February 2023

Working Paper

Algorithmic exclusion: The fragility of algorithms to sparse and missing data

Catherine Tucker

This working paper is available online at: https://www.brookings.edu/center/center-on-regulation-and-markets/

Center on Regulation and Markets at BROOKINGS

The Center on Regulation and Markets at Brookings creates and promotes rigorous economic scholarship to inform regulatory policymaking, the regulatory process, and the efficient and equitable functioning of economic markets. The Center provides independent, non-partisan research on regulatory policy, applied broadly across microeconomic fields.

Disclosure

Catherine Tucker has served as a consultant for numerous technology companies, a full list of which can be found here. The author did not receive financial support from any firm or person for this article or, other than the aforementioned, from any firm or person with a financial or political interest in this article. The author is not currently an officer, director, or board member of any organization with a financial or political interest in this article.

Algorithmic Exclusion: The Fragility of Algorithms to Sparse and Missing Data

Catherine Tucker*

January 18, 2023

Abstract

This paper introduces the idea of 'algorithmic exclusion' as a source of the persistence of inequality. Algorithmic exclusion refers to outcomes where people are excluded from algorithmic processing, meaning that the algorithm cannot make a prediction about them. This occurs because the conditions that lead to societal inequality can also lead to bad or missing data that renders algorithms unable to make successful predictions. This paper argues that algorithmic exclusion is widespread, and that its consequences are significant.

*Catherine Tucker is the Sloan Distinguished Professor of Management Science at MIT Sloan School of Management. She has consulted for many technology companies - please see https://mitmgmtfaculty.mit.edu/cetucker/disclosure/

1 What is Algorithmic Exclusion?

The age of algorithms is upon us. The digital era has allowed firms to collect and store and parse data at far lower costs than ever before (Goldfarb and Tucker, 2019). However, in the past few years the biggest excitement has come from the idea of using algorithms or machines to make better predictions using that data (Agrawal et al., 2016). Algorithms are broadly used in digital advertising. In (Neumann et al., 2019), algorithms make predictions about whether someone who might see an ad is a man or a woman or interested in sports. Algorithms are used in education, where they assess the quality of instructors' education (O'Neil, 2017). In our criminal justice systems, algorithms predict the likelihood that setting bail will be effective (Kleinberg et al., 2018). In HR, algorithms screen resumes to ensure that recruiters focus on the best resumes (Cowgill and Tucker, 2017). Algorithms are used to identify and allocate health spending (Obermeyer et al., 2019). Much of the evidence from the economics literature has been positive, suggesting that algorithms are often superior to the human counterfactual (Cowgill and Tucker, 2019).

However, since the early days of the internet, there have also been concerns that access to the internet parallels existing sources of inequality (Keller, 1995; Servon, 2008). Much of this initial work was focused on access to the internet and then, later, broadband internet (Chiou and Tucker, 2020b). In addition to the question of access, scholars also asked questions regarding usage. While some scholars found positive effects, that poorer households spent more time online (Goldfarb and Prince, 2008), early work also demonstrated signs of their exclusion from the benefits of electronic commerce (Hoffman et al., 2000).

Though much of the digital economics literature has focused on inequality and access and usage of the internet, algorithmic exclusion is a new and important concern for understanding digital exclusion and inequality. Algorithmic exclusion occurs when algorithms are unable to even make predictions because they lack the data to so.

This is a distinct concept from algorithmic bias, which has been discussed extensively in the literature (O'Neil, 2017). Typically, the algorithmic bias literature is focused on the question of whether an algorithm makes predictions that are biased in a way that reflects existing societal inequality (Lambrecht and Tucker, 2019). Usually, the concern is that algorithms learn to propagate existing inequality because they are trained on biased training sets that themselves reflect unjust societal outcomes (Cowgill and Tucker, 2017). By contrast, algorithmic exclusion occurs when algorithms are unable to make a prediction about someone, due to a lack of data about that individual which causes that algorithm to fail to function properly. This paper presents some initial evidence that this lack of data also reflects existing inequality.

To understand the distinction in mathematical terms, it is useful to go to a canonical equation in social science.

$$Y = X\beta + \epsilon \tag{1}$$

In this equation, Y is the variable that needs to be predicted. This might be likely job performance, likelihood of purchase or a risk score. X is a vector of variables that do the predicting. β is a vector of parameters that are estimated from other people's behavior or other settings that help inform how X will affect Y. Much of the worry about algorithmic bias and discrimination has focused on the nature of β. It is easy to think of ways it may be biased in settings where systematic inequality may lead there to be correlations in data that reflect this inequality that are then reflected in β. For example, imagine a hiring algorithm that places a greater likelihood on someone being a successful recruit if they attended a 4-year college rather than a community college. This may well be the product of systematic differences in the economic opportunities of those who attend the two types of college. Alternatively, there may be situations where β is poorly estimated for certain groups due

to missing data in the training set; for example, when voice recognition algorithms struggle with those who are female or non-white because they were not trained with diverse enough voices (Bajorek, 2019). Similarly, machine learning for vision apps has been limited by a lack of data on the behavioral patterns of those with disabilities (Langston, 2020). All of these are concerning, but are not the focus of this article. Instead, the focus of this article is on algorithmic exclusion, which occurs when data that should be in X is missing for an individual, meaning that the algorithmic can not properly make a prediction for that individual. This goes beyond an issue where under-sampled data leads the estimate of β to be distorted. Instead, it reflects the fact that algorithms are unable to make predictions if data is incomplete for that individual.

