FTC PrivacyCon
January 14, 2016
Segment 5
Transcript

AARON ALVA: All right. Our last panel of the day will look at issues around security and usability as it relates to privacy. So I would like to welcome our first presenter, Sarthak Grover. He is a PhD student at Princeton.

SARTHAK GROVER: Thanks, Aaron. Hi everyone. I'm Sarthak, and I'll be presenting our work on the internet of unpatched things. So the main aim over here is to basically look at the current state of devices. We basically ended up studying network traffic from a bunch of smart devices which are really popular. And we want to talk about how these devices may potentially leak user information. My aim is to basically encourage you to think of how we can improve policies to stop this potential leak of information.

So how is the smart home or the IoT environment different from the conventional mobile or computer environment? The problem is that we have a lot of manufacturers, and we have small start-ups coming with their own devices. They may be hiring novice programmers. Apart from that, these devices have no memory. They might not have capable hardware to enforce security protocols which we use in computers and phones.

This makes it pretty difficult to, first of all, implement security protocols on these devices. But apart from that, a bigger issue is that the current smartphone model looks like this. Your devices inside the home send all their information to the cloud to a particular server. In fact, if you have two devices in the home and they want to talk to each other, currently they'll talk to the cloud and the information will get back to the home and something will happen in the home then. So what we have over here is a pretty bad combination. You have hardware, which is incapable, and you have information, which is always being sent to the cloud.

So this combination results in potential privacy problems. Now, IoT devices which are just sending network traffic without security protocols may end up leaking some information about the user. They may end up leaking information about what device is being used inside the home, and they may also end up leaking information about whether the user is home or what he is currently up to. So in essence, what I'm saying is anybody sitting on your network path may be able to find out what you're currently doing inside your home. And this is a big fault.

So what our aim is right now is to basically take up a few devices in our case study and study what kind of personal information or user activity information they leak to the cloud. So what we did was we basically bought some popular devices. We went to Amazon.com. We searched for some very popular home network devices which people are currently using in their smart homes, and we ordered them. What I'm going to show you is results for network traffic analysis for five particular devices-- a camera, a photo frame, a hub, an Ubi smart speaker, which is like an Amazon Echo basically, and a Nest thermostat.

So what we're interested in right now is what kind of information these common devices leak to the network. And the first device I pick up is the digital photo frame by Pixar. So what we found out was that all traffic from this photo frame is sent in clear text. There's absolutely no encryption happening, all right? The cool thing is that this device can actually talk to your Facebook or RSS feeds. So it's downloading photographs in the clear.

And also, whatever action you take on this device-- for example, you press a button; say you press the play radio button-- that'll actually go in a clear HTTP packet, which somebody, again, on the network can read. So if there's somebody sitting outside on the network, like somebody in the ISP or a malicious passive listener, they can see what you're doing through the photo frame. Apart from that, it's also capable of downloading radio streams-- again, in the clear.

So an example of what kind of information we see-- so these are snapshots from Wireshark, basically. And what we saw was that your email which you configured your account with is actually being sent in clear text. What this means is that this photo frame is potentially leaking account data, and anybody on the network path can actually have a look at this email.

Secondly, if you press the button on this photo frame-- say you press the List Contacts button or the Radio button-- anybody, again, on the network path can have a look at what you currently press. So somebody on the ISP can go, this person is currently listening to the radio from his digital photo frame, though I don't know why you would listen to the radio from the photo frame anyways.

[LAUGHTER]

So basically what I mean to say is that you can find out about the user activity, as well as some account information, just by looking at the network information.

The second device we picked up was a shock security camera. It's a pretty common camera which is used for security monitoring in homes. It has motion detection. What we saw was that all the traffic, again, was being sent in clear text. Now, this security camera actually requires a login. So if you want to view the screen, you're supposed to enter a password. But that doesn't mean the stream itself is encrypted. In fact, anybody sitting on the network can still have a look at where the stream is going and what the stream is. Also, if you go to the web interface and you press a button, whatever you did will still go in an HTTP GET packet, again unencrypted.

So videos are being sent as JPEG frames. Also, if you've pressed the FTP button, then all your data is being uploaded to the FTP, again in the clear. And this is an example of what things look like. So the FTP is actually using some really random ports, so you can't really rely on the network to secure you again, because these are non-standard ports which are being used by the device. This is basically private data which is being uploaded.

The third device we ended up looking at was the Ubi. So this, I think, is like a precursor to the Amazon Echo. Basically, it's a smaller voice box which you can talk to, interface with other devices. For example, we have this Ubi interface with the Nest thermostat in our houses. So what

we saw was that all voice you talk to to the Ubi will get converted to text on the device itself, and then text is sent, in clear again, to a server outside. The server here was the Ubi.com.

Apart from that, the Ubi also has certain sensors-- for example, light sensors and temperature sensors, which are still sending the readings in the clear. And the interesting thing over here is when we interface this device with the Nest, it used encryption and spoke over HTTPS. But when it was talking to its own server, it was using HTTP and everything was in clear. So clearly, this device actually has the capability of enforcing security. But somehow, whatever policy they came up with they did not enforce encryption for their own device streams-- only when they're talking to the Google API they enforce encryption.

So this is an example which shows how sensor readings were available. Now, these sensor readings can leak information about whether the light is on in the room or not. In a sense, somebody on the network who's on the path can know whether there's a user inside the room or not based on the luminosity value.

Furthermore, when we were chatting with the device all the text was converted to clear text and then sent to the network. So here, you can see an example of what the chats looked like when I monitored them on the laptop gateway.

The next device we looked at was the Nest thermostat. Now we're actually coming to the more secure devices, and the big ones too. The Nest thermostat from Google actually was pretty secure. All information was over port 443 basically using encryption and HTTPS. What we also found out was that some of the updates incoming were in the clear. And we weren't sure why, so we contacted Nest about this. Turned out, it was a bug and they fixed it.

So here's an example of what we found initially. Outgoing traffic was secured, but incoming traffic, some weather updates, were not secure. They were in the clear text. They had some information regarding the location. And when we told Nest about it, they thanked us and they fixed it.

And the last device which I'm going to talk about is the SmartThings Hub by Samsung-- again, a pretty popular hub from a pretty big company. The good thing was almost all the traffic coming out of this device or going into the device was totally secure over TLS. There was no clear text or [INAUDIBLE] traffic at all. And the flows were all to an Amazon AWS instance.

But the interesting thing is, even though this device is in itself secure-- and in fact, I see this as the model of future IoT devices which are completely secure-- there's still some background information, like three or five packets every 10 seconds going to SmartThings.com, which can somehow let you fingerprint the device.

The good thing is that the SmartThings is a hub. Basically what that means is that you have other sensors attached to smart things over some other protocols, like ZigBee or Bluetooth or Z-Wave. And you don't have a direct view of the sensors. So SmartThings itself makes all the information coming out of the house secure and then sends it out. But a person sitting at the ISP level can still find out that you have a SmartThings Hub inside the home.

So this brings us to my conclusion and some implications on the policy. Basically, I don't want to sound pessimistic or dramatic, but that's what the heading is-- Be Afraid. We know it's very difficult to enforce security standards on smart devices. Inherently, I mean, there are multiple manufacturers. There are only a few big ones, but a long tail of small ones. Smart devices come up on Kickstarter and people buy them. It's difficult to ensure that they all follow the same standard.

