WORLD **PRIVACY** FORUM

Comments of the World Privacy Forum
To the Federal Trade Commission
Regarding Big Data: A Tool for Inclusion or Exclusion? Workshop, Project No. P145406

Sent via https://ftcpublic.commentworks.com/ftc/bigdataworkshop

Federal Trade Commission,
Office of the Secretary,
Room H-113 (Annex X)
600 Pennsylvania Avenue, NW
Washington, DC 20580

October 31, 2014

The World Privacy Forum is pleased to have this opportunity to submit comments regarding the FTC's September 15, 2014 workshop, *Big Data: Tool for Inclusion or Exclusion*? We submitted substantive comments prior to the workshop. We would like to submit these further comments, as they follow up on specific ideas that grew from discussions that occurred during the event.

The World Privacy Forum is a non-profit public interest research and consumer education group. We have published many research papers and policy comments focused on privacy and security issues. Much of our work explores technology and health-related privacy issues, biometrics, consent, data analytics, and many other rapidly evolving areas of privacy. You can see our publications and more information at www.worldprivacyforum.org.

I. Statistical Parity: what it is, and why it is important

At the workshop, panelists discussed legal and ethical frameworks that are applicable to large datasets and issues of discrimination and privacy. The view we articulated was that we should not throw out Fair Information Practices nor existing regulation. However, we acknowledged that something is missing in the area of consumer protection, and we coined the term "Statistical Parity" to describe this issue.

Here is our definition of the term:

> **Statistical parity** means ensuring that all parts of the consumer data analytics

> process are fair: data collection, which data factors chosen and used for analytics, accuracy of the factors and how well the algorithm works for its intended purpose, and then how the final results are vetted and used, and for how long. Statistical parity means finding ways to ensure privacy and fairness in the analytics process from beginning to end, and to ensure that decisions about consumers are accurate and used fairly and in a non-discriminatory way.

We did not mean that each algorithm needs to be seen by consumers or the FTC. This is not feasible, nor a desirable outcome. We did mean that how fair and accurate the underlying factors are, how well the algorithm actually works, how the resulting analysis is used for consumers, and consumer access to the most meaningful analysis is important, and forms a new area we need to pay attention to.

Nowhere is the need for additional work in the area of statistical parity clearer than in the context of consumer scoring and predictive analytics. Consumer scores are simply the outgrowth of a robust and growing market and use of predictive analytics. But as modern statistical shorthand, scores are important. We published a substantial report about this issue in 2014, *The Scoring of America* <http://www.worldprivacyforum.org/2014/04/wpf-report-the-scoring-of-america-how-secret-consumer-scores-threaten-your-privacy-and-your-future/>. We submitted this full report as a formal comment to this workshop, so we will not reiterate the issues at length here.

Briefly, in the *Scoring* report, we outline in depth the policy issues around a lack of privacy protections and fairness in analytics, consumer segmentation, and predictive consumer analytics.

We define a consumer score as follows:

> A consumer score describes an individual or sometimes a group of individuals (like a household), and predicts a consumer's behavior, habit, or predilection. Consumer scores use information about consumer characteristics, past behaviors, and other attributes in statistical models that produce a numeric score, a range of scores, or a yes/no. Consumer scores rate, rank, or segment consumers. Businesses and governments use scores to make decisions about individual consumers and groups of consumers. The consequences can range from innocuous to important. Businesses and others use consumer scores for everything from predicting fraud to predicting the health care costs of an individual to eligibility decisions to almost anything. Dixon/Gellman, *Scoring of America,* p. 8.

Many thousands of scores exist, yet consumers have scant rights to find out what their non-Fair Credit Reporting Act scores are, how the scores apply to them and with what impact, what information goes into a score, or how fair, valid, or accurate the score is. Even if the input to a score is accurate, consumers do not know or have any way to learn how information derived from their lifestyle, health status, and/or demographic patterns is used to infer patterns of behavior and make decisions that affect their lives. Those who create unregulated scores and analytics about consumers have no legal obligation to provide Fair Information Practices or due process to consumers. This in totality is what we mean when we talk about ensuring statistical parity.

II. Examples of Statistical Parity (or lack thereof) in Action

We would like to flesh out this concept of statistical parity with a few specific examples, some of which we touched on at the workshop.

A. The Categorization Paradox: Data segmentation, categorization, and privacy and fairness challenges

When individuals are placed in a category, the very act of categorizing that individual automatically triggers privacy and fairness concerns. This is because when someone is categorized, the very act of doing this can be used for inclusion, or exclusion. At the workshop, we called this the *categorization paradox*. Categorizing or segmenting data may bring **helpful or non-helpful consequences** to the individuals in the data. This is a much-studied and well-known issue, for example, see Cynthia Dwork's work on fairness in categorization, *Fairness Through Awareness*, Dwork, Hardt, et al., arXiv:1104.3913v2 [cs.CC] Nov. 2011.

