TESTIMONY

Committee on Energy and Commerce: Subcommittee on Digital Commerce and Consumer Protection, U.S. House of Representatives

---

Catherine Tucker

November 29, 2017

Executive Summary

This committee hearing is to evaluate the efficacy of current policies and communications with consumers regarding the collection and use of personal data, in the context of the background that algorithms are now often used to determine the content that consumers see and evaluate. My testimony will focus on some of the difficulties of instituting policies surrounding algorithmic bias or fairness, and then talk about some of the unintended trade-offs raised by restrictions of the use and collection of data. To summarize:

- Algorithms may appear biased for many reasons, including economic efficiency. My own research shows that women may be less likely to see an ad for STEM career advice, not because of the usual hypothesized sources of bias, but because other advertisers are willing to pay more for those eyeballs.

- This suggests that, at least in some cases, there may be tradeoffs between correcting bias and economic efficiency when regulating algorithms and their use of data. My prior research suggests that straightforward data usage restrictions impose costs on both firms and consumers.

- In general, identifying an economically optimal approach to data protection is hard because it is difficult to measure what consumers actually want regarding privacy. However, my research suggests that giving consumers a sense of control over how their data is used is welfare-enhancing. Congress should recognize that different types of data have very different types of consequences for consumers, and temper policy to reflect this.

Chairman Latta, Ranking Member Schakowsky, and Members of the Subcommittee: I was honored to receive the invitation to appear before you today to discuss the topic of 'Algorithms: How Companies' Decisions About Data and Content Impact Consumers."

My name is Catherine Tucker, and I am the Sloan Distinguished Professor of Management at MIT Sloan.

1 Algorithmic Bias or Fairness: The importance of the economic context

Since it is the context of the hearing today, I wanted to start by discussing research I have done into what leads 'algorithms' to reach apparently biased results? This was prompted by excellent work done in Computer Science which documented apparent bias in the delivery of internet advertising by algorithms. My recent research has delved into whether there can be reasons grounded in economics that algorithms may appear biased.[1]

We ran a a field test on Facebook (and replicated on Google and Twitter) which showed that an ad promoting careers in Science, Technology, Engineering and Math (STEM) was shown to between 20-40% more men than women. We then investigate why this occurred:

- It is not because men use these internet sites more than women.
- It is not because women 'inflict' this on themselves, by not showing interest or clicking on the ad and the algorithm responds to a perceived lack of interest. If women ever sees the ad, they are more likely than men to click on it.
- It does not seem to echo any cultural bias against women in the workplace. The extent of localized female equality in the workplace is empirically irrelevant for predicting this bias.
- It is instead because other advertise value the opportunity to show ads to female (rather than male) eyeballs. These other advertisers' willingness to pay more to show ads to

[1] See `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2852260` for the full paper.

women, means that an ad that doesn't specify a gender, is shown to fewer women than men. The algorithm is designed to minimize costs, so shows the ad to fewer expensive women than relative cheaper men.

Though this is a case study of a single ad, and a single instance of apparent bias, this research does highlight the following policy insights:

- In this case, it is unlikely that much could have been prevented (or gained) by mandating algorithmic transparency, even supposing it was technologically possible. The apparent bias occurred because of other advertisers' higher valuation of female eyeballs - and this would not have been clear from analyzing an algorithm that was simply intended to minimize costs.

- This bias occurred because of an attempt by the algorithm to minimize costs to advertisers. This opens the possibility that attempts to mandate lack of bias in algorithms can lead to trade-offs if, for example, it prevented all advertisers receiving a 'discount' for showing ads to men. Society may have interest in preventing women from seeing fewer job ads than men, but not in ensuring that women see just as many ads for shoes as men do, and this makes regulating hard.

- It is not clear what the counterfactual would have been. Apparent bias in who sees ads for STEM jobs may happen offline too if employers with job listings shun publications that are more likely to be read by women because ads in such publications are more expensive to advertise in. We only know that this discrepancy occurs online because of the better data and measurement online. This illustrates the importance of knowing the 'but for' world if the algorithm did not exist, but also the difficulties faced in assessing that counterfactual in an offline and less measurable world.

2 Data Protection and Privacy Regulation Tradeoffs

Though it is perhaps stereotypical that an economist would emphasize the need to consider tradeoffs in regulation, I would like to describe some of my recent research which highlights three potential considerations.

2.1 Costs and benefits of privacy regulation

One of the huge benefits of digital data is that it is virtually costless to collect, parse and store. This makes the collection, use and exchange of data for purposes of personalizing the consumer experience both cheaper and easier than a decade ago. However, this lowering of costs has led to evident privacy concerns, as we are now in a world where anyone's data can be viably collated and analyzed by any organization.

One obvious approach in regulation is therefore to simply restrict data collection. As might be expected, such restrictions have real effects on the digital economy which is premised on the use of data. In earlier Congressional testimony I discussed work that I have done into the effects of the EU's e-Privacy Directive which was associated with a 65% decrease in the effectiveness of online advertising for the advertisers I studied.[2] Similarly, within the US my research has shown that the patchwork of state privacy regulations inhibited the adoption of potentially life-saving digital medical records technology.[3]

My most recent research has tried to distinguish between the effectiveness of different types of regulation. One recurrent insight has been that rather than simply being focused on imposing costs or restricting flows of data, regulation appears to be more effective when it focuses on restoring a sense of control among consumers. I have found this pattern both

[2] `https://www.youtube.com/watch?v=meMxH6c1KGE` based on Goldfarb, Avi, and Catherine E. Tucker. "Privacy regulation and online advertising." Management science 57.1 (2011): 57-71.

