

Does Explaining Algorithms Engender Trust?

Krista Mar, Mona Iwamoto, and Michael Amodeo
UC Berkeley, School of Information
MIDS Program
W231 Behind the Data: Humans and Values
W241 Experiments and Causality
Summer 2017

Table of contents

Abstract	2
Introduction	3
Context	3
What does it mean for an algorithm to be fair?	5
How do you explain an algorithm to a non-technical audience?	6
Clarifying our research decisions	9
Novelty	11
Context Background	12
Credit scores	12
Twitter and Online Moderation	14
Risk of Criminal Recidivism	16
Explanations Experiment	18
Methods	20
Survey Design	20
Mechanical Turk	21
Qualtrics	22
Survey Questions	23
Modeling and Analysis	27
Data Preparation and Checks	27
Tests Employed for Statistical Significance	29
Results	30
Responses	30
Twitter Moderation Context	30
Criminal Recidivism Risk Assessment Context	31
Statistical Significance	32
Twitter Moderation	34
Criminal Recidivism	35
Comparison of Contexts for Heterogeneous Context Effects	36
Difference in Order	38
Influence of Other Factors	39
Other Explanations	40
Discussion	41
Conclusion	43
Acknowledgements	44
References	44
Appendices	45

Abstract

Algorithmic decisions are becoming ubiquitous. They can be found everywhere from content moderation, credit decisions, and purchasing behavior to more consequential contexts such as employment decisions and risk of criminal recidivism. The repercussions of these decisions highlight the need to provide protections to individuals. Recent legislation in Europe mandates some degree of transparency into how the decisions are made as well as the ability to opt out. The 'right to an explanation' and its effects remain largely unexplored. This study examines how an explanation of the algorithm affects an individual's perception of the outcome. We hypothesized that providing an explanation increases trust and acceptance of algorithmic decisions. Further, we suspect that explanations have greater potential impact in contexts with decisions of greater personal significance. In this study, the effects of providing an explanation are examined in two different contexts. One context was the flagging of a tweet that was considered offensive by an algorithm and the other was assessing the risk of a repetition of a crime in the court system. Presented with either a graphical representation of the factors considered or no explanation, survey respondents evaluated their satisfaction with the decisions and the explanations. In both cases, we found that providing an explanation improved our respondent's acceptance of the decision significantly. In particular, respondents indicated that the decisions were more accurate and equitable with an explanation. Notably, the respondents perceived the decisions with the explanations as clearer and more meaningful.

Introduction

As algorithms become more commonplace in making automated decisions with significant impacts on people's lives, the importance of explaining these decisions becomes paramount. This automation takes the decisions out of human hands directly. When the time comes to explain a decision, to tell someone they were denied a loan or some other critical service, how can that algorithm deliver that message in a way that the consumer can understand and decide whether or not to trust the decision?

This study explores some methods for identifying the effectiveness of different types of explanations and how the different explanations affect a person's perception or actions related to an algorithmic decision. Does providing an explanation improve perceived value and trust of algorithms by consumers? In different contexts of decisions, does an explanation provide differing impacts?

Research Questions:

- Do certain kinds of explanations of algorithmic decisions inspire different levels of trust and acceptance in their decisions?
- Does the context of the decision affect that trust and acceptance of the algorithm?

Context

The need to find ways to adequately explain algorithmic decisions has become increasingly urgent with upcoming regulations. The European Union (EU) adopted the General Data Protection Regulation (GDPR) in April 2016 and it will go into effect in May 2018. It will replace the data protection directive currently in effect in Europe. The GDPR has been lauded as the most progressive regulation in data protection for consumers. According to the GDPR, in the case of automated decisions, a data subject possesses the right to access "*meaningful*

information about logic involved, as well as the significance and envisaged consequences of such processing on the data subject." This seems to suggest that an individual has a right to some transparency in how an automated decision is made as well as the consequences to that decision. In (*Watcher et al*) they term this as a "right to be informed." In article 22 it states that *"the data subject shall have the right not to be a subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her in similarly significantly affects him or her."* This article seems to suggest that an individual has a right to make an informed decision to opt out of an automated system. In recital 71, which is non-binding, it states that data subjects have a right to "obtain an explanation of the decision reached after such assessment." Recital 71 is where most scholars have taken the right to an explanation.