The categorization of an individual may be based on marketing data, and as such may be highly inaccurate. The categorization may be based on information typically protected in other statutes, such as race, age, and gender, or on proxies for that information. The categorization or segmentation could be offensive in many additional ways. For example, a categorization of someone as mentally ill, or as a compulsive spender could put an individual into a category that is socially stigmatized. If inaccurate, that is problematic in and of itself. And whether inaccurate or accurate, that categorization could result in positive and negative consequences for that individual. We give you a specific scenario for this occurring in the real world.

A major US health insurer worked with an analytics company to determine whether or not publicly available consumer data could enhance the quality and effectiveness of their predictive risk models. They tested approximately 1,500 factors at the household level and found that the consumer information that showed the most value in predicting individual level risk included:

- Age of the Individual
- Gender
- Frequency of purchase of general apparel
- Total amount from inpatient claims
- Consumer prominence indicator
- Primetime television usage
- Smoking
- Propensity to buy general merchandise
- Ethnicity
- Geography – district and region

- Mail order buyer - female apparel
- Mail order buyer - sports goods

Those unfamiliar with predictive models can find it surprising to learn that information about purchasing sporting goods can become a part of a predictive risk score for a health insurer. Yet it is not unusual to find the factors used in this example are in a modern predictive consumer score model. This is actually a fairly short list compared to some models with thousands of factors. This example comes from Satish Garla, Albert Hopping, Rick Monaco,& Sarah Rittman, *What Do Your Consumer Habits Say About Your Health? Using Third-Party Data to Predict Individual Health Risk and Costs. Proceedings*, SAS Global Forum 2013. <http://support.sas.com/resources/papers/proceedings13/170-2013.pdf>.

In today's world, it is unlikely that consumers know or have the ability to know which of their activities, purchases, interests, or inherent qualities has categorized them. It is also unlikely for consumers to be able to know all of the entities that have categorized them. It is further unlikely for consumers to be able to determine what marketplace and life impact such categorization might be having. We understand that there are good examples and negative examples of uses of large datasets. This is part of how the categorization paradox works, and it is important that we move beyond simply citing examples of good or bad uses, and work to extend our understanding into the roots of how fairness is operating at the categorization level.

Among the questions we think are crucial to ask and address with rigor, we include the following:

- How to segment in a fair and accurate way, including removing discriminatory factors used in analyses that are protected in law or factors that serve as proxies for discriminatory practices?
- How to sift segmentation for *hidden* categorization effects and hidden discriminatory effects? (This is often due to mirroring of factors in data sets.)
- How to account for accuracy in each segment, and ensure correct usage of the data according to its accuracy level?
- How to deal with liability issues for someone who acts on incomplete data and gets an outcome deemed undesirable by the individual?
- How to allow for consumer participation and access to categorization results, including removal?
- How to ensure the individual's right to shape their digital exhaust and have control over how they are being analyzed for meaningful decisions that have an impact on their lives or livelihood?
- Categorization based on health factors such as diseases that are not curable are extremely stigmatizing. This is of great concern, as again, the categorization of an individual as having a rare incurable disease can be used to help or to hurt them. How can uses of data in this category be controlled so that the result is benefit for consumers, rather than harm?

B. Modeling / Scoring of individual consumers

Another important scenario to discuss involves the scoring or modeling of individual consumers who are sliced, diced, and analyzed routinely without their knowledge or consent. If one never knows how or when the resulting record will be used, and in what kind of context it will be used, the fairness and privacy concerns should be readily apparent. The trails of our digital exhaust have many potential uses, some beneficial, some not. We understand the lure of research into large data sets; but the data sets with the most impact on privacy are those that can be tied back to an individual. It is the scores that are attached to individuals or even households that we see as having long-term impacts.

Among American adults, each individual with a credit or debit card or a bank account is likely to be the subject of one or more scores or predictive models. Many individuals signed up under the Affordable Care Act have a health risk score. Individuals who buy airline tickets have a score. Individuals who make non-cash purchases at large retail stores likely have a score. Scores such as the medication adherence score, the health risk score, the consumer profitability score, the job security score, collection and recovery scores, frailty scores, energy people meter scores, modeled credit scores, youth delinquency score, fraud scores, casino gaming propensity score, and brand name medicine propensity scores are among the consumer scores that score, rank, describe, and predict the actions of consumers.