One reason why algorithmic exclusion has often been neglected in the literature on social justice and algorithms is that it occurs when algorithms make predictions about individuals on a real-time basis. Many of the most talked-about examples of algorithmic bias involve cases where algorithms are trained on population data to make generalized predictions about individuals. Algorithmic exclusion, by contrast, occurs when an algorithm needs an individual's data to make a prediction.

2 Drivers of Algorithmic Exclusion

In order to make predictions about an individual, algorithms need data. However, there are many reasons to believe that less privileged households produce less of the data used by algorithms, and that the data produced is often more fragmented, limiting its usability.

2.1 Sparse Data

The digital economy is composed of digital footprints. Digital footprints are what users of digital devices leave behind after they have interacted with digital technology. I leave a digital footprint every time I search on a search engine, and every time my phone is tracked to a certain location, whether it be by Bluetooth, a wireless signal, or my having explicitly

opted into being tracked by an app. I leave a digital footprint when I share photos online, upload a video, select what stories I want to read on a news site, pay for a vacation online, or adjust my smart thermostat. In other words, every digital activity consistently produces data.

However, many of these digital footprints reflect economic privilege. Using a computer extensively is a privilege. Owning a smart phone with capacious data is a privilege. Having a smart thermostat is a privilege. Making payments easily online is a privilege. In other words, many of the activities that generate the kind of data that could potentially be useful for prediction are also a function of economic background. Of course, economic privilege has shaped access to data throughout history. Famous diarists whose data has been used by historians to reconstruct every-day life in previous centuries also had to have a certain amount of privilege to be able to read and write and maintain a diary (Alaszewski, 2006). Similarly, the inability of the census to count truly count the underprivileged is well documented (Farley, 1995).

Even thinking about work habits makes this point. The pandemic has revealed a class of jobs whose labor can be conducted remotely without struggle, because it is essentially entirely digital (Greenstein, 2021). However, blue-collar workers more often face restrictions at work on cellphone use (Carlson, 2021) and also have more limited data plans, and have no computer at home (Chiou and Tucker, 2020b). Therefore, there is reason to think that economic prosperity is linked to the quantity of data produced to a household. This in turn affects the ability of algorithms to make predictions about you.

One first illustration of this is the experience of 'Street Bump' in Boston (Crawford, 2013). This app was created to gather data about Boston's streets using smartphone's built-in sensors as a resident drives. The idea was to collect data using an app that Boston residents download on their phones, and then parse it to better identify road problems like

potholes and problematic manhole covers.[1] Through Innocentive, the office of New Urban Mechanics held a competition to best parse and create algorithms based on the data. The competition attracted 700 solvers and 19 submissions (Carrera, 2013).

However, to collect the data requires both possession of a smart phone and ready access to unlimited cell phone data. As a result, however good the algorithms were at identifying road bumps, the fact that poorer households did not have the same level of digital access as richer households meant that there was sparser data about poorer neighborhoods. This could potentially lead to algorithmic exclusion, where poorer neighborhoods were less likely to have their problematic road surfaces fixed than richer ones. In the end, to circumvent the problem, the city of Boston decided to only use city personnel to actually capture data using the app (O'Leary, 2013). This blunt response to the challenges of algorithmic exclusion is suggestive of the breadth of the problems that it causes in a system when properly recognized. It also indicates the challenges of dealing with algorithmic exclusion without human intervention.

A decade ago, policy activists coined the term 'data deserts' to highlight such issues (Castro, 2014). However, unlike food deserts - which is a topic that has received a great deal of well-deserved academic attention (Allcott et al., 2019), there has been little academic attention focused on the difficulties caused by data deserts. This article argues that algorithmic exclusion - which happens when data deserts and algorithms combine - deserves far more attention from academics in terms of understanding causes, contexts and potential solutions.

A more complex example of algorithmic exclusion being initiated by sparse data is provided by Lambrecht and Tucker (2020), who studied the effects of sparse data on the functioning of algorithms in paid search advertising. This paper revisited the findings of Sweeney (2013), who documented a troubling result: Google was more likely to show an ad for a public records check (providing criminal records) when a user searched for a name typically given

[1] `https://www.boston.gov/transportation/street-bump`

to a Black person than when searching for a name typically given to a white person. To explore the mechanism leading to this distortion, this paper's methodology by contrast collected data from advertisers. The authors launched a search advertising campaign on Google that targeted 865 combinations of first and last names that are used either predominantly by Black or White populations in the US. Despite the study being conducted many years later than Sweeney (2013), it again found that ads were more likely to be persistently shown close to Black names.