These devices are also sometimes very low-capability. They don't even have a way to implement TLS on the packets they're sending out, and they also end up using nonstandard ports and protocols. But the good thing is that we are trying to make an effort. For example, I found this handout outside regarding building security in the internet of things. And that's good, because it means we're trying to enforce security at the building block level itself. So maybe the new devices which come up would have security inherent in them.

The second thing is-- OK, so we've fixed devices which are coming up now. What about devices which are already present? How do we get people to patch them up or fix them? So first of all, we want to encourage people to look for bugs. And one way would be bug bounties. But as we've seen in previous talks, bug bounties might work for the big guys but the IoT domain has a lot of small manufacturers coming up, and we don't really know if bug bounties would work for that and whether the device would be popular enough to have users which actually look for vulnerabilities in them.

So the main part is, how do we enforce such kind of things? Who's responsible here? Will the government try to enforce bug bounty programs, or is it the manufacturer which goes, if you find a bug, we'll give you money? And lastly, who pays for this patching in the update part? If this is using your network, is the user responsible for anything which goes wrong, or the ISP or is it the manufacturers?

So I want to end with some of the work which we're currently up to right now. So we've talked about how we can improve future devices in terms of their security and privacy policies. We've talked about how we can improve current devices by trying to find bugs in them and vulnerabilities in them. The approach we're taking right now is how to improve security and privacy policy on the network.

Basically, we're trying to offload policy to the network layer. For example, in case of a smart home, all our information is going to go through a gateway which is inside the house. This gateway might be provided to us by the ISP or it might be our own, but maybe there are parts of security which we can implement at the gateway itself. Maybe we can tell the gateway to enforce certain standards regarding the network protocols which are being used by devices. Or at the very least, this gateway could inform our user that, hey, there's something wrong with your devices, or this device is not using the right security standards.

So what we're looking at currently is, can we offload device security to a gateway or the network layer? And secondly, how much information about the user behavior is actually leaked to outside the home Network All right, thank you.

[APPLAUSE]

AARON ALVA: Thank you. Next we will have Professor Vitaly from Cornell University and Cornell Tech.

[INAUDIBLE]

VITALY SHMATIKOV: Hi. That's my mic?

AARON ALVA: Yes.

VITALY SHMATIKOV: So I'm Vitaly, and I'll be talking about mobile advertising today. So mobile advertising is pretty big these days. If you look at modern app stores, you find that a significant fraction of apps are free to the user, and the way they make their money is by incorporating advertising. So it seems like a very reasonable question to ask, what information about the user is actually available to advertisers? That is, if an advertiser submits an add to a mobile network and that ad gets shown on a user's phone, what can an advertiser find out about the user of the phone on which the ad is being shown?

So that seems like an interesting question, to which apparently so far there hasn't been a good answer. Very few people investigated this. So this is what we decided to investigate in this project, to look at this. But in order to understand this, we first need to understand how mobile advertising actually works from a software perspective. So it requires a little bit of reverse engineering of how mobile software that actually shows ads to users actually works.

Mobile advertising is a little bit similar to web advertising, with one crucial difference. So in web advertising, you typically have a web browser and a web browser is just showing an ad. That has been studied a lot, and even today we've heard a lot of talks and conversation about web advertising. In mobile advertising, things are a little different because there is something in the middle. Mainly, the recent ad library.

So the way mobile advertising works is that apps that are supported by advertising typically include a little piece of code called an ad library, and it's that piece of code that's actually showing ads. It's not the app itself. It's the ad library inside the code. And it's actually very common for modern apps to incorporate multiple advertising libraries, because they make more money that way. So like between a third and half of all apps that are ad supported actually include multiple advertising libraries from multiple providers that are being used to show ads to users.

So the question I'm asking, just to repeat it, what do these ads that are being shown inside these mobile advertising actually know about the user or what they can find out? In order to do this, we need to look at the structure of this whole ecosystem. And I promise I'll try to make it as painless as possible, although investigating it was fairly painful and involved a significant amount of reverse engineering.

But it roughly looks something like this. There are three kind of big parties in the picture. So there is the app which is being shown on the phone. There is the advertising service which is supplying ads to the phone. And then there is the advertiser whose ads are being shown. And there has been a lot of work previously on trying to understand what information is available to the advertising service, but instead we're looking at what's actually available to the advertiser.

And it's not the same question, because there is a big difference between advertising service and the advertiser. The advertising service is typically a reasonably respectable, reasonably reputable company that's maybe owned by Google or Twitter or some kind of recognizable entity. There's large businesses that have reputation at stake. They make a lot of revenue, whereas advertisers-- which is people who actually supply these ads that are being shown-- who knows who they are? I mean, this is dynamically determined. They're fetched in real-time, sold by auction syndication all sorts of ways. These are not necessarily trusted. It's very hard to determine what information they're trying to extract.

And that's why mobile advertising libraries go to fairly significant lengths to protect users from malicious advertising and from snoop advertising and from advertising that stealthily tries to extract information about users. They use a variety of technical mechanisms to achieve this. And I'm not going to go into them. You can read our paper if you want to find out more about this. The short summary is that what they try to do is they show everyday ad that they show to the user inside a little browser instance. There is a little web browser inside every advertising library, and they create a copy of this web browser every time they want to show an ad. And they show an ad inside that thing.

And the good news about it is they can effectively rely on security and privacy protections inside web browsers to protect phone users from malicious advertising. So technically this is known as same origin policy, but you can think of it as just a way of sandboxing untrusted advertising to make sure it doesn't have any access to the underlying phone and cannot learn anything it's not supposed to learn from the phone.

And mostly it works, with one little exception. Mobile ads these days need access to what on an Android phone is known as external storage. And the reason they need to do this is for rich media, because people who view advertising, and especially people who supply this advertising, they want rich experiences, they want video, they went images, and because of that they need to cache a lot of information on the device. So they have access to external storage. But to be safe, they allow ads to load files from external storage but not to read them. So it cannot read it. It can just load it and show it to the user without being able to read it.

So that looks fairly harmless, except that Android external storage is kind of this weird thing. In Android external storage, there is really not a whole lot in the way of access control protections. That is, if there are multiple apps running on the device and they store files in external storage, they can read each other's files.

And that may be not ideal from the security perspective, but this should not really imply a whole lot about security and privacy of mobile ads because as I told you, mobile ads cannot actually

read other apps' files from external storage. They can try to load them and try to show them to the user, but they cannot actually get access to them directly. They cannot look at their content.

So, so far so good. So it seems like this whole way of protecting users from potentially malicious mobile ads is fairly carefully designed and carefully thought through, except that there is this one little weird thing. They cannot read them, but they can try to load them.

Why is this interesting? It turns out that by trying to load a file that doesn't belong to them, mobile ads can learn a little bit. They can learn like one bit of information. They learn if a particular file exists on the device or not. They cannot read it. They just learn if a file with a particular name exists.

