FTC Big Data: A Tool for Inclusion or Exclusion? Workshop
September 15, 2014
Segment 2
Transcript

KATHERINE ARMSTRONG: I'm Katherine Armstrong from the Division of Privacy and Identity Protection. And I have to say, we've been looking forward to today for a very long time. And so thank you all very much for coming and welcome to panel one.

Today, this panel is going to examine the current uses of big data in a variety of contexts from marketing to credit to employment and insurance and how these uses impact consumers. Today, we hope to do one of the things I think the commission does best, and that's to ask questions, to listen, and to learn. Before I introduce the panel, I want to remind everybody that Solon's PowerPoint, or his slides, are available on our website and well worth studying as well as his paper.

So let me briefly introduce our panel, and then we'll begin. Kristin Amerling is the chief investigative counsel and director of oversight for the US Senate committee on Commerce, Science, and Transportation. Danah Boyd is a principal researcher at Microsoft Research and a research assistant professor at New York University.

Mallory Duncan is the senior vice president and general counsel at the National Retail Federation. Gene Gsell is senior vice president for US retail and consumer packaged goods at SAS. David Robinson is a principal at Robinson + Yu, and Joseph Turow is a professor at the Annenberg School for Communication at the University of Pennsylvania. So welcome, and thank you again for agreeing to participate in this panel.

I'm going to start with a question about what is big data. What makes this data unique? Is it the three V's, velocity, variety, and volume? Or does it have something else to do with the relationship derived from making connections among data sets? And you're all free to speak to that, or whoever wants to jump in first.

DANAH BOYD: I'll jump in. So I have a mixed background. I started out as computer scientist. I retrained as an anthropologist. So I look at big data from both of those lenses. And we can look at the technical phenomenon and much of what Solon referred to, get that fact. But there's also a social phenomenon which is in many ways tethered to the hopes and dreams and fears and anxieties associated with big data.

The possibility that we will get to a perfected idea of statistical knowledge, that this will give us a new form of fact that will allow us to make meaning of the world around us, which in many ways obscures the complexity of probabilistic information, right? Which is a lot of what we're dealing with, probabilistic. The data is imperfect, just like Solon was talking about.

And so for this reason, I like to think of big data not simply in its technical sensibilities, but as a socio-technical phenomenon that brings with it a lot of different confusion and chaos. I bring this

up because I think it's really important to remember this, especially in light of the conversation we're having today, because a lot of what goes on is the uncertainty, not necessarily the formalistic mechanisms of data mining, data collection, or data analytics.

DAVID ROBINSON: And if maybe I could just briefly pick up on that, I think to one of the things that Solon mentioned that I think is extremely important that was also central to the FTC's report is that in some of these cases, you have data which was gathered initially for some purpose that didn't require high fidelity, like slightly making more accurate the list of people that you send out a mailer to. And now, in some instances, some of that data is being used for purposes like deciding that certain people are likely to be fraudsters and will not be transacted with by actors in the marketplace.

And I think one of the great concerns that the civil rights community has is to make sure that we're confident-- well, I'll speak only for myself-- I'm confident that businesses are going to do things in ways that are optimal from a financial perspective. That if something helps to make something more profitable, that it'll happen. But I think, what is the harm from a civil rights perspective versus from a business perspective when the occasional minority or otherwise marginalized person is incorrectly excluded from some product that they'd be ready to transact with?

At some level, some amount of that is a cost of doing business. And I think one question is whether the amount of that that's acceptable as a cost of doing business is the same or is different than the amount that is acceptable as a civil rights matter. And I'll just say, our group of technologists that works with civil rights folks released on Friday a new report on big data and civil rights, which you can find at bigdata.fairness.io, which does our very best to inventory these concerns.

GENE GSELL: I'd like to go back for a second to what is big data. Data's been around for a really, really long time. And people have been using it and analyzing it and trying to figure out what it means and what they could do with it. Today, there's just more of it. It's this phenomena, this new thing called big data has existed, it's not something that just came into vogue. It's something that's been around for a long time. And big data, by definition, is more data than your organization can handle.

That's big data. So if you've got more stuff coming to you at home than you can deal with, you have big data. The question really becomes, as more and more data sources become available, more and more data is out there, how do you gather it and make sense of it. I think an awful lot of people give the industry more credit for sophistication than actually exists. Most people, for the most part, are still somewhat overwhelmed and a bit behind the curve on the notion of dealing with all of the new informational data that's coming through.

JOSEPH TUROW: Can I just pick up on that? I agree. And I've talked to a lot of people who say exactly what you're saying in the retail business, for example, that they're overwhelmed and that we're at baby steps now. But it's the beginning of an era. And I would object to the notion that big data or simply the continuation in volume-- because when you start adding velocity and volume and variety, and the notion then becomes predictive analytics. We're in a different world.

We're in a world where hundreds and hundreds of data points are used to come up with conclusions about people that are almost not even intuitive a large part of the time. You come up with the-- you have a key indicator that you're trying to look for. But the notion of which data are going to be used in the end, an example, which may sound crazy, but it's not totally nuts, let's say you're a retail establishment.

And you're interested in trying to predict which people are going to become less valued customers. And you have a definition of a less valued customer. You run your data with your hundreds of thousands of customers. And you find that people who start buying vegetable seeds for planting in an urban environment predict that they are going to become less valued customers in the sense of giving back more stuff, you getting only for sales.

Now, you might say, what does one have to do with another? I could think, and this gets back to what Danah was saying, there are lots and lots of reasons we could think about, and I could give you some, as to why a person buying vegetable seeds would be predictive as a customer that you wouldn't want to deal with the way you deal with other customers, giving discounts and other things like that.

But from the big data standpoint, the key is it's predictive, OK? We may not be sure why it's predictive. And it gets used like that. And the notion of personalizing data that way is a terrific change in the way companies begin to evaluate their customers on many different levels.