Why is the adoption of the GDPR important? Even though most of the large multinational technology corporations are located in the US, they will still need to learn how to comply with the GDPR if they want to operate in Europe. These technology companies will not want to lose such a large market, so they must find ways to comply with new regulations. As seen in a recent example where the EU fined Google \$2.7 billion for preferential treatment of its own shopping results, silicon valley companies are being forced to grapple with regulations that fundamentally change how one of their core products, search, operates (Finley, 2017).

While there is debate on how much power the GDPR will have once enforced, the machine learning community is already concerned about the consequences and has started doing work to sketch out what "providing an explanation" means in machine learning consequences. Explanations are particularly tricky in machine learning applications because there are no clearly defined rules governing the decisions. Decisions are based on weights,

probabilities, and similarities defined by the models that can be difficult to convey in plain language.

While different stakeholders such as regulators and consumers require different explanations on how algorithms make decisions, this experiment will focus on explainability of algorithms to consumers as this is the easiest stakeholder to focus on and the most similar across different decision making contexts.

What does it mean for an algorithm to be fair?

Measuring the fairness of algorithms is a complex problem and it is not obvious how to make an algorithm fair. There has been a lot of discussion among certain subsets of the machine learning, technology law, and social science community on how to have principles of fairness, accountability, transparency, and explainability in machine learning algorithms. Giving clear explanations about how machine learning algorithms make decisions is just as difficult if not more difficult than building an accurate machine learning model. How people are categorized by an algorithm can have huge consequences on people's lives (e.g. whether or not they receive welfare, whether or not they are barred from voting as a potential felon, what gender they are assigned in a system, etc.).

Care must be taken with the inputs to the system. Biased data lead to biased outcomes. When building their models data scientists must consider; how that data was collected, what biases may be baked into it, whether these biases can be corrected for, and, if not, what caveats need to be taken into consideration when using the outputs of the algorithm. There has been interesting work by (Bolukbasi et al) to debias word embeddings like word2vec, which is commonly used as training data for natural language processing models. This research debiased the word embedding for gender bias. While this research is a great example of how

debiasing training data can be done, it is difficult to scale. Rather, it is part of the responsibility of the builder of the algorithm to take these biases into consideration when building the model.

In light of the serious consequences of some algorithmic decisions, fairness of an algorithm must be taken into account carefully since they have a large impact on people's lives. For example in the study by (*Stuart, 2002*) felon exclusion lists had ~25% errors on the list and was racially biased against African American felons. Balancing false positive rates across protected classes such as race is one approach to measuring fairness.

There is ambiguity of whether or not algorithmic decisions should be used at all in certain contexts. Should there be certain situations that need extra scrutiny before using an algorithm in that scenario, similar to how the IRB has protected cases for particularly vulnerable populations such as school children or other protected and under represented classes? In the context of our research, high stakes environments may require additional scrutiny or vetting.

Whether or not people think that algorithms should be used in certain contexts, the reality of today is that they are. Our research is exploring how to make this algorithmic decision making process more transparent to lay people without a technical background so they can gain intuition of when to trust or not trust an algorithmic decision. By giving an explanation, we give people the opportunity to question the algorithmic decision and decide for themselves if they want to trust it and have more substantive reason for objecting to a decision.

How do you explain an algorithm to a non-technical audience?

Interpretability is a problem in algorithmic decision making with all algorithmics from rule based to neural nets. While neural networks are the poster child of a black box algorithm, it is possible to explain how they are trained easier than other algorithms. While linear regressions may seem easy to interpret at first glance, heavily engineered features can prevent these

relatively simple models from being easily interpreted (*Lipton, Z., 2016*). This is why explanations that are possible across different contexts are important. Coming up with real, interpretable, and clear explanations is a continuing area of exploration for researchers.

There are currently several different approaches to explaining machine learning algorithms. One approach is to try to understand how different features influence decisions made by algorithms. A second approach is the use of interactive visual analytics tools. A third approach is to try to understand how the machine learning algorithm learns.

One example of a theoretical framework focused on features is the Quantitative Input Influence (QII) measures developed by researchers at Carnegie Mellon University (*Datta et al*). The guiding research question of these researchers was “How can we measure the influence of inputs (or features) on decisions made by an algorithmic system about individuals or groups of individuals.” QII measures capture the degree of influence of inputs on outputs of the system. The causal QII measures account for correlated inputs, joint influence of a set of inputs, and marginal influence of individual inputs. This framework can be used as a transparency mechanism for black box machine learning algorithms. The developers of QII made the measures such that they can be differentially private while preserving accuracy. QII is especially useful in the case of “black box” algorithms like neural networks.