That seems like, OK, all right. That's fascinating. Why am I talking about this? Because that's really a very small amount of information. So now let's look at how this information might be used by a mobile app. So let's take an application which actually has nothing to do with mobile advertising. It's just a popular application in the Google Play Store that happens to be a drug shopping application. So this allows people to go and look at pharmacies. If somebody's picking a particular medication, they can find a pharmacy nearby where the price is lowest on it.

So in this particular case, you can see there are some medications. These particular things-- actually, the fact that a person is taking one of these might be considered sensitive, because this has to do with anxiety and various psychological disorders. So what this app does, if a person is regular shopping for a particular drug, to make it faster it takes a picture of the pill, the literal picture like I'm showing here, and stores that picture in external storage of the device so that next time it's faster to show this picture.

Now imagine that there is an ad running in a different app on the same device. OK? The app that's showing that would be a totally random app. It has nothing to do with the pharmacy shopping app that I showed you before.

However, as I told you before, an ad being shown in it has the ability to ask a very simple question. Does a file with a particular name exist on the external storage? And in this case, it's asking for a file whose name corresponds to the image of one of the anxiety drugs. So what can a mobile ad-- and this is a question to you guys-- learn from the answer to this question? So all it learns is one bit. If the file with a particular name exists on the device, what can the ad learn by knowing the answer?

SPEAKER 2: [INAUDIBLE]

VITALY SHMATIKOV: If the answer to that question is yes, the only reason a file with this name would have existed on this device is the user used that app and searched for that drug. There is no other reason. So if an ad sees that a file like this exists, it cannot read this file. All it needs to know that this file exists it learns with 100% certainty, because this name is unique, that the person has been shopping for a particular drug.

And this turns out to be pervasive problem. And remember, this ad is being shown in a totally different app. It's not even being shown in the pharmacy shopping app. It's just being shown in some ad which is running on the device-- maybe even later, not even at the same time as the pharmacy shopping app.

And this turns out to be a generic problem, is that if there is an app, that need not even be advertising supported, that puts files under external storage, like a lot of them do, in a way that depends on the user behavior, then an ad shown in any app on the same device can determine that these file exist, and from the fact that these file exist it can infer what user behavior led to the presence of these files. So I showed the example with drugs. And by the way, this violates nothing in the security policy because all the security policy says is that it cannot read these files and it cannot. It does not read the file. It just learns that the file exists.

And actually, turns out that this affects all kinds of mobile apps. Here is another app. This happens to be a mobile web browser which caches visited pages in files with predictable names-- actually, the names of the files are derived from the URL of the page that the user visited. And it's vulnerable to exactly the same attack. A malicious ad running in another app can look at the presence of certain files on the device, and it can figure out which sites the user visited recently because the only reason a file with that name would appear on the device is if it were cached by the user's mobile browser as a result of a previous visit to a particular website.

And in our paper, we have many more examples of other inference that could be done this way. And we actually did an analysis of several very popular advertising libraries, including AdMob and MoPub, which are present in a very significant fraction of Android apps. They all, at least at the time of our study-- I'll tell you in a second what happened later-- had this vulnerability, meaning that a mobile ad shown in any of those libraries could infer information about the user by presence of cached files created by other apps.

We also looked in our study at other issues, like the leakage of location information. I'm not going to go much into detail about this. I'll just show you this picture. And the only thing I want you to admire about this picture is how complex it is, because it shows how in five stages literally in MoPub, information about the user's location can be extracted by an ad.

But it works pretty reliably, and as a result a mobile ad running in MoPub can create very nice trajectories of user movement like this, which immediately reveals a ton of information about the user, including actually user's identity if one of this happens to be like a single family residence where the user lives. So this is really fine-grained information that can leak out through these indirect channels.

OK, so what are the lessons of this study? As far as I know, this is the first reasonably comprehensive study of first, how advertising libraries on Android try to protect users from malicious mobile ads and snooping mobile ads-- with intermittent success, as you can see. It shows-- and this is a slightly more technical result, but nevertheless important-- that standard web isolation policies that are used in web browsers here exactly the same things they used in the mobile context. And they're no longer sufficient, because they no longer prevent leakage of sensitive information. Something more subtle is needed here.

We actually, when we first did this study last summer, we didn't make it public right away because we actually wanted to work with developers of those advertising libraries and companies that deploy them so that they can fix at least the most severe vulnerabilities that we identified. And in fact, some of them did, in particular AdMob, which is the biggest Android advertising service, actually owned by Google. They fixed that in the latest release of their Ad SDK.

Some library developers told us to go away and not bother them anymore. I hope they won't do this after this talk. And if you want more detail, we have our paper online. It's written for a technical computer science audience, but I hope at least the big themes will come across from that. Thanks.

AARON ALVA: Thank you.

[APPLAUSE]

So next we'll have Florian Schaub. He is a postdoc fellow currently at Carnegie Mellon.

FLORIAN SCHAUB: That was correct. Hello, everyone. I'm going to be talking about a project called the Usable Privacy Policy Project. And this is a large-scale project funded by the NSF under its SATC program, and I'm a postdoc in this project. Norman is actually the lead PI in this project. And it's a collaboration with many people at CMU, Fordham University, as well as the Center for Internet and Society at Stanford. And you can learn more about the project at our website, UsablePrivacy.org.

I'm going to give you a short motivation and then give you an overview of what we actually do in this project in different parts. So we look at privacy policies. And privacy policies originally had this promise of service providers would disclose the data practices so users can then make informed choices about which service providers or websites they trust with their data. But the reality looks a little bit different, because privacy policies play different roles, really, for different stakeholders.

So for the service providers, it's not really about informing the users. For most of them, when they draft a privacy policy the goal is to demonstrate legal and regulatory compliance. And this may limit their liability. And regulators are happy about this. They use these privacy policies to assess and enforce compliance. So there's actually a nice and strong interaction between those two players, but that means the user kind of gets left out.

And as a result, these privacy policies are long. They're complex. They're difficult to understand. They're full of jargon. They don't really offer my many choices to users. And I think we all know by now that users mainly ignore them. And this puts us in this really weird situation where these policies outline what companies do with our data and what we allow them to do with our data, but this information is not used by the users or made apparent to them.

And there has been much work on overcoming the status quo here. Proposals like layered privacy policies, showing short summaries of policies, graphical approaches, as well as machine readable privacy policies. But many of these approaches don't go anywhere really, because they

lack industry support and they're not sufficient adoption incentives for companies to actually implement those solutions that have been proposed.

And this is really where our project comes in, because we're looking at semi-automatically analyzing these natural language privacy policies that most websites, most mobile apps, already have. And we analyze them to then extract key data practices out of these policies. And we do this by combining crowdsourcing, machine learning, natural language processing, and this way enable large-scale analysis of privacy policies. And at the same time, we look at modeling users' privacy preferences and concerns so that we can actually provide them more effective notices that focus on those information aspects and data practices users really care about and give them information that is actionable.

And our project has many tightly interconnected threads. And I'm not going to try to untangle this for you right now. Feel free to look at our report to get a deeper insight there. But basically, we have two goals. One goal is we want to better inform users. We want to give them notices that actually inform them and provide them with choices. And we want to inform public policy by showing issues with privacy policies as well showing ways of remedying those issues and also hope that our notices could be provided.