MALLORY DUNCAN: Let me just say a couple more words about the retail industry. Obviously, we operate on a very narrow profit margin. It's about 2% on average. And so it's important for the industry that we're able to find those customers who are going to be long, loyal, valuable customers.

We talk about big data, in a sense, we're really talking about an expansion of what's always been done in the retail industry. If you go back 100 years and you think about how your typical store worked, the store manager was constantly analyzing the shoppers in his store and trying to determine, what is it I have to move in the store in order to attract more people? What is it I have to say to this customer in order to increase the loyalty?

With big data, or what's referred to as big data, is an expansion of that effort. They are new analytic tools in order to accomplish the same thing. If we're not able to bring people into the store and not able to get them to increase what they're spending, then chances are the store's not going to survive.

DANAH BOYD: I think this actually raises a different question, which is tethered to the topic of today, which is how do we even start to measure a sense of fairness? Which is usually what we're starting to think about, sort of the challenges of how big data gets used. Now, in an American historical context, we usually have a battle between equality and equity as our models of fairness, right?

Equality is the idea of equal opportunity. We create an even playing field. Everybody enters the table at the same fair starting point. And that's how we constitute fairness, when we have equal

opportunity. Equity, of course, is saying, guess what? We have a large amount of systemic issues that results in the fact that people do not enter the table at the same level. So how, then, do we think about offsetting or dealing with those structural issues, and how do we think about reconstituting the societal infrastructure so we can think about fairness?

And mind you, we have a long debate in the US on this issue of equity, right? We get into this question of affirmative action. We get into this question of whether or not that constitutes socialism and politics, politics, politics. But there's a third logic that big data brings to bear with what we talk about as fairness, something that is very much coming from the market-driven logic that Mallory talked about, which is the idea that we're trying to optimize out efficiencies, and to think about distribution of limited amounts of resources.

Think about how we allocate in the best way possible in order to either maximize profit, minimize law enforcement officers on the street, in another context thinking about how we distribute resources or maximize opportunities. The challenge with that is that that market-driven logic of fairness often really comes up pretty viciously against our notion of what is equity because of the fact that, as Mallory pointed out, we have these really small margins. And the question then is who bears the responsibility for the fact that we have retailers who need to figure out how to be profitable? And we have the fact that many of our customers are not going to be that profitable element.

We've had this historically. Where do we actually allocate new stores? Do we do it in a way that is near neighborhoods who are not considered profitable? How then do we think about the social ecosystem? The reason I bring this up is because big data is-- when used well, when the predictive analytics are done right, when the data mining is done with some level of statistical accuracy, you can get to a point of all of that unintended discriminatory or unfair outcomes because of the fact that we're trying to maximize profit, minimize risk, and really deal with those efficiencies. And that's part of the trade off in a commercial setting.

KATHERINE ARMSTRONG: And we're going to be following up and circling back to the fairness and ethics as we continue on with this panel. But I think that's an important issue to bear in mind because it resonates to all that we're talking about. I'd like to ask Kristin to also describe a little bit some of the findings of the senate's Big Data Report last year.

KRISTIN AMERLING: Sure, I'd be glad to. And thank you for the opportunity to participate today. Chairman Rockefeller, as chair of the Senate Commerce Committee, recently conducted an inquiry into how consumer information is collected, analyzed, shared, and sold that, I think, shares the goal of this panel today, which is assessing what is the current landscape here.

And just to give you a little bit of background, the inquiry was conducted by reaching out to nine major data brokers to ask what are their practices in obtaining, analyzing, and sharing consumer information. And Jim Rockefeller released findings in a report at the end of last year, a majority staff report. I think there are four major findings that are particularly relevant to the discussion that we're having on this panel today.

First, companies, data brokers that collect information without direct interaction with consumers and often without their knowledge are collecting a tremendous volume of data, and it has tremendous specificity. Second, the companies are collecting this information from a very wide variety of sources.

Third, the result of analyzing this information collected includes products that are lists of consumers that define them by characteristics that include their financial and health status, including groupings of consumers based on financial vulnerability and other vulnerabilities. And they include another set of products that the chairwoman referred to this morning related to scoring consumers, predicting their behaviors based on data that's collected. And some of these products very closely resemble credit scoring tools that are regulated by FCRA, raising questions about how these products that may or may not fall under FCRA are being used.

And finally, the fourth finding I think is worth noting is the lack of transparency that consumers have into data broker practices. And I'm happy to elaborate a little bit more on the four points.

KATHERINE ARMSTRONG: Well, you know what? Why don't you weave them in as we continue the conversation.

KRISTIN AMERLING: OK, sure.

KATHERINE ARMSTRONG: But raising one of the points that Kristin just brought up, I wanted to also throw out to the group whether where the data comes from matters, whether it's coming from internal sources, external sources, third parties, whether it's passively collected or actively collected. Does it matter in terms of use or types of information? Joe?

JOSEPH TUROW: Yeah, I think it matters a lot. But I think we have to be careful to say that just because a store, for example, collects the data, it's not a problem. The example I gave with the seeds, just to push that a little bit forward, could reflect a hidden discrimination.

Let's say a person begins to plant a garden in an urban area because she's just lost her job, has to take care of her grandchildren. Those kinds of subjects can be brought out not in direct discrimination. We know this person has lost her job. We know this person had to take care of her grandchildren. She has no husband or whatever. But rather, the fact that she's buying vegetable seeds.

You see? It's the idea of hidden discrimination even within a particular store. Now, add to that the things that you can buy from third parties that could build even greater profiles about people without anyone knowing that it takes place. People going through stores with loyalty cards, and then the material gets put on top of that, which can lead to many types of discrimination that we have no clue about.

GENE GSELL: So that's certainly a possibility inherent when you do analytics on data. But one of the things that really is driving a lot of the change is the ability to process all of this data. It's one thing to collect it. It's another thing to actually do something with it. And I would contend that the ability to eliminate the need to sample-- so historically, data was so big that you did

samples. And inherent in samples are some of the biases because they're based on how the sampler decides to set up their sample set.