A persistence of ads being shown sounds like the opposite of algorithmic exclusion, but we actually document that this is exactly what was happening. To operate, modern day algorithms are grounded on research on multi-arm bandits. This multi-arm bandit allows the algorithm to make a tradeoff between learning what the underlying X is for a particular piece of content where X is its underlying quality or appeal, and showing the most appealing content (Schwartz et al., 2017). This is colloquially referred to as the 'learn vs earn' tradeoff. A firm faces a tradeoff between showing proven content, and learning about new content that may be better. However, the way that these algorithms have been programmed to work is that they have a greedy initial period where they try and collect as much data on X or how people react to the ad as possible (Schwartz et al., 2017). At that point when they have collected enough data on the ad, they can actually operate and make a trade-off between showing the ad more, or stopping showing the ad. Because of the relative infrequency of Black names in the population (Fryer Jr and Levitt, 2004), the algorithm never learns about the underlying quality of the ad.

The process documented by Lambrecht and Tucker (2020) is a useful example of algorithmic exclusion, where sparse data prevents an algorithm from operating correctly. As a result of algorithmic learning, the platform is more likely to display ads – including undesirable ads – to users who are searching for members of a minority group.

2.2 Fragmented Data

The presence of data deserts, where structural inequality leads to disparate amounts of data being, can obviously lead to algorithmic exclusion. However, we want to highlight another source of potential algorithmic inequality, which is that of unequal data quality stemming from fragmented data.

Every academic has to confront the challenges caused by fragmented data. To take a simple example, imagine a dataset that tracks the fortunes of a firm over time. The firm experiences name changes, address changes, and even misspellings in a way which means that academics have to work hard to ensure they are tracking the same firm over time, rather than inadvertently thinking that they are tracking three or four different firms.

In a world of large datasets, multiple records are often analyzed per person by algorithms. The ability of an algorithm to make accurate predictions about that person is going to be a function of the ability of these multiple records to be merged and correctly identified for a single individual.

To understand why there may be challenges with this associated with inequality, it is useful to understand typically how multiple records are associated with a single person. Usually a dataset will use something called a 'key' to help identify the same person or individual over time. For example, a supermarket might use a rewards number as a 'key' to analyze the same person's purchasing habits over time. Indeed, the difficulty of tracking a single individual without the tool of a rewards number, explains why supermarkets often give discounts and other benefits to people who use rewards cards. However, in many settings a consistent identifying number or code is not available. So in those cases, data scientists would be forced to use less easy 'keys' such as people's names, people's email addresses and people's cell phone numbers. In particular, the use of email addresses and cell phone numbers is particularly common in internet commerce because of the rise of what is called

'deterministic matching,' where data brokers use these forms of data specifically to piece together how the same individual behaves in different digital environments (Brookman et al., 2017).

Some initial evidence of the potential problems of using names to collate disparate sources of data about an individual was identified in a study of credit markets by Bartlett et al. (2022). The purpose of the study was to measure the extent to which algorithms used by financial firms led to worse or better outcomes than human judgment in terms of propagating racial inequality. However, a fascinating part of the study which the authors shared during conference presentations is that their efforts at data matching also reflected disparities in race. Loans taken out by Black people were more prone to have data errors such as misspellings or missing address data fields that meant that it was harder to merge data records for them as well. Though this is an anecdotal by-product of a research study focused on a different research question, it seems likely that this might generalize. Names may well be more likely to be misspelled, and less care may be taken to have a full record, for those who are more marginalized. Therefore, the use of names to try and construct histories of individuals seems likely to cause missing data issues which will reflect inequality.

Similarly, the use of a physical address is likely to be a far less stable key for those who are economically disadvantaged. DeLuca et al. (2019) document how the poorest households in America often move, and when they move they tend to move to places without stable physical addresses - for example, in and out of shelters or short-term housing. Therefore, any attempt to use a physical address as a key to collate and identify records for these poorest households will be fraught with difficulty.

Email addresses and cell phone numbers are likely to also reflect existing inequality which will lead to differential success when they are used as keys. Evidently, both types of key require access to digital technology. Though the cell phone may feel like a ubiquitous technology among older Americans, more than half do not have a cellphone number (Chiou

and Tucker, 2020a). Therefore, any attempts to collate data profiles based on cell phone number will be necessarily flawed for some age groups and lead to differential outcomes for older Americans. Furthermore, younger people who are more likely to change cellphone numbers have themselves reported challenges that result from occasions when they received a recycled phone number, which meant their identity was merged with another person's (McDonald et al., 2021). The Federal Communications Commission has recently set up a database named the Reassigned Numbers Database[2] to help address the issue of almost 35 million reassigned numbers every year. It seems likely that reassignment of numbers is often associated with economic changes and instability, so it also seems likely that lack of cell number persistence correlates with poorer underlying economic opportunities.

