Philip D. Waggoner*, Ryan Kennedy, Hayden Le
and Myriam Shiran

Big Data and Trust in Public Policy Automation

<https://doi.org/10.1515/spp-2019-0005>

Abstract: Big data is everywhere, both in and out of public policy. Though a rich data source, what is the impact of big data beyond the research community? We suggest such that invoking big data-related terms acts as a heuristic for assumed algorithmic quality. Such an assumption leads to greater trust in automation in public policy decision-making. We test this “big-data-effect” expectation using four tests including a conjoint experiment embedded in a recently fielded survey experiment. We find strong evidence that indeed, big data-related terms act as powerful signals of assumed quality where respondents consistently prefer algorithms with bigger data behind them, absent any mention of predictive accuracy or definitions of key terms (e.g. “training features”). As we expect this big-data-effect is likely not beholden to public policy, we encourage more research in this vein to deepen an understanding of the influence of big data on modern society.

1 Introduction

Recently there has been a marked increase in studies that use and cite “big data” across a variety of disciplines, including the social sciences (di Bella et al. 2018; Montgomery and Olivella 2018), organizational research (Tonidandel et al. 2018), business (Chiang et al. 2018), medicine (Weintraub et al. 2018), psychology (Chen and Wojcik 2016), and engineering (Sotiropoulos 2019). The surge of interest in big data appears to be rooted in the desire to leverage the millions of terabytes of data generated around the world to understand more substantive phenomena. This trend is also seen beyond academia where public policy researchers and practitioners, for example, are citing big data sources that inform policy creation (Data and Hub 2018). While these trends are occurring simultaneously across

*Corresponding author: Philip D. Waggoner, University of Chicago, Chicago, IL, USA,
e-mail: pdwaggoner@uchicago.edu. <https://orcid.org/0000-0002-7825-7573>

Ryan Kennedy and Myriam Shiran: University of Houston, Houston, TX, USA

Hayden Le: University of Michigan, Ann Arbor, MI, USA

a variety of fields that tend to avoid each other, Grimmer (2015) points out that not only is the divide between the social, natural, and computational scientific worlds rapidly narrowing, but each field can actually serve and support the other. In short, big data is becoming increasingly a part of the modern analytical ecosystem, both in academia and policymaking, across virtually every discipline.

To date, the focus in most big data studies has been on the many possibilities big data affords. For example, with more data, economists can forecast more efficiently (Hassani and Silva 2015). Or similarly, big data can inform better health-care decisions at lower costs (Raghupathi and Raghupathi 2014). And, in public policy, millions of police traffic stop records inform a deeper understanding of racial bias in policing (Baumgartner et al. 2018). As such, the rise in studies using big data has mostly focused on the value-added from big data as a resource.

With this increased focus on big data, though, comes a second-order question, which focuses not on the size of the data, but rather on the heuristics associated with the term. While big data may inform higher quality research by enlarging the sample size, perhaps scholars have overlooked the possibility of people “trusting” in big data, merely because it is “big.” Bigger must be better, because it is bigger.

An area where such a heuristic trend is clearly seen is in the assessment of algorithms that support public policy decision-making. For example, in criminal justice, scholars have recently uncovered evidence that features of criminal sentencing algorithms may be inflated to engender greater trust, regardless of a lack of improvement in quality and predictability (Dressel and Farid 2018).¹ Dressel and Farid (2018) found that algorithms with only two training features perform as well as the popular COMPAS algorithm with 137 training features. Though the public may not have sufficient information to understand the nuance of creating algorithms or even understand the concept of “training features,” people can generally understand that 137 is bigger than two. And when it comes to criminal sentencing, bigger algorithms must be better, as implied by this case study.

The COMPAS algorithm example is important, because it points to a connection between general public trust in algorithms and public policy decision-making, where, in order for algorithms to flourish in a public sphere, people need to trust them. This is the case especially in a representative democracy, such as America, where decision-makers are accountable to the public, both formally (elections) and informally (public opinion polls). This accountability structure is a powerful conditioning factor in a variety of political and policy decision-making (Cheibub and Przeworski 1999). Regarding “hybrid” decision-making between

¹ But see Holsinger et al. (2018) for criminal justice practitioners’ rebuttal to Dressel and Farid (2018).

algorithms and humans, such trust and accountability is no less potent (Kennedy et al. 2018a). Thus, from the COMPAS case, whether to beneficial or harmful ends, trust in the algorithm tends to be tied to the size of the data used to create and train the algorithm, regardless of the predictive quality or transparency of design, as the COMPAS software is proprietary.