When you have big data and you have the ability to use what I'll call big compute against big data, you eliminate the need for sampling. And when you eliminate the need for sampling, you go against the entire data set. You have a much greater chance of eliminating historic biases that have existed based on the way people have decided that this represents an entire population. You don't have to represent an entire population anymore. With big data and big analytics, you can hit the whole thing.

JOSEPH TUROW: But that's my point. That's exactly what I'm saying. What I'm saying is that increasingly, companies-- and now, it's harder. Five years from now, it'll be easier. Companies will be able to use data in variety, velocity, and volume in such a way as to personalize a model so that if I find that there are 1,000 characteristics that I can bring together and come up with just a couple that make me decide that I should go after you, that may be a discriminatory decision. And you don't even know it because the data that you're using are so part of the person's life in secondary ways that they discriminate even though it's not said that it's a low income person, or a person of a certain minority group. It just shows up that way.

GENE GSELL: I think you give us more ability than actually exists.

JOSEPH TUROW: Last thing I'll say about that is you're right, but what's happening is what is the trajectory of interest. And if you look at what people in the business are saying, that's where they want to go. They don't say they want to discriminate. But they want to say, we want to be able to predict what a person is going to do when that person is walking into a store. Eric Schmidt, at one point, said about Google, we want you to go to Google to find out what your job should be in the future. OK? That's what he said several years ago. We want you to go to Google to find out what your career ought to be. That's quite a statement. They can't do it now.

DANAH BOYD: So one thing that's important to understand is that the data that we're talking about is not just about the data that you may give to a company or data broker or even your interaction purely with them. But in many ways, it's about how you fit within a network of other actors and what else they're doing, right?

Historically, we understood this as categories. And in fact, a lot of our conversation about discrimination is a conversation of how one fits into a protected class or protected category. And you think about categories as a way of bucketing. And this has to do with the fact that we didn't have the whole data set, and we couldn't actually imagine the kinds of personalisation that we're talking about.

Personalisation is only made possible because you actually can position somebody in relation statistically to a whole variety of other actors through networks. Networks that, in many ways, are not intentionally designed by the system creator. They're looking literally for correlations that they can see, or probabilistic connections. But this also means that we're dealing with data sets or people that don't have say over what goes on.

So I think about this, for example, with Facebook. And part of what to keep in mind with all of this is all these businesses have different reasons why they're doing different things. Facebook wants to give you a service that if you've not signed up to their site before, they want when you come in that you don't end up in this weird desert of no friends, no content, no nothing, right? Because that's miserable.

And so one of the things that they have gotten much better at doing is determining before you've even shown up what is the likelihood that you sit within a particular network? Now, they can do this because of the fact that your friends have most likely added your email address to their system. So your friends made decisions to give information about you to Facebook. They can do this because they can also assume that, once they have that basic information, they can make decisions about who else within the network-- what do people like? What are they interested in?

They can start to say, hey, might you be interested in this and give you some channel to start engaging. And this is where we get to this question of, what kinds of data are we talking about? That individual never gave over their information. They didn't give over their list of friends. Their friend gave away them. And the site was able to interpolate. And this is what becomes part of the challenge of a lot of the data analytics techniques that we're talking about. We're not talking about a known trade-off between an individual and a data analyst. We're talking about the way in which an individual is positioned intentionally or unintentionally within this network based on what they have or have not given over, or what's been given over about them without their even realization of it.

KATHERINE ARMSTRONG: So let's follow this up a little bit. So does it matter how this data is being used? Danah's been talking about the social network context. I'd like to take it back a little bit to traditional marketing or eligibility type determinations. Does the use of the data help define how it should be collected or how it should be used?

MALLORY DUNCAN: Models are at best, as I think was discussed earlier, just estimates. And we don't know how reliable they're going to be in every instance. And you can imagine-- and they can be accurate or not.

You can imagine a company trying to sell a very expensive automobile. And it pulls various lists. And it says there's a 30% chance that people will come into your showroom to look at this car versus another list, there's a 20% chance and 5%. So they have the money to send out 10,000 solicitations. And they're going to obviously pull from that first list.

They may not realize until later that that list is 95% men and 5% women. Now, is that a fair determination? Is that accurate for that car? Well, the car happens to be, say, a Maserati Grand Turismo. It may turn out that men are much more interested in a car that is a $200,000 phallic symbol than are women. But you can't really say that the use of the analytics was inappropriate in that case.

DAVID ROBINSON: I think one thing that is so important and is not yet part of what we're often talking about but is sort of under the surface of what we're talking about is the desire that consumers and historically the regulatory regimes have to understand why decisions were

reached. So one of the big things that happens in the Fair Credit Reporting Act context is that if an adverse decision is reached, the consumer has this right to have explained to them why the decision was reached, which means that if new kinds of data are being used to reach debt-recovered decisions, there needs to be this ability to spell out in some fashion how did that decision arise from that data?

And relatedly, in the Equal Credit Opportunity Act context, a model that has a factor in it that's correlated with protected status, which of course many of the key factors are that predict credit worthiness, sadly because credit worthiness is itself not uniformly distributed across protected status groups and the majority, so how do you decide whether, not withstanding the fact that it correlates say with race, a factor can still be used in a credit model? And it turns out there's a two factor test.

One is that the factor has to have a statistical relationship to credit worthiness, which is unsurprising. And the other requirement is that the factor has to have an understandable relationship with credit worthiness. So under existing ECOA precedent, if buying seeds at the store predicts that you are a bad credit risk and someone wants to use that in a credit model, even if the prediction is stronger than lots of other more intuitively financial related factors, it may nonetheless turn out that this use is not acceptable because the relationship is not, in the words of the financial regulatory guidance, understandable.