Researchers at NYU (*Tamagnini et al*) have developed a visual analytics interface called Ravelo that enables analysts to understand causes behind predictions of binary classifiers by interactively exploring a set of instance-level explanations. This tool is suggested as a way to assist analysts in understanding how the machine learning model made a decision, spotting potential issues, and possibly deriving insights on how problems can be solved. While this model can be generalized to various machine learning models and contexts, it currently only works on binary classification and binary feature sets. This tool would be useful for data savvy

individuals, but may not be as useful for the general public. As shown in Figure 1, the interface is complicated and not immediately clear for a non-technical audience.

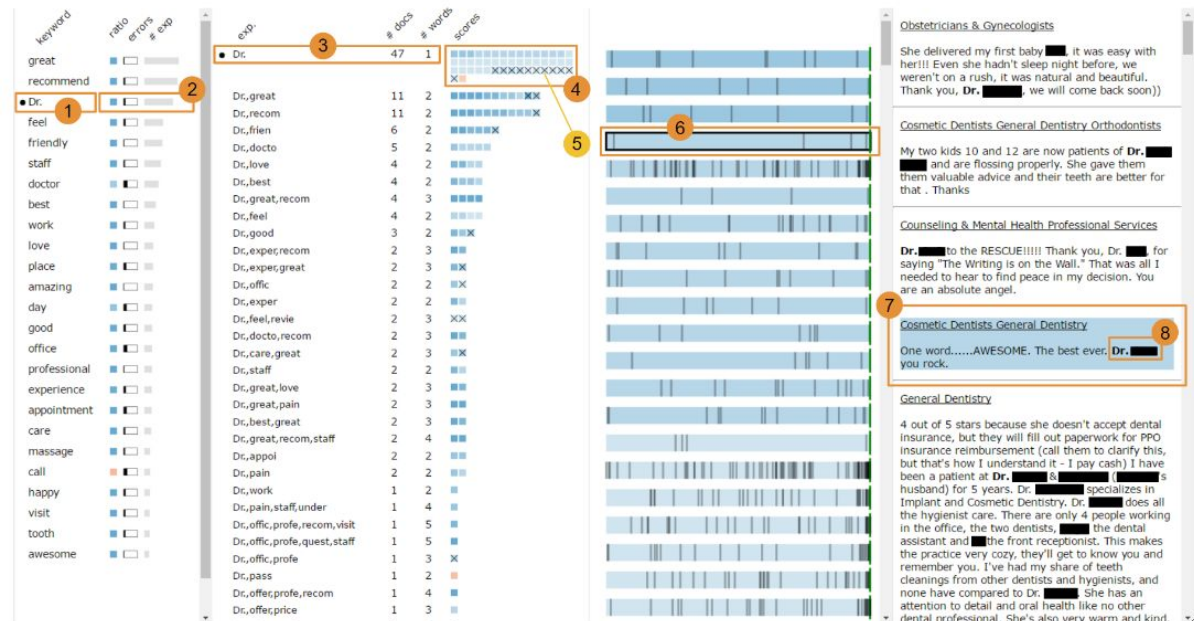


Figure 1: Rivelio Sample Output

A third approach developed by researchers at the University of Washington (*Ribeiro et al*) proposes a technology named LIME (Local Interpretable Model-agnostic Explanations), which is a “novel explanation technique that explains the predictions of any classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction.” Additionally, the researchers propose a method to explain a model by presenting representative predictions and non-redundant explanations. These models were tested in the context of support vector machine, text classification, and image recognition algorithms. An interesting angle that this approach took was using cases in which a person should not trust a classifier and showing a drop in trust once the explanation was given. A goal was to give subjects insights into classifiers and when not to trust them and why. This approach was simple and gave non-experts intuition into why not to trust a classifier. Interestingly, subjects did not notice

serious problems when looking at many mistakes in the raw data. An example is shown in Figure 2.