In this paper, then, we are interested in exploring this second-order question, which is centered around that which people think matters in algorithmic design as it relates to public policy decision-making. Though people may or may not be good at evaluating algorithmic efficiency or internal validation, for example, the “big data” heuristic likely bucks this trend in that it provides a signal of complexity and thus, assumed quality. As such, we focus on uncovering the features that people tend to care about as they assess algorithmic credibility. We suspect people take algorithms more seriously when the algorithms claim to use big data or simply cite big data characteristics in the design (e.g. citing a large number of features used to train the algorithm). In so doing, to our knowledge, we are offering the first quantification of this “big-data-effect,” where big data equals quality in peoples’ minds, regardless of their ability to fully understand and digest algorithmic processes and functions. As such, we are interested only in whether such a big-data-effect exists, saying nothing of the normative implications of such a finding (e.g. whether this effect is “good” or “bad”).

We proceed as follows. First, we motivate our expectations surrounding this “big-data-effect” by briefly reviewing relevant work at this intersection of big data, algorithms, and the notion of trust in automation. Then, we introduce our empirical strategy which includes four approaches to assessing this effect based on a recently fielded original survey experiment: natural language processing, descriptive ranking of the “most important features” that should go into an algorithm, a conjoint experiment simultaneously evaluating preferences across a number of design features, and finally process-tracing, or “click analysis” that illuminates what people tend to prioritize as manifested in their interaction with the survey instrument. Ultimately, we find strong support across most tests for the big-data-effect, where respondents consistently cite the size of the data and number of training features as most important in algorithmic design. Further, we find that respondents tend to strongly favor algorithmic and code transparency, and also the inclusion of humans at the point of the public policy decision.

2 Big Data and Trust in Public Policy Automation

As noted in the introduction, practitioners and scholars from a range of fields have been drawn to big data’s potential. Given the tremendous amount of data

collected by various institutions at any given point and technological advancements allowing for efficient storage, organization, and analysis of these data, there seems to be no end to the potential of big data (Boyd and Crawford 2011), especially in matters of public policy. For example, Google and the Centers for Disease Control and Prevention (CDC) were able to predict the spread of the winter flu in the US by leveraging location and search term data from users (Ginsberg et al. 2008). Further, New York City's analytics team improved its efficiency in responding to nuisance complaints by over a factor of five after creating and analyzing a dataset that compiled data from 19 different government agencies (Fox 2013). In these instances, big data empowered analysts to tackle policy problems with which they may have otherwise struggled. However, a dataset's quantity does not translate to any guarantee about its quality (Boyd and Crawford 2011). For example, though Twitter data is often used to evaluate public sentiment, 40% of active users sign in, not to participate, but rather to observe (Twitter 2011). About 10% of users create 80% of the content, and Twitter users are younger and more likely to be Democrats than the general public (Wojcik and Hughes 2019). This does not even count the estimated 48 million Twitter users that might not be human, but rather "bots" designed to post prolifically to achieve commercial or political ends. As such, though beneficial for some applications, simply using and citing big data does not automatically equate to higher quality.

Proponents of automating certain public policy-related processes, such as automatic facial recognition of wanted criminals via video surveillance, argue that algorithmic techniques would not suffer from human biases such as learned discrimination. However, if given current data to train, these algorithms would necessarily inherit biases persistent in current society (Barocas and Selbst 2016; Hacker and Petkova 2017; Dressel and Farid 2018). For example, if an algorithm is predicting the probability of a future arrest using past arrest records, and the criminal justice system exhibits racial bias in who gets arrested, the racial bias can be reflected in the algorithm's output, even when race is not used as an explicit input (Dressel and Farid 2018). Similarly, facial recognition algorithms trained largely on White and Asian students have been found to do a poor job at differentiating Black faces, and algorithms to target advertising have been found to systematically exclude some demographic groups (Rahwan et al. 2019). Thus, the distinction between "biased humans" and "unbiased algorithms" is not nearly as clear as often assumed.

While some work has explored factors related to the presentation of an algorithm such as how "anthropomorphic" it may be (Forster et al. 2017), how easy the algorithm is to use, or how transparent it is (Hoff and Bashir 2015), political institutions will never be able to effectively or fully adopt these automated

techniques in the absence of public trust. As society evolves in the age of big data and as decision-makers adopt these automated algorithms as supportive aids, there remains an associated need to understand the factors that influence individuals' trust in automation (Kennedy et al. 2018a).

Further supporting the motivation to understand the contours of trust in algorithms as decision-makers, previous work has found a strong bias toward algorithms. This bias is far reaching from trusting search engines even when information is already known (Wegner and Ward 2013) to relying on algorithms to solve logic problems (Dijkstra et al. 1998), even when the algorithm is known to have made an error (Dijkstra 1999). This all points to a general algorithm bias (Logg 2017). Building on these findings in the literature, we suggest that a “big-data-effect,” which should engender greater trust in algorithms, may be reasonable to expect.