And to identify data practices of interest, we approached this really from different perspectives. So part of our research team looks at legal analysis. Joel Reidenberg and his team analyzed privacy harms and litigation cases to see what issues come up the most. We conduct user studies where we determine what are privacy preferences, concerns, and expectations of users. And Ashwini this morning talked about some expectation in that context.

And we also look at the policies themselves. So how are they written? How are data practices actually expressed in those policies? And we have some work going on right now that looks at quantifying the ambiguity and vagueness in privacy policies. And to analyze these policies, we started by building an annotation tool that basically allows us to give policies to crowdworkers, or the annotators. And this kind of tool shows them the policy on the left hand and then a question on the right, and we ask them to answer the question but also mark text that basically provides the evidence for their answer.

And that's really important, because this text selection, in combination with the answer, then helps us build machine learning models and train machine learning classifiers. And by showing these questions or tasks, annotation tasks, to multiple annotators, we can actually get quite robust results.

However, just giving this to some untrained crowd workers and saying, oh well, ten people say that's OK, is not really a good idea. So we conducted studies to compare the annotation performance of experts who either write policies or have long experience in analyzing policies, graduate students in law and public policy and untrained crowd workers recruited from Amazon Mechanical Turk. And we asked those people to annotate different privacy policies. The crowd workers, or skilled annotators, the grad students, annotated 26 policies. And then six of those policies were also annotated by experts.

And I'm not going to go too much into the details for the sake of time, but one of the interesting results is that even the experts don't always agree on the interpretation of a privacy policy. And one reason for that is that the policies are vague, but also that they're sometimes contradictory and there are just too many different contexts handled in a single policy.

Good news is that for data collection practices, those are relatively easy to identify and to extract. They're usually in one part of the policy. But data sharing practices are a bit more complicated. They're spread out throughout the policy. Sharing is mentioned in many different contexts and parts of the policies. So it's kind of difficult to extract finer nuances reliably.

Now when we compare the performance of the crowdworkers who are skilled annotators, we actually find quite encouraging results. So when we hold the crowdworkers to a certain quality standard-- 80% agreement, which means eight out of 10 crowdworkers need to come up with the same interpretation-- then we actually find that in a large number of the cases, these crowdworkers agree with the interpretation that our grad students find, as well. So they come up with an accurate interpretation.

And in almost all of the other cases they don't reach agreement, which, means they don't give us wrong answers. This dark bar is a percentage where they come to a different conclusion than the skilled annotators. So that's great. Either we get an answer from our crowdworkers, with a high likelihood it's actually correct, or we don't get an answer. And that tells us the policy might be vague on that particular issue we're trying to analyze.

So this shows that accurate crowdsourcing of privacy policies is feasible, but privacy policies are still long and complex. So we look at leveraging machine learning and natural language processing to further enhance those extraction tasks and make it easier for crowdworkers to complete these tasks faster without loss of accuracy.

And one approach we tried here or we've been developing here is predicting and highlighting relevant paragraphs. So we take the answers we have from our skilled annotators, and we use that to train logistic regression based relevance models for different types of data practices we want to extract. And then we highlight the top five, top 10, paragraphs that most likely contain answers or information about the data practices we want to extract. And what we find is that really helps the annotators to come to conclusions faster without affecting the accuracy.

And we did additional experiments or analysis where we looked at, do they actually just focus on those five paragraphs or do they also read other parts? And they do read other parts of the policy, but it helps them to focus their search and find parts in the policy again.

Another thing we do is we split up this relative complex task of reading a privacy policy by splitting the policy up into smaller paragraphs and then giving a crowdworker only a single paragraph. We can further split those tasks, as well. So rather than asking them multiple questions at once, we first asked one set of crowdworkers to label what category of data practice is described. Is this a sharing practice? Is this a collection practice? Is this maybe about user access? And then in follow-up questions, we can ask more details that are particular aspects for

that kind of category. And that means that the task interfaces we can show to crowdworkers are a lot more compact, and they can complete those tasks faster and with lower errors.

And based on that, we've developed an annotation scheme that really makes use of this approach. This is an interface not for crowdworkers. We're using this with law students. But the next step is to then break this up again with a project just outlined. But there's a very fine-grained invitation approach, and we're currently collecting data from law students. We already have over 100 policies annotated.

And this provides a really, really rich picture on how information is represented, how data practices are represented, in the policies. We're going to release a data portal to allow exploration of this data on privacy day this year, January 28. So visit our website towards the end of the month. And the nice thing about this data is it's really helpful to train machine learning and natural language processing models, and drive research in this area.

Ultimately what we would be hoping for is that we can actually automate the extraction. And one approach we've been working on here is paragraph sequence alignment. So if I have a paragraph in one policy, in the Amazon policy, and I know that this one's about collection of contact information, and if I compare that paragraph to other paragraphs in other policies, there's a high likelihood that I can find similar paragraphs that also describe the collection of contact information. And this way, we can basically reduce which paragraphs we might even have to show to crowdworkers, and in this way automate some of the annotations and analysis.

Now, once we have all this data we want to provide notice to users and here really focus on making sure the information we give you is actually relevant. So we highlight unexpected practices, practices users care about, and information should be actionable. If users can't make a choice, then there's no point in showing them information because they're just going to be helpless. We heard about this this morning, about users becoming resigned because they can't make any choices.

So what we do, or what we want to do, is we want to show the available choices that are made available in the privacy policy. There are not that many. But because we can scale up this analysis to many websites, we can also show more privacy-friendly websites as alternatives to users and this way offer them choices that go beyond what the policy of a single website might offer.

And we're currently in the process of developing a browser plug-in to basically make this technology available to users. And the idea is that we display a limited set of relevant practices-- and we're going through an iterative design process at the moment with focus groups and online studies-- but hope to be able to release this plug-in to the public this summer.

So in conclusion, what we do in this project is we semi-automatically analyze privacy policies. And we do this with crowdsourcing, with natural language processing, and machine learning. And the goal of our project is really to enable large-scale analysis of these privacy policies. So at the moment we're annotating 100 policies. In a year, we're hopefully annotating 1,000 policies, and we're doing it at the same cost or even cheaper. That's the idea.

And at the same time, we're really interested in understanding what users care about so we can on the one hand, focus the analysis, but also help regulators focus their activities potentially to look at those issues users care about or are concerned with. And at the same time, we want to show ways to effectively inform users about the data practices that are currently lost in those policies. No one's going to read the policies, so if we want to make those policies usable we need to extract information that is really relevant to users and show it to them in a format that actually makes sense to them and actually allows them to act on it. All right. Thank you.

[APPLAUSE]

AARON ALVA: And our last presenter of the day will be Norman Sadeh. Norman is a professor in the computer science department at Carnegie Mellon.

NORMAN SADEH: Well, good afternoon. I think very few people in this audience probably appreciate how much progress we've actually been able to make over the past few years in both modeling and predicting people's privacy preferences. And so this talk is about sharing some of these results with you, and showing you also how this effectively supports our vision of developing personalized privacy assistance, in particular the success-- or at least, the early success-- we've had with mobile apps in particular, and how this, we believe, can be extended to IoT.