When it comes to email addresses being used as keys, there are two potential challenges - email addresses being linked to an individual and then email addresses becoming redundant. Older research (Markman and Scott, 2005) on email addresses identified a division between two email address norms. The first set of email addresses tended to be connected with work or university or school and have naming conventions that are more easily identified with a person. For example, my email address at MIT, cetucker@mit.edu, reflects my actual name. The second set of email addresses tends to be more anonymous and use nicknames or jokey references and be associated with free email providers. Though this research is older, it does suggest that people who tend to have a corporate or university alumni address will be easier to profile. In addition, the possession and consistent use of email addresses is itself a reflection of frequent interaction with digital technology, and so therefore may reflect inequalities in access.

[2] `https://www.reassigned.us/`

3 Effects of Algorithmic Exclusion

3.1 Missing Predictions

The first question is one of simply errors in the prediction process which are generated due to algorithmic exclusion leading to missing predictions. In recent research, Neumann et al. (2019) illustrated the extent to which predictions are simply missing about many individuals in our data economy. This study attempted to figure out general degrees of accuracy in terms of predictions made by data brokers. 'Data brokers' are large firms, often attached to credit bureaus, that aggregate data in the economy and make predictions about individuals. One of the most startling things about this study was just how limited predictions made by data brokers were. In Neumann et al. (2019), 35,000 panelists were recruited and data brokers were asked to make predictions about the age and gender that were associated with these panelists cookies as they arrived at a website. These were often the largest data brokers in the world. However, very frequently no prediction was made. Indeed, it was actually very common for the data broker to only make a prediction about a very small sliver of the potential population.

This data is surprising because typically the premise of the data economy is that data is ubiquitous and widespread about each individual. This shows that in general, there are actually far fewer predictions that can be made about an individual than might be expected. This was due to both the difficulties of matching cookies and also limited digital footprints.

3.2 Inaccurate Predictions

The other consequence of missing data is that there can be inaccurate predictions as a result of it. This is explored in Neumann and Tucker (2021). This paper follows up on Neumann et al. (2019) to investigate what causes data brokers to have so few and so inaccurate predictions about individuals. The evidence suggests that incorrect predictions are highly correlated with wealth, education and home ownership. These drivers were particularly pow-

Table 1: Data Brokers Often Do Not Have Predictions About Individuals

Data Broker	Prediction existed
Vendor A	1396
Vendor B	408
Vendor C	1777
Vendor D	495
Vendor E	527
Vendor F	480
Vendor G	562
Vendor H	1016
Vendor I	2336
Vendor J	14342
Vendor K	346
Vendor L	547
Vendor M	456
Vendor N	5099

Source: Neumann et al. (2019)

erful for predictions about background or demographic data and less predictive for people's interests. One potential interpretation of this pattern is that the inaccuracy in this case was driven more by fragmented data than by limited data. This is because the scope of digital data created should be more useful for predicting interests than anything else. On the other hand, fragmented data might limit the ability of the data broker to accurately merge the data profile with voter records or other sources of truth that are useful for establishing background data. Though it is on the face of it ambiguous whether it is desirable that a data broker can indeed merge voter records with other information, their inability to do so could have consequences for individuals if that merger is then used for credit decisions or risk profile assessment.

A recent study by Kaplan et al. (2021) describes some of the challenges that the issues of algorithmic exclusion can cause. In this study, the authors attempted to match voter files from North Carolina with the records of data brokers. The voter files contain self-reported race, and the authors asked major data brokers to append a birth year to them. What is