In sum, there has recently been much discussion on the revolutionary nature of big data, especially in public policy. Further, there are an increasing number of projects addressing and leveraging big data. Yet, to date, there remains a lack of understanding of whether and how the stated use of big data in an algorithm (and features related to it) affects general levels of trust at the mass level. With this paper, we explore this expected “big-data-effect,” where bigger features such as the size of training data, are associated with higher quality, regardless of predictive accuracy. Do people embrace big data to the same degree as scholars and practitioners? In other words, is big data something that affects the mass public's levels of trust and credibility in a similar way as scholars? These ideas, which link trust and big data in a public policy context, allow for gauging the generalized levels of trust and broader effects of big data in society.

3 Empirical Strategy

We are interested in addressing the linkage between the mass public and that which influences trust in algorithms in an effort to understand the breadth of big data's role in society, beyond academic and governmental research. In short, we are interested in quantifying the “big-data-effect,” which we expect is a heuristic that engenders greater overall trust in algorithms, regardless of levels of understanding on the part of the respondent. To do so, we recently fielded an original experiment, which allows for the exploration and unpacking of the contours of big data and mass trust in public policy automation.

We embedded four different tests and measures in our experiment to corroborate findings, ensure consistency, and strengthen external generalizability:

(1) natural language processing of an open-ended survey question; (2) an explicit rank ordering question where respondents selected preferences of the “most important” features to be included in the design of a public policy algorithm; (3) a conjoint experiment allowing us to experimentally compare the explicitly stated choices and tradeoffs made by the respondent; and (4) process-tracing, where interaction by users with the survey allows us to analyze the precise expression and timing of respondents’ interactions with the survey tool. Process-tracing or “click analysis” is yet another way of assessing that which respondents consider to be most important, and thus worth their time.

4 Study Design, Data, and Sample

To explore the expected big data effect, we recently fielded a survey with an embedded experiment using a recruited sample of adult respondents utilizing Amazon’s Mechanical Turk (MTurk) (Clifford et al. 2015), who were paid \$1.00 for successful completion of the survey. The study ($n=661$) was in the field from January 16, 2019 to January 28, 2019.²

Upon answering consent questions, we asked respondents a few basic questions, which included verifying that they were operating in the US, as well as an open-ended question asking respondents to define an algorithm: “In your own words, how would you define an algorithm?”. This question gets at the big-data-effect in a different way by assessing the degrees to which they are familiar with algorithms, as well as assessing consistency in important features that should be included in algorithmic designs from their perspectives. While it is not an attempt to directly measure or capture the big-data-effect, the value of this part of the analysis is that it provides a window into the importance placed on certain algorithmic design features in the minds of respondents.

Next, we asked all respondents to rank algorithmic features by importance using a drag-and-drop user interface. This allows the respondents to explicitly rank order their preferences, and thus allows us another opportunity to observe that which respondents consider to be important design features, which is another way to assess the expected big-data-effect. The value of such an explicit measure of design preferences is in its pairing with the conjoint experimental design,

² Though originally recruiting 878 total respondents, several were eventually dropped either due to not completing the survey, or were screened out given their use of a virtual private server (VPS) meant to disguise their location. To screen out such problematic respondents, we use the recently released software, “rIP”, which scores and locates the use of VPS (Kennedy et al. 2018b; Waggoner et al. 2019).

where the ranked preferences balance the experimental findings to deepen an understanding of that which respondents prioritize most in algorithmic design. If patterns are similar across both experimental (conjoint) and non-experimental (rank-ordered preferences) measures, then this would strongly point to the presence of a substantive big-data-effect, rather than random noise coming from a single measure or test.

Then regarding the conjoint experiment, respondents were given a pair of possible algorithmic designs, which mirror those that would be used by a judge in determining criminal sentencing, and then they were asked to select their preferred design. Respondents were shown six algorithmic design features, the order of which varied randomly across respondents: (1) a human role in the algorithm design; (2) the location from which the data for the algorithm was collected; (3) the number of defendant characteristics (or “factors”); (4) the size of training data; (5) the source of algorithm designer; and (6) the transparency of code that went into the algorithm design. This design is called a conjoint experiment, which has its roots in marketing studies (Green and Srinivasan 1990), and has recently increased in prominence in political science research (Hainmueller et al. 2015). Our conjoint experiment allowed for simultaneous comparison of hypothetical algorithmic designs, thereby offering us the ability to empirically assess the tradeoff in algorithm design features that either increase or decrease the algorithm’s favorability. Effect sizes by the level of each design feature are displayed as marginal means and average marginal component effects, which are discussed and presented in Figures 3 and 4.

Importantly, we did not include a battery of post-treatment demographic questions, given the recent findings that such post-treatment controls can bias the main effects (Montgomery et al. 2018). Further, controls are not needed in an experimental context, where any differences between subjects in their responses should be a function of the treatment, with any confounding factors sufficiently accounted for in randomization.