To create a consumer score, the score modeler feeds raw information (factors about consumers) into an algorithm designed to trawl through reams of data to detect consumer behavior patterns and to eventually sift consumers into a ranking by their scores. Each score generally has a name and predictive or descriptive function.

Today scoring models are easily built from data that is **extrinsic** to the final score. No nexus may exist between the **input** to a score and the **output**. In the financial scoring area, companies can now build financial scores from social media, demographic, geographic, retail purchase history, and other non-traditional information that may not be included in the formal credit file. In the health arena, analysts can now build health risk scores from mere wisps of demographic data, without any actual patient records.

In this new world of scoring, where analysts use factors extrinsic to the purpose of the score to build scores, that a person has red hair can be used as a factor. And the more factors, the better. Instead of using 30 factors, why not 3,000?

How to specifically bring fairness and privacy protection into each step and aspect of this process is an important area for inquiry and work. It is not feasible to have human review of each algorithm creating modeling or consumer scores – there are too many scores, too many models, and it happens too quickly in too many places. What, then, is the answer to bring statistical parity to individual consumers? We do not have all of the answers yet, but we are working toward them, and encourage the FTC to focus on this issue.

III. Individuals' Digital Exhaust and Big Data

As individuals live in an increasingly sensor-rich, data-rich world it becomes less and less possible to control the transactional and other data flowing from us as we simply live our lives. Connected cars, biometric identity controls, daily financial transactions typically recorded via debit or credit cards, and many other digital byproducts add to these data flows. There are legitimate uses for much of this data, much of which ends up being part of "big data" or large datasets.

How this fundamental issue of rights to shape our digital exhaust is navigated will have long-term impacts on privacy, and on certain questions related to large datasets, when those datasets contain personally identifiable information, or can be linked to or impact an individual. While some large data sets are genuinely not able to be tied back to an individual, some are. It is the datasets that contain individually identifiable digital exhaust that we are concerned with here.

Individuals need rights to shape these identifiable or re-identifiable digital exhausts, in particular those that have meaningful impacts on their lives. Otherwise, individuals lose much personal choice, and this can ultimately be limiting in a democracy, and much worse in non-democratic states. We are not saying that this will be easy to navigate – this issue is complex, and it is likely that overlapping and layered rights will need to be employed to be effective. In some situations, it will not make sense for people to shape or control their digital exhausts. In other situations, it will be crucial.

To cite an everyday example of digital exhaust: when we make purchases with credit or debit cards, retailers can sell or share our transaction histories with third party data resellers or data brokers, who can then use the results to build categories and segment that data with ever increasing levels of refinement. This is part of an important fact-finding that came from the FTC's Data Broker report. Scores applicable to individuals such as the Consumer Prominence Score can be the result. This score has been used to help determine health risks for health plan rates, as we documented in our *Scoring of America* report.

Currently, consumers can control some of this kind of digital exhaust by using cash. It is possible that some digital wallets may also help in this regard. But what actual rights do consumers have to restrict the sale of their retail purchases to third parties? Should that right exist, or should other frameworks protect consumers? There is need for much fact-finding and discussion here.

We especially see the need for consumers to have rights to shape and control their digital exhaust in meaningful areas. Some fragments of this is legislated; for example, HIPAA now gives patients a right to restrict disclosure of treatment information to a health plan when services are paid for in cash up front. The FCRA gives consumers the right to dispute inaccurate information in certain eligibility contexts. These are the types of remedies that will be welcome as data and its uses increase. However, these rights are scant, and fragmented across sectors. As strong authentication and identity increase in use, it will be extremely challenging for the average consumer to hide their digital exhaust tracks. We would like to work ahead of this curve to see how problematic inclusion/ exclusion issues can be avoided.

IV. Protection Frameworks

Digital exhaust and Statistical Parity are challenging problems, and in the workshop, we suggested several approaches. We would like to articulate the “layered” model we discussed at the workshop again here, and discuss some other approaches that came up in discussions at the event.

A. The layered protection model

We have hypothesized that a layered approach that provides multiple overlapping protection regimes is going to work best going forward. We would like to see Fair Information Practices plus statistical parity, for example, working together as a model. Perhaps a responsible use framework and an ethical framework could be added on top of this, in addition to, **but not replacing** the other frameworks. This way, existing law and other frameworks could be used as a more complete net of protection.

At the workshop, we mentioned the Nuremberg Code as a human rights ethical code of great value in regards to human subject data research. It is this kind of ethical code that could be part of the mix that is helpful in achieving results in consumer protection and fairness. The Nuremberg Code would be helpful in some sectors, while other well-established and genuinely vetted ethical codes may be helpful in other sectors.