So this is joint work with a large team that will be acknowledged on the very last slide. I don't think I'm going to have to work very hard to convince the audience that people care about our privacy. And yet, as we all know also, people are often very surprised when you tell them, for instance, what sorts of apps they've downloaded on their mobile phones and what information is being collected or shared by these apps.

This is just an example of an early study we conducted in this space. The biggest offender in that case was an app that some of you at the FTC are quite familiar with called Brightest Flashlight. 95% of people who had that app were extremely surprised and very upset to find out what information that app was actually collecting.

And so as we all know, and as Florian just emphasized again, very few people read privacy policies, obviously. And that's certainly part of the reason why we have this level of surprise. Also, as I think we are also realizing, many of us have tons and tons of settings and just don't have the time to configure all these settings.

For instance, if you are a smartphone user-- and as most smartphone users, you probably have somewhere between 50 and 100 apps on your phone-- these apps typically will require between three and four permissions. These are permissions to access some of your more sensitive information. If you do the math very quickly, you realize that this will require people to configure somewhere around 150 different settings. How many people are willing to configure 150 settings on their cellphone? Not that many.

And so with this in mind-- and obviously, with a recognition of these challenges both already on the fixed internet and in the mobile space-- the natural question is, well, if this already doesn't

work on the fixed web, if this already doesn't work on the mobile web, what are the chances that it's going to work in IoT with the Internet of Things. And so our vision in this space, as I said, is this idea that perhaps personalized privacy assistance could be developed that will actually reduce the burden and implore you to manage your privacy better across these different environments.

And so the idea is that these personalized privacy assistance in particular will learn over time your privacy preferences and will be able to semi-automatically configure many of those settings based on various correlations between how you feel about sharing your information with one app versus another app, based on also understanding what your expectations are, going back to the presentation that was given this morning by Ashwini Rao, who's been looking at these issues in particular.

For instance, if you think, as Florian also mentioned, about privacy policies, when you read these privacy policies they tend to be very long, very verbose. But very often, at the end of the day there's only a very tiny fraction of the text in that policy that matters to you, and perhaps even a tinier fraction of the text that pertains to things that you didn't already expect.

And so perhaps this personalized privacy assistance could help us by highlighting those elements of policies that really would be a surprise to us, that perhaps would lead us to modify our behavior as we enter a smart room, for instance, in an IoT context. Perhaps this personalized privacy assistance could also help motivate users to revisit some of their settings, to verify that they still feel the same way. Privacy preferences are not fixed. They might change over time based on experience, based on what you learn.

And so again, what I'd like to do is I'd like to share with you some of our success at actually supporting some early elements of this functionality. What you're seeing here is effectively an early model that we built about how people felt sharing their information with various mobile apps for various steps of purposes, whether the app required this information for internal purposes, for sharing with advertising networks, for profiling purposes, or for sharing with social networks.

I'm not going to describe this chart in great detail because time is limited, but effectively what we're supposed to see here is that people don't always feel the same way on average when it comes to sharing their information. There are clearly differences between sharing your location information at a fine level versus sharing it at a coarse level. There are differences when it comes to sharing, for instance, access to SMS functionality, and certainly depending on whether you're going to be doing that for advertising purposes versus using it purely for the purpose of the app that you're trying to download. People are going to feel very differently.

What this figure, however, doesn't show is how difficult it is to actually configure settings. And the reason why it's difficult to configure settings is that this chart here, as you see, is not the whole story. The whole story actually comes out when you start looking at this other chart, which shows you the standard deviation when it comes to these preferences.

And so the story here, and the reason why privacy is so complex, is that we don't all feel the same way about these issues. If we did, then it would be simple to come up with defaults and use these defaults for the entire population and it would be done. And perhaps even the FTC could jump in and say, well, nobody feels comfortable about this. Therefore, we're going to outlaw it. Clearly, that's not the way we operate.

And so the reason why this is complex is because we have this diversity in preferences. Some people are quite fine with their fine location being shared with advertisers, and others object. The good news, however-- and this is a result that has come out of our research over the past years-- is that very often, it is possible to organize the population and their preferences into fairly small groups of people that feel very much the same way about these issues.

And so what I want to share with you here is, again, an early example of our work in this area, where again, looking at these mobile app permission preferences we're able to organize a population of users in just four groups. And just based on these four groups and what we're able to predict based on the preferences within each one of these four groups, we're able to show that it might be possible to predict somewhere between 75% and 85% of their privacy preferences when it came to configuring their permission settings.

And so this is very, very simple technology. I'm going to show you that we've been able to go much farther than that. But that gives you a sense already for how easy it is, actually, to predict many different settings that perhaps people would want to have.

So this next chart here shows you the next step in our research in this area, where we looked at actually a population of 240,000 users. I should actually say a population of 3 million users, but we had to clean up the data quite a bit. And we've actually zoomed in on the fraction of the population that was most engaged with their permission settings. So these were LB users LB is a variation of the Android operating system. It was an early version of Android where users could actually configure many different settings.

And we were able to show that through profiles, but also through personalized learning, we could, just by asking people a very small number of questions, effectively predict most of the settings that they would need to configure on their smartphones for the apps that they were going to download. So for instance, if you were to ask them just six questions you could effectively reach a level of accuracy of about 92%. If you're willing to double the number of questions you're asking, you're getting close to 95%.

Now, we are not suggesting in any way that you should fully automate the setting of privacy permissions. We strongly believe in dialogs with users. But there are situations where it's extremely clear how the user feels about some settings. And there are situations where you can determine that actually, your model is not good enough to predict what those settings should be. And that's where you should ask the user. And so that's effectively what we're advocating.

And so we've gone one step further this past summer, and we actually piloted this technology with real users on their actual cell phones. And so we develop profiles-- in this case, I came up with seven different profiles-- and ask people to download this very early version of the

personalized privacy assistant. This assistant would ask them between three and five questions based on the actual apps they had on their cell phones. And based on their answers, it would recommend a number of different settings, as you can potentially see on the right hand side of the slide in front of you.

And so we ran this, and to make a long story short, we ran this for effectively a period of 10 days. The last six days of the study, we actually tried and see if we could nudge users to modify the settings that they had adopted based on recommendations made by these assistants. We tried very hard, with nudges like the one you see here. These nudges are very affected, by the way. So when it comes to getting people to rethink their privacy preferences, when it comes to motivating them, we've actually got an entire study that shows that those types of nudges work very well.

And so here's what we found. So we found that among the recommendations made by our personalized privacy assistants for their mobile apps, about 3/4 of recommendations were adopted by users. And we also found that even after they had adopted these recommendations and modified their settings based on a recommendation, even though we were trying very hard to get them to revisit these settings, they would not change them. That means that in this case, about 5.6% of those recommendations were later modified despite technologies that we were sending them to revisit and rethink their settings.

Now, how do we know? You might say perhaps they were just lazy. Perhaps they ignored your nudges. Well, we had intentionally come up with recommendations that were ignoring a number of other settings, and so the nudges also covered settings that we had not covered in our recommendations. And those settings, users were actually modifying. So we know that they were actually truly engaged, and so this suggests to us that these recommendations are actually pretty close to how people feel about these issues.

