Navigating the "Trackless Ocean":
Privacy and Fairness in Big Data Research and Decision Making
Keynote Address at the Columbia University Data Science Institute
Symposium on "Data on a Mission: Transforming Privacy, Cities, and Finance"
April 1, 2015

Good morning. Thank you, Steve, for your very kind introduction. And thanks to Columbia University and the Data Science Institute for inviting me to speak with you today. The Data Science Institute is poised to examine some of the most challenging and exciting problems in a way that combines scientific, social, economic, and business perspectives. We need all of these perspectives to understand the role that big data is playing, and will continue to play, in our society. Law and policy need to be part of this equation, and I am pleased to have the opportunity to share my thoughts about how policy and legal concerns relating to privacy and data security can be integrated into your work.

You in the audience enjoy a commanding view of these issues. You are the faculty and students at one of the world's leading centers of research and experimentation on the fundamental questions of data science and its application in areas ranging from genetics to planning cities of the future. You work and study in a city that is a leader in civic uses of big data. New York City publishes more than 1200 data sets on a seemingly endless variety of topics, from pothole complaints to school-level SAT results, and makes them freely available to the public.[1] New York is also encouraging engineers, developers, and designers to turn this data

planning and delivery of government services[9] are all areas being transformed by data analytics. These efforts truly represent data "on a mission." And we are only at the very beginning of these developments.

But with these opportunities come challenges. Big data challenges some of our notions of privacy and creates data security risks on a large scale. And some uses of big data could undermine principles about the fair treatment of individuals. These are issues that I would like to discuss with you today. Concepts like privacy and fairness are hardly uncontested. I can't summarize them in a tidy checklist or set of rules that you can immediately translate into code. So asking you to make privacy and fairness part of what you do day to day in your labs and companies may seem like a big request.

One of Columbia's most illustrious alumni, Supreme Court Justice Benjamin Cardozo, encountered a similar feeling when he first became a judge in New York State.[10] In most cases, Cardozo said, the law was well settled and the facts were clear, and the decision in those cases was obvious.[11] In a smaller number of cases, the law was clear, but the facts were less so. In those cases, Cardozo took comfort from knowing that the general rules were clear, even if he had to take a deep dive into the facts to reach a decision. He also believed that these cases could lead

justice that would declare itself by tokens plainer and more commanding than its pale and glimmering reflections in my own vacillating mind and conscience."[14]

I'm guessing that some of you data scientists may feel the same way about privacy questions. Or maybe you are now wondering whether you ought to. That is not my intention; I don't mean to trouble your spirits. In fact, what I want to suggest to you is that you're not on a trackless ocean. There are some well settled principles of privacy and fairness that you can turn to as you develop data driven research projects here at Columbia or other research institutions, or within your companies. This isn't to say that most big data privacy questions fall into Cardozo's first category, in which there's only one reasonable answer. Nor is it to say that there are no disagreements about what basic values we should seek to uphold in the era of big data.

But you should not resign yourself to drifting on a trackless ocean. There is a wide expanse of solid land for applying principles of privacy and data security protections to your work, and being familiar with this broad landscape will serve you well. You may need to take a close look at the specifics of a research proposal or algorithm to spot and address privacy issues. The answers may not be unique or self-evident; reasonable people may disagree about what specific steps best address the issues that you find. But whether you're running a company or doing research supervised by an institutional review board, your work depends on the trust of those who provide data to you. Thinking carefully through the issues and putting reasonable protections in place is critical to building this trust.

To help you on your journey, I'd like to show you how to navigate the trackless ocean of privacy and data security for big data research and decision making.

Solid Land: Section 5 of the FTC Act

Let me start by explaining the role that my agency, the Federal Trade Commission (FTC), plays in protecting consumers' privacy and data security, which are among our highest priorities. We enforce several sector-specific privacy and data security laws, such as those dealing with financial information, children's information, and credit reporting. We also enforce Section 5 of the FTC Act, which prohibits unfair or deceptive acts or practices, to address practices that these more specific laws do not cover. Over the past 15 years or so, we have brought nearly 100 actions under Section 5 protecting millions of consumers – in the United States, Europe, and elsewhere – from deceptive and unfair data practices. We have used this authority to bring enforcement actions against well-known companies like Google, Facebook, Twitter and Snapchat.[15] We have also brought cases against companies that are not household names, but

[14] *Id.*

[15] *See, e.g.*

which we believed violated the law by spamming consumers,[16] installing spyware on their computers,[17] failing to secure consumers' personal information,[18] deceptively tracking consumers online,[19] violating children's privacy,[20] and inappropriately collecting information on consumers' mobile devices.[21] Most importantly, the broad reach and remedial focus of Section 5 allows the FTC to protect consumers from harm as new technologies and business practices emerge. I'd like to spend a moment or two explaining how my agency has done this, because Section 5 enforcement is an important part of the "solid land" that you should know about when considering privacy and security aspects of data science.

The FTC has also used Section 5 to address data collection irrespective of specific representations to consumers. In 2013, for example, the FTC brought an action against a firm that developed software for rent to own companies to install on computers they offered to consumers, to disable the computer if the consumer failed to make t

So the "solid land" of privacy and data security under Section 5 extends pretty far. The FTC's enforcement actions make it clear that the law's prohibitions on deception and unfairness, which have been in force since 1938, apply to personal data collected offline and online, from "old" technologies like PCs and laptops as well as from apps, smartphones, and connected devices.