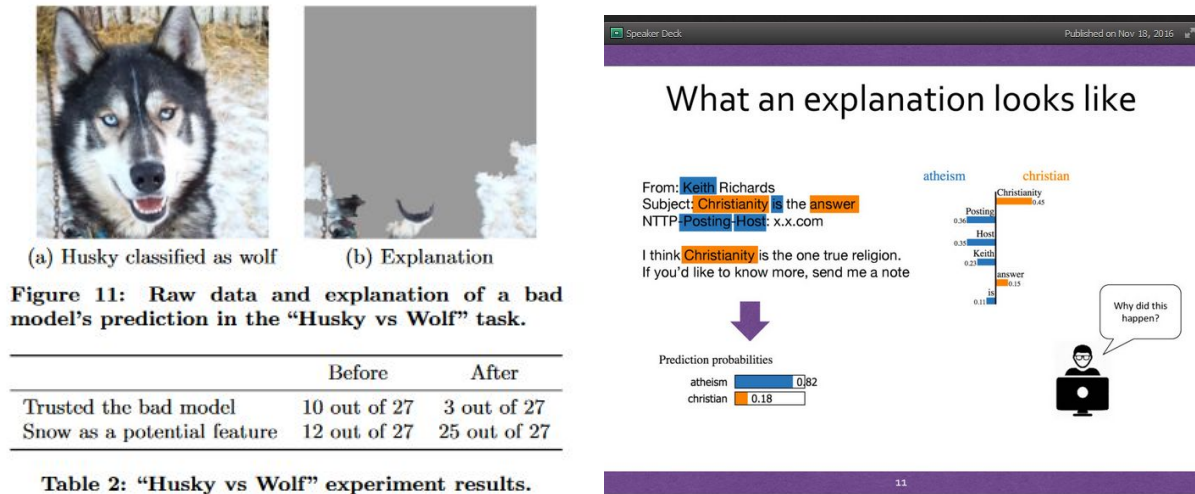


Figure 2: LIME Sample Output

For the purposes of our study, we adopted the simple, visual approach of the LIME model for one of our explanations in the pilot since it easy to interpret as well as possible to implement and appropriate for non-technical audiences.

Clarifying our research decisions

While initially there were many research questions, contexts, and explanations we wanted to explore, we limited the scope of our research in order to have meaningful conclusions and statistical power from our survey results. This meant that we had to make hard decisions on what to focus on and what to leave to future researchers. We decided to use specific attributes to operationalize trust and acceptance, to focus on Twitter and recidivism algorithms, and to provide a single treatment explanation instead of multiple explanations.

Trust and acceptance are concepts that are hard to operationalize and quantify into survey questions. We used principles from the Fairness, Accountability, and Transparency in Machine Learning(FATML) community and ideas from previous research from Ischool students

on moderation to think of relevant survey questions to operationalize trust and acceptance. To operationalize trust, we asked if the decision was fair, accurate, and satisfactory. We explicitly asked if the decision was useful, clear, and meaningful to operationalize acceptance of the explanation.

In order to test to see if there was significance of effect of providing explanations between context, we needed to choose some contexts to research. Algorithmic decision making is becoming a reality in basically all facets of our lives. In selecting contexts that would be the most useful, we thought about contexts in which there are the highest potential impacts on people's lives. We also attempted to identify contexts where machine learning algorithms are currently being used or implemented, as these are harder to explain. Finally, we included contexts that have had contentious stories inciting debate about the use of algorithms in these contexts. After receiving feedback to simplify our survey design by David Reiley, we decided to focus on medium and high impact contexts because these contexts have a higher impact on people's lives. Additionally, there has already been some research on the effectiveness and user experience of recommender systems (Knijnenburg et al), which included research on explanations within that context.

In an effort to simplify our study design, we chose one explanation to test against the control. As a control, we stated that an algorithm had made the decision instead of a human, since we wanted to explicitly focus on the effect of providing an explanation to an algorithmic decision rather than focus on the use of algorithmic decisions. Since we decided to only use one explanation as treatment, we ran a pilot study to determine the explanation that performed the best. We then used that explanation for our full study.

We are assuming that the algorithms are operating correctly and accurately. We did not want to introduce questions about whether the algorithm was making just decisions, because

we were testing specifically if we could increase trust through use of an appropriate explanation. This relies on the assumption that people do not automatically trust an algorithm, which we feel is valid based on public perception of algorithms and particularly negative press within the criminal recidivism context. Other scholars have focused on trying to get humans to understand when an algorithm is not accurate. Because our focus was on the possibility of changing the perception in a positive matter, we picked cases that were somewhat clear that the algorithm was operating correctly and accurately.