[3] Miller, Amalia R., and Catherine Tucker. "Privacy protection and technology diffusion: The case of electronic medical records." Management Science 55.7 (2009): 1077-1093. and Miller, Amalia R., and Catherine E. Tucker. "Can health care information technology save babies?" Journal of Political Economy 119.2 (2011): 289-324.

in responses to very personalized internet advertising,[4] and also in the realm of personalized medicine and genetic data.[5] Other researchers have confirmed that the level of perceived control may also positively affect consumer's appreciation of the use of algorithms.[6]

Of course these costs in terms of efficiency need to be set against the potential for benefits for consumers.

2.2 Difficulties in Establishing Consumer Preferences over Data Use

We also ran an experiment which investigated whether undergraduates at MIT would be willing to release what might be considered very personal data regarding their friends' contact information. We found that on average many of them were willing to release the data. There was a subset of students who stated a preference for privacy and did not release the data. However, if this set of students were offered a slice of cheese pizza in exchange for this data, then they were as willing as the rest of the student population to share this information.[7]

There are two ways of interpreting this study. One is that there is often a discrepancy between an individual's privacy preferences as stated in surveys and what they do with their data when faced with very small incentives (or benefits) of giving that data - the so-called privacy paradox. Another is that if MIT students (who I hope are very well informed about data, privacy and algorithms) behave in a way which is so inconsistent with their stated preferences, then we may need more consumer protection. Regardless of interpretation, though, this study emphasizes the extent to which it is hard to use survey-data or stated-preference data to pinpoint exactly what kind of privacy regime might best benefit consumers.

[4]Catherine E. Tucker Social Networks, Personalized Advertising, and Privacy Controls. Journal of Marketing Research: October 2014, Vol. 51, No. 5, pp. 546-562.

[5]Miller, Amalia R., and Catherine Tucker. "Privacy Protection, Personalized Medicine, and Genetic Testing." Management Science (2017).

[6]Berkeley J. Dietvorst, Joseph Simmons, Cade Massey (2016), Overcoming Algorithm Aversion: People Will Use Algorithms If They Can (Even Slightly) Modify Them, Management Science, forthcoming

[7]Athey, Susan, Christian Catalini, and Catherine Tucker. The Digital Privacy Paradox: Small Money, Small Costs, Small Talk. No. w23488. National Bureau of Economic Research, 2017

2.3 Differences in Potential Harm of Data

The other issue I wish to emphasize is that it is easy in a discussion regarding data to treat all the collection and parsing of data as potentially injurious (or not) to consumers.

There are three criteria I use in my own work to consider the potential 'harm' of data.[8]

- Could the use of this data lead to negative economic consequences for the consumer?
- For how long could there be potentially negative economic consequences for the consumer associated with this data?
- Could this data also have potential negative consequences for others?

To understand these three criteria, let me contrast two potential types of data: 1) 'data that I have been searching and researching flowers as a holiday gift for my mother'; and 2) 'digital genomic data capturing the makeup of my genome.'

The data that I have been browsing for flowers as a holiday gift for my mother is unlikely to have huge economic consequences for me. Instead, the most likely consequences of the release of this data to third parties is that for the next few weeks I receive ads that invite to purchase her flowers, and that may even contain discounts in order to entice me to do so. On the other hand, the public release of my genomic data could lead employers to decide not to employ me if there were reasons to fear for my long term health, and similarly could lead insurance companies to not offer me long term care insurance. Releasing my genomic data has far larger economic consequences.

The data that I have been browsing for flowers as a holiday gift for my mother is unlikely to have much permanent value. I presume there are many people out there who have similarly uninspired gift ideas, so the data is unlikely to have any uniquely identifying value.

[8]See Miller, Amalia R., and Catherine Tucker. "Frontiers of Health Policy: Digital Data and Personalized Medicine." Innovation Policy and the Economy 17.1 (2017): 49-75.

Similarly, the data has little permanent value: In thirty years this data is likely to have little consequence. On the other hand, my genetic data precisely identifies only me, and in thirty years the data will continue to have the same value it has today.

The data that I have been browsing for flowers as a holiday gift, does not really affect anyone else or have informational value about anyone else - except that perhaps it might be possible to piece together my mother's preferred colors. On the other hand my genomic data does have huge spillovers for my siblings, and my children, in that if I am found to be genetically susceptible to something like Huntington's disease, this is a hereditary trait that also elevates their perceived risk levels.

This framework emphasizes that different types of data can have different consequences, and that any regulation, rather than treating all data the same, needs to distinguish between what kinds of data may be actively harmful to consumers and what data may not be.

It also emphasizes that it is tricky to regulate data use by algorithm without consideration as to the economic consequences of the use of that data. There are certain narrow spheres where algorithms and their use of data can have huge consequences, such as employment opportunities and health. However, many uses of algorithms (and data) lead to inconsequential and potentially beneficial increases in personalization of services for consumers and cost-savings for firms.

Thank you for the opportunity to share these thoughts and I look forward to answering your questions.