After the definition, feature ranking, and conjoint experiment, we asked the respondents to compare three algorithmic designs and then asked whether they would trust the algorithm to make criminal justice decisions in their states. Specifically, the task description read, “For the next part of the survey, we are going to have you look through three different descriptions of algorithms like those you have been evaluating so far, and we will ask you whether you would support the use of that algorithm in making decisions about criminal sentencing in your home state.” During this task, we tracked the timing and interaction with the page by the respondent using the MouselabWEB tool developed by (Willemssen and Johnson 2009). This allows us a more indirect look at that which respondents consider to be the most important features to include in algorithms in the

explicit context of criminal sentencing. The results are discussed and presented in Figures 5 and 6.

5 Study Results

We begin by analyzing responses to the open-ended question to provide an informal look at that which respondents think matters most in algorithm design, as seen in respondents' unfiltered definitions of algorithms. Then, we present the rank ordered algorithmic design features, which offers an explicit look at respondents' explicit algorithmic design preferences. Third, we analyze the conjoint experiment by presenting and discussing the results in two ways: marginal mean values and average marginal component effects, fit using a generalized linear model. Finally, we analyze and discuss the Mouselab process-tracing.

5.1 N-grams for Open-Ended Question

To analyze the open-ended question,³ we leveraged n-grams, which is a widely used natural language processing (NLP) technique allowing for plotting densities of combinations of words used together. This question gave respondents space to respond however they wished, with no length limitations. We then preprocessed the responses (e.g. stripping whitespace, removing extraneous characters, removing stopwords, etc.), and used visual densities of n-grams to assess patterns in respondents' responses. Specifically, we calculated the frequency of the most used two-word (bi-grams) and three-word (tri-grams) sequences. This offers an informal look at that which respondent consider the most important aspects of algorithms. The n-gram plots are presented in Figure 1.⁴

In Figure 1 we see the most commonly used pairing of words (seen in the upper bi-gram plot) is “set/rules,” “computer/program,” and “solve/problem.” This suggests that respondents consider algorithms to be computer-based problem solvers that follow some “set of rules.” The tri-gram in the lower panel

³ “In your own words, how would you define an algorithm?”.

⁴ The preprocessing of text documents, including open-ended survey responses, often requires removal of “stopwords,” which are typically articles and filler words, such as *a*, *and*, and *the*. The n-grams plot shown in Figure 1 is with stopwords removed. To demonstrate the value of preprocessing messy text data, we present the n-grams plots in the Appendix with stopwords included, in Figure 7.

Figure 1: Bi- and tri-grams plots for open-ended question.

provides slightly more clarity on these terms, with the most frequently used trios being “set/rules/followed,” “set/rules/used,” and “rules/followed/calculations.” This is more revealing of respondents’ levels of understanding of algorithms, where a common theme is that algorithms follow some set of rules to perform calculations. Interestingly, the term “computer” was used in only one of the tri-grams, which was used only five times. This ultimately suggests respondents may have a general sense that algorithms help make calculations based on some set of rules, however defined. Yet, interestingly, there remains no consistent theme of algorithmic performance, training data, or even the size of data. This look at respondents’ levels of understanding of algorithms offers a window into that which they think is most important in an algorithm on the surface (hence it is included in their definitions). Yet, whether such definitions have any bearing on the level of trust placed in algorithmic designs is a question we unpack in the subsequent sections, starting with respondents’ explicit rank-ordered algorithmic design preferences.

5.2 Rank Ordered Algorithm Design Preferences

Next, we asked respondents to rank order the features of algorithms that they considered to be “most important” in assessing its quality, based on a list of six provided design features mentioned in the previous “Study Design” section. Respondents were able to interact more closely with these features using a drag-and-drop interface. This approach, which includes *no experimental manipulation*, offers respondents an opportunity to explicitly order that which they considered most important in defining and assessing the quality of an algorithm, which is particularly revealing in light of the previous results in Figure 1 suggesting respondents likely have little direct experience interacting with algorithms. These results, which are displayed in Figure 2, complement the subsequent conjoint experiment, whereby we will be able to directly compare *though experimental manipulation* that which respondents consider to be of greatest importance in designing a “high-quality” algorithm.

In Figure 2 we can clearly see that the most important feature contributing to higher quality algorithms is the size of the training data, which offers a first hint of the likely presence of a big-data-effect. This is seen in the upper left plot, where over 200 respondents (a clear majority across all other algorithmic features) selected training data size as most important. Similarly, the second most important feature of an algorithm is yet another big data heuristic, which is the number of input features. This pattern is in line with the findings from (Dressel and Farid 2018), suggesting the sheer number of input features (i.e. defendant characteristics in a criminal sentencing algorithm) acts as a signal for higher algorithmic quality. The respondents in our survey seem to agree in this ranking, which is behind only the overall size of the training data. Interestingly, respondents seemed mixed on the importance of transparency, as well as the inclusion of a human in the design, as seen in the lack of unanimous ranking of these features.