While non-technical audiences might think they want to see source code or a very detailed explanation of the algorithm, we assumed that this would produce an explanation that was too difficult to interpret. This is why we adopted explanations that were aimed at being clear and simple rather than overly technical or complicated. Also, because we were dealing with hypothetical algorithms that we have not designed, we could not create specific explanations on the rules of these theoretical algorithms.

Novelty

Our research is about perception. Through providing a simplistic, yet non obvious explanation to participants, we give the participants an opportunity to think critically about the algorithmic decision, to question the decision, and to understand the decision making process.

Previous work has focused on creating explanations, biases in training data, and studying responses in different contexts. We are building on the previous work by focusing on moderation and recidivism, which are rapidly evolving contexts for adopting algorithmic decision making, using explanations that are founded on current research.

Context Background

Before further describing the contexts for our study, we wanted to highlight credit scoring as a case where algorithms have been used for a long time to determine credit scores. This has a real impact on people's lives, and, unlike our other contexts, this is a context where the Federal Trade Commission (FTC) has set regulations. After exploring how regulation works in the credit scoring context, we will explore the contexts highlighted in our study, which have not seen interventions from the FTC yet. While there are many contexts we could have chosen for our study, we chose moderation as it is a highly contentious context where companies like Facebook and Twitter are actively working on shifting more moderation to algorithms. Additionally, we chose the courts as a context because court decisions can have drastic impact on people's lives and because there has recently been heated debate about the lack of transparency and potential bias in algorithms that determine risk for defendants.

Credit scores

Credit scores are one context of algorithmic decision making that is regulated by the federal government. Credit scores affect both an individual's ability to get loans like mortgages and the interest rate of those loans. There are a few companies that provide credit scores, but one that is common is FICO. We'll take FICO as an example. Since their algorithm is proprietary, they are not forced to reveal the exact formula. However, there is some transparency in how it calculates scores. The infographic below shows the breakdown as explained by FICO: 30% amounts owed, 35% payment history, 10% new credit, 15% credit history, 10% credit mix. Their website also explains what is meant by each of these terms. In a

more generalized sense, the FICO score shows you what goes into the score and the relative weights on each of these features.

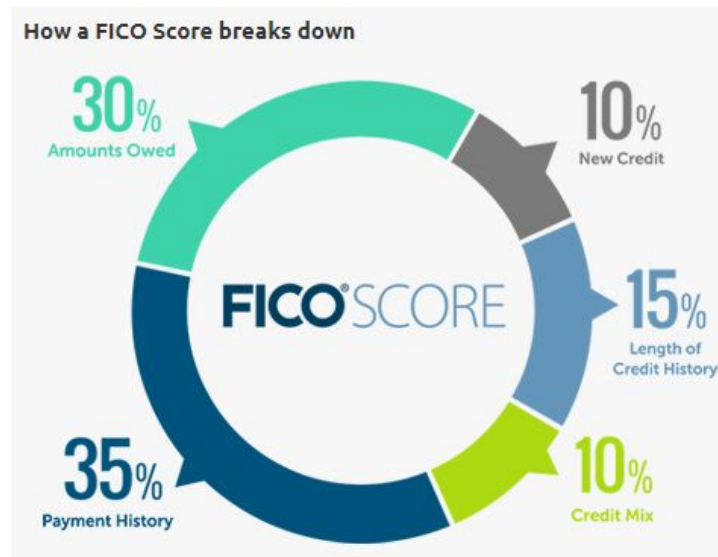


Figure 3: FICO Scores Calculations

The FTC regulates credit scores. The Fair Credit Reporting Act gives you the right to get your credit score from the national credit reporting companies. If there are adverse actions that are taken against a consumer because of their credit report, there are also regulations as to how this is reported to a consumer. Consumers also have redress in case of errors on the credit report to dispute errors. Consumers receive a free yearly report, but can pay extra to have more frequent reporting. The Equal Credit Opportunity Act (ECOA) prohibits credit discrimination on the basis of race, color, religion, national origin, sex, marital status, age, or because you get public assistance.

A similar framework for other proprietary algorithms could be developed. Such a framework could take key practices and standards from credit reporting including being transparent about the inputs to the model, being transparent about the weights on the model, and providing the score to the consumer.

While it is useful to think about how credit scoring and reporting is regulated, the dynamic and rapidly evolving nature of machine learning algorithms might make it hard to develop a similar system, though similar principles can still apply.

