typically include your IP address, which pages you visited and some persistent identifiers, so cookies, browser fingerprinting, other ways of building a profile of you across all of your web browsing. And in some cases, it can include how far down the page you scrolled, what data you typed into forms. And so, on a hospital webpage that could be particularly revealing if you use a form to schedule a doctor's appointment or something like that. Next slide please.

And so, as I alluded to in our prior work of tracking across health systems, we found ubiquitous tracking. So, the co-founder of our initiative who, anyway, Tim Libert who wrote webXray showed in 2015 across an enormous swath of health-related pages that 91% had some tracking. Probably through the passage of time that has become more ubiquitous, although we can't directly compare. And we found that for medical journals, for hospitals, for abortion clinics, which should be particularly troubling in light of Hiba's finding that this is just not on patient's radar and certainly not something that they want, and for kind of a broad swath of Covid related pages, that tracking is essentially ubiquitous.

Notably, we always find in whatever element of the healthcare system we study that a few health system entities are managing to make do without tracking. They're providing the same healthcare services as everybody else, but they're kind of leading the way in terms of there not being any tracking on their page. And also notably, the number of domains that data is transferred to is quite high. It's not just Google that's receiving this data, there are other entities almost always. Next slide please.

And so, we ran that same exercise using webXray to document tracking on this sample of 100. We found that 96% had a third-party data request and there was a median of nine parties. So slightly less tracking than we'd seen previously, although the sample is different. So maybe some beginnings of a signal suggesting that the FEC and OCR guidance is beginning to influence behavior, which we would hope. And then in this sample, both having two reviewers look through the entire webpage, do searches for privacy, the word privacy, and other terms that might identify the privacy policy, manually search the entire web page and then do web searches with that restricted to that domain to see if they could find a privacy policy. 29% of hospital Web pages in our sample did not have a privacy policy. Next slide please.

Of those privacy policies that we identified, the 71 out of 100, we found that the average reading level was a college reading level, that 97.5% had at least the freshman college reading level, and that the average word count, essentially all of them had at least 2,000 words contained in their privacy policy. So, these are not approachable documents in general. Next slide please.

We then looked at the contents of the privacy policy and codified those. We found that 20% did not disclose that there was automatically collected information about the IP address. Nearly a quarter, more than a quarter, did not disclose that pages was visited within the site were automatically collected information. Nearly half didn't discuss the date and time of visit, which timestamps are very standard. All those three elements are kind of the minimum that one would typically collect with the tracker. We found... that's fine. So, there are lots of other elements here, but in the interest of time, next slide please.

And then looking at what that information was used for and how that was disclosed, we found that most discussed that data was used for tracking and analyzing the site use. We found that three quarters talked about marketing and advertising communications, which is a very typical use of tracking data. And then there's a variety of other elements here that I think all interesting, but we can talk about it later in the discussion if you like.

And then notably on the right here, we found that very few mentioned specific third-party companies. So around half mentioned a specific third-party company. The majority of those mentions were for Google, or Alphabet, which is indeed by far essentially every website that has some tracking has Google or Alphabet on it in our experience within the health domain. But the number of specific third-party companies disclosed did not reflect the universe of third-party companies that we see in this, on hospital web pages. Next slide please.

And then we found that only about 80% talked at all about any rights that the user had, and those specific rights are kind of poorly represented individually. So, two thirds talked about you have the right to disable site cookies and it kind of drops from there, in terms of what they're telling users about what they can do with their data on the web tracking. And then we look for specific populations. So, we looked for is there any mention of do you have additional rights if you're disabled? And we find that only 3% do. We find that children who have a law specifically dedicated to their privacy, about two thirds mention. We did not find any focus on cognitively impaired or elderly or other kind of special populations that health system entities might be particularly concerned about protecting from web-based tracking. Next slide please.

And so, thank you for being here. It's really an honor to be here. I just want to call out that it takes a village to do all research. And so, our co-leads of the Penn CME Digital Health Privacy Initiative, which is digitalhealthprivacy.org if you want to look at our previous work. So Lujo Bauer, who's at CMU. I'm at Penn. Matt McCoy is an ethicist at Pen. All of us are kind of instrumental in leading this. Tim Libert helped us found it and wrote webXray. And then our three law students, Sam Burdyl, Yungjee Kim and Angela Wu really did an enormous amount of lawyerly work that I can't even begin to understand, as I'm not lawyer, but it was very impressive work from the meetings and they put an enormous amount of thought into making this work consistent and to align with both construct validity and with what's legally interesting. So, thank you for being here.

Crystal Grant:

Thank you, Ari, for that great presentation. Our final paper is from Tofunmi Omiye, who will speak about his paper *Beyond the Hype: Large language Models Propagate Race-Based Medicine.* Tofunmi?

Tofunmi Omiye:

Yeah. Thanks for the introduction. So, I'll shift gears to talking about AI and how that affects medicine and healthcare in general. Next slide. Yeah. So, I'll just be introducing some terminologies and then will quickly talk about large language models in general, and then I'll talk specifically about my paper and what the impact looks like. Next slide. Yeah. So, in the past maybe two years, but mostly in the past year, there's been a lot of buzz about generative AI. It started off with ChatGPT, which was introduced by OpenAI, a company based out of here in San Francisco. And then before we knew it, people started implementing some of these generative AI models in healthcare systems. So, there were some medical schools that were using like ChatGPT to teach students while some places like Stanford and the UC system [were] also trying to pilot this new model, this new AI model, for either patient care or just instruction in general. Next slide.

And so that got us thinking about are these models rigorously evaluated and how could they be really useful in healthcare setting and what could be the drawbacks of that. Before I talk about large language models in general, I'm just going to introduce some terminologies. So there's something called deep neural networks, which was before the time of large language models, we had all these really powerful AI models that were modeled after the neural system in the brain, so just be able to process very large amounts of information. And they've been used for prediction tasks like Google search recommendations, YouTube recommendations, Netflix. All of them are powered by the deep neural networks, and also even the medicine. Some people have started using them as part of clinical decision support in healthcare.

But around 2022, large language models [inaudible 04:26:32] become the next phase of this new AI models. And you see what happened was that you just take very advanced neural networks. They got trained on a lot of data, like Reddit, Facebook posts, Twitter posts. And then we realized that they could be able to converse in this very human-like format. Next slide. And so, it's important to really understand how these models are trained in order to be able to see how you could [inaudible 04:26:58] them and just understand their behavior because they're still very new. There's a lot of research in trying to understand what goes on into actually creating these models. So, I'll just quickly take us through some very simple steps.

So, the first thing that every data from the internet is collected. So, we have data from GitHub, which is a code repository. We have data from all the books that have been published in all of human history. We have data from all the web pages, news sites, like New York Times, Washington Post and Twitter posts and just everything online. And there's some filter mechanism that goes on. So, you're trying to just remove either names from that or remove... Just some form of language filtering. And then after that you just go to some more pre-processing steps to make that data ready for training by this new AI model.

So, before that, the data, like the text and everything that is collected, is usually converted to numbers because computers only understand numbers. And the image on the far right is the new architecture that was released out of some researchers at Google that really brought in this era of large language models. So, you feed all this data that you've collected. This is billions, trillions of data. And then you feed them and just train them in all this new model. Then we go to the next slide. So after that, when you do that first step, which I discussed in the previous slide, something you get is a base model. So for anyone that's tried ChatGPT or any of the other large language models out there, that base model that is first produced, you cannot interact with it. It's just a random model that would produce nonsensical outputs sometimes. But there's a process they call fine-tuning phase in which people would ask this models questions and then whatever responses it gives is going to get fine-tuned with those responses. So you have someone labeling the responses as either good or bad.

And then the next step is because it's easier to be able to give different options of good, bad ,or bad to worse and allow the model to just get trained on that and repeat this process over and over and over again. At some point, you have another model, which is called a reward model, to also penalize this large language model based on what outputs it brings out. And so, at the end of this process, that's when the... We call it a fine-tuned model, which is the one that powers ChatGPT or any of the other generative AI model out there. So, this is the useful model. This is the one that the users get to see. Next slide.

So, when this model was created, so we started seeing a lot of papers talking about medical applications, but also a lot of health systems starting to implement these types of generative AI models. And again, because they converse in very human-like form, you could imagine them being used for answering patient questions, chatbots, and also from generating radiological reports and many of other things. People are starting to look at them for clinical trials and just every aspect of how you think patient care can be delivered. So, I just put these two [inaudible 04:29:57] there to show a paper talking about using them for chatbots and people already doing that, and another paper also talking about generating reports. In the case of generating reports, you also have the vision, the visual aspect, which the models have advanced into where they can also take not just textual data, but they can also look at visual data and textual data and be able to generate something coherent. Next slide.

With every new technology, there's also a lot of challenges and it's just trying to balance how do we use these technologies in the right way, especially in a way that benefits all of humanity. And that's what we're going to be talking about for the rest of this presentation. So, some of the early challenges I was realizing this model is you can see for each of the texts on the left were just prompted acts of this model, like, "Oh, what would you give me as an example of an attractive person?" Now you try this multiple times, and you see there's a very particular stereotype and demographic that it continues to produce. And when you talk about terrorists, you also see that. So you start to see the biases is very obvious that these models produce, and you can see how they can start to amplify these biases. But then our question was this is just social biases. How does this translate to healthcare settings where these models are already being used? Next slide will talk about that.

Yeah, so question is does this bias affect healthcare? So, we took all the publicly available large language models at that time. So, this was ChatGPT from OpenAI, Claude from Anthropic, another company also based out in San Francisco, and Bard from Google. So the reason why we tested all of these models was because some of them were talking about being non-biased and all the marketing gimmick was like, "This is a very transparent model," and those types of things. So we took some questions from literature that has been some of these landmark papers on racial bias, false beliefs about differences between Black and White people and also just general debunked biased questions that we know in the health systems that are present and we ran these prompts multiple times. Again because of there's a model stochasticity that occurs, so you need to run multiple times just to generate an average response. And then we realize that most of the responses or many of the responses I would say were based on debunked race-based content. So again, we see that these biases translated very easily to the healthcare system.

And then the next slide would show just a representation of that. Next slide. Yeah so, this is just a chart. So, we had all these physicians from different health systems rank the outputs of these models to see how race-based, wrongly race- based, they are. And we know race is a social construct, but a lot of these models, some of them showed the genetic basis of race. When we ask questions about how they calculated the glomerular filtration rate, which is just a kidney function test, for a Black woman, that question is important because in the past few years we realized that these equations have been negatively affecting minority populations, especially Black people, and also even preventing them from getting the type of care that they needed. And we saw that these models continue to encode those types of biases.

And it was asked about pain threshold, which is something that we know has also been debunked, and your skin thickness, differences lung capacity. So, I guess the interesting thing is some of the models that marketed themselves as the more less biased, more transparent types of models are even likely to propagate more race-based medicine. And those are some of the models that also being used in the healthcare system. So, we realized that, again, the bias translates very easily. And that was really concerning. Next slide. Yeah so, I put this slide about automation bias because sometimes some people think that, "Well, why do we need to try to have these very high standards for some of these models?" But the thing we need to understand and remember is that human beings inherently trust machines. We think when a computer says something is very accurate. And that's why a lot of people have been scammed. That's why when things [inaudible 04:33:54] do Google search and Google says something, people inherently believe that. And that's just because we think computers are very right and we almost blindly follow computers' instructions.

And this was a news article from a time where a tourist drove down a boat ramp because the GPS told her to do so. And we've seen many times in the past where AI models that was the more traditional statistical models have again supported or just worsened healthcare delivery to minority populations in this country. And now that we're going into a new generation of models, I think it's just important to start rigorously evaluating them, start checking and seeing if all this biased content that is unfavorable to minority populations are present just because they're not just biased, but they also amplify bias. It's just because of the automation bias that exist in a lot of machine models. Next slide.

So, I think my takeaways from this presentation is we're still in the early stages. Like I mentioned, these really came out just two years ago. So, we have a very beautiful opportunity to shape its direction, everyone listening to this and just people on panel, and accuracy verification perpetuates legal issues. So there's a lot of open challenges in this field of how do we use these language models in the right way within health systems and how do we make sure they continue to support every population, and also which contributing to the data sets for training these models. I talked about the training process. A lot of data is extremely important beyond the architecture.

And again, total model evaluation is extremely important, just before you deploy them, especially in a very sensitive situation like medical care where it's life or death situation sometimes. And I think the last thing I would just say is we cannot be conservative here. I think we need an interdisciplinary approach. We need different domain experts that need to come together to see how we can just fine-tune this technology such a way that it benefits all of humanity instead of taking us back decades behind, but we should just continue to move forward as a society. Next slide.

So just acknowledging all the people that have worked on this paper, Dr. Daneshjou and Dr. Zaba from Stanford, Dr. Rotemberg is at Memorial Sloan Kettering Cancer Center in New York, Dr. Lester and Simon Spichak from UCSF. So, this was really a multidisciplinary and also multi-institutional effort. And I'm just grateful for my co-authors here. Next slide. Yeah. Thank you so much for listening to me. Feel free to scan the QR code to just get more resources that I use for this presentation if you just want to reach out to me to get a better understanding of AI bias and language models in healthcare setting. Thank you so much.

Elisa Jillson:

Thanks, Tofunmi, for a great presentation. And thank you to all of our speakers. We'd like to ask a few questions about everyone's research. And let's start with one for Hiba. Hiba, how do period tracking apps compare to other health apps in terms of privacy concerns? And what can consumers do to protect themselves?

Hiba Laabadli:

Thank you for your question. I would say that mobile health applications in general, including period tracking apps, raise importance privacy concerns. So these concerns primarily revolve around how much control the user has over their own data, how long data is stored and the accessibility and clarity of privacy policies. As Ari mentioned, even when privacy policies exist, they often require at least a college level education to be understood. Also, again, as Ari's work pointed out regarding the privacy policies of hospital websites, these policies often do not provide clear information about which specific parties the data is shared with. And to answer the second part of your question, our study revealed that nearly half of our respondents believe that the responsibility for privacy protection primarily lies with the users themselves, so what steps can an individual takes to safeguard their privacy.

One key recommendation that we have is to be cautious about sharing personal information. Users should look for apps that don't require signing up for an account or that allow the use of pseudonyms for more privacy. And as I mentioned during the presentation, our participants also reported being widely uninformed about the privacy practices of their period tracking app, which aligns with broader research showing that only a small fraction of people thoroughly read privacy policies. However, understanding these policies is important, particularly to know who your data is being shared with and how much control you have over it. If navigating these policies feels daunting, we recommend that users turn to consumer protection organization's websites for more digestible summaries. Finally, once the app is downloaded, we recommend users to explore settings to look for options to opt out from practices they're not comfortable with.