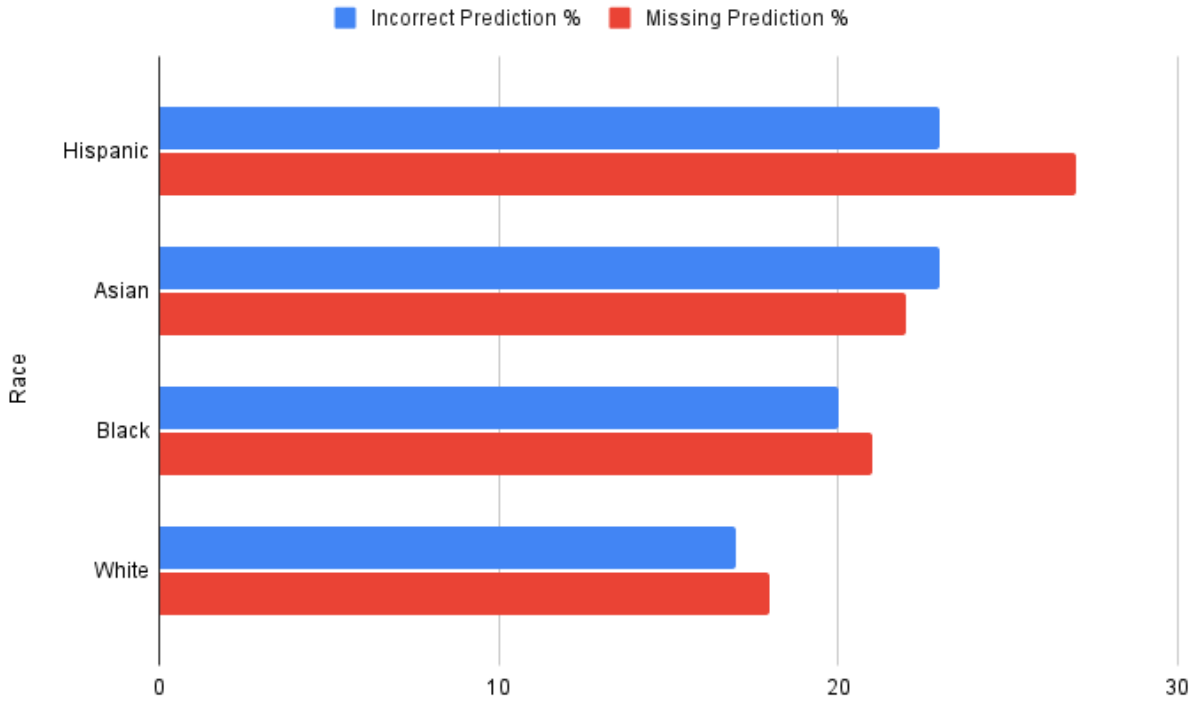

Figure 1: There are effects from both missing predictions and incorrect predictions from Algorithmic Exclusion

Source: (Kaplan et al., 2021)

striking, as shown in Figure 1, is that the inability to actually match these files with data brokers' records was as problematic both in terms of incorrect birth years and ability to merge the files, leading to missing predictions, for some races. This is first-order evidence that both inaccurate and missing predictions can reflect the existing societal divide of race.

3.3 Are there ever positive consequences of algorithmic exclusion?

In general, the examples of algorithmic exclusion this paper has discussed so far have something in common - they all assume that the subject benefits from accurate prediction. This may well be the case for credit reporting, school admissions, hiring and housing background searches. However, there are some instances where perhaps accurate prediction is not always desirable. For example, when we think about government surveillance, or predictive policing

software, it is not always clear that algorithmic exclusion will necessarily be a problem for the individual. If ultimately, an individual's desire is to not be successfully surveilled by the government, then the government lacking data to make predictions about that individual would be desirable. This taste for not being surveilled does not have to be related to a propensity to commit crime - it could merely stem from a classic taste for privacy as defined in the original legal literature (Warren and Brandeis, 1890) which defined privacy as the right to not be intruded upon. It also could be the case that people in certain high-risk groups would benefit from algorithmic exclusion if it enhanced their ability to access health insurance. In this instance the person concerned would benefit, even if it would lead to inefficient pricing of risk. Take the example of Romani and Sinti people who are concerned about biometric surveillance in the EU, given the history of Nazi efforts to use biometric markers to identify and eliminate them. At least historically, the unusualness of Romani and Sinti naming practices and locational instability saved many lives.[3]

Looking forward, the fear of future governments and firms using today's digital information adversarially may lead low-income households happy to remain relatively digitally anonymous. This research is at least somewhat reassuring that while data is fragmented and sparse, and algorithmic exclusion occurs, that this presents at least some type of protection. In general, it useful to dwell on these examples, however, as they show that different prediction domains give rise to different benefits and costs of data for consumers, which in turn affect our assessment of the real welfare consequences of algorithmic exclusion.

4 Potential Policy Approaches for Addressing Algorithmic Exclusion

This paper has emphasized a new policy issue in the form of algorithmic exclusion. It has explained how algorithmic exclusion is different from the more commonly discussed

[3]See `https://edri.org/our-work/roma-rights-and-biometric-mass-surveillance/`

algorithmic bias, and why algorithmic exclusion is likely to reflect a lack of privilege in society. In this section, we discuss policy implications of algorithmic exclusion that stems from missing or fragmented data.

4.1 A rethinking of the privacy debate

Since the advent of the advertising-supported internet, the question of digital privacy has received increasing policy attention (Goldfarb and Tucker, 2012). The focus of this debate has been on the large scale collection and parsing of data for individuals and whether or not it represents unwarranted intrusion into private life. Multiple regulatory efforts have attempted to deal with these concerns, in particular the General Data Protection Regulation and the California Consumer Privacy Act.[4] In some sense these policy efforts have been focused on the question of too broad algorithmic inclusion, rather than the question of algorithmic exclusion which is the focus of this paper. It seems the mirror image issue of data deserts and data fragmentation for low income and historically disadvantaged households, and consequent algorithmic exclusion, has received far less attention. One provocative interpretation is that because overly broad algorithmic inclusion is felt most keenly by those who are richer and more privileged in society, algorithmic exclusion has received less attention. Or even more provocatively, that the current focus of regulators on privacy regulation is more likely to have positive effects mostly for those who are better off.