And I actually think that it's a central tension in big data because when you think about the promise of it, it's to service relationships that weren't intuitively obvious to us in the first place, things we didn't already know, but that nonetheless are useful in the marketplace. But I think that to the extent that the payoff from these new technologies is to tell us stuff that we couldn't intuitively have figured out, by the same token, it's a double edged sword. By that same logic, you have the problem of it being very difficult potentially to explain either to consumers or to make visible to regulators what the relationships are, or even for the decision makers in business themselves to understand what are the reasons why certain factors are standing up in these models.

GENE GSELL: But there's also a tendency to have big data be more inclusive than exclusive. And I'll give you a quick example. We work with the state of North Carolina, their education system. And one of the things that has been determined to be very important about education and going through education is the ability to take algebra in the eighth grade.

Now historically, the way you got into eighth grade algebra was teacher recommendations. We've been able to work with North Carolina around analytics to analyze test scores, just pure test scores, from the fourth grade through the seventh grade actually, to determine that there is a group of the population that is normally not considered for eighth grade algebra based on combinations of things that are beyond just the test scores, or things in the test that are more than just the actual answers.

And as a result of this, we've identified for the state of North Carolina, the schools have identified 20% more students who were not eligible for eighth grade math based on teacher recommendations. And of those 20% more students, 97% of them go through eighth grade

algebra without a problem. So they would have otherwise been excluded. But through big data and analytics, they're included and they succeed. And it's a huge win for inclusion, not exclusion.

KATHERINE ARMSTRONG: OK, does anyone else have some examples of how big data has been inclusive or solved a problem similar to what Gene has laid out in either the traditional credit or marketing advertising context?

So all right, then let's take this a slightly different way. But I would like the panelists to be thinking about real examples that they have because one of the goals of this panel is to sort of lay the landscape of current usage.

GENE GSELL: I have lots more, but I figured I wanted to let other people talk. So I'll hold them and work them in.

KATHERINE ARMSTRONG: How about you do another one, and then we'll see if that triggers?

GENE GSELL: So along the lines of credit scores and how people are included or excluded, through the use of better data and better analytics, one of the large auto companies that issues credit on a regular basis has been able-- and historically, they're very conservative, which is we want our risk profile against our consumer loan base to look like this. They've been able to use big data to actually include more people in the sample set than it excludes.

They actually have a mantra, which is how can we be more inclusive, turn down less people, if you will, so that we can tease out the people who historically don't have a good FICA score, but they are, in fact, still good credit risks? So working with them through the analytics, we're able to find the people who are normally excluded, include them back into the population to give credit to. And again, the historic default rate on the incremental people that we bring back into the population is lower than historic credit failure rate across the entire data set.

KATHERINE ARMSTRONG: So I think that weaves into one of the comments that David's paper that was released earlier last week noted, that 70 million consumers do not have credit scores. And that alternative data can often be a positive way to include people that previously aren't part of that mix. So Gene, without going into the special sauce, can you tell us what is it about the scoring and analytics of credit that allows non-traditional data to be used in such a positive way?

GENE GSELL: I'm not a credit expert. I should preface this by telling you that. There's an ability to get more sophisticated modeling across a larger data set. And the more information I have-- it's a classic statistical problem-- the more information I have from a statistical forecasting perspective, the better able I am to predict. So by bringing in more data, different vehicles, different data vehicles, I'm able to, if you will, tease out the most likely to be successful credit worthy people. But I can't tell you what the algorithm does.

DAVID ROBINSON: Just to go specifically to additional data and credit worthiness, the big sort of frontier there that has shown signs of statistical strength has to do with the payment of utility

bills. So cellphone bills, power bills, things like that. And on the one hand, there may be people for whom traditional FICA score data does not exist. Nonetheless, they've been paying their power bill on time for many years. Turns out that's a good predictor that they would be a good loan risk.

And so by including that data, there is the potential to expand the group of borrowers for whom the lender can have confidence that they're likely to repay. Nonetheless, when you change how data is used from one purpose to another purpose, there are also social justice risks. So in this context, for example, with utility payments, in New England, there are many states that have assistance programs where if you are unable to pay your power bill, they will keep your heat on in the winter. But what they require you to do is to show that you're delinquent in the payment of your power bill in order to receive the needed assistance.

They say, you don't have to skimp on food. You can buy your groceries and not pay your power bill, and then we'll come in and help you. Of course, if the world changes in such a way that the power bill now becomes also the key to accessing credit, then that conflicts with that assistance program in a way that may lead those people to have a really difficult choice where the state assistance program ends up in effect saying that you have to commit some kind of credit self-harm in order to end up getting help keeping the heat on in the winter.

Now of course, the possibility exists to revise those programs in ways that resolve that concern. But I guess what I'm really saying is that the benefits that are there, I think, are best realized when we tread particularly carefully with the repurposing of data that was gathered in one context for use in another. And I would again say that the use of data to lock people out of transactions that was at first gathered for marketing purposes, where errors were much less of a concern, is a serious social justice concern.

DANAH BOYD: You'll notice that one of the things that happens is that we're often going to public sector examples. And part of the reason why we do this, even as corporate sector working with public sector, is the fact that many of the decisions that are made within private enterprises are not visible. And so this becomes this trade off. Do you assume that the private sector actors are inherently evil? Or do you assume that they're actually trying to do the right thing? And we can agree or disagree on a whole variety of that.

That's actually where it becomes really difficult because these same techniques that can be used to increase different aspects of fairness can also be used to create new kinds of complexities. And it's that tension that becomes really difficult because it's often not visible. And it's not only just not visible to outsiders, it's often not visible to the actors themselves as they're trying to do a lot of the predictive analytics that they're working on it.

We're working with complex learning algorithms. Do the engineers even understand what's going on? This is where we get back to this question of scoring as an example there. Now, the other thing is that when you do this kind of work, what do you do as the intervention? So I'll give an example.