Crystal Grant:

Thank you, Hiba. Ari, I have a similar question for you. How does hospital website tracking compare to tracking on other websites? And what steps can individuals take to reduce the tracking that you've observed?

Ari B. Friedman:

Yep, so yeah. I mean, the good news is that when we've studied, so for instance in our COVID-related paper, we looked across all COVID-related pages in a variety of ways and then we broke that down by academic sources versus government, sources versus academic sources probably primarily being hospital education based ones versus commercial sources, a lot of which were news organizations. There's a lot, a lot of tracking on the general web and there's less on the health-related web. The problem is that you have less, but you have tracking, still have some tracking on entities that have a higher obligation we think to patients and visitors, as well as are much more revealing of your health than browsing the general web. So, I think in general the health system is doing better than everyone else. It's just the status quo of web pages is tracked to a degree that's almost impossible to understand unless you've installed Privacy Badger or Ghostery and browsed the web and just looked at all of the different places where your data is going.

Crystal Grant:

Thank you, Ari. So, in addition to implications for consumers, the research presented today also has implications for companies. Tofunmi, what are some ways AI companies can use these findings to improve their models?

Tofunmi Omiye:

Yeah. Thank you for your question. I think there are a couple of ways to think about these. So the first one is the AI companies need to be aware of the bias that exist in some of those models because sometimes the data set issue. So it's just we take the efforts and just acquire more representative data. Another one is also in a methodology issue and there's just a bunch of people working on that and that's something also of interest in our lab. So I think just the AI companies being aware and starting to penalize the models because when I talk about the training process, there are ways to penalize the models for non-accurate and false race-based responses.

And I guess the interesting thing is since we released our paper, one of the company that actually had the worst performance on this set of questions, I noticed had released a paper on even getting more questions similar to some of the ones we had created and they're just building this data set of questions for other AI companies to evaluate and they have tried to retrain their model. So those are type of the rapid response that we want in the community and it's good to see that they're doing that. And I think all the AI companies just do more rigorous evaluation and, again, at the ideation process, get more domain experts to start asking this question even before the models are created. So I think it's just collaboration and also data sets building and just getting more representative data set. Thank you.

Elisa Jillson:

Our next question is for the whole panel. What are the next steps for research in the area in which you focused in your current research? And Hiba, perhaps you can start us off.

Hiba Laabadli:

Of course, I'd say we have two main areas we're focusing on. First, we want to look into how the recent advancements of generative AI is changing the way period tracking [inaudible 04:42:27] health apps in general work and what it means for user privacy. With apps like Flo using AI to offer more personalized services like chatbots, we believe that risks are higher than before for consumers. So for instance, if consumer data is being used to train AI models, there's an inherent risk associated with the potential leakage of this highly sensitive training data. So we want to understand these risks and the user's attitudes towards them better and see what protections are in place.

Our second area of research is... well, our current study focused only on participants who identified as women. And now we want to expand our research to include the broader demographic. Some apps, for example, lets partners track menstrual cycle together by using a partner mode. So we want to investigate attitudes towards sharing such intimate data, the consent mechanisms in place and the overall impact on privacy expectations.

Elisa Jillson:

Thanks, Hiba. And same question for you, Ari. What are the next steps for research in this area?

Ari B. Friedman:

Yeah, thanks. So, we have, I think at this point, and our group and others have documented that there's tracking everywhere within health-related websites. Our active interest right now is trying to look through various mechanisms see what the implications of that tracking are. So we observe the data's flowing into this black box and we're trying to look into how when you browse the web, if you are a person who has different health characteristics, are we able to see that the ads that you're shown change and things like that and therefore be able to see that actually there's a pretty big implication here in terms of not just this data flowing in, but the models are even without a human in the loop changing to try to sell you more health-related things in various ways.

And then I agree with Hiba. I think it's fascinating the implications for LLMs and the risks LLMs increase. So, we've had some discussions around what are the implications going to be there. And the more you know about somebody when your options are showing them ads from a marketing perspective, that's one thing, but if you can write custom ads for people on a personalized basis based on their individualized profile, the risks there are certainly heightened. And so, we're keeping an eye out for that and anticipate that at some point unfortunately somebody will do that and then we will try to go study it, so.

Elisa Jillson:

Thanks, Ari. And Tofunmi, same question for you. What do you think are the next steps for research in this area?

Tofunmi Omiye:

Yeah, I think there are three ways I think about this. So, the first thing is from the data set perspective, because we've seen time and time again that the data really matters in the training of this models, trying to build very more representative data sets. So something we're looking at is an international concerted effort and trying to get different data from just different parts of the world because, again, as you're trying to build this new form of technology, you want to make sure everyone is carried along and we can be able to do that. So that's one.

The second is just exploring new methodologies to penalize the models for non-accurate responses or biased responses. And there's a lot of people starting to look into this because these models are still very new. We don't understand how they work. So, the methodology side is a bit tricky, but that's something that's still very interesting to explore. And I think the [inaudible 04:46:09] but not the least thing I think about is also trying out some of these experiments on the newer generation of this model. So, we talk about large language models, but now we have something they call vision language models, which just means there's a vision component and a language component. And you can see that very applicable in medicine because medicine is inherently multimodal. We take different forms of input, not just clinical history. So just trying all those experiments and seeing some of these biases persist and, again, do the loop again of trying to build better data sets, better methodologies, and hopefully you can have better models that have less bias and can just have just better healthcare for all types of human beings.

Crystal Grant:

Thank you. And we have another question for the whole panel. Some of you have spoken of the policy implications of your research. What are the big takeaways for policy makers? And Tofunmi, maybe you can start this time.

Tofunmi Omiye:

Yeah, thank you. That's an interesting question. I think one thing to think about is before deployment of those models, just mandating rigorous evaluation, either on the health system policy angle, even state or national policy angle, because these models be part of the decision-making process, and I want to make sure that we're providing the best healthcare for every types of person. And I would say the White House released some AI statement. This was I think a couple months back. It has been well received by some people and not as well received. But I think something like that is a good step, a first step in the right direction, just trying to put out regulations and making sure we're mandating just some form of regulation before deployment of those models.

And another thing is even recurrent reevaluation because there's something they call distribution shift and also just data set shift. So, these models can also change over time. And there's been some people documenting that. So just having policies that also mandates just the rigorous evaluation of them over time and just make sure that they consistently hold up to standards that would want for healthcare delivery in the health system in general. Thank you.

Crystal Grant:

Thank you. Ari, can you speak to big takeaways for policymakers?

Ari B. Friedman:

Yeah. So, this work [inaudible 04:48:37] on privacy policies, and I think we show that privacy policies on health system sites are no less complicated than privacy policies elsewhere. So the best thing [inaudible 04:48:52] would be if we don't have to rely on privacy policies at all from my perspective and we don't have to rely on individual action at all, both of which are, I think, imperfectly protective, if we had a baseline level of expectation of privacy that matched what the average American had in terms of expectations of privacy in 1990, and we could return to that time through policy. And so, to me that calls for additional legislation that there's only so much that can be done with existing legislation.

But also I think that the OCR and FTC discussions and enforcement actions around viewing this as a HIPAA problem is I think potentially highly effective because we can use one of the few highly effective privacy laws that we do have to give patients what I think they expect, which is that when you go to a hospital web page, you don't read the privacy policy or you can't even find the privacy policy because it doesn't exist after two dedicated researchers looked for quite some time to try to find it, that you're still going to have protections and that your seeking care from your doctor is protected just like it would be if you walked into your hospital and filled out a form that said, "I had a new patient appointment with an HIV specialist," or care for any other kind of stigmatized or not stigmatized disease.

Crystal Grant:

Thank you. And to Hiba the same question. Any big takeaways for policymakers?

Hiba Laabadli:

Of course. Well, I want to quickly go back to our results. So our study revealed that users are particularly alarmed when their data related to menstrual tracking location or mental health or intimacy data is shared, which they see that as relevant and invasive. And currently there is, let's say, a gap in the legal frameworks, both in the US and globally, HIPAA, the DDPR, in explicitly protecting reproductive health data specifically. And as I mentioned in the presentation, there has been new policy efforts like the My Body, My Data Act, which aim to address these issues by controlling the sharing of reproductive health data. However, these efforts still lack clear definition of what constitutes necessary data practices and leave too much room for interpretation. So I would say one big takeaway is that we need clear, unambiguous legislation that specifies what data collection is essential for the functionality of period tracking apps, limiting unnecessary data collection and sharing.

Another important finding of our research is the lack of awareness regarding privacy practices and risks in the context of period tracking and fertility. So, for example, our respondents were mostly not concerned about the implication of search history data. So we believe that there should be initiatives and campaigns to [inaudible 04:51:57] consumers about the potential risks with reproductive health data.

Crystal Grant:

Thank you, Hiba. And I'd like to thank all of our panelists for their great research and presentations today. Our next panel, which will begin momentarily, will address artificial intelligence and machine learning. Thank you.

Julia Horwitz:

Good afternoon, everyone. My name is Julia Horwitz, and I'm here with my colleague, Ronnie Solomon, to moderate this afternoon's panel on artificial intelligence and machine learning. Ronnie and I are both attorneys in the Bureau of Consumer Protection's Division of Privacy and Identity Protection. And we are very pleased to introduce our talented panelists. We have Patrick Gage Kelley from Google, Umar Iqbal from Washington University in St. Louis, and Batul Yawer from Arizona State University. Thank you to all of our participants for being here this afternoon. And Patrick, please feel free to kick us off whenever you're ready.

Patrick Gage Kelley:

Excellent. Thank you. So really excited to be here this afternoon to talk to everybody across the live stream. This is some work that we have been doing at Google looking at what attitudes towards privacy look like. So, this is a paper we published called T*here will be less privacy, of Course: How and why people in 10 countries expect AI will affect privacy in the future*. Next slide. So to get at some of these questions, we really want to know what people think are going to happen broadly. And we asked them, "In the next 10 years, what do you think will happen in your country," where we replace your country with the name of their country, "Because of artificial intelligence?"

Next. So we give them choices, like for example, on the privacy question, they get to say there'll be more privacy, less privacy. They can say no change, or they can say they don't know. And next slide. And so we did this in summer of 2021 in 10 different countries, and this was part of a larger survey where we were looking at broader attitudes towards AI, the things they would expect. But today we really want to focus the results for PrivacyCon on what we found about privacy. Next slide. And so for example, in Germany, as just an example country, we saw that about 10% of respondents said they would have more privacy, about 56% said they would have less privacy, and then there were another 20% who said no change and 14% who said they didn't know.

But if you focus on the respondents who did express some sort of directed sentiment, how they thought privacy would change because of AI, next, we can show how it actually... There's quite a few more here who said less privacy than more privacy or about a ratio of 0.18. Next slide. And we can look at this across all 10 countries. And so if we look at just that privacy question again, we see that privacy overall gets a ratio of about 0.45, where about twice as many people think privacy is going to get worse, they're going to have less privacy in the future, than that privacy will get better.

And, next, we're focusing on privacy here because, one, we care a lot about privacy. This is PrivacyCon. I'm a privacy researcher, but also because privacy was actually in some ways at the top of this list and not at the top in a good way. Privacy is one of the things that people across all 10 countries thought was going to be worsened or lessened the most because of AI. You can see there are some things where they were really excited and quite optimistic about it, things like quality of life overall, healthcare, transportation, education. And then there were really three things that they were very worried about, privacy, personal relationships, and job creation or job loss. Next slide.

So we wanted to know more directly about these privacy concerns. Why do people think privacy was going to get so much worse? So we separately asked an open-ended question where we said to all of our participants in all 10 countries, "In what ways will artificial intelligence affect privacy in the future?"

And they could say anything they wanted. They could say something positive. They could say something negative. This is all the context they had. Next slide. And so, across our almost 10,000 responses, we developed a code book working with our partners at Ipsos, a global marketing firm that helped us analyze this data, and we coded each of those responses, so each of the open-ended responses that people had, with a code book of over 368 codes,

Patrick Gage Kelley:

... codes, which also means we had about a hundred thousand words in that corpus, and on average, about 10 words. And so these were really quite a detailed, quite a rich data set. Next slide. To give you a sense of how this response quality looks, because again, with surveys, large surveys like this that are mostly conducted online, there's a concern that people will sort of be off topic, that they won't give any sort of good answer. And we had about 75% of our responses across those 10,000 express some sort of privacy sentiment. Another 18% didn't know. And again, that's a completely valid response. Many people may not know how AI is going to affect the future. And so they said things like, "I don't know," or things that were blank or no real comments. And then we had about 8% that were sort of unrelated, things where they were talking about robots or productivity, and we couldn't draw a clear tie to privacy.

Next slide. So there's four themes that I want to talk a little bit about today that we saw in the privacy responses. The first is data at risk, where respondents generally felt that AI was just going to need lots and lots of data. That was going to be gathered across multiple devices. That was going to be cross-linked and aggregated. And because we were gathering all this data to build AI systems, it was going to be vulnerable to misuse and hackers. For example, we had participants who said, and I'm going to read some quotes here, "Just AI requires a lot of human data. For machines to think, they need to analyze and base their decisions on data. Data will be a hot commodity, and companies will look for all the ways you can to give them your data and use it as inputs for AI." That was a respondent from Kenya.

Another respondent from the Philippines said, "It's already here. Every time I use my phone or the internet, AI is at work gathering all information passing through my devices." A respondent from Australia said, "Facial recognition, other ways to track people will negatively affect privacy. AI will allow a lot more information to be gathered and collated." So you can get this sense that people really understood that bringing all this data together was going to put it at risk. And some of these respondents directly talked about how this would be exposed. For example, a respondent from Japan said, "No matter how secure it is, I think it will make it easier for the leakage of personal information to occur." A respondent from Brazil said, "More and more, our data will be in databases, exposed to strong by increasingly skilled hackers."

Next. Our second theme is that this data was seen as highly personal. Overall, our respondents really felt it wasn't just that AI was going to be collecting lots and lots and lots of data, but really, that that data was going to include a lot of really sensitive, highly personal, extremely precise information. And that could be used potentially to influence or manipulate people, or for other nefarious purposes. One participant said from South Korea, that “they would be collecting our words and actions down to the smallest detail.” Another respondent from Australia said, " Much more private data will be collected and used, contacts, places you've been to people you call, websites you visit, what books and articles you read, what products you buy and use, tracking via face recognition, masses of information to be used to predict behavior." One participant from Brazil said, "Our privacy will be totally affected because technologies will capture data about our conversations, our consumption habits, our medication, our contacts, and everything else to create a database of things and predict things that may interest us."