More generally, the empirical literature in economics on privacy has focused on quantifying the effect on firms and technology adoption of privacy regulation (Goldfarb and Tucker, 2012). This can be seen in the recent empirical privacy literature, which focuses on the consequences of GDPR, such as its effects on firm's measurement (Goldberg et al., 2019), market structure (Johnson and Shriver, 2019), data collection practices (Adjerid and de Matos, 2019) and venture funding (Jia et al., 2018). The difficulty of quantifying consumer-sided effects of privacy regulation has led the consumer-focused literature on privacy to rely largely

[4] `https://gdpr-info.eu/` and `https://oag.ca.gov/privacy/ccpa`

on survey methods (Martin and Murphy, 2017). This paper contributes to this literature by investigating likely effects on consumers of the collection of data surrounding them in terms of its accuracy. This analysis allows insights into the differential effect of privacy regulation by people's background, and suggests that privacy regulation is likely to have the most welfare-enhancing outcomes for those who are more privileged. For those who are less privileged, the questions of data inaccuracy and data sparsity may be more pressing.

4.2 Approaches that are unlikely to be effective

If policy makers were to start to focus on the question of how to tackle algorithmic exclusion, it is important to first emphasize what policy approaches are unlikely to be successful.

One popular policy instrument that has been suggested is the idea of algorithmic transparency. Indeed, the upcoming Digital Markets Act and Digital Services Act in the EU suggest that algorithmic transparency may be a useful policy tool for dealing with hypothesized policy issues resulting from large technology platforms. The basic idea of algorithmic transparency is that if technology firms explain their algorithms directly, or allow policy bodies to inspect their algorithms, then this will help fix algorithmic policy challenges (Pasquale, 2015).

However, such a policy approach does not help address algorithmic exclusion. This is because the problems of algorithmic exclusion stem from a lack of data or fragmented data. Neither of these problems are easily identifiable by looking at algorithmic code. Instead, they can only be identified by understanding the nature of the data that the algorithm uses as an input and the extent to which missing or sparse data itself reflects existing societal issues.

Another approach that has been suggested in the algorithmic policy literature is algorithmic auditing (Shellmann, 2021). In this policy approach, teams of auditors study algorithmic outcomes or outputs for signs of bias (Landers and Behrend, 2022). However, this policy

approach has been focused on understanding the way an algorithm operates. Of the twelve components of an algorithmic audit suggested by Landers and Behrend (2022), only one is focused on input data, and there the focus is on non-representativeness of the data - rather than whether data is scarce or missing. In general, algorithmic audits are set up to understand bias in β - not the absence of X.

Of course, having your data harvested by an algorithm may indeed be convenient for the purveyor of the algorithm, but algorithms themselves may not be designed with the interests of the data subject in mind. In fields such as resume screening, credit scoring and risk assessment, there may be benefits for the subject of having an algorithm make an accurate judgement about you. On the other hand, depending on your view of the state, there may be advantages to the individual if the state cannot render algorithmic judgement. This is especially applicable in the policing and criminal justice context, but can also extend to child and family services and the administrative state more generally (Schenwar and Law, 2020).

In general, this discussion suggests that we need to stop focusing just on outputs and process, and should also consider missing inputs and missing outputs in algorithmic policy. The influence of background characteristics leading to different data availability, which in turn affects the ability to accurately offer data on background information on individuals, should be of particular concern to policy makers. Many of the data brokers supplying marketing platforms with data also provide demographic information for credit decisions, insurance decisions and other types of background checks and risk assessments. Our research provides first empirical evidence that the accuracy of digital profiles depends on who you are - with strong differences between the 'poor' and the 'rich' - so regulators should consider which policies may help leveling the playing field and creating fairness in business decisions. In particular, often it is very hard to obtain detailed information about one's own profile attributes that data brokers have created, making it cumbersome or nearly impossible to correct wrong information about a profile attribute (Miller, 2017).