Line of Sight Navigation: Fair Information Practice Principles

As researchers and members of industry working on cutting-edge problems, you are probably thinking beyond the bare minimum of what you should do to stay on the right side of the law. You might wonder, for example, whether you're dealing with sensitive information; or, if you know that you are, what privacy safeguards you ought to put in place as you work with sensitive data.

A set of Fair Information Practice Principles (FIPPs), which have been part of the privacy landscape for at least four decades, are invaluable in analyzing the challenges that come up in big data analytics today. The question we need to ask is not *whether* they apply, *how* but they apply.

In 2012, the FTC set out a framework for thinking about how these principles can be applied in our data-intensive,

stronger protections than other kinds of personal data, including more rigorous security and more robust notice and choice before collection. For instance, the FTC recommends that companies obtain affirmative express consent before collecting sensitive information. This means, at minimum, giving consumers a clear

mobile apps and found that they transmitted information – some of it relating to sensitive health conditions – to seventy-six third parties, including ad networks and analytics firms.[37] For example, one app transmitted health-related search terms, such as "ovulation" and "pregnancy," to third parties. In many instances, third parties received information about consumers' workouts, meals, or diets identified by a real name, email address, or other unique and persistent identifiers.[38]

When it comes to wearables and health apps,

the amount of linkable data in the hands of researchers can allow them to personally identify their data subjects. For example, a study performed by Yaniv Erlich, a computer scientist who is now on the faculty of Columbia, and his fellow researchers, showed that you can identify a man based on part of the DNA sequence from his Y chromosome, his age, and his state of residence.[42] Studies involving credit card transactions,[43] geolocation,[44] hospital discharge data,[45] and other types of data have reached similarly jarring conclusions. This is why appropriate technical measures are only one part of the FTC's recommendations governing deidentification. We also recommend that the companies that control deidentified data publicly commit to refrain from attempting to reidentify the data, and that companies ensure any downstream data recipients also agree to not reidentify the data. Together, these technical and accountability measures will reinforce the trust that consenting data subjects place in companies and researchers.[46]

Now let me turn to data minimization. Data minimization is often presented as an obstacle to some of the most tantalizing portions of the big data program. Big data sets can allow us to see interesting correlations, or make significant predictions from seemingly humdrum data. Critics of data minimization argue that data scientists won't be able to find correlations or make predictions unless they have free rein to explore data. They argue that just collecting data without using it any further is not harmful at all. Furthermore, in the critics' view, data analysts will be able to distinguish between beneficial and harmful uses, so there is no need to worry about how much data companies collect. Their answer to data minimi

giving up on data minimization could start to chip away at othe

and online), and our interests. But they can also contain inferences about more sensitive attributes, such our race, our health conditions, and our financial status. Data brokers may describe us as "Financially Challenged" or perhaps having a "Bible Lifestyle."[53] They may

I believe that we need to begin with more transparency and accountability, and everyone who participates in the very broad range of settings that I have discussed – companies, technologists, consumers, and policymakers – has a role to play.

Let me begin with consumers. Consumers should be able to exercise appropriate control over information that goes into the pipelines that feed the algorithms that end up having an effect on their lives, particularly where the pipelines are not visible to consumers. I have long urged data brokers and similar firms to give consumers tools so they can tell consumers that they do not want to have their information used for marketing purposes. Consumers should also have the ability to correct information that is used for risk mitigation and other comparably substantive decisions. And these tools should be immersive, with intuitive UIs, so consumers can easily exercise this control. The FTC's data broker report, as well as the White House's big data review, included my recommendations along these lines.[57]

Some in industry are taking steps to provide greater transparency and control to consumers, but we have a very long way to go here. Ultimately, I believe we need legislation to address these issues, but industry can and should do more right now to make these tools available to consumers.

Of course, transparency is not the whole answer, because consumers cannot navigate this complex ecosystem themselves. Responsible data brokers and analytics firms should recognize that they are the Pole Stars in this ecosystem, and they should strive to ensure accountability throughout their data supply chain. They should examine carefully their own practices, as well as the practices of the companies that feed into their pipeline, and those customers who use the output at the end of the pipeline.

Companies should also do more to determine whether their own data analytics result in unfair, unethical, or discriminatory effects on consumers. For example, what if a company analyzing its own data, in an effort to identify "good" versus "troublesome" customers, ends up tracking individuals along racial or ethnic lines. A *Harvard Business Review* article argues that this kind of result isn't just possible, but inevitable.[58] I believe legislation will be necessary to ensure that all companies – the most scrupulous ones and those that would rather remain in the shadows – are held to the same standards. But companies should not wait for legislation to start tackling these issues.

Of course, you data scientists and technologists in the audience have a key role to play. You have the technical insights that are necessary to determine whether specific analytics practices pose risks of excluding, or otherwise placing at a disadvantage, groups defined according to sens I hav(You ha-(pr4, an)]TJ-19.33 -1.to)]TJ-21.