Overall, the sense we get from our participants here is that it's not just, again, that there's this huge volume of data, but that these insights can really be used. One participant from Australia said, "Large corporations will ruthlessly use AI to market their products or services to a wider demographic by sharing clients' private information with each other." Next. For our third theme, I want to talk about state and surveillance. Here, our respondents were specifically drawing out ways that AI supported surveillance, not just from companies, like we saw in some of the personal ways, but also from governments through the use of just constant monitoring, and also identification. One participant from Japan said, "I feel like I'm being monitored at all times." One participant from Australia said, "AI will be the ultimate spy." One participant from Brazil said, "It'll be everywhere and in everything we use, being able to monitor us." One participant from Japan said, "I think that even faster and more accurate identification of individuals is progressing. And in some cases, I think that constant observation is also possible."

And a participant from Germany said, "It'll be possible to spy on the population even more easily than it is now. It will be even easier to control our lives." And for our last theme, next, we also wanted to bring up how our participants raised questions about consent. Respondents felt, broadly, that yes, they were going to want to use AI systems, yes, AI systems would need all of this data, but respondents felt that this scaled personal data collection would occur without any sort of meaningful consent, sometimes without awareness, and that it was going to be required to get access to these really important beneficial systems and the use of AI services that would do all the positive things that we saw earlier in this report, where we saw them increasing education and healthcare and all the good stuff that would come with it would be impossible to give up all their personal data to get. We had one participant from Kenya say, "People's data will be invaded whether they know and give their consent or not."

One participant from the United States said, "It has already invaded households beyond what the majority of people know. There is no privacy now." And finally, one participant from China said, "AI requires people to reveal themselves while exposing their privacy." Next slide. So just in summary, we saw these four large themes emerge in our data set. The data is at risk, the data will be highly personal, the data will happen without consent, and the data will also be used for constant monitoring including by governments. If you want one takeaway slide, this is the summary. So what do we do with this? Next slide.

We have a couple of reflections from recapturing this data. The first is that people overall had really thoughtful responses. Everything you saw along the last slide, the quotes that I read, and you can read more in the paper that it's linked on the website, this narrative is really aligned quite well with both the press, and also with expert opinion. These are real privacy concerns, these are not fictional things. These are not things they made up. This is not sort of existential concern that isn't applying today, but these are real concerns that our participants have with AI, and they're really quite thoughtful. They're really quite well reasoned. Next. This means that we can potentially engage a lot more with people with solutions. There's a lot of spaces where people's privacy concerns and desires seem sort of nebulous or a bit confused, or maybe their actions don't totally align.

But in this world, people have a really good sense of what AI might be doing, and this seems like a really great opportunity to engage with them, to help better educate how they can avoid some of these spaces, to understand how they can help influence product and technology, and potentially lead to policy change. Next. Third, we hope we can use these themes as a roadmap for people's privacy concerns with AI systems. So, as we're building and developing new AI systems, we can look at each of these themes and say, is there a way where we can collect less data for this one? Is there a a way for an actual consent moment to occur? Can they sort of limit some of the uses of this technology? How will these systems be used by governments? And how will that be disclosed? And so, we can use these themes as a way to really think through the development and the release of AI systems.

And then finally, next, our fourth finding here was... I do want to note that 22% of our respondents across all of the countries thought they would have more privacy. And this isn't just an optimistic halo effect. People overall are quite happy with privacy, but some of our participants really did describe ways that they thought AI might help make their data more secure and might lead to enhanced sorts of encryption or security, and there was a sense that they might really be able to have some real privacy gains as a result of this. Next. So I just want to thank you all so much for having us here today and for sharing this, and I look forward to questions after the other talks.

Julia Horwitz:

Patrick, thank you so much, Umar, you're up next.

Umar Iqbal:

Thank you. So yeah, my name Umar, and I'm an Assistant Professor at the Washington University in St. Louis, and I co-lead the privacy and security research lab at the university. In today's talk, I'm going to discuss our research on assessing the security of large language model-based platforms like ChatGPT. And this is a joint work with collaborators from the University of Washington. Next slide. So my goal that this talk is to convey three key things. First of all, I want to show you that large language models, or LLMs, are being increasingly extended as full-fledged computing systems or platforms that support third-party apps. And second, I want to convince you that some of these LLM platforms are emerging without a systematic consideration for security, privacy, and safety. And I'll do that by demonstrating that existing third-party apps have the potential to exploit the LLM platforms and harm the users. And third, I want to motivate that we need to systematically study the security of these platforms before they get even more widely deployed and adapted. And to address that problem, I will also describe our framework, which aims to lay a foundation for secure LLM platforms and systems. Next slide. So large language models are extremely capable, but they still have limitations for a number of use cases. And broadly, we can group these limitations into two classes. First of all, LLMs cannot automatically leverage up-to-date data unless it is explicitly provided to them by the users. And second, LLMs cannot act on their recommendations. For example, they're very good at giving you step-by-step instructions to make a flight reservation, but they cannot book that flight for you of day one. So to address these limitations, platforms like OpenAI and Google are increasingly extending or packaging their LLMs as systems. And these systems have the capabilities to maintain a persistent memory, write and execute code, and connect to online services.

So ultimately, the goal by giving LLMs these capabilities is to make their functionality similar to other computing platform, like mobile, IoT, or even desktop devices. And as a positive outcome of these abilities, LLM platforms are able to support third-party applications, which is further extending their capabilities. Next slide. So the third-party apps and LLM-based systems differ in how they're developed as compared to apps and other mature computing platforms like mobile. And what I mean by this is that the third-party LLM apps are defined in natural language, and they also interact with each other, the LLM and the user through natural language instructions. And there are opposite advantages to it. For example, someone without deep technical knowledge will be able to easily develop it.

But on the other hand, the natural language is not as precise as programming language or programming languages, which have been typically used to develop. So as a result, the apps developed in natural language can have ambiguous and imprecise definitions and interactions, which could potentially lead to inadvertent security and privacy issues. On top of all of that, these apps are developed by third parties who cannot be implicitly trusted. And this trustworthiness of third-party apps has been an issue in almost all prior computing platforms. Whenever they have integrated third-party apps, they have brought in a number of security, privacy, and safety issues. Next slide.

So now, considering that third parties have introduced security issues in other computing platforms, and on top of that, the natural language definitions and interactions can be imprecise, you would imagine that the LLM platforms make security and privacy a key consideration in their design. But unfortunately, that does not seem to be the case. So in the short span of roughly two years where third-party apps have existed, researchers have already identified a number of potential mechanisms through which third-party apps can be exploited to launch attacks and harm users. And in addition to that, there is anecdotal evidence which suggests that the apps are not even properly reviewed by some of the LLM platforms. And what is even more concerning here is that when the security and privacy issues have been highlighted to the platforms, they have not given them much serious attention.

So overall, these concerns highlight that at least some LLM platform app ecosystems are emerging without a systematic consideration for security, privacy and safety. And if these systems are widely deployed without these key considerations, third-party app integrations could result in harm to not just the users, but also to other apps and the LLM platform itself. Next slide. So in our research, we try to address this problem by proposing a framework which aims to lay a foundation for secure LLM platforms with third-party integrations. Our framework is essentially a threat modeling process where we systematically uncover potential risks and attacks that the stakeholders in LLM platforms could pose if they were compromised, or if they were malicious. And we organized these risks in a taxonomy and also include potential mechanisms that could be used to carry out the uncovered attacks. And ultimately, with this exercise, our goal is to inform the LLM platform designers to triage and eliminate vulnerabilities by following a structured taxonomy. Next slide.

So to formulate our framework, we use OpenAI's ChatGPT as a case study because it has the most matured app ecosystem. And specifically, we use ChatGPT plugins. Plugins is basically one of the terms used by OpenAI to describe apps. We start our process by surveying the capabilities of stakeholders in the LLM platforms, which includes the apps, the users, and the LLM platform itself. And our goal with this surveying is to assess how malicious stakeholders could leverage their abilities to attack each other. And while we are doing this exercise, we want to make sure that our assessments are grounded in reality. In other words, we want to make sure that the attacks we uncover are realistic and not in mere hypothesis. So we analyze the code of the apps or plugins hosted on the OpenAI platform, and then also use them to see if they have the potential to implement adversarial actions that we enumerate in our taxonomy.

And in addition to uncovering the attacks, which could occur due to the presence of an active adversary, we also include attacks or risks that might inadvertently occur in LLM based platforms because of the ambiguity and imprecision of natural language. And we consider both the attacks that uniquely apply to LLMs or LLM platforms, and also the attacks that have existed in prior computing platforms but apply to LLM platforms. Next slide.

So after exercising our framework, we uncovered a number of attacks and we gripped them by stakeholder combinations, which includes the attacks between the apps and the users, the attacks between the apps and the LLM platforms, and the attacks between the apps themselves. So on this slide, I have listed some examples of attacks as sub bullets for each of the categories. And I've highlighted the attack categories which uniquely apply to LLM platforms, and I've also added the numbers in the brackets, which represent the apps that are present on the OpenAI's marketplace, which have the potential to cause these specific issues. Next slide.

So the first attack that I would like to discuss is the hijacking of a user session with the LLM platform. Basically, LLM platforms allow apps to alter their behavior, and apps can simply do this by using instructions in their functionality descriptions that direct the LLM to behave a certain way. So this feature has some obvious benefits if it is implemented properly. But unfortunately, based on our assessment, we noticed that when apps instruct the LLMs to change their behavior, it persists beyond the context of using the app. For example, we found an app which directed the LLM to always respond in English when the user interacts with the app. But based on our interactions, we noticed that even in the instances where the app's not used, the LLM still responds in English, which is a deviation from its typical behavior where it always responds in the language in which you ask it a question. Next slide.

We also found instances where the apps could hijack other apps installed on the LLM platforms. The main technique that these apps used was that they made their descriptions similar to other apps or the added keywords or phrases which could trigger other apps or the keywords and phrases which could be related to popular online services. And we noticed that even in the instances where the LLMs could access both the apps or an online service or the app, we noticed that in several instances, the LLM called the app, which was potentially squatting or copying the functionality of an other. And this is one example where the ambiguity and imprecision of natural language could inadvertently lead to security, privacy, and safety issues.

We also found an app which contained instructions basically to demonstrate this risk. The app basically contained instructions that instructed the LLM that it can assist with shopping from Amazon. And when we use the app, we noticed that when the user explicitly specifies that it needs help from amazon.com and it directs the LLM to not use any third party service, the LLM still ended up routing the user query to the third party app. Next slide. So the last deck that I want to discuss today is about the harvesting of user data. So we found a significant number of apps which collected excessive user data. And in many cases, this was perhaps more than what they needed to resolve the user query. For example, we noticed a number of app that expedited full user prompts. But interestingly, we also noticed a couple of instances where the apps were careful and they specifically instructed the LLM to not collect and share user's personal data.

But unfortunately, in both of these instances, we noticed that these directions were not enforced for data which was collected before interacting with the app. For example, we found a travel management app which instructed the LLM to not collect and share user's email address, but in the cases where LLMs already had access to this data, they shared it with the app. And next slide. So overall, in terms of the key takeaways, I have demonstrated with examples that, unfortunately, security, privacy, and safety do not seem to be the key considerations in [inaudible 05:18:54]

Julia Horwitz:

I apologize for interrupting, but we're at time and we'll need to move on to the next presentation. I apologize.

Umar Iqbal:

Totally understand. And the remaining takeaways that are on the screen. So that concludes my talk. Thank you.

Julia Horwitz:

Wonderful. Thank you. Batul, please, whenever you're ready.

Batul Yawer:

Hi, all. I'm excited to talk with you all today. My name is Batul, and today we'll be discussing the work I had done for my master's thesis, *The Reliability and Validity of a Widely Available AI tool for Assessment* *of Stress Based on Speech*. This is an example of the fallacy of AI functionality. I know that was a lot of words there, so essentially what we're asking is, can AI really detect psychological stress from speech? As a consumer, my bias is to trust tools that involve AI, and I imagine I'm not the only one. In this presentation, we'll see if a publicly deployed AI tool that measures stress from speech actually works. Next slide.

So I would like everyone here to think about if they have felt stressed in the last six months. Most of you, I'm sure, are likely responding, saying yes or nodding your heads, and everyone has experienced psychological stress at some point in their life. Stress can lead to various health disparities, including neurological disorders, such as anxiety and depression. Since we know stress is a factor of many health disparities, stress monitoring is a crucial aspect of preventive care. Next. So how do we go about monitoring stress? So one of the ways of monitoring stress is known as the perceived stress scale. This is kind of our clinical standard. This is a self-report questionnaire that was created in 1983. An individual will go ahead and read these questions and rate the question from being zero, which is never, to four, which is very often. At the end there is a scoring guideline, and those score ranges dictate if you're at low, moderate, or high levels of stress.

This tool has been validated with high internal consistency and construct validity, and it's described as a classic stress assessment instrument. When I say high internal consistency, this means that the questions within the questionnaire have been tested to ensure that they are all equally contributing to measure psychological stress. Overall, the PSS-X or the perceived stress scale has shown that it is a consistent measure of psychological stress within itself. Cortisol levels, as we know, are known to increase with continuous psychological stress. And in previous research, the PSS-X has a positive correlation with serum cortisol levels. So these studies have measured high construct validity for the PSS-X, and having high construct validity means that the PSS-X measures what it states to measure, which is psychological stress. So the PSS-X has been validated for over four decades and has shown to be the clinical standard for measuring individual psychological stress.

Next, we have the Cigna Stress Waves test. This is a new AI tool. This has an AI tool that analyzes 60 seconds of an individual's speech to measure their psychological stress. You can find this tool on the Cigna Healthcare website. And an individual will click on Create Your Stress Portrait, and then they'll go through the prompts and record themselves responding to the prompts for 60 seconds. After the 60 seconds is up, the individual will receive a score of low, moderate, or high levels of stress that is associated with being at risk for high blood pressure or cardiovascular disease. This tool was released without any publication data, validation data, and it was presented as a clinical-grade AI technology. So on our next slide, that has us wondering, is it possible to detect psychological stress from speech? When we think about speech under stress, we see that there are limited findings in literature for classifying psychological stress from speech, and there's a couple of reasons for that.