Table 1
Components of AI System to Be Audited

Component	Questions to ask	Applied to focal example
Components relating to models		
1. Input data	How were input data collected in terms of population and research design? How did these factors affect data quality?	Were the input data collected from job incumbents? How are incumbents different from applicants? Is there range restriction on variables of interest?
2. Model design	What drove initial model choices (e.g., criterion, predictor set, algorithm)? Were they informed by theory or empirically derived? If empirically derived, from what data (and of what quality were those data)?	How is performance defined, and how are we sure of its validity? Why are word choice, answer content, voice patterns, and visuals being included as predictors? Are these decisions theoretically supported?
3. Model development	Once the initial model was created, how was it refined? What approaches were taken, and what likely effect did these approaches have?	Is every decision about every model created during the development process fully documented? When were models discarded and why?
4. Model features	How were the raw input data engineered into model features? Was this process conceptually or empirically driven? What alternative feature engineering approaches were explored?	What specific natural language processing techniques were used? What evidence is there of quality speech-to-text conversion? What facial features emerged from analyzing video? What biases were explored from these engineering processes?
5. Model processes	How does the model use inputs to generate scores? How were alternative approaches explored and evaluated?	What stress tests were conducted, and what types of bias were investigated? Did these tests result in changes to the model, and if so, how and why?
6. Model outputs	How was the quality of predictions generated by the model evaluated, such as for psychometric reliability and validity? How was cross-validation conducted, and was it appropriate given claims about model generalizability?	Are scores consistent over time and upon multiple administrations (i.e., reliability evidence)? Is there evidence that they reflect the predicted constructs they claim to (i.e., validity evidence)? Do they show differences among classes of interest (e.g., race, gender, color, national origin, religion, disability, age) and combinations of classes?
Components relating to information and perceptions		
7. First-party interpretation	Does all messaging from the algorithm developer logically, honestly, and transparently follow from answers developed elsewhere in this audit?	Does the developer claim the model predicts job performance? What evidence in the audit forms the basis of this claim? Is anything exaggerated? Are important details left out?
8. Second-party effects	Who is directly affected by the use of the algorithm, and how have their outcomes and reactions been assessed? What is the relative impact of acting upon false positives versus false negatives on second parties?	How do nonselected applicants react to the news that the algorithm did not assign them a high enough score to be selected? What information is communicated to them, and how do they evaluate that information?
9. Third-party understanding	How have perceptions and evaluation by outside observers been assessed and incorporated? Have outside regulatory groups and community organizations been consulted?	How do experts in employment law view the documentation and performance of the algorithm? How does the public view this use of algorithms?
Meta-components		
10. Cultural context	Has the broader cultural context in which the algorithm will be used been considered? Have members of the community participated in the design of systems that will affect them?	Do power differentials exist between designers, employers, and job candidates? Have cultural assumptions been made? Will development decisions in one culture be applied to another, and if so, how has the development process been adjusted to prevent cross-cultural application challenges?
11. Respect	Is the algorithm being used in a way that conforms to generally accepted ethical standards, such as those in the *Standards*, the SIOP *Principles*, the OECD *Principles* and the UGAI?	What ethical standards do the developers claim to have followed in development? Is there evidence of decisions made following that ethical framework? What evidence is there that individual fairness has been a priority in development?
12. Research designs	How do the research designs (including sampling, experimental design, variable choices, analysis, and interpretation) of any studies conducted to support a claim affect the validity of conclusions?	For every claim that appears to be based upon empirical observation, does the study design support the claims made? Were all design decisions defensible from the perspective of modern methodological research? What impact might they have had on the validity of drawn conclusions?

Note. AI = artificial intelligence; SIOP = Society for Industrial and Organizational Psychology; OECD = Organisation for Economic Co-operation and Development; UGAI = Universal Guidelines for Artificial Intelligence.

Figure 2: Algorithmic Audits are not focused on Algorithmic exclusion
Source: (Shellmann, 2021)

References

Adjerid, I. and M. G. de Matos (2019). Consumer consent and firm targeting after gdpr: The case of a large telecom provider. *Mimeo, Virginia Tech*.

Agrawal, A., J. Gans, and A. Goldfarb (2016). The simple economics of machine intelligence. *Harvard Business Review 17*.

Alaszewski, A. (2006). *Using diaries for social research*. Sage.

Allcott, H., R. Diamond, J.-P. Dubé, J. Handbury, I. Rahkovsky, and M. Schnell (2019). Food deserts and the causes of nutritional inequality. *The Quarterly Journal of Economics 134*(4), 1793–1844.

Bajorek, J. P. (2019). Voice recognition still has significant race and gender biases. *Harvard Business Review 10*.

Bartlett, R., A. Morse, R. Stanton, and N. Wallace (2022). Consumer-lending discrimination in the fintech era. *Journal of Financial Economics 143*(1), 30–56.

Brookman, J., P. Rouge, A. Alva, and C. Yeung (2017). Cross-device tracking: Measurement and disclosures. *Proc. Priv. Enhancing Technol. 2017*(2), 133–148.

Carlson, R. (2021). Cellphone bans in the workplace are legal and more common among blue-collar jobs – they also might be a safety risk. `https://theconversation.com/cellphone-bans-in-the-workplace-are-legal-and-more-common-among-blue-collar-jobs-they-also-might-be-a-safety-risk-173741`. (Accessed on 04/29/2022).

Carrera, F. (2013). By the people, for the people: The crowdsourcing of 'STREETBUMP': An automatic pothole mapping app.

Castro, D. (2014). The rise of data poverty in america – center for data innovation. https://datainnovation.org/2014/09/the-rise-of-data-poverty-in-america/. (Accessed on 01/18/2023).

Chiou, L. and C. Tucker (2020a). In times of crisis, digital businesses need to think about 'technology laggards'. https://sloanreview.mit.edu/article/crisis-digital-business-technology-laggards/. (Accessed on 05/01/2022).

Chiou, L. and C. E. Tucker (2020b). Social distancing, internet access and inequality. *Mimeo, MIT*.

Cowgill, B. and C. Tucker (2017). Algorithmic bias: A counterfactual perspective. *NSF Trustworthy Algorithms*.

Cowgill, B. and C. E. Tucker (2019). Economics, fairness and algorithmic bias. *Working Paper, MIT*.

Crawford, K. (2013). The hidden biases in big data. https://hbr.org/2013/04/the-hidden-biases-in-big-data. (Accessed on 05/01/2022).

DeLuca, S., H. Wood, and P. Rosenblatt (2019). Why poor families move (and where they go): Reactive mobility and residential decisions. *City & Community 18*(2), 556–593.

Farley, R. (1995). Looking for the last percent: The controversy over census undercounts.