And so we strongly believe that this is the way to go for mobile apps. The question is, could we go one step further and could we generalized this to IoT? And so we have started to work in this area. The vision here is that you would extend this to deal with smart spaces.

And so what we're doing right now is we're building an infrastructure where owners of different resources, or resources that are going to be sensing different aspects of your behavior-- cameras, location, presence sensors and the like-- have to be defined in a register by the owners, the people who own these various resources. You know if you enter a room like this, there actually are a number of different people who could actually have deployed different resources already today that collect some of your information.

For instance, it could be the case-- I hope that's not the case-- that the Wi-Fi routers in this room perhaps collect some of your information. These Wi-Fi routers are not necessarily owned by the people who operate the building. Perhaps they're owned by the FTC. Perhaps they're owned by a third party. I don't know, and perhaps it's better not to ask. But on the other hand, the HVAC system in this building might be owned by an entirely different entity. And that HVAC system might be collecting some information, too.

And so the idea is that the owners of these resources should be able to very simply declare where these resources are deployed and what information these resources collect, and all the other sorts of attributes that you would ideally want to see in a privacy policy. So we're developing an infrastructure where, through a series of dropdown menus, people can specify different elements of their resources without requiring them to do any programming, and looking at what it takes to turn this information into machine-readable privacy policies.

The idea is that users then, with their personalized privacy assistance, would be able to enter this space, discover relevant resources. Their assistants would determine, based on their expectations and their preferences, what, if anything, they need to be warned about or informed about. And if there happens to be settings-- in an ideal world, we would like these personalized privacy assistance one day to also configure these settings. We're not there yet, but that's effectively what we're aiming for. So this is roughly how this is hopefully going to work one day.

So let me try to quickly recap, and also make some connections with our public policy in this space. So we truly believe that this approach to effectively leveraging machine learning, in particular building personalized models of people's privacy preferences and expectations, is one way of making notice and choice practical, right?

Today, the number of systems that you're encountering, especially in an IoT context, is just way too great for anyone to imagine that users are going to be able to read policies or configure settings. There's really a need to help users, and to really do so by, number one, building models of what they care about, how they feel about different sets of issues, try to effectively alleviate a burden in that context, and also make it much easier for the various owners of different elements of the infrastructure in the IoT to participate within this infrastructure.

So as was pointed out by Sarthak, I think, in the first presentation on this panel, one of the challenges of IoT is a diversity of players. If you think about the way you interface with the fixed internet, most of your interactions are mediated by the browser. And so it's sufficient in principle to just configure settings in your browser. On the mobile web, by and large the cellphone mediates your interaction, or Android.

And so it's sufficient to configure a number of settings at that level. In IoT it's a very different story, where you have a number of different players that might contribute different elements of the infrastructure. Many of these suppliers might also be smaller entities that don't have the sophistication that Google or Microsoft or Facebook might have, and so we really need to move towards an open environment with open APIs, where effectively people will expose settings that will enable one, through personalized privacy assistance or equivalent technology, to effectively configure many settings on behalf of the user.

And so that's really our vision in this space. You can think of two different ways of deploying this personalized privacy assistant technology. One is to effectively rely on companies like Google or Facebook, each one of them potentially developing its own personalized privacy assistant, building models of the users. You can imagine also some potential tensions and potential conflicts of interest when it comes to that so this would clearly have to come with very strong guarantees.

Or you could imagine a more ambitious effort, where you might say, well, after all, there are actually some interesting correlations between the way you feel about your settings on mobile apps when it comes to sharing information with mobile apps, and perhaps your settings on Facebook, and perhaps your settings in your browser. And so rather than asking you these five or 10 questions in each one of these environments in order to determine what your privacy preferences are, how about just asking these questions perhaps just once, then using a personalized privacy assistant that cuts across all these different environments, interacts with these open APIs to effectively configure many of these settings on your behalf.

So that's our vision in this space. It's not guaranteed that these APIs will be made open. In fact, today they are not. They're very much part of the strategy that some of these larger entities have when it comes to building their ecosystems. But we would like to effectively build an effort towards perhaps convincing these larger players that they would all benefit from opening up these APIs. And perhaps people will ask me questions later on so I get to say more about this, but I think I've run out of time. So thank you very much.

[APPLAUSE]

AARON ALVA: So we'll conclude today with our final discussion of the day. So unlike previous sessions that have focused mostly on privacy, this session has focused on security and usability research as it relates to privacy. So Sarthak discussed security issues related to IoT devices and how they may affect privacy in the home. Vitaly presented on ad libraries and how the lack of tailored security controls in some contexts could result in disclosure of users' information through shared external storage.

For usability, Florian shared about an entire line of research going on around using machine learning, crowdsourcing, and other methods to make privacy policies more usable and for consumers, for businesses, as well maybe for regulators. Finally, Norman presented new ways to understand and manage users' privacy expectations through personal privacy assistance. So overall, this session has provided some new views into different strands of privacy research to consider.

And with that, all of those will add to the policy conversation here. I want to welcome Geoffrey Manne, the Executive Director of the International Center for Law and Economics, as well as its founder, and Davi Ottenheimer, who holds many hats in the security community, including authoring a book on big data security. Geoffrey and Davi will provide some thoughts on this session as it relates to privacy for a few minutes each, and we'll start there. So, Geoff.

GEOFFREY MANNE: So I thought the papers presented some really interesting things, as did the papers throughout the day. And since this is the last session and I have you here, I'm going to talk a little bit more broadly, at first anyway, than just about the papers today-- but in a way that's consistent with what Aaron was saying, which is to say that the papers are interesting, there's some really important stuff here, but as is so often the case, the problem is deriving the appropriate policy implications from it.

One of the things I would say is that it's a little bit unfortunate we don't have more economists and engineers talking to each other. As you might have gathered from the last panel, an economist will tell you that merely identifying a problem isn't a sufficient basis for regulating to solve it, nor does the existence of a possible solution mean that that solution should be mandated. And you really need to identify real harms rather than just inferring them, as James Cooper pointed out earlier. And we need to give some thought to self-help and reputation and competition as solutions before we start to intervene.

Now, it is certainly something in the nature of a conference like this-- and for that matter, the kinds of papers that people are writing, because journals don't publish papers saying there's nothing wrong. They publish papers saying there's a problem, and perhaps suggesting solutions to them. So we've talked all day about privacy risks, biases in data, bad outcomes, problems, but we haven't talked enough about beneficial uses that these things may enable. So deriving policy prescriptions from these sort of lopsided discussions is really perilous.

Now, there's an additional problem that we have in this forum as well, which is that the FTC has a tendency to find justification for enforcement decisions in things that are mentioned at workshops just like these. So that makes it doubly risky to be talking even about these things without pointing out that there are important benefits here, and that the costs may not be as dramatic as it seems because we're presenting these papers describing them.

So think about the potential vulnerabilities that we talked about on this panel. The question to me becomes, should they leave the FTC to any kind of enforcement if companies don't engage in the type of security that was recommended in some places, or even any security at all? And again, this is an FTC workshop, so counselors out there are actually going to have to wonder if their companies are now on notice, and if the very selection of papers for presentation here perhaps indicates anything about the FTC's enforcement agenda.