One, there are numerous ways of defining stress. For example, psychological stress can be acute, it can be environmental, it can be chronic, and this ambiguity causes issues in pinpointing specific speech-based markers for psychological stress. And then speech is variable. So speech that may sound stressed for one person may be relaxed for another. There's no universal speech signature for psychological stress. Next, we have a quick visualization that represents speech under psychological stress. So on one side, we have various ways of defining stress. And then on the other side, we have various ways of measuring speech under stress. And that means that speech changes in the context of psychological stress are complex and variable. So building an AI tool that analyzes these two complex and variable concepts made us question if this tool actually works. So with this background, we prospectively validate the Cigna Stress Waves test. Next slide. We did this by recruiting 60 participants. Each of the participants completed two of the Cigna Stress Waves test for repeatability purposes, and then they also completed the PSS-X. They completed all of these tasks within the same session. And the Cigna Stress Waves test was done twice so that we could see if the test is consistent within itself. The PSS-X test was done to see if the Cigna Stress Waves test has construct validity in comparison to the PSS-X. So we did this to see if the Cigna Stress Waves test actually measures psychological stress. Since all of this was done within the same session, this was all conducted and finished per participant within 15 minutes. And it's important to acknowledge that the Cigna Stress Waves tests, the two trials were done back to back. All the participants used the same monitor and the same microphone so that the data quality that was going into the AI model was relatively similar. Next slide.

What we found for test-retest reliability is that on the X-axis on this graph, we have the trial one for the Cigna Stress Waves test. And on the Y-axis we have trial two for the Cigna Stress Waves test. Sorry, that's a mouthful. But what we see here is that the data is pretty spread out, and that's reflective on the correlation of -0.1. So this means that if you do the test twice within the maximum five minute interval, you get very different results. So participant one on the first trial can receive a result of seven, and that correlates to a result of low psychological stress. And then a minute later, they can take the test again and receive a result of 25, which is related to moderate psychological stress. If the test was consistent, we would see that the spread of the data points would fall closely to that red line as their psychological stress levels did not change within that maximum five minute interval.

Next, we look at the conversion validity, which compares the average Cigna Stress Waves test trials and the PSS-X score. Here, again, we see that the correlation is 0.2 and the data points are very spread out. This means that the Cigna Stress Waves test gives a very different value than the PSS, which is, again, that clinical standard for psychological stress. If the Cigna Stress Waves test measured stress, we would, again, expect those scores to fall near that red line to display some kind of a relationship with the PSS-X. However, for the Cigna Stress Waves test, that is not the case here. So overall, what this means is that the Cigna Stress Waves test is not able to provide consistent scores if taken numerous times, and it does not measure what it states to measure which is psychological stress. Next.

So as we remember, the Cigna Stress Waves test made a claim that it is clinical-grade AI technology. And to establish a clinical-grade algorithm, reliability metrics, such as ICC, Cohen's, Kappa, and R should be close to the value of 0.75 to one. And for the Cigna Stress Waves test, that is not the case. The repeatability value is at -0.1, and the construct validity value is at 0.2. These values are much closer to zero than they are to 0.75 or to one. This highlights a critical need for the claims made about AI-based health tools to align with robust evidence. Next. So the reason why we wanted to validate this tool was because we found a significant gap in the literature regarding data related to speech production. Under psychological stress, this lack of foundational knowledge or any published data about this tool raises concerns about the basis on which the clinical-grade claims about the algorithm were made. They claim to be clinical-grade, but I'm not finding a scientific justification for how the AI tool is interpreting speech to classify psychological stress. Next.

This all falls under the fallacy of AI functionality. This is when consumers often assume that AI ensures functionality without verifying its reliability. This is a psychological bias that leads individuals to trust AI technologies at face value, overlooking the importance of thorough validation prior to deployment. And by falling for the fallacy of AI functionality, there are impacts of this deployment. A micro-impact of premature deployment can be seen in the Cigna Stress Waves test. A nocebo effect, which is an adverse psychological response due to a negative outcome, can occur when doing the Cigna Stress Waves test. So for instance, I have low levels of stress, but I come upon the Cigna health insurance website. I take the Cigna Stress Waves test, and it tells me I have moderate levels of stress and I'm at risk for high blood pressure or cardiovascular disease. Now, I'm starting to feel stressed out. And on a macro level, AI tools that are prematurely deployed within the healthcare system have potential consequences of misdiagnoses, misinformation, and misinterpretation of data.

Batul Yawer:

Next slide. The Cigna Stress Waves Test has some validation issues. They made claims of clinical-grade performance without publishing any validation studies, and our study was actually the first external validation, and this revealed poor algorithm performance. The Cigna Stress Waves Test has since adjusted its online marketing post our publication from "clinical grade AI technology" to "proprietary AI technology". This indicates the importance of validation and repeatability.

As researchers, we do have some actionable steps that we can take. We can take some steps by emphasizing the basic statistical validations, such as repeatability and construct validity, and this can contribute to the development of reliable AI algorithms. By doing so, this also increases public awareness about AI limitations, encouraging a critical evaluation of AI reliability. For instance, when I see the verbiage "clinical-grade AI technology", my psychological bias is amplified to trust the system under the fallacy of AI functionality. However, when I see the verbiage "proprietary AI technology", I'm more inclined to critically analyze my results and not take them at face value.

We can also establish guidelines, and that makes it very essential for maintaining the integrity of these AI tools, and, of course, collaborating between researchers and regulatory bodies, such as the FTC, strengthen the development and oversight of these AI tools. Next slide. These are the references that I had used throughout my PowerPoint presentation. Then, next slide. Thank you all for listening.

Ronnie Solomon:

Thank you all for those wonderful presentations. Really fascinating research. We're going to jump into the Q&A portion of our panel now. I'd like to start with Patrick. Patrick, unlike some complex technologies that many people don't seem to understand, like private browsing modes, VPNs, encryption, quantum anything, people seem to have a pretty strong awareness of the privacy concerns around AI specifically. How should this change our approach to privacy and by ours? Feel free to answer as to any group or perspective, whether that's regulators, developers, the press, etc.

Patrick Gage Kelley:

Thanks for the question, Ronnie. I think that this actually should be seen as really optimistic work. I think that one of the points here is that people are really interested in AI, they're certainly excited in AI. That's why we see it being used as a marketing tool, as in the last presentation. I think we really do have more opportunity here to engage and to really say, "You have a good sense of how some of these privacy concerns work, what would meaningful consent look like?", and that means we can really do some workshops and bring people into the public policy process, the regulatory process, to say, "What kind of consent would you want in the AI space? What kind of data collection minimization would help?" I look forward to having more of those conversations, especially as we see this newest wave of generative AI technologies, again, really excite people about the potential for AI.

Julia Horwitz:

Thank you, Patrick. The next question is for you, too. We understand from your paper that some people thought that AI would make privacy better. Why did they think this, and is there more that you can share about privacy optimists?

Patrick Gage Kelley:

Yeah. Overall, again, as we mentioned briefly, people think AI is going to make everything better. Right near the top of that list, again, is healthcare, so when you see people advertising AI as a solution for stress testing and for other things, people have really been sold that AI is going to improve drug delivery and can fold all these proteins. In the healthcare space, it's going to be really, really helpful. Some of this is probably thinking, "If it can make all of that better, it can make my privacy better too."

We saw answers in the data where it said AI would be able to develop new encryption schemes, it would be able to somehow protect data more In that way. We saw that AI would be able to be used as a tool to prevent hacking. People could deploy it upfront to say, "What would the hackers do?" and have AI be the hacker first so that the systems couldn't be hacked. There were some real examples, but there's also just a lot of optimism around AI, and I think this is a great call for all of us to say, "How could we use AI to make privacy better? What can we actually do?", and then go and do some of that and actually have it work in some of those ways that people were hoping it would.

Ronnie Solomon:

One more follow up for Patrick. It sounds like your work was done before the explosion of consumer facing generative AI products and technologies. How do you think generative AI changes the landscape, if at all?

Patrick Gage Kelley:

I think this is one of the questions we can't answer from the data, but it's something that we are definitely still committed to watching. Not just our team at Google understanding how this will change things, but the whole world is watching as people have more attention paid to AI. I think we will have to see in future surveys and future data work, and also in some ways once the hype has died down a bit. I think there was a moment where it was kind of like, "Wow, AI is now accelerating. AI can do anything at all", and the question for some of these privacy concerns is going to be how does that end up working out? Is it privacy concerns? Is it other concerns that generative AI really raises? It may not shift things much, or it may just further engage people with thinking about what outcomes could come from this change.

Julia Horwitz:

Thank you, Patrick. The next question that we'd like to ask is nominally for Umar, but I would encourage any panelist who has thoughts about this to weigh in as well. Umar, we would like to ask you, what can large language model developers do to build secure LLM platforms? What would that platform look like?

Umar Iqbal:

Great question, and I have a compartmentalized answer, I guess. The first thing they can do is that they have policies and guidelines which currently do not seem to be imposed. I think, as a start, they can start strictly enforcing those guidelines. The second thing is that they should do a better job at acknowledging the issues reported by developers and users, which is also missing. That's a small thing that LLM platforms can do. The other thing is to basically design the platforms from the ground up in a way that they have strong security and privacy guarantees. When you say that, I think the obvious thing that comes to mind is that these techniques need to be researched or designed from scratch. Even if we look at the techniques that have been used in existing platforms, like sandboxing, compartmentalization, monitoring, some of the execution flows and all that. Even if you start from there, you can substantially improve the privacy and security of the systems. I'll stop here.

Ronnie Solomon:

That's very interesting. Next question, again, is for you Umar. To the extent you didn't already describe this in your presentation, can you talk about what kind of risks are users currently exposed to when they interact with large language models? For example, can you describe any of the plug-in vulnerabilities that you referred to in your presentation?

Umar Iqbal:

I guess the first type of issues that users are exposed to are the ones that exist because of the bad design or the lack of enforcement. These are the cases where the plug-ins or apps, they're not inherently malicious, they're not trying to do anything wrong, but because of the flaws in the design, how the platforms have implemented the apps, they have been exposed to these issues. The other instance is that since the platforms are not designed with security, privacy, and safety in mind, so it leaves more opportunities for adversaries to exploit the platform.

For example, if there's a motivated adversary, they can hijack the whole conversation history of the user. They can hijack the user's session with the platform, they can steal the credentials that are needed by some of the apps and which are present in the form of the text, which basically any app in that session can extract. These would be the harms that immediately come to my mind. I guess another important point here is that this space is very new, and as people are figuring out different mechanisms to use these systems, attackers are also figuring out new interesting mechanisms to do harms. Some of them don't currently exist yet, and as this space evolves, there will be more and more complicated attacks and risks.

Julia Horwitz:

Thank you. I think a follow-up question is what, if anything, can users do now to protect themselves against these risks?

Umar Iqbal:

I think one good thing that platforms are doing right now is that as they're launching these applications, and some platforms are doing a better job than the others, they have very clear warnings, or the programs are in beta, they're not open to all the users. Users have to make an active choice that they use an app or give explicit permission that they want to use an app. As someone who does not have a lot of tech-savviness, I think my suggestion would be to use the services which they're familiar with, and if they notice something that's abnormal, they should immediately report it to the LLM platforms. There is a lot that needs to be done to secure these platforms. Some things are needed to be done by the platforms, but I anticipate that in the future there will also be client-side solutions which users can use without relying on the platform to protect their security and privacy.

Ronnie Solomon:

One final question for you, Umar. What are the takeaways that you want the audience to be aware of from your presentation? I know we had to cut you off at the end of your presentation, so just want to give you a quick chance to,

Umar Iqbal:

Quickly, three takeaways. First, I think these platforms are getting more complex. Their functionality is going to be on par with other computing systems, like mobile or maybe desktops, so they're going to be very, very capable, which means that the harms will be more pronounced. These harms, in addition to from adversarial actors, they're also going to be because of the ambiguity and imprecision of natural language. That's something unique to the LLM platforms. LLM platforms basically need to design their platforms since they're just starting out with security and privacy as a key consideration. They can do that by using threat modeling, using the processes that we propose in our research and others within the future.

Julia Horwitz:

Thank you, Umar. Batul, we have some questions for you as well. The first of which is, you touched on this in your presentation, but there was so much interesting information we wanted to make sure that we got to this specifically. Is it possible to detect psychological stress based on speech? In other words, is this a problem with a solution?

Batul Yawer:

From the lit review that we had done, it seems like there aren't many significant correlations between speech features and detecting psychological stress. There are a few, but speech in itself is pretty variable, so to be able to pinpoint something that specific... I had briefly mentioned in one of the slides when we're talking about a concept like speech, and then we're talking about a concept like stress, they're both so complex and so variable. Then, to pinpoint both of those complex and variable ideas is really difficult. As of right now, the answer is no, there is no scientific backing to figuring out how you can detect psychological stress from speech.

Ronnie Solomon:

Batul, just a quick follow up there. You mentioned that the app in question was released without validation data. Can you talk a little bit about what you would've liked to have seen? In other words, what does strong validation data look like in this context?

Batul Yawer:

That's a great question. I will say probably validation compared to the PSS-10, or a psychological stress assessment tool that has been validated itself. I want to point out that they did claim that the AI model was trained on 150,000 samples of speech, but there's no publication of that data., I don't know. It's like, even if they used that many samples, how is the AI tool classifying psychological stress from speech? First, I would like to see how they do that, and then second, I would like to see them compare it to a robust measure for psychological stress.

Julia Horwitz:

Is it fair to assume that including more data would not fix the issues with the AI model?

Batul Yawer:

That is exactly correct.

Ronnie Solomon:

One more question for you, Batul. To the extent you didn't address this in your presentation, can you tell us how the AI model interprets voice data to classify speech?

Batul Yawer:

On the website itself, it does claim to use pitch, tone, word choice, and pauses to evaluate speech. Outside of that, to understand how the AI tool interprets all that information, I'm not sure. It's kind of like a black box, and they don't have any additional information about it. To add on to Julia's previous comment, adding more data definitely would not fix the issue because, as we saw with the repeatability results, even if I took the test twice back-to-back, I'm getting inconsistent results.

Patrick Gage Kelley:

Can I jump in on your question?

Julia Horwitz:

Yeah, please.

Patrick Gage Kelley:

Batul, I think one of the things that's so interesting here is that for people, again, who've heard that AI is going to do all these great things for healthcare, that it's going to do it, you could also imagine that some AI was trained to listen to our speech and that there were things that changed in our speech when we were more stressed, and that there was this sense that AI, who gets to know ourselves better than we know ourselves, could find something that maybe we wouldn't know. I guess I'm wondering how do you want consumers to more critically think about some of these things? Because it seems plausible that with the right amount of data of speech and collated with ground truth around stress, maybe it could find something, but who knows, there's no paper, they haven't done that. Their current test clearly isn't going to do that. How do you want people to skeptically consider these things when it's just saying this new AI could do everything?