The days and era of having a single-silver bullet solution to privacy are long gone, and it is much more helpful to find effective ways of layering protections and building in individual consumer choice – meaningful choice – wherever possible. Perhaps there will be a big–data-specific idea that comes into practice. This could be added where needed layered on top of the baseline fairness frameworks.

We emphasize, however, that any model that leaves out FIPs will be incomplete in approach.

B. Responsible use frameworks and context

A focus on responsible data use has become an important part of the privacy conversation. What comprises “responsible use” in one setting may not be so in another setting. Trying to quantify and weigh both benefits of data use and potential risks and harms is a helpful exercise. However, while this approach seems to make sense, it is not clear how it could possibly incorporate all the various and changing perceptions of benefits and harms that individuals might have.

Because of the inherent problems here, we urge research and more inquiry on specific scenarios to tease out a variety of underlying patterns – and we also urge an approach that views responsible use from an individual point of view, including choice mechanisms that individuals should have. We also urge a study of privacy risks or harms from the point of view of the individual, not just the organization that will be benefiting from the use of the data. Privacy-related risks to a company or a government agency are very different from risks or potential harms as seen by the individual, and it is the risks to the individual that privacy rules should be designed to mitigate.

Going back to our retail sharing example, would sales of consumer transactions constitute "responsible use" in a framework, and if so, on what basis? Is there a role for FIPs in a "responsible use" framework? Are consumers notified about uses under such frameworks? Do they have access? Or do fundamental information rights get taken away in this model? In a use-based model, often the FIPs-based rights (or at least some of them) get removed in preference to back-stopping the privacy challenges at the point of use. We would like to see research that looks at layering a responsible use framework with FIPs-based frameworks, and possibly other frameworks for maximum effectiveness.

Some of the questions around responsible use models include:

- How are FIPs incorporated in a responsible use framework?
- Are consumers notified about data uses under responsible use frameworks? Do they have access? Are any fair information rights taken away in this model?
- How does statistical parity plus FIPs plus use protection work together compared with other single-model approaches?
- How can a responsible use framework be designed to mitigate risks to the individual?

C. The Role of Individual Consent in Large Datasets

Granular consent, or consent with many detailed choices for the consumer to exercise about how their data is used or shared, remains an important area to think about. Granular consent is not appropriate in every instance, especially when thinking about very large data sets. However, granular consent deserves much more attention as a possible helpful tool. We admit that granular consent can be controversial in many respects, and fact-based discussions are much more helpful than conjecture in shaping policy.

We offer a generic comment about the role of consent. Consent should not be required on an all-or-nothing basis as a means to protect privacy. There can be and there should be considerable variation in the nature of data subject consent. Too often, data subjects are presented with a single choice: agree to preset terms that unfairly favor the person who presented the choice, or go away. In many circumstances, consent should offer data subjects more choices, and the choices should be presented in a meaningful way. Consent should be unbundled to the extent practicable. We need to know more about how to make consent more meaningful, especially in the area of large datasets. We should not just assume that it is an impossibility – it is not, as has been demonstrated in the field of health.

Specifically, granular consent has been a hotly debated issue in the electronic health records space for over a decade. It is useful to briefly look at this area for ideas to apply to other types of large datasets. Electronic health records (EHRs) have provided very large data sets in the health sector. EHRs give health care providers, insurers, and other data users expanded ability to transfer patient records for lawful activities. However, because electronic health records collect so much information in one place (or make it available through one system), the need for granular consumer consent expanded in this area. For example, patients often do not want their dentist or eye doctor seeing or having access to their psychological records, but they nevertheless may want to share some records with some doctors.

A large body of academic literature exists in the area of consent and EHRs. In an important and well-researched study, the study results were unambiguous. (*Patients want granular privacy control over health information in electronic medical records,* National Institutes of Health, Journal of the American Medical Information Association, Kelly Caine and Rima Hanania. J Am Med Inform Assoc 2013;20:7–15. doi:10.1136/amiajnl-2012-001023 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3555326/pdf/amiajnl-2012-001023.pdf>

No patients reported that they would prefer to share all information stored in an electronic medical record (EMR) with all potential recipients. The study found that patients' sharing preferences varied by type of information and by recipient (eg, primary care provider versus dentist). Further, overall sharing preferences varied by participant. Patients with and without sensitive records preferred less sharing of sensitive versus less-sensitive information. The authors wrote:

> Patients expressed sharing preferences consistent with a desire for granular privacy control over which health information should be shared with whom and expressed differences in sharing preferences for sensitive versus less-sensitive EMR data. The pattern of results may be used by designers to generate privacy-preserving EMR systems including interfaces for patients to express privacy and sharing preferences.
>
> To maintain the level of privacy afforded by medical records and to achieve alignment with patients' preferences, patients should have granular privacy control over information contained in their EMR. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3555326/.

