MSI at the ARF
MARKETING SCIENCE INSTITUTE

MSI Webinar: Data Deserts and Algorithmic Exclusion

July 13, 2023 | Virtual | 12:00 pm – 12:30 pm EDT

Speaker:

Catherine Tucker – *Professor of Marketing, Massachusetts Institute of Technology, Sloan School of Management.*

Overview:

In this MSI webinar, Catherine Tucker (MIT, Sloan School of Management) examined the emergence of data deserts and algorithmic exclusion leading to biases and discrimination. Drawing on her research at MIT, she touched upon the privacy paradox where people say they care about privacy but are willing to relinquish private data quite easily when incentivized to do so. Tucker then pointed to an FTC Privacy conference which marked a shift from the term privacy to algorithmic discrimination and algorithmic bias which connote more of a potential for harm. While it is important to address algorithmic bias, the issue of algorithmic exclusion is a significant issue as well because it addresses missing data created by "differences in privilege" of the user. Tucker defined algorithmic exclusion as "instances where machine learning, artificial intelligence and algorithms get things wrong in a way that causes us disquiet because data is missing." Algorithms need data to apply their predictions effectively. Data sparsity can lead to inequality when economically privileged individuals have greater access to technology that generates digital data, leaving others behind in "data deserts" (by analogy to urban "food deserts"). Data fragmentation results when data brokers lack sufficient information to correctly identify less privileged individuals by gender, race or other demographic characteristics.

Takeaways:

Algorithmic Exclusion

- **Algorithmic exclusion is "when algorithms err because data is missing due to differences in privilege."**
    - This may occur because the data does not exist or the firm in charge of the data capture must piece together fragmented data which means that data is incomplete for less privileged individuals. Missing consumer data leads to people being excluded.
    - Ultimately in the algorithmic bias debate, there is concern regarding how to estimate statistical relationships if the data (coefficient of how to weight a particular piece of data) is biased.

In equation form (it may be Friday night but this is MIT):

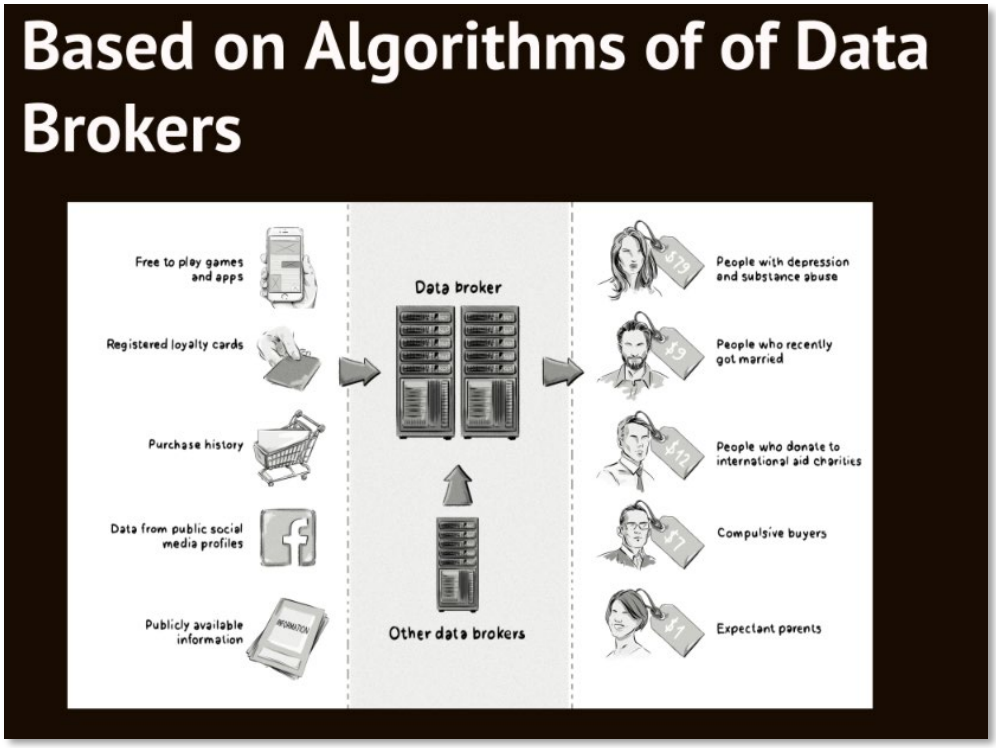
$$Y = X\beta + \epsilon$$

Sparse Data

- In an example provided by Tucker's research, she examined data sparsity via an app that was created to address infrastructure issues (potholes) in Boston, leveraging technology from handheld devices, which required a great deal of data.
    - The technology unintentionally benefited wealthier neighborhoods while neglecting poorer ones which stemmed from issues like a lack of access to unlimited data plans for cell phone owners in poorer neighborhoods (data sparsity).

Fragmented Data

- Fragmented data is more nuanced than sparse data in the sense that when you're feeding data into an algorithm it is usually derived from several datasets (cell phones, email, addresses, names, etc.) to match them to create predictions.
    - Methods used to match data and to create these predictions can also lead to issues of algorithmic discrimination.
- Field study research on leading data brokers found that data purchased on audience segments (Third party consumer profiling) varied greatly in quality and are often inaccurate in their predictions.



- Predictions from the results of the field study indicated that accuracy regarding gender based on data from data brokers was only correct approximately half of the time (i.e., no greater than chance).
    - Results from the study found wealthier, more educated home-owning people were more likely to be accurately profiled because they were more likely to have accurate demographic information.

Results

Data Broker	Number of Cookies	Gender Accuracy
A	1396	27.5
B	408	25.7
C	1777	35.2
D	495	56.4
E	527	48.8
F	480	47.9
G	562	46.8
H	1016	33.2
I	2336	33.6
J	14342	42.4
K	346	30.6
L	547	51.9
M	456	49.1
N	5099	62.7

- In terms of race, predictions from data brokers found incorrect predictions and missing predictions to be higher among people of color, particularly among Hispanic and Asian populations.

Conclusions

- Privacy is a 'rich' person's concern.
- Data inaccuracy is a bigger concern for less privileged groups, creating exclusion and leading to poor predictions (data deserts).
- Algorithmic transparency or auditing doesn't address the missing data issue.
- More focus needs to be placed on data deserts and the way underprivileged populations are represented.

Sources:

The Digital Privacy Paradox: Small Money, Small Costs, Small Talk (Working Paper No. 23488).
Source: Athey, S., Catalini, C., & Tucker, C. (2017). *National Bureau of Economic Research*.

Algorithmic Exclusion: The Fragility of Algorithms to Sparse and Missing Data.
Source: Tucker, C. (2023). *Brookings Institution*.

Frontiers: How Effective Is Third-Party Consumer Profiling? Evidence from Field Studies.
Source: Neumann, N., Tucker, C. E., & Whitfield, T. (2019). *Marketing Science*, *38*(6), 918–926.