So in Microsoft Research, which is the academic arm of Microsoft, which is nice because it means that researchers publish a lot of their experiments, and so you can see certain attempts to try to figure these things out. And I'll give an example that's not focused on discrimination, but it shows the challenge here. Eric Horvitz is a researcher at Microsoft Research.

And he's at the point with Bing data where he can predict, with a high level probability dependent on somebody searches, whether or not they're going to be hospitalized within the next 48 hours. That's a really interesting puzzle. Now, the question is what do you do with that information? If you are Microsoft and you're running Bing, does that mean you send a warning sign, like, you're about to be hospitalized? That's creepy, right? What's going on with that?

Does that mean you figure out a subtler way, a slick advertisement, as a way of suggesting that they might think about it? Again, where do we get on the Minority Report creepy zone of it all? Or do you not do anything because you don't want to deal with the liability? Those are ethical questions that become part of it, things that companies struggle with all the time when they're doing this. They start to see a trend. They start to realize a correlation. And they go, OK, how do we intervene in an appropriate way.

Now of course, this also becomes a challenge when companies have to think about the responsibility they have beyond their particular domain. So for example, JP Morgan & Chase does amazing analytics work to predict, with high levels of probability, whether or not somebody is engaged in trafficking of humans, particularly for sex. And they can do this based on a whole set of different financial patterns that become obvious.

OK, so their response, because they're a company, they don't intervene in human trafficking. Why should they? So of course they're going to work with law enforcement. But that sometimes is a good idea and sometimes not. And a lot of people who work on trafficking issues have identified why often law enforcement is not the best intervention point where social services is. So how, then, do we think about the ethics of those responses?

And this is where we've got this big challenge with corporations. What are they choosing to look at? Are they choosing to do it in a way that we deem to be ethical or appropriate? What do they do with the information that they get? And when and where or should they make this information public?

And it's not easy to work things out. So I don't want to assume that just our silence and failure to give examples is not that companies are engaging always in bad acting. A lot of it is that these things aren't visible for a whole variety of ethical concerns.

KATHERINE ARMSTRONG: And I think that's one of the points, Kristin, that the report showed last year. Would you care to elaborate on that?

KRISTIN AMERLING: Yes. We ran into this lack of visibility issue in a number of ways when we were looking at the practices of the representative data broker companies. First, the companies are gathering information largely without direct interaction with the consumer. So the

consumers themselves aren't really aware that other companies are using their information, or that the companies necessarily even exist.

And then, in looking at the contractual provisions provided to the committee, we saw that many of the companies perpetuate this secrecy by including contractual provisions in their contracts with their customers that say, you're prohibited from disclosing what your data source was. And then, even when a number of companies do provide-- a number of the companies we surveyed do provide some rights of access for consumers to look at the data that they have on them. And in some cases, they provide some rights of correction if the consumer feels the data is inaccurate. But even when those rights are provided, and not all companies do provide them, they don't have much value when the majority of consumers aren't even aware that the companies exist or are collecting this data.

And then we, in addition, ran into several large companies that out right refused to provide to the committee who were their specific data sources and who were there specific customers. So those were all obstacles in trying to understand how this information is being used and analyzed.

MALLORY DUNCAN: We're in a very interesting situation right now, especially in the retail community because we're in a transitional period. For a long time in the world, there existed the online community, which a great deal of information tends to be gathered. And then, there's the in store community, where it's a lot more meager. And we've seen behaviorally in stores and consumers where they want to view this as omnichannel. And they want to buy it online and they want to return it in the store.

Well, that means there have to be data flows back and forth between those two markets. And so the folks who are running the store have to figure out, how far can we go? And what we find happens, and this may explain some of the information shortages that you're talking about, what happens if that they look at correlates to what consumers expect in terms of the use of information in the store, and that's the model they use. So they tend to be very conservative in terms of expanding the use of the data or the expansion of that data in a store.

KATHERINE ARMSTRONG: Could you give an example of that?

MALLORY DUNCAN: Sure. There may be cookies that are used online that will travel from location to location. In a store environment, we're uncomfortable with that kind of movement. We would say consumers are comfortable being observed in the store. And so information may be gathered and used within the store context. But they're very reluctant to go beyond that because that violates the store's expectation and the consumer's reasonable expectation.

DANAH BOYD: Let's be clear. Mallory's hinting at the fact that there are actually a lot of startups out there that are actually trying to track mobile phones in the stores. And there's a big tension within the retailers as to whether or not to implement that because it parallels the cookies issue. It allows you literally track the unique identifier of a phone, see whether you've seen that person before, see what their patterns are, see how they're navigating the store. All of that is technically feasible. The question is whether or not retailers want to implement it, or what the challenges are of doing so.

JOSEPH TUROW: Well, I've spoken to a couple of people who say they do exactly that now. And all you have to do is think about loyalty cards, loyalty cards which are kept by virtually everyone here who goes to a supermarket, probably uses a loyalty cards. It's like 90% of Americans who go to super market that give out loyalty card use them because otherwise you lose a lot of money if you don't. They track everything you do.

Until the last few years, they haven't been able to do much with it. They haven't, for lots of reasons, done any big data analyses. That's changing totally. And there are companies, for example Kroger which owns part of Dunnhumby, which is a company that's designed just to do these sorts of analytics. The idea now, companies like Macy's and others are putting pods of these beacons in stores that look at you when you reach a certain point and then give you specific blandishments, like discounts based upon your shopping habits.

Catalina Marketing, for decades, had been giving people these long coupons as you check out based upon 52 weeks of looking at your shopping habits anonymously. Now, they're beginning to do stuff in the store in a digital sense and outside the store. So in fact, you're absolutely right. What's happening out is stores are getting so nervous about the online environment that physical stores are bringing the internet to the store.

And the big data are extremely a part of that in ways that Danah mentioned and in other ways, as well. And that's exactly what's happening. It's a fascinating trajectory partly because of the growth of big data in the online world.