But here's the thing. Having a possible vulnerability and acting unfairly under Section 5 are not the same thing. And by the way, that's essentially, I think, the holding in the ALJ's decision against the FTC in the LabMD case. Also, in terms of the desirability of enforcement, I think it's important to note that a couple of papers in this session and elsewhere throughout the day have suggested either that self-help is or can be working.

Norman's paper most obviously and immediately suggested a version of that. Or that despite the potentiality of all of these problems, something is actually preventing these vulnerabilities from being dramatically exploited. Self-help has direct legal implications, say, for a deception claim where it matters if it's available. But both self-help and the limited exploitation of risks are important in the economic calculus of the desirability of enforcement.

So I want to end really quickly by saying-- I have more specific questions and comments about the papers when we discuss, but overall I'd just like to say that I think that last point is an area in which we're lacking in research. And I would like to see significantly more research on the implications of the availability of self-help.

And what are the incentives for consumers themselves? We spend all our time talking about the incentives of firms and the implications of legal liability on firms, but what about the consumers themselves? What about self-help? And how does and should the FTC take account of those?

AARON ALVA: Thanks. Davi?

DAVI OTTENHEIMER: All right. Well, I feel like someone's given me a big basket of balls to juggle here at the end of the day, and I'll try to make sense of it all and put on a little bit of a show.

Teeing off what Geoff just said, the idea that there are sort of these experiences we can have and we can learn from and then there are these things we can discover through hard science is a fair split. And I'll apply it now to the talks we heard today, the four talks. I think that goes back to the question, should you study computer science or should you study social science? Should you have an applied approach to risk or should you have an academic approach? And a lot of times, people forget that there's something in the middle. There's a fair balance between the two.

So it was interesting to me to hear the first speaker talk about one end of the spectrum, which is essentially unit tests of these devices, these IoT devices. And then the second speaker took us through an integration test scenario, where what are these devices like in the wild? Let's look at how they're used by people, the economics essentially, the social science of how they're used. So those are two ends of the spectrum, essentially.

And so then the third fourth speakers brought in the middle ground, where you have somebody saying, well, maybe we can use this analytic exercise to help people make small rational decisions. So you reduce the decision set criteria so people can choose from something realistic so you're not forcing people to make big analytic analysis. It's really small. And that's kind of the two ends that I see.

And then the fourth speaker, even more interestingly, has a shared model where not only are you making things easier to decide, accuracy and choice, but you're encouraging, nudging people. So you're bringing an economic model towards the middle, towards simpler decisions with nudges. So that's kind of how I see all the four put together. And I guess I have a ton of questions for all the speakers, but we don't have that much time so I'll hand it back.

AARON ALVA: Thanks. So I wanted to ask, since we're running out of time, a general question across all of the presenters, if there's one policy message that you think currently your research is engaging in as you discussed in your presentations but is lacking in technical measures that would actually help you implement a policy goal you'd like to see? What are those shortcomings, and how would you like those shortcomings addressed? And it's open to any of the presenters.

NORMAN SADEH: OK. That's a tough one. Clearly one has to be realistic about what can be done and how much room for maneuver, I guess, the FTC has in this space. But I suspect that the FTC can play a role in bringing together key stakeholders, encouraging dialogs. And so for instance, the issue to I was alluding to at the end of my talk, for instance in terms of opening

APIs, clearly this will never be something that one would be able to mandate. But perhaps efforts can be encouraged by bringing together key stakeholders.

At the end of the day, when privacy is presented the right way and when people are looking at this rationally, everyone can benefit from better privacy, including vendors that are sometimes presented as if they didn't care about privacy. I think that if you look, for instance, at what is happening today in the mobile space, it's very clear that everyone has come to realize that they don't want to be seen as the people who don't care about privacy. And that creates strong incentives for them to rethink the way in which they've been approaching some decisions in that space.

So I think that perhaps the FTC can, on the one hand, continue to do what it's been doing very well, I believe, which is to encourage best practices it has done, for instance, for mobile apps, as it has done more recently when it comes to IoT security. And perhaps also convening meetings and encouraging efforts where people look at opportunities for perhaps developing common standards-- not trying to impose any standards.

And standards are very challenging and very tricky efforts, but at least trying to bring together key stakeholders and getting them to think about where they've got effectively common interests and where they might benefit from perhaps developing some open APIs.

VITALY SHMATIKOV: I think transparency is very important. Better understanding and better disclosure of how information is collected and shared between various players in the picture is crucially important, because what we have in mobile space today is these old permission models. They capture something about security of the devices. They capture virtually nothing about privacy. There is a lot of information collection and sharing and information used between all kinds of parties-- platform operators, ad libraries, ad builders, advertisers-- that simply exist outside the existing permission models that a lot of privacy work focuses on.

So to the extent FTC can help shed light on this and ask for more disclosure of information collection practices and information flows in this massive mobile ecosystem, that would be an extremely useful service because that is not happening today.

SARTHAK GROVER: So I would totally agree with that. Transparency is the big issue. And maybe the FTC can go, in terms of, say, IoT devices or even mobile apps, unless you follow a certain set of policies, we won't allow you to sell these devices to others.

But the problem comes back to a point Norman mentioned, that in terms of IoT devices they're not really open APIs. I mean, who basically sits there and looks that all of this? Who does the analysis when you don't really have access to the code? And the software and the hardware are basically integrated. You don't have choices in case you feel like something is wrong in the software. You aren't really in a place if it's something else.

So transparency is the main issue, and it should be encouraged. But quite frankly speaking, I don't know how to go about it.

GEOFFREY MANNE: But one of things-- I mean, there's always tradeoffs. And it may not surprise you all to learn that I wrote a paper called "The Costs of Disclosure." So I agree, transparency tends to be a good way of achieving these things. But it's not costless. As Norman had on his last slide, he pointed out that if we have open APIs, we're going to be empowering the groups that are collecting these massive amounts of information through open APIs with an enormous amount of information that creates perhaps even greater vulnerabilities than the ones we're protecting.

And there may be other examples like that, too. So my question really is, before we settle on transparency even as the right sort of optimal kind of solution here, we should be aware that there are costs to that as well, and that again, potentially we're creating more risks than we're solving.

DAVI OTTENHEIMER: That's right. I put it as transparency to whom? So you're building a trust relationship, so it's transparency to somebody that you essentially trust to give you the right answer, given that they have the information.

And so I've done audits over 20 years, and I can tell you-- just being able to see into something doesn't mean you're in a position to make a decision on it, which is sort of what the presentations were about to some degree. We give people the information. The people aren't positioned in a way that they can digest it, because they don't have the analytic capability at the time they're given the information.

That's why I'm saying balance-- if you take sort of the unit tests, you can say that's inadequate because you have a compliance checklist. If you take the environmental or the integration test, you can say, well, that's not fair because that's not a typical use case. And so if somewhere in the middle is proper use of device prepared for use case, then that's, I think, a good fit.

FLORIAN SCHAUB: So I think concerning transparency, an interesting point to think about is also that the privacy policies we have right now are not written for users, and they're not meant to provide transparency for users. And we need to realize this. And I think this needs to be more clear in regulation as well, that if we want to inform users and achieve transparency for users, then we need to come up with user-facing notices that are actually made for users.