Batul Yawer:

In my personal opinion, I relate it to just fake news. How do you go about navigating fake news? Just seeing things on social media isn't completely real, and dealing with AI isn't completely real most of the time, and it's specifically in the case of my study. AIs researchers, I feel that validating these kinds of AI tools, especially in the healthcare setting, is kind of the first step. Ensuring that they don't use verbiage or make these claims, like clinical-grade, because that does cause this psychological bias around that AI tool. Those to me are the first steps on our end to help consumers look into this and realize the AI hype is kind of over and I need to make sure that these results are valid before go into it.

Julia Horwitz:

I think I have one final follow-up question for everyone, which is, what can the FTC do to help?

Batul Yawer:

For me, I really like doing research, so if there are more tools that you find that might be a little skeptical or that might seem too good to be true, I would love for those to be sent to researchers so that they can see if there is any data behind it. If there isn't, they can conduct some validation studies.

Umar Iqbal:

From my side, I think FTC is doing a great job with events like PrivacyCon, of course, and I've been sending emails and stuff to different people at FTC and they've always been very responsive. I think one thing where FTC can contribute more is to maybe actively engage with researchers, maybe have some small seed grants where you can collaboratively work on a project and explore problems. That's one thing that I would wish happens in the near future.

Patrick Gage Kelley:

I think from the work we're doing here, I just really want to see more opportunities to engage the public. Even having events like this where researchers and the FTC come together to share findings, I think having these things be available so that people can really take a more active role so that they can say, "Here's what I'm concerned about with AI, or with this psychological testing that I might use, or with this LLM that I might be having some developer work on." There's a lot of range for people to get involved, and I think having more channels where the FTC can help facilitate that is really wonderful.

Ronnie Solomon:

Thank you all so much. With that, this concludes our panel on AI and Machine Learning. I want to thank all of our panelists. We will take a 10-minute pause for our afternoon break, and we'll resume at 3:05 PM for our Mobile Device panel. Thank you all.

Jamie Hine:

Welcome back from the afternoon break, everyone. As a follow-up to some comments in the last panel, want to remind everyone that we have the privacycon@FTC.gov email address. If there's any questions for the panelists, please send those. That's an email address that operates all year, so if you have any research that you think might be interesting to the FTC, you're welcome to send that at any time. If you have any ideas for next year's PrivacyCon, we welcome those as well. We have two more panels, we're going to move right into the Mobile Device Security panel six.

Madeleine Varner:

Thank you so much. Good afternoon. My name is Madeleine Varner and I'm a Senior Technology Advisor at the FTC. I'll be co-moderating today's panel on mobile device security with my colleague Andy Hasty, who's an attorney in the Division of Privacy and Identity Protection. Today we'll be hearing from three researchers, Abbas Acar of Harbor Labs, formerly of Florida International University, Allan Lyons of University of Calgary in Canada, and Sumanth Rao of UC San Diego. To get started, I'm going to turn it over to Abbas.

Abbas Acar:

Hello, my name is Abbas Acar and I'm currently working as a Senior Research Scientist at Harbor Labs. This work has been done when I was a postdoc at Florida International University. In this work we also collaborated with Dr. Gülistere Etuncay from Google. In this work, our work is on analysis of Android security updates. This work also published and presented in NDSS. Next slide, please. Android is by far the most popular operating system with over 3 billion active mobile devices. Like any other software, vulnerabilities are also found in Android as well, and patching regularly and timely these vulnerabilities is critical and essential for end user security as well as privacy. However, with the new devices and models introduced every day, it becomes more and more complex to be able to distribute these patches to all of the devices in the market.

Therefore, users are experiencing issues and irregularities, such as delays in the updates or missing updates. Users experiencing these irregularities express their opinions and complaints in the community forum, as we can see some of these in these examples. These issues are sometimes specific to the models, sometimes specific to the carriers or regions. Next slide, please. Security updates rollout process in Android consists of several steps. In the first step, Android Open Source Project, AOSP, collects the CVEs and patches affecting devices running Android. Next, AOSP publishes the multi-security bulletin containing these CVEs and patches. After AUSP, OEMs, like Samsung, apply patches and do the customizations and finally create the final firmware containing the security updates. After OEMs, carriers do the test and approves the final firmware. Finally, end users accepts the security update containing these patches and installs it through the over-the-air update. Next slide, please.

Abbas Acar:

Some prior studies in the literature look at the issues in this process. However, these prior studies are either limited to specific type of vulnerabilities, such as kernel vulnerabilities, or they have limited number of security updates or firmware images. And in addition to being limited in the scope, there are also some simple but highly relevant questions that these studies do not answer. For example, what's an average support duration of an Android device? And what are the factors impacting the distribution of security updates? And also, these studies do not perform an analysis on the unpatched devices. For example, when does it become unsafe to use unpatched devices? Next slide please.

For our analysis, we put a lot of effort into data collection, and we collected security updates from top three OEM and as well as Google. And our data collection ended up with 354,000 security updates from Samsung, 2000 security updates from Xiaomi, 9000 security updates from Oppo, and 900 security updates from Google.

And as you may have noticed, Samsung has a lot more security updates than the other OEMs. The reason for that is Samsung is the biggest player in the market for a decade, while Xiaomi and Oppo are recently catching up, and Google has limited number of devices in the market. The security updates that we collected has been used in hundreds of different unique devices, as well as tens of different unique countries. And it almost covers the entire history of these OEMs. In addition to security updates, we also collected support list to be able to better understand what's available for end user to check whether their devices are supported or not. Next slide please.

And let's look at the security updates during the supported period of these devices. Our analysis showed that Samsung devices received around 16 security updates on average, and they are being supported for a duration of two years. However, during this analysis we noticed that these devices stay in the support list for almost six years' duration. Therefore, we conclude that the stay in the support list is not equal to the duration that devices will receive security updates in practice.

In addition to the count and duration, we also found that Samsung devices on average receive 50 days of frequency security updates, and 140 days of delays. On the other hand, Google devices receive around 36 security updates for a fixed duration of three years with a frequency of one month and essentially with zero-day delays.

On the other hand, we also look at the Xiaomi and Oppo devices. We found that these devices receive relatively fewer security updates with a better delay performance. Our conclusion in this part is that Google devices, Google provides regular multi security updates for a fixed duration of three years. While Samsung security update behavior varies significantly depending on the factor that we will see in this presentation. And Oppo and Xiaomi of relatively fever security updates for a shorter duration. Next slide please.

So far we look at the security update behavior of different OEMs during the supported period. However, these devices stopped receiving security updates after a while. And in this part, we wanted to look at the behaviors of these devices during unsupported period. And our first results show that 36% of the pairs have at least one impacting CVE and they haven't received any security updates in the last three months. And we also wanted to look at the CVE accumulation of these unpatched devices over time.

Our findings show that in the first quarter, these unpatched devices are likely to receive around 76 CVEs on average, and three of them are critical. And the number of CVE can reach up to 380 CVEs in two years, and around 600 in five years. And we also look at the other factor, other metrics of these CVEs on the unpatched devices. And we found that 89% of these CVEs do not require user interaction and 86% of them can be exploited with minimal effort on the attacker's part, while 27% of them can be exploited remotely. Next slide please.

And so far, we look at the different behaviors during the supported and unsupported period. And whether saying devices support or unsupported is not a binary decision, we found our analysis showed that support types could be monthly, quarterly, or bi-annually. And we identified that these different support types can have different support behavior. For example, a device during a monthly support is receiving around 36 security updates. While during the quarterly support, these devices are receiving around nine security updates. And during the biannual support, these devices are receiving four security updates.

And these results align with some prior study showing that 25 of delays for Android devices. As we can see, there are different support types other than monthly support type. Therefore, this shows the dataset and comprehensiveness of our analysis. Our conclusion in this part is that timeliness and availability to have security updates varies significantly for different support type. Next slide please.

The second factor we wanted to look is the geolocation. And for this one we grouped and sorted different countries and we look at the top five and bottom five countries. And please note that in this analysis we only look at the monthly supported devices. So all of the devices we used in this analysis are supposed to be getting monthly support. And we found that top five countries received three times more security updates than the bottom five countries.

And to better understand the scale of this impacting factor, we also mapped the number of security updates into the real world map. On the right side, you can see the number of security updates received by devices in each region, as well as the delays for each region. And you may have noticed the irregularities between different regions. Therefore, we conclude that there are significant variations in support across different regions and countries. Next slide please.

And during our research, we identified some key issues that could be immediately fixed. For example, we found that there are significant variations in models and pair support behavior, and these two pairs that we found that one of them has stopped receiving security updates in October 2020, while the other pair belonging to the same device is still receiving security updates as of today. And we also identified some discrepancies in the support list. We found that for example, Galaxy A7 2018 model was in the support list until December 2022. However, this specific pair stop receiving security updates in January 2019.

And we also found some discrepancies in the partnership agreements. For example, AER-certified pairs, Android enterprise recommended certified pairs have guaranteed support date today or sometime in the future. However, we identified 200 AER-certified pairs have not received any security updates for at least a year. And sometimes announcements also could be misleading. For example, an announcement made by Samsung in February 2021 and saying at least four years of security updates for eligible devices. And we checked those eligible devices and we found that 343 pairs out of 8,000 pairs already stopped receiving security updates. Next slide please.

Our key takeaways from our research. We identified significant variations in updates due to OEMs, geolocation and device type. And we found that being in the support list does not guarantee that these devices will receive security updates. And we also found that unpatched devices pose immediate risks to the end user for publicly known remotely exploitable and simple attacks that require no user interaction. And in general end user is left with unclear, inconsistent or minimum information across different sources, and sometimes even from a single OEM. And in this matter, OEMs can adopt guaranteed support date, end of support device list, or model-based support list approach to address these issues. And our general conclusion from this research is that there is room for accountability and improvement in which OEMs should publish the support duration of all device models and pairs and maintain the timelines of their security updates. Next slide please.

Thank you so much for listening my presentation. Our code and data is available on GitHub, and this work has been approved by Artifact Committee of NDSS. Available, functional, and reproduced.

Madeleine Varner:

Thank you so much. We're going to now turn to Allan?

Allan Lyons:

Thank you. My name is Allan and I'm going to talk to you about some of the excessive logging that we found on Android phones. This paper was published last year at Usenix. Next slide please.

Who remembers contact tracing? The Google Apple explosion notification framework was intended to be privacy preserving, but there was a problem. The key idea behind this framework is a concept of securely produced random looking rolling proximity identifiers that would make it impossible to attract individuals while still providing a way to notify others of their potential exposure once there was a positive diagnosis. Given the popular understanding of COVID at the time, the framework was based on the solid algorithm, but on Android there was an implementation problem. See the key advertised privacy of this framework, we relied on these rolling anonymous identifiers to remain secret. Next slide please.

However, we found that regardless of the Android device that was running on the framework with log, these supposedly anonymous identifiers in the system log. So, in the top section we see an example of the log message generated each time that a phone generated a new identifier. And in the second section, we see the log messages that were generated whenever another device was heard. So taken together, these log entries provided a picture of all anonymous identifiers in an area, and anyone who was able to correlate the log files from multiple devices would be able to reconstruct social graphs describing about who met with who, et cetera.

For example, if you saw the two different devices recorded here in the same identifier, you would know that these two devices had been in the same area. The bottom section is something slightly different. In the original version, if you recorded a positive diagnosis, this message box would display asking you to share your random IDs. The corresponding log entry, however, is from the Android system indicating that the prompt was accepted. This is built into the OS. This data probably should never have been logged in the first place. And then when this was disclosed to Google, they quickly fixed this specific case. Next slide please.

So what is ending up in the Android logs? The potential for security problems stemming from sensitive data has been known for a very long time and Google has developed some very clear policies surrounding the logging of sensitive data. Sensitive information in the logs is only a problem if someone can read it and the simplest solution is to not log it, hence the prohibition against the routine logging of potentially sensitive information. But how did that end up in the logs? And secondly, who has access to these logs?

See if all developers followed Google's development guidelines, there wouldn't be anything to see. However, that is not the case. And we found several examples of potentially sensitive information ending up in the logs. So we decided to examine whether the different types of identifiers were ending up in the log files? And types of identifiers that could be used to identify an individual? So as noted in earlier today, that location is particularly sensitive since if you'd know [inaudible 06:16:31] someone lives and where they sleep and where they work, you likely have identified a unique individual. So next slide please.

So what particular is landing up in the logs? So we manually tested several new phones or newly imaged phones and examined their log files for identifiers such as phone number, email addresses and whatnot, and non-resettable devices, IDs such as the Android ID, the IMEI of the phone, the serial number, the MAC addresses, and several other identifiers that uniquely identified the device. We found the identifiers in all of the phones that we tested regardless of the model of the manufacturer. Some even logged to Bluetooth payloads.

So as is shown in prior work, there is a concerning lack of control over the privileges in the system about which apps have which permissions. And even though in Android for more than a decade, there has been a really restricted permission that allows apps to read the log data and apps have to be pre-installed by the manufacturer to even to get that permission. But several in our set had lots of apps that could read the log data and we found even one phone that had 95 different apps with the permission that would potentially allow to read the data. Next slide please.

But that left a question of all the PIIs in the log files or the new phones that we found. Is that true for all the phones out there in the world? So inspecting log devices, devices at scale is challenging. We can't purchase every handset model from every manufacturer around the world. And regardless, even there's even the regional variations as pointed on the previous session that it's different. You get different settings depending on where you are in the world.

So what we did was we developed a field study that where volunteers could go to our website and with an accompanying app analyze their own phones for the presence of identifying information on their own system. And I should note that none of the data, we didn't actually collect the data, all the analysis was actually done on the volunteers' own devices and we never saw the actual data.

And after filtering out incomplete reports and devices that obviously appeared to development devices, we were left to data from 1,214 unique devices representing 529 model variants. And we found that sensitive information was found across the board. Next slide please.

So that gave the question, some of the identifiers and logs were found on so many phones, where was it coming from? So one of the prevalent ones was related to Wi-Fi. So we checked on Google Pixel 3 running a factory build of Android 12 and examined the log entries generated while the phone was establishing a Wi-Fi connection. And we found a lot. On this slide we have found several MAC addresses. So you can think of a MAC address as a unique identifier to either the device, or to where you're connecting to. And since Wi-Fi, access points, and cell towers don't move around, if you know the MAC address of the Wi-Fi access point, you know the location.

So for those who are trying to geolocate my office using these MAC addresses? I anonymized these slides, so that doesn't actually point out to my office anymore. And since Android is open source, we tried to find out where's the code that's actually generating these log entries. Next slide please.