A study of Indiana patients on this issue found similar results:

> In a study of the granularity of individual preferences among 30 Indiana patients, not one agreed to share all of their information with all providers on a list of provider types. Primary-care physicians were trusted most and fared best, but still only about 75% of patients surveyed agreed to share all their records, even with them. <http://www.modernhealthcare.com/article/20140916/BLOG/309169995

By furthering more factual discussion on granular consent and widening the inquiry into other areas of privacy, there are many important privacy advancements that could be made applicable to, for example, data brokers and data resellers.

We recognize that giving people too many choices can be overwhelming and may not produce the best range of outcomes. That consideration should be a factor in designing consent protocols. The needs for more granular consent cuts across many types of data and many technologies. Again, a better technical architecture will allow the implementation of more nuanced and more appropriate outcomes. We observe that data users have a stake in this debate too.

Some of the specific questions that need to be answered about granular consent include the following:

- How to create granular controls that are flexible through time and can be revoked without causing problems throughout complex systems?
- How to ensure data preferences that are expressed stay with the data as it travels?
- How to design granular consent so that it is not overly burdensome for consumers or data holder and can take place at the time when it is salient and meaningful for consumers?

V. Assessing privacy in a large dataset environment

The workshop discussion touched at times on the issue of how to assess privacy and fairness outcomes in a big data environment. We are interested in finding way(s) of fairly and impartially assessing end results and, therefore, ways of designing systems and products so that the end results can be achieved. We have observed the increasing processing of personal information by government and private entities. Much of the increased processing is the result of technological advances. The same technologies and advancements should be employed to protect privacy and to support data subject rights and interests.

We have several broad suggestions as to some of the promising candidates.

A. Data Provenance and MetaTagging

In some circumstances, it will be appropriate to tag data so that **the source** of the data is transferred whenever **the data** is transferred. In other words, we suggest that metadata be employed to show the provenance of the data. This is technically possible and already happens to some extent in Electronic Health Records. It is also being used in some other contexts. It has great potential for privacy. We observe in passing that better provenance of data has great value to researchers and other data users and data subjects as well.

Metadata will help individuals determine how their data has been processed, who is responsible for any given data field, when the data was created or added, and how to exercise rights of access and correction. Finding practical and appropriate ways to tag personal data is an architectural issue that needs research and attention.

Another area for research and policy attention that would go far to ending some of the worst cases of information abuse is how to ensure the meta tagging is kept with the data, and how to allow for audit trails based on the data provenance that consumers could track. Instead of focusing on opt out, which has its own problems and challenges, a focus on how users could manage rights through managing metadata could potentially go far to address challenges with data reselling, data use, or data alteration that is not authorized, desired, or is harmful.

B. Information tech that supports data subject rights

Information technology can directly support data subject rights. Electronic Health Records provide one example. Giving patients robust access to their health records is a *feature* of EHRs. That same type of access is appropriate for other kinds of electronic records. However, access and other data subject rights must be baked into the technology from the beginning. This is yet another architectural issue worthy of research and policy attention.

C. Ad tech potential

Finally, the same technology already used to deliver advertising to individuals with a particular set of characteristics can be used to deliver notices and requests for consent to individuals under specific circumstances and across platforms. This idea has been discussed by Esther Dyson in a number of contexts. See, e.g, her article <http://www.huffingtonpost.com/esther-dyson/release-90-user-managed-p_b_650383.html>. We do not have copious research to share here, but we see this as having good potential and we would like to see more ideas developed here.

VI. Conclusion

It is still early in regards to certain aspects of large datasets. Nevertheless, we can be sure of some things even now. It is sure that data collections will increase, and that the digital exhaust of consumers will comprise a substantive portion of this increase. It is sure that data analysis – including predictive analysis – raises challenging problems that are not yet fully addressed in current regulatory regimes. It is sure that large datasets contain potentials for help and harm, inclusion and exclusion, sometimes all at once. And it is sure that all of us have much work to do to discover the right questions, answers, and approaches that will find a good balance.

We have suggested that statistical parity is important, and that a multi-layered approach that includes this plus FIPs plus other codes and approaches might be the right answer. We look forward to testing these ideas against facts and groundtruthing what might work as solutions. We appreciate the FTC's work on this issue, and appreciate the thoughtfulness with which it has conducted itself.

Respectfully submitted,

Pam Dixon

Pam Dixon
Executive Director,
World Privacy Forum
www.worldprivacyforum.org