MALLORY DUNCAN: And then, if I could, it's also because the consumer expects that seamless experience. And it presents the retailer with a bit of a dilemma. You want to treat the consumers in the way they like to be treated. But you want to be sensitive to the privacy implications and the use of the data at the same time. And how you square that circle depends on the reputation of each retailer.

KATHERINE ARMSTRONG: But is it a transparency issue? Do you think we're at a-- in 5, 10 years it'll be totally different because the consumer's expectation of privacy or not sort of being their purchases or their behavior being followed? I almost hear you saying that it's sort of expected online but not in a store. That seems like a little bit of a disconnect to me.

GENE GSELL: To some extent, it's generational. I am high on the creep factor on some of those things. But my kids, they have no problem. They expect that. To your point, they expect the same kind of offers and services, interaction, online. When they walk through the store, they expect the same experience.

DANAH BOYD: I want to sort of [INAUDIBLE] because young people, there's a lot of self-delusion. Young people are actually just as self-deluded about a lot of this as we adults are. There's not this big difference between young people. They want privacy, too. They're focus very heavily on the people who hold immediate power over them.

I want to just think through an experience that all of us had. We came in here this morning. We knew it was going to be recorded. We knew people were going to take pictures. We're at a public

event, right? You saw the webcast notice. And yet, when we heard this morning the list of details, like if you object at any moment to a photograph being taken, as Tiffany went through this morning, you said, I want to leave, right? This is really creepy.

And even though you know it, part of it is that you had to put it down. You had avoided it. You hadn't thought about your hair in perfect quaffed form. This is one of the challenges that we run into all the time, which is that notice and information is not always the best way to actually create a meaningful relationship. And there's a lot of self-delusion on both sides.

The reality is that also we collect a lot of videotape that we never look at, right? My guess is that most of us are never going to look at the videotape of how badly our hair looks on that camera, right? Part of it is this interesting challenge of, how much do we purposely put this information aside and navigate through. But I would not put this as a generational issue. This is not a generational issue. And Chris Hoofnagle in particular has done phenomenal work looking at the consumer side of it. Young people feel the same way as adults, but their trade-offs look different.

KATHERINE ARMSTRONG: This is an education issue, then. It's easy to suggest that it could be a generational thing or not. But I wonder, how do we educate people, not just adults, not just children or younger people to expect that, or to know, that their transactions will be recorded or collected?

DANAH BOYD: [INAUDIBLE] asking to educate them about the fact that they are powerless, right? That's what the education ends up being about. Either you opt out of this room, or you'll be recorded, period. But you have no say. And that's one of the trade-offs that happens all the time online or in commercial environments.

You want to go and buy something from Best Buy. You will be recorded. Get over it. Otherwise, don't go into Best Buy.

DAVID ROBINSON: Just to pick up on this transparency and something that Danah earlier said about how we go to these public sector examples because we don't know what's going on inside of these private enterprises, I think that's absolutely true and is central, really, to the FTC's future decisions about what to do in this area. Education about the fact that a practice happens in general does really little if any help to try and figure out whether that practice manifests in a discriminatory fashion for particular people.

And Dr. Sweeney's work on the discriminatory delivery of online ads is indeed a unique example available in the public discussion, which is why the chairwoman mentioned it this morning and we've come back to it here. And I think what I'd like to see is a world in which you don't have to be a world leading data scientists who also happens to personally be the victim of discrimination in order to have the tools that are necessary to check that that's happening and address it.

And certainly, after the study came out, Google changed its practices with respect to the delivery of ads opposite names in general in order to avoid the discrimination harm of these disparaging arrest-suggestive ads. But that's an extremely unusual case. And I think we would all like to see a world in which if harms like that are happening to people who are not academics and data

scientists with all of the resources that it would take to be a personal scholar of that discriminatory harm, when that harm befalls someone who's in a different position who's more in a marginalized position, I think what we would all like to see is for those harms to be treated with equal seriousness. But I think the fear that the community has right now, which I think is an extremely well-grounded one, is that when harms of that sort do befall someone who's in a marginalized position, they really don't have the tools today to not only solve but necessarily even to diagnose those problems.

KATHERINE ARMSTRONG: Sorry. I was going to say that some would argue that the Fair Credit Reporting Act is a mechanism in the credit context because it's doing exactly the sorts of things you're talking about, which is when adverse action-- if an adverse action is taken, you're provided a notice that the adverse action was a result of something in the credit report, and you're given the opportunity to dispute that information.

So I wonder whether the expectation in the credit world is a little bit different because they know they have this mechanism in place. And whether that's a metric that's useful in another context.

MALLORY DUNCAN: I think we have to make qualitative differences. When we're talking about credit or insurance or education, we may have very different expectations than when we're talking about marketing. Let me go back a moment ago to the example of the sports car. One solution would be to say, no, you must send that offer to come in and test drive the car to more people.

Well, the consequences of that is that people receive the offer who have no interest in it. That's depleting the funds that the dealership has for sending it out. Or people rush in to test drive it who have no ability to purchase the car, thus tying up the service folks at the auto dealership. So you really have to look at the quality of what you're doing as opposed to just saying, let's take the credit reporting structure and apply that more broadly.

DANAH BOYD: I also don't want to dismiss the credit reporting. I think it's an important intervention, and I think I'm very excited to see that being a regulatory intervention. But also, let's be realistic. Many of the people that are most hit by it have not the time, not the connections, not the understanding, not the literacy, not the wherewithal, and they don't feel a sense of power to be able to actually fight it in many cases.

And so when we actually look at that, it's also a question of who has all of those resources, those soft resources, to be able to do the thing that they're supposedly protected for. And that's where this interesting tension emerges of, where are we trying to get marginalized voices, whether we're talking about youth, whether we're talking about protected classes, to raise up and try to be powerful against systems of power that are meant to actually challenge them? Or where are we trying to think about the role of different kinds of advocacy groups or different kinds of actors who work on their behalf?