Fryer Jr, R. G. and S. D. Levitt (2004). The causes and consequences of distinctively black names. *The Quarterly Journal of Economics 119*(3), 767–805.

Goldberg, S., G. Johnson, and S. Shriver (2019). Regulating Privacy Online: The Early Impact of the GDPR on European Web Traffic & E-Commerce Outcomes. *Mimeo, Northwestern*.

Goldfarb, A. and J. Prince (2008). Internet adoption and usage patterns are different: Implications for the digital divide. *Information Economics and Policy 20*(1), 2–15.

Goldfarb, A. and C. Tucker (2012). Privacy and innovation. *Innovation policy and the economy 12*(1), 65–90.

Goldfarb, A. and C. Tucker (2019). Digital economics. *Journal of Economic Literature 57*(1), 3–43.

Greenstein, S. (2021). Remote work. *IEEE Micro 41*(3), 110–112.

Hoffman, D. L., T. P. Novak, and A. Schlosser (2000). The evolution of the digital divide: How gaps in internet access may impact electronic commerce. *Journal of computer-mediated communication 5*(3), JCMC534.

Jia, J., G. Z. Jin, and L. Wagman (2018). The short-run effects of GDPR on technology venture investment. Mimeo, University of Maryland.

Johnson, G. and S. Shriver (2019). Privacy & market concentration: Intended & unintended consequences of the gdpr. *Available at SSRN*.

Kaplan, L., A. Mislove, and P. Sapiezynski (2021). Measuring biases in a data broker's coverage. *AMC IMC conference proceedings*.

Keller, J. (1995). Public access issues: An introduction. In *Public access to the Internet*, pp. 34–45. MIT Press.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics 133*(1), 237–293.

Lambrecht, A. and C. Tucker (2019). Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science*.

Lambrecht, A. and C. E. Tucker (2020). Apparent algorithmic discrimination and real-time algorithmic learning in digital search advertising. *Available at SSRN 3570076*.

Landers, R. N. and T. S. Behrend (2022). Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes ai predictive models. *American Psychologist*.

Langston, J. (2020). Shrinking the 'data desert': Inside efforts to make AI systems more inclusive of people with disabilities - the AI blog. `https://blogs.microsoft.com/ai/shrinking-the-data-desert/`. (Accessed on 05/01/2022).

Markman, K. M. and C. R. Scott (2005). Anonymous internet? Examining identity issues in email addresses. In *Annual meeting of the International Communication Association, New York, NY*.

Martin, K. D. and P. E. Murphy (2017). The role of data privacy in marketing. *Journal of the Academy of Marketing Science 45*(2), 135–155.

McDonald, A., C. Sugatan, T. Guberek, and F. Schaub (2021). The annoying, the disturbing, and the weird: Challenges with phone numbers as identifiers and phone number recycling. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–14.

Miller, C. R. (2017). I Bought a Report on Everything That's Known About Me Online.

Neumann, N. and C. Tucker (2021). Data deserts and black box bias: The impact of socio-economic status on consumer profiling. *Mimeo, MIT*.

Neumann, N., C. E. Tucker, and T. Whitfield (2019). Frontiers: How effective is third-party consumer profiling? evidence from field studies. *Marketing Science 38*(6), 918–926.

Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science 366*(6464), 447–453.

O'Leary, D. E. (2013). Exploiting big data from mobile device sensor-based apps: Challenges and benefits. *MIS Quarterly Executive 12*(4).

O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Pasquale, F. (2015). The black box society. In *The Black Box Society*. Harvard University Press.

Schenwar, M. and V. Law (2020). *Prison by any other name: The harmful consequences of popular reforms*. The New Press.

Schwartz, E. M., E. T. Bradlow, and P. S. Fader (2017). Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science 36*(4), 500–522.

Servon, L. J. (2008). *Bridging the digital divide: Technology, community and public policy*. John Wiley & Sons.

Shellmann, H. (2021). Auditors are testing hiring algorithms for bias, but big questions remain. `https://www.technologyreview.com/2021/02/11/1017955/auditors-testing-ai-hiring-algorithms-bias-big-questions-remain/?utm_medium=search&utm_source=google&utm_campaign=BL-ACQ-DYN&utm_content=categories&gclid=Cj0KCQjw37iTBhCWARIsACBt1Izc1xWN-sa_zEp0FDdRv9NklApcv4rtJpA1_Ee5Sd1akbm_IWYL2iwaAmbKEALw_wcB`. (Accessed on 05/01/2022).

Sweeney, L. (2013). Discrimination in online ad delivery. *ACMQueue 11*(3), 10.

Warren, S. D. and L. D. Brandeis (1890, December). The right to privacy. *Harward Law Review 4*(5), 193–220.

The Center on Regulation and Markets at Brookings provides independent, non-partisan research on regulatory policy, applied broadly across microeconomic fields. It creates and promotes independent economic scholarship to inform regulatory policymaking, the regulatory process, and the efficient and equitable functioning of economic markets.

Questions about the research? Email communications@brookings.edu.
Be sure to include the title of this paper in your inquiry.