Twitter and Online Moderation

The prevalence of online harassment is increasingly coming to the attention of users of social media platforms. A 2014 Pew survey of Americans found that 73% of adult American internet users had witnessed harassment online, and 40% had personally experienced harassment. (Duggan, 2014)

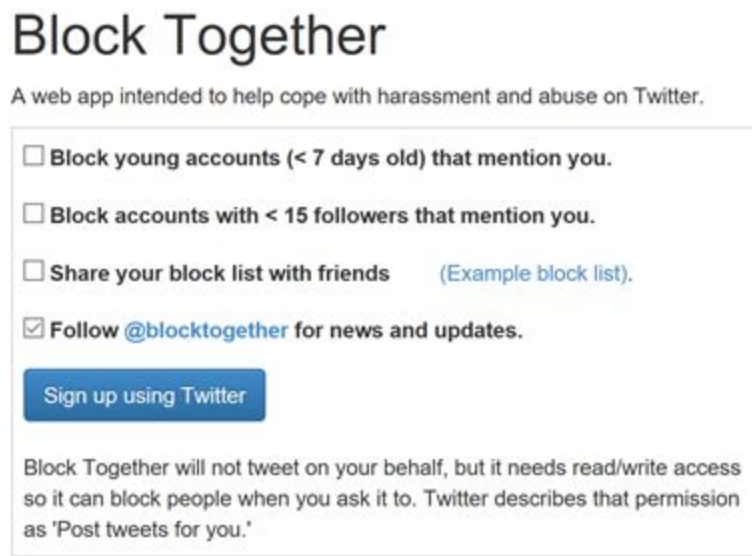
Broadly termed harassment, we include hate speech, incivility, trolling and threats in our definition of offensive content. Sites that host user-generated content including Facebook, Twitter, Flickr, YouTube, Instagram, and Foursquare, as well as in the comments sections on blogs and news sites, face challenges in both protecting speech as well as protecting the rights of those marginalized by intimidating content.

However, since knowledge of publishing offensive content can render the site open to liability, there is often little incentive for sites to take action prior to user feedback. Therefore the primary mechanism for controlling offensive content is user-initiated flagging. For example, Twitter relies exclusively on these reports to trigger the process of reviewing content. By flagging a post, individual users can express their concerns regarding violations of the platform's "community guidelines".

Additionally, communities and algorithms can contribute to flagging offensive content. Users can subscribe to third party bot-based collective block lists. Similar in functionality of ad-blockers, these applications use and maintain curated lists of blocked accounts and can screen posts from a user's Twitter feed. Many are open-source and can share lists across

applications like as Randi Harper's ggautoblocker (<http://twitter.com/ggautoblocker>) (Geiger, 2016).

One organization called BlockTogether.org provides options to block accounts meeting various criteria as shown below. In under one second, by using Twitter's streaming API, its algorithms auto-block possible offensive posts. Figure 4 shows some options of Block Together



The screenshot shows the 'Block Together' web app interface. At the top, the title 'Block Together' is displayed in a large, bold, sans-serif font. Below the title, a subtitle reads 'A web app intended to help cope with harassment and abuse on Twitter.' The main content area is a light gray box containing four checkboxes with corresponding text: 'Block young accounts (< 7 days old) that mention you.', 'Block accounts with < 15 followers that mention you.', 'Share your block list with friends (Example block list).', and 'Follow @blocktogether for news and updates.' The first three checkboxes are unchecked, while the fourth is checked. Below the checkboxes is a blue button with the text 'Sign up using Twitter'. At the bottom of the gray box, a disclaimer states: 'Block Together will not tweet on your behalf, but it needs read/write access so it can block people when you ask it to. Twitter describes that permission as 'Post tweets for you.'

Figure 4: Block Together Sample

With the increasing use of black-boxed teams of humans and algorithm that enforce these site rules, users may demand explanations of flagging or censorship.

Colleagues at the UC Berkeley School of Information have studied the flagging of online harassment. These projects were presented in a MIMS students capstone project called the [Moderation Machine](#) (Danker, Glenn, Mahar, Witt Advisor, & Mulligan, 2017) and a MIDS students capstone project called [Empathization](#) (Chan et al., n.d.). The findings of the Moderation Machine project compared people's perceptions of human vs. algorithmic flagging of harassing content. Humans were perceived as more accurate, fair, and trustworthy than

algorithms in their ability to flag content. In the Moderation Machine study, the text read “an algorithm has flagged this message as potentially harassing.”