And I think we have to be realistic about how we're dealing with this. This is the challenge with education. I think a lot of our education narratives go back to consumers without actually thinking about the lack of other resources that they have to make sense of, or feel, agency or

power in light of what's going on. And I think that's a difference between how we think about it theoretically and what we think about in a regulatory context versus what I see on the ground when I deal with a lot of marginalized people who are just like, I don't feel like I have any sense of power to do anything about this, so don't tell me about it.

KATHERINE ARMSTRONG: So what is the solution? What are your recommendations for empowering those people?

DANAH BOYD: This is what I do believe. I believe strongly in the role of advocacy as a mechanism to be speaking on behalf of groups. This is one of the reasons why Dave and I spend a lot of time talking with different legacy civil rights groups for this reason. Those folks need to be educated on behalf of populations as opposed to necessarily-- and they need to have the transparency and the tools and the mechanisms with which to hold systems of power accountable without always going direct to consumer as the right direction there.

DAVID ROBINSON: So these are groups that have unique-- you know, that hold the franchise and have earned the franchise to speak for these communities in policy settings. There are people whose job that is. There are people who do it for everything down to migrant farm workers, and really the most marginalized people in our country have people who are there.

But I think that making the practices transparent enough to give hand holds to advocates in those cases in which there's a role that they do need to play I think is a role that the FTC itself has often successfully played. And certainly, I think the FCRA is a good model for the things that it applies to and it certainly has played a role in making underwriting a relatively conservative area in terms of the applications of big data as compared to these unregulated marketing practices, although as the chairwoman noted in the case of these thinly aggregated scores the may be used to lower credit limit that are putatively outside of FCRA, I think it becomes difficult.

And frankly, I think there are legislative and ultimately constitutional questions about how far the FCRA style model could be extended into the marketing world that I think really do force us too-- also, law and regulation have a valuable role to play. But so does corporate citizenship, potentially. I think people who say we're doing stuff in a way that we would like to be responsible, and we would like to take affirmative steps to make sure that we're not inadvertently having disproportionate adverse impacts, I think there's a role actually there for collaboration with advocates. Because right now, it's not clear what the signposts are, what the benchmarks are for making sure that you're not doing these things inadvertently.

And I think that if I were to project forward five or 10 years, my recommendation, my hope, and also my prediction would be that there are going to be some practices that emerge. And my guess is that they are going to emerge probably in a collaborative fashion that's probably outside of the legislative process.

MALLORY DUNCAN: David, I want to be very careful I think here because access to credit is essentially a fundamental right in this country. Access to a high end men's fashion catalog is not. And we ought not to conflate the two in this discussion.

KATHERINE ARMSTRONG: Go ahead, Kristin.

KRISTIN AMERLING: The kinds of products that we saw in our review of data broker practices that involved marketing did go beyond products designed to promote the most appropriate car or reach the people who are most interested in cooking magazines. There are a wide variety of groupings of consumers based on their financial and health status that includes lists of people who have diabetes, Alzheimer's, are suffering from depression, that consumers may not be as happy to find that they're on as finding out that they can be targeted for the best car that's most tailored to their needs.

And there's actually an interesting article that just came out last week by Bloomberg on widespread sale of health ailments lists that goes right to this point where they reported that just with simple Google searches, the reporters were able to find lists of consumers with their names and addresses that were identified and associated with specific diseases. And they interviewed some of these consumers. And one, who was associated with the diabetes list, was surprised and not at all happy to find out that he was on this list and said he didn't have diabetes and nobody in his family had it. So there are some sensitivities raised by some of these products that I think are a little more in a grey area than just these are the best products to tailor to these needs.

KATHERINE ARMSTRONG: So we're about to run out of time. But I'd like to give everybody on the panel an opportunity to say some parting remarks. We have some question cards from the audience that raise some issues that I think would be worth mentioning. And that is the level of trust that may appear to be missing in the big data context, the relationship of marketers, a person that goes to a store, may choose to go to the store. There may be a level of trust there. But the invisibility of big data disperses that trust a little bit, perhaps. But I would each of you-- and I feel terrible, in a way, because we have ended this panel talking about what the last panel is going to be talking about more, which is sort of the path forward. So as you provide your final remarks, if you would also remember that we were laying the landscape, and if you can bring it back to what's happening now as we wrap up, that would be fabulous.

JOSEPH TUROW: OK, I had a path forward. I'll try to make it a now.

KATHERINE ARMSTRONG: Lay the landscape.

JOSEPH TUROW: The now part of it reminds me about the-- I think it's shameful that in a Commerce Committee hearing, when a senator asks a representative of the data industry whether he could name his clients, he refuses to do that. These are areas of life that impact all of us. And the collection of information about us and their use, I think, should be required. I think companies should be required to say which-- data brokers should be required to say who they get it from. What are the categories?

Because these affect us every day. In terms of education, I think most people learn about credit cards and loyalty from Jennifer Garner on TV commercials than they learn from anywhere else. We have no learning about this stuff anywhere. People are-- it's totally obscure. And I would suggest that's purposeful.

I think the idea of big data is a continuity. There's an element of continuity between that and the quantification of the individual that has gone back 30, 40 years. But we are in a century now that I think will be looked at as the century of data, the century of pinning numbers on people and trying to figure out where that leads people. And we're only at the beginning.

So I think we have to realize that this stuff is important not just for now. And it's going to get much stronger with greater processing and the kinds of things that people are saying today, we can't do it, are going to be done. So the issue is not is this going to happen because it's too futuristic, but when it happens, are we going to have the conceptual tools to deal with it?