5.3 Conjoint Experiment

Next, we come to the focus of our study, which is the conjoint experiment. Building on the previous non-experimental rankings of prioritized algorithmic design features, our goal in this phase was to assess respondents’ preferences about algorithms through the experimental manipulation of comparing two randomly assigned algorithm designs. In all designs we randomly varied the level of the feature across the six main features, as discussed above: (1) a human role in the algorithm design; (2) the location from which the data for the algorithm was collected; (3) the number of defendant characteristics (or “factors”); (4) the size of

Figure 2: Respondent ranked preferences of algorithmic design features.

training data; (5) the source of algorithm designer; and (6) the transparency of code that went into the algorithm design. For example, the levels of the “size of training data” feature were: “1000 defendant records,” “10,000 defendant records,” “50,000 defendant records,” “100,000 defendant records,” “500,000 defendant records,” and “1,000,000 defendant records.”

We display the effects of the conjoint design in two ways: marginal mean values (non-parametric and descriptive) and the average marginal component effects (AMCE). To get the values, the preferred choice was regressed on each feature variable. These results are shown in Figures 3 and 4, with different shades of gray corresponding to each variable, and individual points reflecting levels of

Figure 3: Marginal mean values.

Figure 4: Average marginal component effects.

each variable.⁵ Regarding interpretation, marginal mean values reflect the mean values of preferences for a given design feature, holding all other feature levels constant at their mean values. When the choice is forced such as in our case (i.e. “forced conjoint designs”), the marginal mean value is normalized at 0.5, allowing marginal means for individual feature levels greater than 0.5 to reflect positive algorithmic design favorability, and values less than 0.5 reflect negative design favorability for each level of the design feature. For ease of interpretation, we place a cut point at the value of 0.5 in Figure 3. Values to the right of the cut point in Figure 3 suggest respondents view the given level of the design feature as positively contributing to the overall algorithmic design, and values to the left of the cut point suggest the given feature level is negatively contributing to the design of the algorithm. Then, for interpreting the AMCEs in Figure 4, values are in relation to a baseline feature (e.g. baseline = “2 factors” for the “number of defendant characteristics” feature). Feature values greater than 0.0 suggest that, relative to the baseline feature, the given feature increased the likelihood of selecting the overall algorithmic design profile, and values less than 0.0 decrease the likelihood of design selection. Similarly, we place a cut point at 0.0 for reference and ease of interpretation.

First, regarding the marginal mean values in Figure 3, note that the two big data-related design features, “number of defendant characteristics” and “size of training data” are all positively skewed, with only the highest values for each crossing the 0.5 threshold discussed above. This suggests that algorithms with the most features (137, like the COMPAS algorithm) and the largest training data sizes (500,000 and 1,000,000) are the strongest predictors of algorithmic design profile favorability across all other possible features in the algorithm’s design. This is a strong point of evidence suggesting, indeed, the big-data-effect is a prominent conditioning force in assessing the assumed quality and trustworthiness of algorithms.

Next, regarding the AMCEs in Figure 4, there are similar supporting patterns, where the “training data size” and “number of defendant features” that go into the algorithm design exert the strongest effects on the likelihood of algorithmic design selection in this forced context. Here again, this suggests that the big-data-effect is influencing the degrees to which respondents trust, and thus favor algorithms with big data features in comparison to all other features, including transparency, humans in the design loop, data location, and algorithm source.

⁵ We validate these findings in the Appendix by displaying the MM and AMCEs by age with two cohorts split by median age value from the experiment in Figures 8 and 9. We found no substantive differences across age suggesting the results are robust across generational cohorts.

Respondents consistently prefer the big data versions of all randomly assigned algorithmic profiles, and thus responding in expected directions to the big data heuristics, thereby quantifying the big-data-effect.

5.4 Process-Tracing

Finally, we relied on process-tracing to lend a final measure of the salience of algorithm characteristics. One of the concerns about self-reported measures of characteristic importance, such as the explicit algorithmic design features above in Figure 2, is that respondents may respond by giving answers based on what they think should be important. Process-tracing bypasses this threat by tracking the time and patterns of interaction with the survey instrument to offer a back-door look at that which respondents care most about, as seen through their physical behavior. Moreover, process-tracing gives us a slightly different measure of trust from the conjoint experiments, focusing on respondents' interest in design characteristics, as opposed to their preferences when confronting trade-offs.

To conduct the process-tracing, or "click analysis," we utilized the MouselabWEB software (Willemsen and Johnson 2009). Respondents were shown a screen laying out what each algorithm attribute means and were asked to read it carefully. All attributes and their values remained consistent from the earlier conjoint experimental design. Respondents were then given an ID and asked to link to a MouselabWEB generate website. Once they reach the site and sign in with their ID, they were shown a screen like that in Figure 5. Each box listed the name of the algorithm attribute described in the earlier screen. Only by clicking on the box (or tapping in the case of a mobile device) were respondents able to see the value of that attribute for the algorithm (in Figure 5, for example, the "Transparency of algorithm" box is clicked). Respondents had 15 seconds to click through the features of the algorithm before they would be unable to view any more information, and then they were asked to choose whether they would support the use of that algorithm in their state. On average, respondents took about 11.5 seconds to review the characteristics and choose a response. The order of attributes on the screen varied randomly by respondent, and each respondent was given three algorithms to evaluate, with the first being a trial run for them to get used to how the system worked.