So the phrase, "Own MAC address," is pretty obvious and easy to find in the log entries, and so it didn't take us long to find the code. And with a bit more digging, we found out that there's a build time constant in this part of the code that if config no standard or debug was defined when they were building the code, then none of these log entries would've shown up. So one thing I'd like to point out again is the part of that name where it says, I'd pronounce it, standard out. And if you remember the context and the era where these kind of software was originally developed way back when my hair was dark, you would know where that's coming from.

Traditionally, in the code you might have had a lot of print line statements that would just print out messages of the code run. And since Android doesn't have the same type of console like your old desktop, it appears that the log messages are being used to the same sort of effect. And so enabling these debug messages in our production code is something that we referred to as debugging during deployment.

As a comparison, we also examined the logs from, Xiaomi Redmi Note 9 and tried to find equivalent logging. And that was one model that we had that we cannot find these log entries in the log. And the explanation we have is that at least there's one developer out there that notice that if you enable that constant, that this location information would not end up in the log. So there's at least some developers doing the right thing. Next slide please.

Another example that we've looked at was apps. So we performed case studies on manually executing a few apps in a more realistic way and then examine the logs. So this is an example of a log from the Shoppers Drug Mart app. We've also looked at the CVS app, which is a popular pharmacy in the U.S., and it worked out about the same way in this particular location, this bit of data that I've pulled out of the log, it was highlighting that the particular person was shopping for daily low dose aspirin. Now the only thing that you would use that for is if you were a high risk of heart attack or stroke. So, health data is in the logs. Next slide please.

So where's that coming from? Well, we found in the logs that the Adobe Experience, SDK was being used. And it turned out that in all these apps, even though in the documentation for this library it says, "Don't set it to debug mode," in the shipping apps, it was actually set debug. If you set that debug to not be true, it would not log. Next slide please.

So do apps read logs? So what we did was we linked our field study we referenced earlier in all the models that we saw there with a data select that was collected for in a previous paper. And we found that given those models, we would've expected to find about 1,319 apps with the read logs permission that would've given the whole access to read the entire log. And you can read that paper if you see what all can happen with the log files. And another example is with a Google feedback tool when if an app crashes, it does upload the entire log. Next slide please.

And so just as a conclusion, we know that logging of potentially sensitive information is prevalent despite Google's recommendations to protect end users, including identifiers, location, health-related information, and that many pre-installed apps on these phones can, and some do read these log files. Thank you.

Madeleine Varner:

Thank you so much. Sumanth?

Sumanth Rao:

Thank you. Hey folks, my name is Sumanth, I'm a PhD student at UCSD. This work was led by Alex Liu, who's also from UCSD with folks from NYU, Cornell Tech, and was presented at PETS last year. Next slide please.

So the context here is that stalker-ware is increasingly being used every day for the purposes of stalking and monitoring phone users. The adverse result of this is that you have victims of spyware abuse. I want to draw your attention to one statement from a survivor of intimate partner violence who says, "He's tracking everything. Whatever I do, he sees," which is referring to them being actively monitored by their partner. Now to enable all of this, you have a market out there for consumer spyware apps where you have dozens of vendors and competitors who provide apps with different functionality which are easy to install for the purposes of spying. Next slide please.

So these spyware apps market themselves in clever ways. So here in the U.S., both the Computer Fraud and Abuse Act, as well as Provisions of the Wiretap Act prevent you from actively monitoring somebody without their consent. But these apps are, the way they market themselves is by saying they're either for parental control, or employee monitoring where the laws often blurred. And each of these apps provide and advertise a bunch of other functionalities.

For instance, monitoring SMS is some of the basic functionalities they produce along with more advanced functionalities like WhatsApp messaging, or iMessages, or Snapchat. So while the community has an understanding of what these apps are and what their capabilities are, we don't yet have an understanding of how it is that they achieve these capabilities. And so this motivated our first research question. Next slide please.

So to see how they use the Android APIs, we take a reverse engineering approach. We focus on Android, since this is where majority of the spyware apps are found. So we start by taking the source files of these spyware apps and decompiling them into a human-readable JavaScript file. And we study the API usage by trying to find functions that these apps advertise and then trying to link it in the Java source file that these are the particular API calls in Android that these apps are calling. And how is it that they're implementing this? Next slide please.

So in this paper we analyze 14 leading android spyware apps and we study how it implements a broad range of capabilities. I think the note here is that all of the apps that we studied are off store apps. That is, they're not found on the Android or Google Play Store because they violate their terms and conditions. But because Android allows side loading, it is very easy to install these apps from a third party vendor and it only takes minutes for them to be installed.

So in this paper, we reverse engineered 14 apps and we categorize them based on their capabilities. So some of them offer basic features like call logging, text, and some others offer quite advanced features, things like hiding the app icon and obscuring the process of uninstalling these apps, or even getting access to the phone camera in a very subtle way. The takeaway is that most of these are using Android APIs in a creative and a novel way just for exploiting their purposes. Now in the interest of time, I'll present just an example of this creative approach documented in the paper. Next slide please. So one of the capabilities is called invisible camera access. A spyware can secretly take a picture of a victim who's using the phone and send it to the cloud where the attacker can then see what the victim is doing. Now, Android API today does not offer an API, which says, secretly take a picture. So we found that the spyware apps achieve this in a subtle way. So one of them is using the preview option. So a normal app, such as a camera app, uses the preview concept to make the users see their own picture. For instance, if you're taking a selfie, you'll see a small preview of your front camera at the bottom left part of the screen. But the spyware apps takes the same concept of a preview and shrinks the size of the preview to a 1 x 1 invisible pixel, which it then embeds into the browser. And this is very hard for a normal user to spot. Next slide please.

Okay, so now that we have given a glimpse of how technically sophisticated these apps can get, unfortunately they aren't very secure. Just in the news last year, LetMeSpy, which is again, one of the spyware apps here was hacked and around 200,000 devices and millions of data points which users had uploaded to these servers were leaked. And much previous around four years ago, FlexiSpy, which is another app, also got hacked by hackers and a lot of user data was exposed. Next slide please. All of these breaches led us to ask the second question, which is what are the measures in place taken by these spyware vendors to safeguard the data that is being collected? And to do this, we've involved an end-to-end approach. So this figure shows how the architecture of these apps are, they collect data from a device and they stream it to the backend server, and then they provide a UI or a web portal for the attacker to log in and try to see what the text messages and calls and events are, which have been streamed from the victim's phone device to the backend server. And so, we focused our protections trying to see what sort of protections are available on the client side, on the server side, and along the network, which connects both of them. Next slide please.

The overall result is that we identified five classes of privacy vulnerabilities in all of the 14 apps that we study. Again, we classified these based on basic protections and more advanced one. For instance, along the network, one of the most common issues we found is that sensitive information is being leaked. Versus along the server side, one of the most common problems is that data which is supposed to be deleted is actually retained. So data retention policies are pretty bad, or you can actually get unauthenticated access to victims' data stored on the cloud user.

For instance, you can do a cross account request forgery where you take the tokens associated to one account, swap it with another account, and try to get access to a different person's credentials and data. And to enable all of this, we purchased subscriptions from each of these 14 apps, used our own test accounts and tried to play around and see if we're able to get access to a lot of data. Again, for the interest of time, I'll present just a few examples. Next slide please.

So in this scenario, an adversary along the network, or a man in the middle, so they're in between the phone and the backend server and trying to collect information, which is put onto the network by the phone app. And so we observe that in practice a lot of apps transmit data through plain text, which is HTTP. And this data can even include sensitive information like the passwords, which the user uses, as well as all of the data which the app is monitoring. Things like SMS message, or call logs, or phone, or your OTP messages. Everything gets sent over the wire in plain text. Another example of this is that when the attacker asks to take a subtle photo of the victim, the phone app then takes a photo using one of the many approaches I talked before, and then puts it onto the network as packets not encrypted in a plain text HTTP format. And these packets can be reconstructed back to get the actual image. Next slide please.

So a second scenario of this is in SMS commands. So SMS commands are a feature offered by certain advanced spyware where if the victim is not connected to the internet example, they're roaming outside or they are not connected to the Wi-Fi, an attacker can send an SMS message to the phone and the spyware app would see the format of the SMS message. It would be a special format with a command in it, and then it would process the instruction and send it to the server. But in this case, we realized that the SMS command is virtually unauthenticated. For instance, anybody could send an SMS command pretending to be the attacker and the spyware app would execute this. And so I can present a bunch of other results in a similar space.

But the key takeaway here is that there is not enough effort in securing the sensitive data. And so there might be a few reasons here. The first is that the incentives aren't aligned because it is the attackers who are procuring these apps, but it is the victim's data that is ultimately being stored in the cloud. And from the spyware vendor's perspective, it is a much more harder and complicated problem to work around the Android APIs and it is more important for them to prevent being detected versus implementing more secure protections to the data which they collect. Next slide please.

So in summary, we studied the spyware capabilities of 14 leading spyware apps. We documented the creative ways in which these spyware apps abuse the Android ecosystem, and then we identified a range of privacy deficiencies in which user data is ultimately exposed on the web. Thank you.

Madeleine Varner:

Thank you so much. Those presentations were all fantastic. So now we're going to move into the Q & A portion of the panel. And so to start, I'm going to kick it off with a question for everybody. Mobile devices are so prevalent. I'm wondering what can everyday consumers take away from your findings? And let's start with Abbas.

Abbas Acar:

Thank you. In my presentation, in our work, we show that there are issues delivering the patches for the known vulnerabilities. And we know the, so the vulnerability, the attackers know the vulnerability and however they're not being patched on the end user device. And for this one, the users have very limited information that they can check, and they can basically go check whether their devices is supported or not. Sometimes as we have seen it could be inconsistent between different sources. When your support list says, "Okay, you're going to get the support, but you're not getting any support." And they do post on

Abbas Acar:

On the community forums, these companies. So, my conclusion here is that it looks like more responsibility of OEMs rather than the end users to be able to get the update. Some technical users can go update their devices sometimes with the customer operating systems and they can check and they can go to the technical details and configurations of their phone and they can see which update they don't get and what kind of vulnerabilities are fixed in those specific updates. However, the general users have no way to verify or to know whether the device will be supported or will receive the next secure update or not. So therefore, I think there are things that the technical users can do, but for general user, I think it's the responsibilities on OEMs, I think.

Madeleine Varner:

Allan, same question.

Allan Lyons:

Yeah. And in some ways, I'd almost say it's a similar answer in the sense that it's up to the OEMs because, and even put it on Google as well, I tried to point out there that Google has a policy, before NAP can be published on the app store. It's got to go through certain things. Disabling logging, extra logging is already there. They already have prohibitions against logging. And the examples I pointed out, none of those were malicious. They were either oversights or mistakes in the programming or something. And then furthermore, now that this stuff is in the logs in, like I pointed out there, even on brand new phones that apps that can get the permissions to read the entire log file are put there by the manufacturers. And if those have bugs, that can be exploited by other apps to leverage their way into the log file. Well, if that gets back onto the updating of the Android phones in the first place, because if that OEM doesn't update the apps that they pre-installed, they probably won't be getting updated. They won't get updated through the regular play store thing. Right?

Madeleine Varner:

Yeah. Thank you Sumanth, same question.

Sumanth Rao:

Yeah, I think at least with that, the spyware stuff, it's a couple of reasons. So a company like Android can definitely implement more protections in place, especially because one of the most simple way all of the spyware gets installed is by side loading. And this is a much harder problem in the case of iOS because the most simplest step, which involves installing the spyware app is 21 steps long versus Android where it's much simpler to just enable the setting and install this in a span of five minutes. Now the other approach is also depending on the sophistication or the technicality of the user, some of these apps don't do the greatest job in concealing themselves and a more tech-savvy user would be able to spot that versus if let's say a random app accesses your location over a span of five minute intervals, then you know that it's doing something another indicator. And so it's a mixture of both, both from the OEM side and from the user side.

Madeleine Varner:

Thank you.

Andy Hasty:

I'm going to ask maybe a harder question, piggybacking off of Maddie's question, if you were to distill down the takeaway from your research into one, maybe two sentences for everyday consumer, what would it be? And I'm going to start with you, Abbas.

Abbas Acar:

So everyday users, I think they should go check whether their devices in support list or not today. And if they're not, at least some of them will be able to notice that they're not getting any support and it's out there. And then as we have seen, the number of CVs vulnerabilities is accumulating over time. Some of them are critical and very, very sensitive, both in terms of security as well as pricing. So therefore, they should. And the second thing they can check is the configuration. They can see the level of their security of their devices. It's called security patch level. So they can go check how up to date is their device and depending on how is the result looks like, they can update their phone or not, I think.

Andy Hasty:

Thank you. Allan, same question. Trying to distill down the takeaways from your research into a single sentence, maybe two sentences for everyday people.

Allan Lyons:

Okay. The key thing I'd say to take away is that there's way more personal information on your phone than even if you are expecting the app to, that they did a good job. There's so much more information that reveals everything about you, your location, apps you use. Our earlier session was talking about period tracking apps. Probably all the interesting information out of those apps is probably in the log. Even if they do have a good privacy policy, it's all there.

Andy Hasty:

Thanks, Allan, how about you, Sumanth? One or two sentences takeaway distillation for normal everyday consumers.

Sumanth Rao:

Sure, sure. That's a tricky question. I think my answer would be something similar to the advice I give my grandparents, which is keep a balance on all the apps you install on your phone. If there is something which you see which is out of the ordinary, stick it out, try to open it, try to see what app it is and uninstall it if it's not needed. For an everyday user, it's very easy to just keep, just lose track of all the apps which are on your phone.

Madeleine Varner:

So I have a question specifically for Abbas. I'm curious, why does the distribution of security updates vary so widely from country to country and even for the same device model?

Abbas Acar:

That's actually a great question and considering the data that we have, we collected all the data from public resources. So we have a limitation of being able to access to the internal resources of the OEMs decision process. But like you said, two models, and they're being used in neighbor countries and one of them is still receiving security updates as of today, one of them stopped years ago. And we have no way to verify why they stopped in one of them and the other one still receive. Maybe there is no more users that in that specific country, but maybe they're overloaded in that specific country. We can only speculate and say some technical factors, but right now with the data that we have actually we have no way to verify. So yeah, that's the limitation of our, we actually study.

Madeleine Varner:

Thank you.

Andy Hasty:

Allan, I'm going to ask you a question here. In your view, are developer guidelines effective enough here? How effective are these guidelines about not logging sensitive information and preventing these practices?