DAVID ROBINSON: Just to pick up on the question about trust and where things are today, I think there's an unrealized opportunity to create greater trust with consumers in terms of how these technologies are being used. And I think that the tools that we have from prior regimes about notice that your data is being collected, the notice and consent regime, frankly I don't think offer the tools to create that trust because, as Danah was saying, the data is collected in a way that you don't have fine grained awareness. And you certainly don't have fine grained choice about what's going to happen. And I think that the tools that we need in order to be able to have practices happen that game the predictive pay off from these analytics, but at the same time give consumers good reason to trust that things are being done in a way that they can feel comfortable about, I think those tools have really not been perfected yet and that we're in a place, we're in an exploratory initial place now of needing to build new tools for accountability and trust consistent with the business leveraging of these tools.

GENE GSELL: I guess what I'd say is the genie is out of the bottle. Stuffing it back in is not going to happen. Data is a part of what's going on. There's more of it than there ever was. And there will continue to be more than there was last year or this year. I think that for the most part, the uses of it are much more positive than negative. There are enormous examples of big data being applied to solve big problems, big worldly problems, big human problems, in health care, in genetics, in disease control, in commerce, in terms of how to minimize fuel consumption across airlines or UPS or people like that. But for the most part, it's really very, very positive that we can now compute on data that wasn't even available 2, 3, 5, 10 years ago.

From a consumer perspective, again, I think the economic model still will drive most of the thought process around this. A retailer doesn't want to do something that creeps you out. And the minute they cross the line, they get the worst thing possible for them, which is you opt out. And the worst thing for a retailer is a fair amount of opt outs. They want to keep you in the [INAUDIBLE]. They want to be relevant to you. They want you to be responsive. And their only notion is to give you something more relevant to you so you don't have to filter out all of the noise that's out there. And I think that there are clearly some privacy things that need to be monitored and watched, but on balance, I think most consumers are electing to opt in as opposed to opt out.

MALLORY DUNCAN: I think Gene said it well. There are a lot of retailers out there, several million. And so there's a lot of choice and opportunity for consumers. And trust in that context is more than just one element, such as sharing this data flow or another. It really is about developing loyalty with the customer. So the customer trusts the retailer and wants to return and

maintain that loyalty. One easy example, there are companies out there that gather, like Amazon, huge amounts of data. And yet, consumers know this because they see the sign that says, if you liked this item, you may like that item. They appreciate that, and they go back and shop again and again because they trust Amazon to do what's right by them. And that's what other stores are aiming for.

DANAH BOYD: This space is actually extraordinarily complex. And it's not there are inherently good actors and evil actors. It's about everything is a lot of gray zone. And the other thing I think is important to highlight in this is that we often talk about companies that we're thinking about as high level brands, brands that we can hold accountable and recognize.

But then, we also deal with data brokers whose names nobody recognizes, who are holding on to data, who are buying data at bankruptcy situations, who are capturing things and pulling together data sources that we don't even know about. And this is one of the reasons why this space gets very murky because we often talk about it within specific silos rather than the complexity of it. And Washington's been talking a lot about data supply chains, which I think is a way of interestingly thinking about it. It's a metaphor. It's not a perfect metaphor. But it's a really interesting metaphor to start thinking about that.

How do we start thinking about holding supply chains accountable when we're thinking about these data issues? Not just in terms of the data brokers that the FTC's looking at, but in terms of all our own behaviors around this. The other thing I think is really important to highlight is that many of the companies, especially the big names, are really trying to do their best. They're trying to figure out how to hold this stuff in a responsible way.

But as David pointed out, they don't always know what the best practices should be. And this is why there's tremendous opportunity for meaningful cross-sector collaboration to try to figure these things out. Regulation is one approach. It's a very power strong armed approached. But collaboration is another approach to start thinking about how do we evolve the best practices, and how do they differ per sector? Because as Mallory pointed out, it's different when we're talking about retailers than versus what we're talking about in terms of finance and credit.

What does it look like, and how do we pull things together? Finally, I want to end with a philosophical point, which I think is also about the state of being. The notion of fact in a legal sense emerged in the 1890s. It's a really modern concept. And anybody who lived through the last election in this country saw that we're kind of in a post-fact state. But for better or worse, one of the things that's sort of coming up as a new equivalent to fact is rethinking probabilistic understandings. This is the big data element. This stuff is here to stay. Part of it is understanding what probabilistic systems mean for our whole ecosystem because in understanding probabilistic systems, you realize it's not cleanly fact. It's about trying to figure out how to deal with this.

And how do you hold probabilistic systems accountable? And how do you think about their role in things like rule of law? It's going to be very, very messy. And this is where I say this because a lot of what we're dealing with in terms of the systems that we're trying to hold accountable are probabilistic systems, which are not intended or designed to be discriminatory in a traditional sense, in the narrative of a fact. But they're done in this way that ends up unintentionally doing

so. That goes back to Solon's comment. I think it's really important to understand that philosophically because that's one of the things that we need broad spread literacy on before we run into the systems, or we just assume to treat these things as facts.

KRISTIN AMERLING: I just want to go back to the issue of transparency and visibility. That's a theme that emerged from our inquiry. It's emerged in many of the comments today. The chairman has proposed legislation to provide consumers access, the right to correct their records, the right to opt out if they don't want their information being used for marketing. And this is kind of a baseline for transparency. And it's very interesting to hear about these additional non-legislative tools. We recognize this is a complex and evolving issue and are looking forward to continuing to be part of the dialogue about the impact of big data on consumers.

KATHERINE ARMSTRONG: I want to thank everybody for participating in this panel and bringing the different perspectives that you have. I think one thing that seems fairly clear is that there is no single solution, or there's not even any single way to look at this, that it's very much something that we must look at through a multifaceted lens when we're talking about marketing, credit, social media, and all these other topics.

I hope we were a little successful in laying and assessing the current environment. But I know that the panelists here could have actually participated on any of the panels today because it all does, as David said, a lot of gray areas. So thank you very much, everyone. And you need to return--

[APPLAUSE]

Audience members, you need to return here at 11:00. You have about a 10 minute break. There is a cafeteria, but you can't bring any food in here.