Because the system required respondents to link to an outside utility, we experienced substantial drop-off in participation in this part of the survey. In total, we had 231 respondents who successfully completed this task (linked, entered their ID, and clicked on an attribute), producing 432 completed scenario

Within each box is a characteristic of the proposed algorithm. Click on the box to reveal that characteristic.

Location for which data for algorithm is gathered	Details of algorithm proprietary	Human role in final score
Size of data used for algorithm	Number of defendant characteristics	Source of algorithm

Timer: 6 sec

Would you support the use of this algorithm to assist in sentencing decisions in your state?

☐ strongly oppose
☐ somewhat oppose
☐ neither support nor oppose
☐ somewhat support
☐ strongly support

Next Page

Figure 5: Example of click tracking experiment

evaluations, in comparison with the 661 respondents who completed the other parts of the survey.

We tracked three measures of the importance of the algorithm attribute as revealed by the process-tracing. First, we looked at which feature of the algorithm was clicked on first. The expectation is that, in a time-constrained environment, users will try to get information on the feature they consider the most important first. The second looks at the number of times that a feature was clicked on more than once, with the expectation that more important features will be viewed multiple times by users. The third analyzes at the amount of time the user left the box visible, with the intuition that respondents will spend more time reading and reviewing the information in boxes they consider more important.

The results are shown in Figure 6. For the attribute clicked first, there is not a great amount of variation, with users seemingly clicking at random or in the order they appeared on the screen. Notably, the only feature that stands out is the role of humans in determining the final score, which received about 10 more first clicks than the others. Those who clicked on this first also tended to rank the importance of a human contributing to the final score higher than those who did not ($r=0.14$). Second, we noticed no real pattern in the features that were clicked more than once by users, with most attributes receiving about the same amount of multiple clicks. Finally, in terms of time the boxes were open for the respondent to view the information, while we again do not see major differences, here the “size of the training data” had the most time open among these respondents,

Figure 6: Plots of results from process-tracing click experiment.

followed by the transparency of the algorithm, the human role in the final score, and the number of defendant characteristics.

At first glance, most of the respondents appear to have selected information to view randomly or in the order of appearance on the screen. The fact that they did so is, however, nonetheless revealing. While our earlier results suggest relatively strong preferences for algorithms developed using “big data,” or the big-data-effect, these process-tracing results suggest that, unlike how we think of “single issue” voters in the political context, the preferences exhibited might be a reflection of the information environment, as opposed to a strongly ingrained decision-rule held by members of the public. In other words, while some attributes are more highly valued in making choices like in the conjoint experiment, they are not valued to the exclusion of other information. The possible exception to this from these results is that maintaining a human in the design loop, which also had a substantial effect in the conjoint experiment and may be more salient for its supporters. Future research should probe the human/algorithm tradeoff in algorithmic design and implementation, especially as it relates to levels of mass trust in algorithms.

6 Discussion and Conclusion

In this paper, we were interested in exploring that which engenders trust in automation and algorithmic advice in a public policy setting, and specifically in quantifying what we expect to be a big-data-effect, where big data-related terms should act as signals of assumed algorithmic quality. Exploring this in the context of criminal sentencing algorithms, the use of which are on the rise, we were particularly concerned with the big data aspect of this nexus. Specifically, we expected that when people are asked to assess algorithms, regardless of their individual levels of understanding of the complexity that goes into algorithmic design, respondents tend to rely on heuristics of quality, which are most clearly and simply related to big data. Across a variety of measures, tests, and an experimental manipulation, we found strong evidence pointing to the conditioning role of big data in engendering greater trust in the use and design of algorithms as well as their level of quality in public policy decision-making. In the process-tracing, however, it was revealed that, while these preferences are strong, respondents did not maintain the heuristic so strongly that they ignored other information.

Importantly, the value of our approach is that each of the four tests we used were all tapping the effect of big data on trust and algorithmic approval at the mass level, but each in distinct ways. There are subtly different aspects of trust being explored in our study, which are essentially different ways in which algorithms are evaluated. These are captured in all of our measures: trust within a simple definition of an algorithm, trust within preference formation about algorithms themselves, trust within the type of information that would be important to highlight, and also trust in direct interface with algorithmic-related content. By evaluating variance across each of these four tests, we have offered a profile of trust in algorithms, as conditioned by big data. Such a profile offers a unique, new window into the general public's mind on the effectiveness and uses of big data and algorithms in the context of public policy.