Allan Lyons:

Well, in a way it's kind of like you may have experienced in junior high. If it's not for marks, you're not going to do it. And even though it's on paper, the policies are there, not supposed to log it. But if nobody's going to check, for example, when you upload a new app to Google Store, say you're going to publish your new app, there is a bunch of checks that Google has to do just to make the thing published. It shouldn't be that hard for them to at least check some of these other things as well. They won't get sneaky cases, but like I said, a lot of these are just straight up, either they missed it and it's really obvious as soon as you start running the app that "Hey, they're logging way too much stuff," and that could be automated.

Andy Hasty:

I guess a follow-up here, are there any changes that you would like to see to reduce this problem?

Allan Lyons:

Well, it's more like if you take the existing policies and actually enforce them, enforce the existing policies, that would probably go a long ways.

Madeleine Varner:

Thank you. Sumanth, I actually have a related question for you. In your view, what sort of countermeasures or defenses can app stores and platforms implement?

Sumanth Rao:

So there's two parts I think to the answer. From the app store perspective, none of the apps which we used in our study were from the app stores. Again, because most of the Google and Android app stores have decent policies in place to check regularly for these sort of apps and then prune them from the app store. But the side loading part, which I mentioned, Android kind of makes this whole process much easier because you can go to a third party website or even just Google the results for getting the best spyware and you're hit with 10 pages of results, all of which have links where you can download a spyware app and install it easily on your phone. And the second part is from the Android perspective. So Android 12 has a privacy dashboard in place where it sort of shows a timeline of every permission which an app uses regularly over time. And this makes process much easier for you to sort of triage and see if there is an anomalous app, which is asking for an extra permission unauthenticated access than it normally does. But how a regular user uses this sort of privacy tool and Android, is an open question when it comes to stock web.

Madeleine Varner:

Thank you.

Andy Hasty:

Abbas, back to you. I have a question about support timelines. Is there a guarantee that CVE is found before devices' end of life will be patched?

Abbas Acar:

That's a great question. Actually, the results that we have here showing the best case scenario because we are assuming that let's say the secure updates being sent and customized and why OEMs are not missing any, the CVs that is known before even the end of life, this could require more research to check if those CVs are actually being patched even before the end of life the devices. So therefore, and another, for example, in addition to that CVs being fixed, non CVs at least being fixed, there could be another scenario. For example, the support durations are mostly we are giving is from the devices release date. However, some people are buying the devices later than the release date. So those people will actually receive a shorter duration of support than actually that we have seen in our dataset. So therefore our results are actually showing the best cases than the actual case in practice could be even worse. Yeah.

Andy Hasty:

Thanks. And a quick related question since you brought this up. Any reactions to the end of life duration? Does it seem reasonable in light of when people get the devices and how long devices are actually in use?

Abbas Acar:

Yeah. One interesting result that we are, when we are doing this research, we found that the OEMs do not publish the end of life for these devices. So for our analysis for example, we use the last secured update date that the devices receive for the end of life the information. And instead of that, some Android enterprise recommended devices, they have guaranteed support date so that they will receive the support until end of that date. However, that date also sometimes was inconsistent and depending on the device and/or the other factors that I presented. But yeah, that's a great question as well, to bring it up. At least they could publish the end of life so that the users can know until what date their devices will be supported and they can stop using after that specific date, yeah.

Madeleine Varner:

Thank you. I have a question for Allan. What changes would you like to see to reduce the number of developers logging sensitive info or exposing sensitive info in logs to other apps?

Allan Lyons:

Yeah, I think really to actually change that, there's going to have to actually be probably penalties or some regulation to actually enforce it because like I said, in junior high, if it's not for marks, nobody's going to do the homework. The policies exist, the idea about security problems being too much data in the logs being a problem. A lot of app developers weren't even born when that problem was pointed out in the, I think late eighties, early nineties. So yeah, so I think unless there's a regulation, it's probably not going to change. Or some enforcement, yeah.

Andy Hasty:

Allan, is there a technical, so even if information is being written to the logs, is there a way to cabin who has access to those logs to the app that wrote it, or is this really just the junior high problem?

Allan Lyons:

Well, for a normal app, if you installed an app out of the app store, those apps can only ever read their own log entries that they wrote. The problem is that it comes with all the pre-installed apps, like I pointed out that the kind of lacking control over who gets to be on there. Pre-installed apps, they could come from anywhere, whether it's Facebook partnering with Samsung to say that, "Oh, we want to have the Facebook app or Twitter app or on all the new phones," those pre-installed apps, those are the ones that can end up with permissions to read the entire log file. On the newer version of Android, as of 13, so that will pop up a message to say, "Hey, certain app is trying to read the entire log file, allow or deny." That's in the generic version of Android. But the enforcement of that, again, is then back on the manufacturers to say, they could provide a back door. So they could still have a way for a business partner to gather data as part of their business processes. So when their policy says, "Oh, we're sharing it with their business processes to improve product reliability," or whatever. So yeah, it's kind of a hard question that way.

Andy Hasty:

Thanks, Allan. We are running out of time, so Sumanth, this may end up being the last question, we'll see. But I'm curious, did you disclose the issues that you found to the vendors? And if you did, what sort of responses did you get?

Sumanth Rao:

We did, as a part of the research, we had to sort of disclose it to the vendors, and long story short, we didn't hear back from any of them. And I think the problem there is that their incentives don't match and none of them actually care about the effect that the data is not being protected. So from a regulatory standpoint, I think we need more interventions from both the industry, the government or the research community for making spyware more restricted out on the internet.

Andy:

Are there research questions that you still want to pursue or you would encourage other folks to explore?

Sumanth Rao:

Sorry, is that for me?

Andy Hasty

Yeah, you mentioned research and I thought I'd ask, do you have questions that you would like to answer or you'd like other researchers to explore?

Sumanth Rao:

Yeah, I think that's a good question. I think at least in the Android ecosystem, over the last just one year, Android 12 and 13 has come out, which does have a lot of different techniques to make it evident to the end user that they're either being monitored or surveilled. And the way these apps work is that an attacker would have physical access to your phone to go and install it, and they would enable all the permissions in one shot. But then I think under 1213 sort of asks again after a while that, "Look, this permission is being repeatedly monitored as a popup." And so there is a user perspective question as to what sort of technical skills does a user need to detect that they're being stocked. And so something like a user study might be a useful research in this space according to me.

Andy Hasty:

Excellent. Thank you very much. And thanks Allan and Abbas as well and Maddie, I think we are out of time. So I am going to hand the floor over to the last panel of the day on deep fakes.

Spencer Jackson-Kaye:

Good afternoon, everyone. My name is Spencer Jackson-Kaye, and I'm an attorney in the Division of Advertising Practices. I will be co-moderator this session along with my colleague Leah Frazier, who is an attorney in the Division of Privacy and Identity Protection. We're really pleased to introduce our talented panelists. We are joined by Mehrdad Saberi from the University of Maryland and Yan Ju from the University of Buffalo. Thank you to both of our panelists for being here with us this afternoon. With that, Mehrdad, please feel free to begin whenever you're ready.

Mehrdad Saberi:

Okay, thank you, Spencer. Hi, I'm Mehrdad Saberi, I'm going to present work today on robustness of AI image detectors. So in this work, we're basically trying to evaluate the robustness and reliability of detectors that have been or will be proposed for detecting AI generated images. Next slide, please. Okay. So why we need to detect these AI generated images? If we're not able to detect these images in a reliable way, they could be abused in a lot of ways. You probably have seen some cases in the news, but they basically can be used by anyone to spread misinformation on social media, deceive other people or can be used to create fake evidence that can be used against some individuals and some other cases like impersonation and blackmail. Next slide please.

So in the presentation today, we'll cover two most popular ways to detect these AI generated images. The first way is to insert watermark into these AI generated images. So this is basically injecting some binary or text message into your image in a way that it doesn't change the image that much, but later on you can decode this message from your image using your decoder technique or if it's a neural network based model or anything. So in this way, if you have an AI generated image that you have inserted watermark in before, then you can, after decoding, you can be sure if this is generated or a real image that doesn't have that watermark. Another way to do this would be just to use binary classifiers and train them on a dataset containing labeled phage and real data and then use that classifier to detect these images.

Next slide please. So the questionnaire is that are any of these methods, categories of methods reliable or not? So first we will discuss image watermarking and then we'll briefly discuss classifier based method. For image watermarking, we included... So we are evaluating some of the existing watermarking methods to see if they are actually robust or not. So we categorize the existing methods, some of the existing methods into two categories based on the amount of perturbation that they're adding to an image to make it watermark. So if they're adding a large perturbation to the image, we will call them high perturbation. And if they're adding a low perturbation, we'll call them imperceptible watermarks.

And this in our paper, we decide if a watermarking technique fits into low or high categories by the L2 distance of the watermark that they're applying. Next slide please. So for imperceptible watermarks, you basically have an image and then you inject a watermark to it. And your goal is an injective watermark that is imperceptible to that image. So the image wouldn't change that much. And your goal is that your attacker cannot remove this watermark or add watermark to your real images easily. So when you have a fake image, like AI generated image that has watermark, you want it to not be easy to remove this watermark. So our first attack that we propose is on imperceptible watermarks. This attack is really simple. So we basically take any image, apply Gaussian noise to it, and then use the denoising process of diffusion models, which is the backward pass of diffusion models if you're familiar with it, to denoise those noisy images.

So as you can see, the images on the second column have some noise, Gaussian noise, and then in the third column, they are denoised. So what this does is that, for example, if you have two images that one of them is watermarked and one isn't watermarked, and when you add Gaussian noise to them, you will get two Gaussian distributions with different means. But by adding Gaussian noise, these distributions will have some intersection between them. And when you denoise these points, the points in the intersection area will all go to the same denoised image. So this is the basic idea of the theory that we prove. So basically there is no way for any watermark detector to be able to have high accuracy in this case after we perform this attack on every image. Next slide please.

So our theory here, for more details, you can check our paper, but the theory basically says that your error, the error of any watermark detector would have a lower bond. So your error can't go lower than the term on the right here. And this term depends on the amount of noise that you're applying, which would depend on how good your denoising diffusion model is, and the Wasserstein distance between the distribution of watermarked and non-watermarked images. So this would basically mean that if your watermarks are imperceptible and are applying low perturbation, then this distance would be low and you will have a higher lower bond for your error. So your model can't get an error better than this. Next slide.

So we also perform this attack empirically on some existing watermarking techniques, and we see that this attack is able to break watermarks that have imperceptible perturbations. So as you see here, the AURC of 0.5 shows a random detector. So if this attack is performed on the imperceptible watermark, it's able to break them down to the level of just a detector that does random guess, but it doesn't work as well on high perturbation watermarks. Next slide please. So for high perturbation watermark, we propose another attack, which is an adversarial attack. I'll try to explain it fast. So how this basically works is that for high perturbation watermarks, we expect the watermarked and non-watermarked distributions to be separable using a classifier because their distance is high from each other. You can consider Wasserstein distance here, for example.

So basically, if we have a set of watermark and non-watermark samples for a watermarking technique, and you can gather them from internet or anything, you can basically train a classifier that is able to classify them with high accuracy. And then you can perform adversarial attacks on this classifier to, for example, turn a watermarked image into non-watermarked and vice versa. And we train this substitute classifier and then format are adversarial attack on this classifier. And after we get the results of those adversarial attacks, the good thing here is that they can be transferred to the real watermark detector. So if you perform adversarial attack on the substitute classifier, they can actually break the real watermark detector too. And yeah, next slide please.

So we performed this attack on two of the existing watermarking methods that have high levels of perturbation, and both of them are breakable to 0.5 AUROC, which is a random detector, with some level of noise. And here we are performing L infinity adversarial attack, if you are familiar with it. And we also show in the picture on the left that it does not apply that much noise to the images, but it is able to break down the accuracy a lot. Next slide please. So another attack that we have in our paper is a spoofing attack. I won't explain it in details here, but basically a spoofing attack is able to take an image and insert watermark into it without having access to the watermarking method. So you can, in case of deep fake detection, you can get a real image, for example. And then using a spoofing attack, even without having access to the watermarking method, you can inject some watermark to it. So it'll be considered as deep fake. Or you can also insert the watermark of, for example, some company like Google into your image and then claim that this image is generated by Google or anything like that. So it can be used in a lot of ways, and we show that a lot of existing methods are not robust against this attack. Next slide please.

So that was it about watermarking methods. And we basically show that imperceptible watermarks are provably breakable. And for high perturbation watermarks, we propose a strong adversarial attack that can break existing methods. So the other category of detectors would be classifier based detectors. For that, we propose some theoretical analysis in our paper that what this basically says is that if the distribution of fake and real images are close to each other, then these classifier-based detectors cannot be robust. And that is very intuitive. So if the distribution of fake and real images are really close to each other, if you have a classifier, it won't be that robust. By applying some noise to the data, it can easily misclassify them. And that is kind of the goal of the generative models, to have the distributions of their generated data close to real data. And as the generative models are improving, this distance is being shown to, it keeps decreasing, so you have closer and closer distributions. And we also show some empirical analysis on some existing detectors in our paper, but I didn't include it here. But yeah, basically what we say is that in the future, as the generative models are

Mehrdad Saberi:

... having better qualities and more realistic images. Having good and robust classifier-based detectors would be more and more impossible. Next slide, please. So that's it. Thank you for listening. You can also view our code on GitHub and view our paper too for more details. Thank you.

Spencer Jackson-Kaye:

Thank you so much, Mehrdad. Next up we have Yan Ju, and you can start whenever you're ready.

Yan Ju:

Okay, sure. Okay. Hi everyone. I'm Yan Ju and I'm a PhD student at the university in Buffalo. And my advisor is Professor Siwei Lyu. And today I'm going to share our paper titled Improving Fairness in Deepfake Detection, which was accepted by [inaudible 07:09:56] 2023. Next slide, please. Okay, so deepfake refers to the combination of deep learning and fake and encompassing any fake content generated by deep learning models. So deepfake can include generated images, videos produced by identity swap or face reenactment in audio and more. So in this paper, we primarily focus on fairness issues in video-based detectors. And with the development of deepfake generation, more and more deepfake detectors are proposed to mitigate the impact of such fake content. However, recent works show that current deepfake detectors exhibit unfairness, so they may demonstrate inconsistent performance across different groups such as gender, age, or ethnicity.

And here is an example. So figure one illustrates that false positive rate, meaning that misclassification of a real image as fake among various groups such as male Caucasian, female African, and female Asians, et cetera. However, this figure shows that disparities in FPR suggesting that female Asians or female Africans are almost three times more likely to be mistakenly labeled as fake than male Caucasians. So this indicates that bias against millions of people in large-scale commercial applications. So the question is how can we make current deepfake detectors fair for different groups? Next slide, please.