And that could include requiring user evaluation of those notices. Are they actually effective at communicating what they're supposed to communicate? And we've been doing a lot of those studies at CMU, and we find most notices are not effective. And it's really hard to design an effective notice.

DAVI OTTENHEIMER: Here's an interesting counterpoint. If you make more information available-- the more information that becomes available, I should say, the more behavior changes. So if you actually-- I could show you exploits, for example, to your model that show as you get this in position where your machine learning algorithms are working and you're actually getting the answers you want, the people writing the policies will change them just so you can't see them anymore.

So the transparency has to be in concert with the right trust model where people want it to be shown in the way that it's comfortable for them. Otherwise, they adapt and your transparency backfires.

AARON ALVA: Norman, did you want to address the transparency with the--

NORMAN SADEH: I'd like to respond to the last comment. So I think it's clear that privacy is an arms race. I think that-- and I worked together with Florian on the project that he described. But the day that site operators, for instance, start modifying their policy based on our technology because of the success of our technology would be a very good day.

We're not quite there yet. If that day happens, we will actually have the ability to probably identify that. And that might potentially be something that the FTC would be interested in. Whether the FTC would necessarily be able to do very much about it or not, I'm not sufficiently versed into the legal ramifications of that but I suspect that it would have something to say if you can establish effectively a pattern where once you effectively are able to capture some practices that are not necessarily putting these companies in good light, they start modifying the way in which they're presenting the text, I suspect that something could potentially be done.

FLORIAN SCHAUB: And it's also quite imaginable that it would go the other way so that companies actually improve their language to be better presented by these independent mechanisms. And we have conversations with many different companies that say they would actually welcome having such kind of technology out there, because they do invest a lot of money and time in having privacy policies that are descriptive. But it's basically in vain at the moment, because this information is not used and it's not clear to users that this is the case. So I think this could go both ways, but it's going to be interesting to see how it plays out.

GEOFFREY MANNE: My sense would be that the primary reason for the unintelligibility of existing disclosures privacy policies is the legal risk-- and for that matter even, regulatory enforcement. So if the problem is we don't have disclosures that actually inform the users, then to me we've largely identified a really important disconnect between how we're regulating and the power of users, which goes to the point I was making before, which is why I really liked what you were describing. The sort of app that you guys created seems to me like it has amazing potential.

But once we have something like that, think of what that does to the need for additional forms of regulation. You might still need some deception regulation, but you've done a really good job now of actually giving users what they want. And then because users are so heterogeneous, because types of data are so heterogeneous-- I think, by the way, on your paper, there's a big difference between an email address being accessible and the content of a communication, even with a computer device.

A real problem with over-general-- and this may be actually partly reflected in the bad privacy policies-- like a network-level response to the problem you were identifying, is that-- well, I don't know enough about the engineering, but at first cut I would say it doesn't differentiate. It

just imposes a single policy on everyone regardless. And that's really unlikely to be the right outcome.

But that is a problem with the relatively blunt policy tools that we have. So again, I think there's real value in empowering the users, as long as that leads to a reduction in the incentive of these more blunt tools to come in.

AARON ALVA: So we have about 45 seconds left. I wanted to ask the presenters, if you have the ideal privacy agenda in your research, what would it be, in one or two sentences, going forward?

NORMAN SADEH: I think I've outlined our agenda. And there were three presentations today, so I strongly believe in this vision of personalized privacy assistance. It's clearly not something where we're entirely there yet, but we've got some very promising results. If I can take another 30 seconds--

AARON ALVA: No, sorry.

NORMAN SADEH: All right.

AARON ALVA: But thank you, though.

FLORIAN SCHAUB: OK, so I think what's also important, what we're starting to look at is providing information and integrating these dialogs into the user's interaction flow. So rather than having a privacy notice, a privacy policy somewhere else when the user interacts with it, make it part of the interaction.

The mobile platform developers are doing a good job doing this already, or starting to do this already. You have those just-in-time dialogs that pop up, and they don't disrupt the interaction flow. They actually help it, and they actually encourage the app developers to build dialogs around it that tell you why this notification is going to pop up and why they want your location.

So that's great. I think that's a good direction we're going. And I think we're doing quite interesting research to [INAUDIBLE] to the Internet of Things.

NORMAN SADEH: And we're not biased.

AARON ALVA: So I'll stop you there. I encourage the audience to ask Vitaly and Sarthak after this, but I wanted to conclude by-- oh, there you are. OK.

[LAUGHTER]

So the FTC's new Chief Technologist started on Monday. And so I wanted to welcome Lorrie Faith Cranor from the FTC, and we also thank Carnegie Mellon for allowing her time on leave for her to be here with us.

LORRIE CRANOR: Thank you. I will keep my remarks brief since we're over time. First of all, I wanted to thank all of the FTC staff who did such a wonderful job organizing this event. Can we give them a big round of applause?

[APPLAUSE]

Yeah. So this is my fourth day, so I had nothing to do with it. But these guys did a really great job. I also want to thank all of you for coming and for participating.

A few notes on some things that I heard throughout the day-- it was a lot to absorb, and I was busy scrawling notes and trying to synthesize what I heard. So I think that some of the key areas that I heard, there's a lot of really interesting empirical research that is being done. And some of the areas of this being done in that we heard about-- we've heard about survey and interview research about what consumers understand, and especially what they expect and what they desire. We also saw that some of this research is then being used to find ways to actually assist consumers, figuring out ways to reduce the number of notices that they need to see, and configure their settings automatically.

A question that came up in almost every panel, I think, was a question about how we can make transparency and notice and choice more effective. We heard over and over again how ineffective it seemed to be, and we heard some ways forward, some paths to maybe making it more effective.

We also heard about measurement research that looked at a variety of things. We heard about measurements on the extent that people are being tracked, and what technologies are tracking them. We also heard about statistical and machine learning research to understand how algorithms impact users, and our speakers observed that in order to have algorithmic transparency it's not enough to just know what the algorithms are, because that doesn't really tell us very much. What we need is systems that help interpret the results of the algorithms and show us the impact of those algorithms.

I saw some research that built models and investigated the impacts of different approaches to privacy protection and can help shed light on the effectiveness of different approaches. We saw research to understand the impact of incentives in a purchase of cybersecurity.

We also saw that many of the researchers who spoke here had developed some tools that had been very useful in their own research, and many of them had actually offered to make their tools available to other researchers who could also use them. And I think the community is developing a tremendous tool set that should enable a lot more research to happen going forward.

We also heard from research an eagerness to partner with companies to do empirical research. Some people noted that in order to do the research they wanted to do, they need information that only the companies have. And so there was an invitation to partner with them.

So those were kind of the highlights of what I heard today. I'll be very interested in hearing from all of you about what you found useful. We're also interested in getting feedback on this event.

Should we do it again? If so, should we do it exactly the same way? What should we do differently? We'd be very interested in hearing that from you.

One of the things that I would like to do while I'm at the FTC is to try to better bridge the gap between academic research and policymakers. I think the privacy area is an area where there's a real need to inform policymaking with research. And so as such, I look forward to continuing the discussions that we started here throughout the year. Thank you.

[APPLAUSE]

[MUSIC PLAYING]