Though beholden to the public policy realm, implications of our study likely extend to other fields with studies at the nexus of big data and trust in automation, especially those in which respondents may be more likely affected by the outcome of the decision. For example, in medical studies interested in whether patients will agree to experimental treatments that involve automated tests, researchers should account for the use of big data-related heuristics in attempts to “sell” patients on proposed treatments. Or in developing autonomous weapons systems, engineers who highlight the use of big data techniques would likely be more successful at convincing funders and lawmakers of the quality of the technology.

Another limitation of our study is that it was conducted with only American respondents, such that inferences are beholden to the US context. As such, we suspect that further research conducted in non-US contexts would positively extend our analysis. Such studies would not only deepen the findings presented here, but also allow for broader generalizations to be made about human behavior beyond the US. In sum, the big-data-effect we have uncovered in this study is likely present across a variety of domains and other countries. Thus, given these limitations, we encourage future research at this intersection to continue to deepen an understanding of the contours surrounding the burgeoning presence and influence of big data in society.

Finally, our experiments cannot draw normative conclusions about whether this “big-data-effect” is good or bad – there simply is no “correct” or “incorrect” answer to the experiments structured above. As with any heuristic, this is likely to be useful in some cases and problematic in others – the key is to be aware of the heuristic and when it is likely to cause problems. The increase in the scale and resolution of data available is important for a range of reasons (Lazer and Radford 2017), but, at the same time, research suggests that “big data hubris” has led some scholars to view big data as allowing them to ignore issues like the validity and reliability of their data (Lazer et al. 2014). Our research suggests that these attitudes may also be affecting the general public. There is also plenty of evidence that big data is not a substitute for theory and careful measurement (Hosni and Vulpiani 2018). Yet, the particular situations in which this heuristic is good or bad for any number of downstream questions is a task with which we must leave for future research.

Acknowledgements: This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), funder id: <http://dx.doi.org/10.13039/100011039>, via 2017-17061500008. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

Appendix

Figure 7: N-grams plot with stopwords included.

Figure 8: Marginal mean values by age.

Figure 9: Average marginal component effects by age.

References

- Barocas, Solon and Andrew D. Selbst (2016) "Big Data's Disparate Impact," *California Law Review*, 104:671.
- Baumgartner, Frank R., Derek A. Epp and Kelsey Shoub (2018) *Suspect Citizens: What 20 Million Traffic Stops Tell us about Policing and Race*. New York: Cambridge University Press.
- Boyd, Danah and Kate Crawford (2011) "Six Provocations for Big Data," *Computer* (Long. Beach. Calif), 21:123.
- Cheibub, José Antonio, and Adam Przeworski (1999) "Democracy, Elections, and Accountability for Economic Outcomes," *Democracy, Accountability, and Representation*, 2:222–250.
- Chen, Eric Evan, and Sean P. Wojcik (2016) "A Practical Guide to Big Data Research in Psychology," *Psychological Methods*, 21(4):458–474.
- Chiang, Roger H. L., Varun Grover, Ting-Peng Liang and Dongsong Zhang (2018) "Strategic Value of Big Data and Business Analytics," *Journal of Management Information Systems*, 35(2):383–387.
- Clifford, Scott, Ryan M. Jewell and Philip D. Waggoner (2015) "Are Samples Drawn from Mechanical Turk Valid for Research on Political Ideology?" *Research & Politics*, 2(4):1–9.
- Data, IBM: Big and Analytics Hub (2018) "How Twitter Data Helps Shape Public Policy." <https://www.ibmbigdatahub.com/infographic/how-twitter-data-helps-shape-public-policy>.
- di Bella, Enrico, Lucia Leporatti and Filomena Maggino (2018) "Big Data and Social Indicators: Actual Trends and New Perspectives," *Social Indicators Research*, 135(3):869–878.