So there are several works that have exposed the unfairness issues in deepfake detection. Among these, only one work attempted to solve this issue at the data level. So they assume that the key reason for the unfairness is the imbalance of different demographic groups in their training data. So they deconstructed balanced training data set where every demographic group has an equal number of training data, but collecting a large balanced data set can be costly and labor-intensive. So to our knowledge, no existing works solve unfairness issues in deepfake detection and the algorithm level. So developing more effective bias-mitigating deepfake detection solutions remains an open challenge. Next slide, please.

So we propose two fair deepfake detection methods: demographic agnostic fair deepfake detection, shorted as DAG-FDD, and the demographic aware fair deepfake detection, shorted as DAW-FDD. So these two methods designed for training fair deepfake detection models with or without demographic information. These two methods can be combined with existing deepfake detectors. So here we present an example showing the demographic annotation of a training sample. It means that we know the gender, age, and race attributes it belongs to. Unfortunately, most of the deepfake data sets don't provide such details. So if such annotation is unavailable, we can adopt the proposed DAG-FDD. And if we have these annotations, we can utilize the proposed DAW-FDD to get better performance. The next slide, please.

So I will talk about the brief idea of these two methods. So, for DAG-FDD, it's an application of the current fairness method to the fair deepfake detection task, and it can be applied when demographic annotation of training data are unavailable because we don't know which demographic group each training sample belongs to. So the goal of our method is to ensure that all groups with at least a specified occurrence probability have low error. Specifically, we assume that each group occurs with probability, and then we can minimize the overall loss on these samples to ensure that all or no latent groups have low error despite us not explicitly knowing these latent groups.

And DAW-FDD can be applied when training data has demographic information. The goal is to ensure that the losses achieved by different user specified groups such as different races or gender are similar to each other, and the losses across all groups are low. So the key idea behind this method is we design a two-level loss function in which intergroup loss is used to address the imbalance among demographic groups. And the inner group loss is used to address the imbalance in real versus a fake class in each group. Next slide, please.

As for the experiments, we utilized four commonly used deepfake video datasets to evaluate our methods and the number of samples and attributes of each dataset are shown in this table. We employed two groups of metrics to evaluate the performance. The first group is fairness metric, including three metrics. And these metrics mirror the gap between groups where smaller values represent a fair method, but sometimes, although each group has similar performance, the overall detection performance is poor making it less ideal for real-world applications. So there is typically a trade-off between fairness and detection performance. Therefore, we also report four commonly used deepfake detection metrics, and higher values denote that they have better detection performance. Next slide, please.

We conducted extensive experiments. Here is a qualitative result of a detector on two datasets. And from this figure we can see that for original detector, the maximum FPR gap is 17.93 between the male other group and the female Asian group. But after using our methods, the difference dropped to 9.65 and 6.61, which are much fairer than the original methods. Next slide, please.

And here are some quantitative results. We applied our methods to two popular deepfake detection models, Xception and RECCE on the FF+ dataset. And comparing them with original method without any fairness constraints and several conventional fairness methods and data level fairness methods. As show in these two tables, our methods achieve good fairness without sacrificing detection performance too much. Next slide, please.

We also applied our methods to a detector on the other four different deepfake datasets and four different detectors on the same dataset for a thorough comparison. And these two tables show that our methods outperform the original method without sacrificing the detection performance. And this also demonstrates that our method can generalize well to existing datasets and existing detectors. Next slide, please.

To sum up in this paper, DAW-FDD and DAG-FDD are proposed for training fair deepfake detection models with or without the help of demographic information. And we conducted extensive experiments and from the experiment result, it shows that these two methods can be combined with existing detectors improving their detection fairness on various datasets distinctively. And one limitation of our methods is that they rely on the assumption that the loss function of the original detectors can be decomposed into individual terms. So allowing us to combine our methods to the classification term in their original laws. And for future work, I think it would be very interesting to explore if our fairness methods can generalize to same deepfake datasets or deepfake detectors. Next slide, please. you can scan these barcodes for more details about our paper and code if you are interested. That's it. Thank you.

Leah Frazier:

Thank you, Yan, and thank you, Mehrdad, for your presentations. We'll now move into some questions and answers. So this first set of questions has to do with the applicability of your work in other contexts. It presents a lot of fascinating possibilities in terms of improvement in deepfake detection. So we'll start off with you, Yan. Your work focused on improving fairness across various gender and racial groups. To what extent is it possible to adapt that to include or address additional attributes or different types of data or demographic issues?

Yan Ju:

I think our methods can be adapted to other attributes and other types of data very easily because currently we only conduct experiments on deepfake video datasets and consider only two attributes, race and gender and their intersections. But this is because of the limitation of the current deepfake dataset annotations. So I think maybe in the future if you have other types of data such as image or audio and you can label your data with the you are interested in, then I suppose our method can also adapt well to such datasets or attributes.

Leah Frazier:

How far off do you think that is?

Yan Ju:

I'm not sure because our method is only designed to loss functions, so it's very easy to adapt it to other datasets because it's not related to any specific datasets or specific attributes. So you can easily add it to your original classifier and try to improve the fairness.

Leah Frazier:

And, Mehrdad, your work addressed two different categories of detection methods, classifier-based and watermarking. What other detection methods are there, and to what extent can your learnings be applied to those methods?

Mehrdad Saberi:

So I think there are some other categories of methods that people are working on. I know about some people trying to detect these deepfakes using image retrieval-based methods and making them decentralized using blockchain and those type of things. So the evaluation that we provide here cannot be directly used in those methods. For example, watermarks that we are covering here, they can also be used in those cases for image retrieval to check if an image exists in your set of deepfake images or not. But it's not a direct application, so it has to be studied more.

Spencer Jackson-Kaye:

Kind of in the same vein, Mehrdad, your research addressed deepfake detection in the context of images. Could your findings be applied to things like video deepfakes or voice cloning?

Mehrdad Saberi:

I think the methods that we propose here can also be applied to those cases. For example, if you consider text domain, it would probably be harder because texts have a discrete nature with a discrete word so it would be harder. For those cases I know that other types of words exist using rephrasing the text, other types of attacks. But for voice and video, I believe that the methods that we proposed here might be able to be applied in those domains.

Spencer Jackson-Kaye:

Kind of similar, but, Yan, could your findings be applied to other contexts, maybe beyond faces? And then also, I know that you talked about limitations of annotation, but are there other limitations of the proposed methods?

Yan Ju:

Yeah, so for the first question, I think, as I mentioned before, our methods can adapt to other contexts such as images with other objects, it doesn't need to be the face object. So we conduct experiments only on face related deepfake datasets because our currently existing work show that unfairness occurs on face-related deepfake detection so we are trying to address it. And face is also sensitive because it has private information such as your ID, etc. So fairness in the face-related dataset will cause large impact. But if you find any other types of data or objects you are interested in, and it also shows unfairness between different attributes, I think our method can be added to their model training.

And for the second question, the limitation of our method, I think the limitation of our method is that they can be used when the loss function of your deepfake detector can be decomposed into several terms. Because the input of our method is the binary classification laws. So we need such binary classification laws at the input of our method, and then we can manipulate these laws to be more fair among different groups. So I think that's the basic limitation.

Leah Frazier:

I think that one question that a lot of people may have as they're considering the work that both of you have done is to what extent is it possible to improve deepfake detection methods? So both of you have focused on various limitations of detection methods, be it through, Mehrdad, your work on the trade-off between evasion error rate and robustness or, Yan, and the case of your work on fairness. So, Mehrdad, I was wondering, is there really any hope for robust detection techniques for AI-generated images? Is it possible that there is going to be some detection method that is truly effective?

Mehrdad Saberi:

Yeah, that's a good question. So in our paper, in some cases, like for imperceptible watermarks we are providing a provable attack that can guarantee you that you can break these type of watermarks. But in some cases like high perturbation watermarks, we are only providing empirical attacks. So there might be some methods in the future that can be robust to these attacks. So I can't answer that with confidence, but one interesting point that I was thinking about after publishing this paper was the meaning of deepfakes in the future.

So for example, we are starting to use AI tools even on our real and authentic images now. For example, when we are taking a photo with our phone camera, we might do some post-processing with AI tools on them or might do some edits and some effects on them. So at some point maybe all of the images might have some AI artifacts in them. So the meaning of detecting these AI generated and real images might be different. And in that case, I feel like we might not have any reliable detectors that can actually detect these images. We might have some detectors, but they won't be reliable on anything they can be used in some important cases like some law cases or things like that. That's just my opinion.

Leah Frazier:

And, Yan, what do you think about that?

Yan Ju:

Yeah, actually I don't know the answer because recently a lot of new generative methods are emerging every day, and it's hard to see if there is any general effective method that can address all these fake contents. So I can see that with the development of the generation models, the detection will be better and stronger. And I think for the development of the fake detector, we need to figure out how to evolve with the emergence of current generation besides retraining their models every time new model-generated data occurs. So I think it is quite to answer for now.

Leah Frazier:

A corollary to that is with respect to your work in fairness and detection methods, to what extent do you think that we'll be able to get to a place where the detection methods aren't just comparatively fairer, but that they're actually fair? Do you think that's possible?

Yan Ju:

I can say that we just started exploring this topic, so I think we may have a long way to go. So far, our work has been focused on framing this issue within a very narrow scope. Specifically, we only consider attributes like race and gender in visual images and testing our approach on the publicly available datasets. But the data in the wild is quite different, and the attributes you care about are different so the definition of fair is different. And moreover, the metric you use to evaluate fairness issue is also dependent to your own task and to your application. So yeah, I think it's just started. And I think maybe in the future, if we have a perfected detector that can handle all the fake content with 100% accuracy so the unfairness does not exist because I can detect everything with 100% accuracy. But I don't know, it's very difficult. So I'm not sure if that's going to happen.

Spencer Jackson-Kaye:

That's actually a good transition. Do you think that there are some potential methods, methods for improving on your work in the future? Any areas that may be added?

Yan Ju:

Yeah, I think one possible way to improve our current method is, as I mentioned in the conclusion, is to explore if our method can generalize to the unseen data that the detector has never been trained on. Because the generation ability, it's very important for a deepfake detector because a new generative model occurs every day. And so I think, yeah, it's a possible way to improve in the future.

Spencer Jackson-Kaye:

And. Mehrdad, the same question to you. What do you think would be a way to improve on your work in the future or build on it?

Mehrdad Saberi:

So I think it would be interesting to look into the other categories of detectors as I mentioned, like the image retrieval ones maybe. And also I think in the future, after our work there have been some papers that propose some other types of attacks against these detectors. So I feel like there's still room for having different types of analysis and different types of attacks. And of course, researchers and companies will come up with new detectors and new watermarking methods and things like that. So we have to see how the fight between them will go.

Leah Frazier:

This is a more general question and we'll start with you, Yan. Aside from the issues that you addressed in your research, what other pressing issues do you see arising in terms of deepfake detection?

Yan Ju:

My previous works are only focused on detecting the generated content passively. So I'm quite interested in exploring more proactive detection method such as adding watermarks during the generation of such content. So I want to find out if we can make the detection easier with the help of the generative models. So the generative models can play a very important role in this way. It would be better if the large model can embed some traces before releasing. So the important thing is how we can design very robust traces for the downstream detection, and, at the same time, such traces cannot affect their generation viral quality. I think that's very interesting.

Leah Frazier:

Same question to you, Mehrdad. What other pressing issues do you see on the horizon, aside from the one that you mentioned about people voluntarily manipulating images of their own?

Mehrdad Saberi:

So I haven't thought about this that much, but I agree with Yan that being able to combine these different techniques might result in better understanding, like combining these detectors with some watermarking methods that inject some artifacts at generation or with other types of technique. Because on their own, they haven't resulted in a real-world practical technique yet I feel like.

Spencer Jackson-Kaye:

I think we have time for one more quick question, but, Yan, what takeaways or implications do you think there might be for consumers who encounter deepfakes in the wild?

Yan Ju:

I'm sorry, can you repeat that question?

Spencer Jackson-Kaye:

What practical takeaways are there from your research for the average consumer who comes across deepfakes out in the world?

Yan Ju:

Got it, got it. I think my insight is that it's important for us to stay alert and not rely only on a single detector's results because they might perform well in certain scenarios, but not in all scenarios. So I think the detection outcomes could be biased toward different testing samples. So I may say always keep critical about the detection results.

Spencer Jackson-Kaye:

Do you have anything to add, Mehrdad?

Mehrdad Saberi:

Yeah, I agree with Yan that people shouldn't trust any image that they see everywhere and don't trust the results of the detectors. But it won't have that much of a problem for everyone. It imposes some problems, for example, for artists and their artwork can be easily copied if they can't add a watermark to it or people that are generating images as art. But I don't think it'll have that much of a problem not being able to detect these images for a normal user, unless you have some neighbor that creates a deepfake of you robbing their house or something if they really hate you.

Leah Frazier:

That is our time for today. So thank you, Yan and Mehrdad, for your presentations. And I'll turn it back to Jamie Hine for some closing remarks.

Yan Ju:

Thank you.

Mehrdad Saberi:

Thank you

Jamie Hine:

... for an excellent day. We've learned so much here and we hope that you did as well. So just a few quick reminders: all of the materials are on our event page, all of the papers are linked there. If there's any updated materials, we'll add those. There's biographies where you can learn more about all the people who presented today. In addition, in about a week, we'll put up transcripts and videos so you can see anything that you might've missed or that you want to watch again. In fact, all of the old PrivacyCons are also all archived there if you're interested in learning more about research that's been presented at previous events.

I want to also invite anyone to send us ideas for topics for next year. Are there things you're interested in hearing about research that people are doing? You can send those to privacycon@ftc.gov. Also, if you have any ideas about how we can make the event better to connect better with our audience, also please send emails to that address.

So the program today wouldn't be possible without so many people at the FTC and I just want to thank a number of them, and that includes all of the attorneys and the economists and the technologists. It includes folks from our Bureau of Economics, from the Bureau of Consumer Protection, from our Office of Technology. It also involves a number of other units at the FTC, our Division of Consumer and Business Education, our Office of Public Affairs, and our events team who helps put all of this together, our partners at Open Exchange who help make it all work on the web. But last and most important, there are two very, very special people who really made everything happen from the very first day until the very last moment, and those are the two paralegals that make everything happen at PrivacyCon. The first is Johana Mejia-Portillo, and the second is Ryan Zwonik and I want to personally thank them because the reason why today was such a great event was because of the two of them. So with that, I will say thank you. We're looking forward to seeing you at PrivacyCon 2025. Take care.