- Dijkstra, Jaap J. (1999) "User Agreement with Incorrect Expert System Advice," *Behaviour & Information Technology*, 18(6):399–411.
- Dijkstra, Jaap J., Wim B. G. Liebrand and Ellen Timminga (1998) "Persuasiveness of Expert Systems," *Behaviour & Information Technology*, 17(3):155–163.
- Dressel, Julia and Hany Farid (2018) "The Accuracy, Fairness, and Limits of Predicting Recidivism," *Science Advances*, 4(1):55–80.
- Forster, Y., F. Naujoks and A. Neukum (2017) "Increasing Anthropomorphism and Trust in Automated Driving Functions by Adding Speech Output," In: *2017 IEEE Intelligent Vehicles Symposium (IV)*, pp. 365–372.
- Fox, Mark S. (2013) "City Data: Big, Open and Linked," *Municipal Interfaces*, 19–25. https://www.researchgate.net/publication/262674890_City_Data_Big_Open_and_Linked.
- Ginsberg, Jeremy, Matthew Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark Smolinski and Larry Brilliant (2008) "Detecting Influenza Epidemics Using Search Engine Query Data," *Nature*, 457:1012–1014.
- Green, Paul E. and Venkat Srinivasan (1990) "Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice," *Journal of Marketing*, 54(4):3–19.
- Grimmer, Justin (2015) "We are all Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together," *PS: Political Science & Politics*, 48(1):80–83.
- Hacker, Philipp and Bilyana Petkova (2017) "Reining in the Big Promise of Big Data: Transparency, Inequality, and New Regulatory Frontiers," *Northwestern Journal of Technology and Intellectual Property*, 15:1–22.
- Hainmueller, Jens, Dominik Hangartner and Teppei Yamamoto (2015) "Validating Vignette and Conjoint Survey Experiments against Real-World Behavior," *Proceedings of the National Academy of Sciences*, 112(8):2395–2400.
- Hassani, Hossein and Emmanuel Sirimal Silva (2015) "Forecasting with Big Data: A Review," *Annals of Data Science*, 2(1):5–19.
- Hoff, Kevin and Masooda Bashir (2015) "Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust," *Human Factors The Journal of the Human Factors and Ergonomics Society*, 57:407–434.
- Holsinger, Alexander M., Christopher T. Lowenkamp, Edward Latessa, Ralph Serin, Thomas H. Cohen, Charles R. Robinson, Anthony W. Flores and Scott W. VanBenschoten (2018) "A Rejoinder to Dressel and Farid: New Study Finds Computer Algorithm Is More Accurate than Humans at Predicting Arrest and as Good as a Group of 20 Lay Experts," *Federal Probation*, 82:51–56.
- Hosni, Hykel and Angelo Vulpiani (2018) "Forecasting in Light of Big Data," *Philosophy & Technology*, 31(4):557–569.
- Kennedy, Ryan, Philip Waggoner and Matthew Ward (2018a) "Trust in Public Policy Algorithms," Available at SSRN 3339475.
- Kennedy, Ryan, Scott Clifford, Tyler Burleigh, Philip Waggoner and Ryan Jewell (2018b) "The Shape of and Solutions to the MTurk Quality Crisis," Available at SSRN.
- Lazer, David and Jason Radford (2017) "Data ex Machina: Introduction to Big Data," *Annual Review of Sociology*, 43:19–39.
- Lazer, David, Ryan Kennedy, Gary King and Alessandro Vespignani (2014) "The Parable of Google Flu: Traps in Big Data Analysis," *Science*, 343(6176):1203–1205.
- Logg, Jennifer M. (2017) "Theory of Machine: When do People Rely on Algorithms?" *Harvard Business School Working Paper*, No. 17-086. <https://dash.harvard.edu/handle/1/31677474>.
- Montgomery, Jacob M. and Santiago Olivella (2018) "Tree-Based Models for Political Science Data," *American Journal of Political Science*, 62(3):729–744.

- Montgomery, Jacob M., Brendan Nyhan and Michelle Torres (2018) "How Conditioning on Post-treatment Variables can Ruin your Experiment and what to do about it," *American Journal of Political Science*, 62(3):760–775.
- Raghupathi, Wullianallur and Viju Raghupathi (2014) "Big Data Analytics in Healthcare: Promise and Potential," *Health Information Science and Systems*, 2(1):1–10.
- Rahwan, Iyad, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, Nicholas A. Christakis, Iain D. Couzin, Matthew O. Jackson, et al. (2019) "Machine Behaviour," *Nature*, 568(7753):477–486.
- Sotiropoulos, Fotis (2019) "Hydraulic Engineering in the Era of Big Data and Extreme Computing: Can Computers Simulate River Turbulence?" *Journal of Hydraulic Engineering*, 145(6):02519002.
- Tonidandel, Scott, Eden B. King and Jose M. Cortina (2018) "Big Data Methods: Leveraging Modern Data Analytic Techniques to Build Organizational Science," *Organizational Research Methods*, 21(3):525–547.
- Twitter (2011) "One Hundred Million Voices," <http://blog.twitter.com/2011/09/one-hundred-million-voices.html>. Accessed: 2011-09–12.
- Waggoner, Philip D., Ryan Kennedy and Scott Clifford (2019) "Detecting Fraud in Online Surveys by Tracing, Scoring, and Visualizing IP Addresses," *Journal of Open Source Software*, 4(37):1–5.
- Wegner, Daniel M. and Adrian F. Ward (2013) "How Google is Changing your Brain," *Scientific American*, 309(6):58–61.
- Weintraub, William S., Akl C. Fahed and John S. Rumsfeld (2018) "Translational Medicine in the Era of Big Data and Machine Learning," *Circulation Research*, 123(11):1202–1204.
- Willemsen, Martijn C. and Eric J. Johnson (2009) "MouselabWEB: Monitoring Information Acquisition Processes on the Web," Accessed March 7:2009.
- Wojcik, Stefan and Adam Hughes (2019) "Sizing Up Twitter Users," <https://www.pewinternet.org/2019/04/24/sizing-up-twitter-users/>.