As an extension to this project, we explore the effects of providing an explanation for an algorithmic flag. We include in our analysis the context of moderation and examine whether explanations increase or decrease perceptions of accuracy, fairness, and trustworthiness. As algorithmic flagging of harassing content becomes more common, so will the need for platforms to provide explanations that are satisfactory to customers.

Risk of Criminal Recidivism

A second context in which algorithms are being used to make decisions is in risk assessment of criminal recidivism. The practice of using statistical models to predict risk of recidivism has been in place in some contexts in the United States since 1923 (US Courts OPPS, 2011). These models evolved over time, and federal regulations even began to require probation officers to classify individuals with a category of supervision required. The use of various tools to perform classification proliferated across many systems without a standardized set of rules or data inputs. This has naturally begun to include sophisticated algorithms, many of which are proprietary, being implemented by various counties, states and the federal government.

These algorithms have come under heavy public criticism in the past several years for many reasons. First, some algorithms have been shown to include inherent biases. A Pulitzer Prize-winning article in 2016 by ProPublica examined the COMPAS algorithm used in parts of Florida and showed that although the algorithm had the same accuracy for white and black defendants, it identified false positives and false negatives at extremely different rates (Angwin et al). The metrics that the algorithm was using were biased against people who come from

communities with higher rates of criminal incarceration, and many viewed this as a proxy for race. Other recidivism algorithms have been shown not to contain such inherent racial biases, particularly the federal government's Post Conviction Risk Assessment (PCRA) (Skeem). However, public reaction to the ProPublica piece seems to show that people are unaware that these sort of algorithms are being used. Increasing awareness of these algorithms is important, as is increasing awareness of their accuracy or inaccuracy.

It is also worth noting that the algorithms have come to be used for more than what they were originally intended. The PCRA and many of the historical risk assessments are focused on determining appropriate supervision for those on parole or probation. Now, COMPAS and others are being used to assist in determining sentencing. Even in cases where judges are supposed to use them as guidance, there are instances where judges rely on them too heavily. The Supreme Court declined to hear a case in 2017 about whether it was appropriate to use these algorithms in sentencing (*Loomis v. Wisconsin*). The Court declined because the risk assessment was only supposed to be a guide to the judge, and it was not what determined the sentence.

A major deficiency with the risk assessments that makes it especially relevant to our study is that the decisions often come without clear explanations. COMPAS produces an overall risk score that is a summation of 23 different metrics on which the subject is rated on a scale of 1-10, normalized against other subjects. The scores on these other metrics are usually given as explanation, but there is no clear explanation of what the underlying metrics mean or how the questions the subject answered turned into the score in the metric (Angwin et al).

All of this has several implications for our study. First, we want to assume that the theoretical algorithm is working properly, or at least with a high level of accuracy. We also want

to design our questions to be using the algorithm in the proper context. More accepted contexts are probation supervision and setting bail before a trial.

Explanations Experiment

This investigation extends the work of our colleagues and addresses the question of how providing an explanation of the algorithmic decision improves the subject's acceptance of its outcome. We break down the subject's acceptance into several components, views regarding fairness, accuracy and transparency of the decision and their opinion about the usefulness, clarity and meaningfulness of the explanation given.

For this study, the contexts of moderation and recidivism were chosen to represent differential levels of impact to the individual. The context of a taste-based recommender system such as Netflix, which we deemed low impact, was also considered. However, to maintain the power of our study, the latter was eliminated.

Subjects were recruited from Mechanical Turk and directed to a survey created in Qualtrics. Once in the Qualtrics environment, subjects were randomly assigned an order of contexts, some experienced the Twitter questions first and conversely some received the recidivism questions first. This design was chosen to account for any difference in exposure to similar questions over time.

To choose the type of explanation that would be effective, a small pilot study was conducted. Three types of explanation and a control were tested. The first explanation expressed 'why' the algorithm flagged the content or deemed the defendant to be high risk of recidivism. The second explanation stated that the decision was similar to decisions made by humans and the last explanation showed a graphic breakdown of the factors considered by the algorithm. See Figure 5.

The results of the pilot indicated that the graphical representation showed the highest potential in affecting a respondent's acceptance of the algorithmic decision. The results of the pilot study are discussed below.

For the design of the main study, each respondent was assigned both contexts in a randomized order and then randomly assigned treatment or control in each context. Treatment consisted of the graphic explanation of the algorithmic decision and control was no explanation.

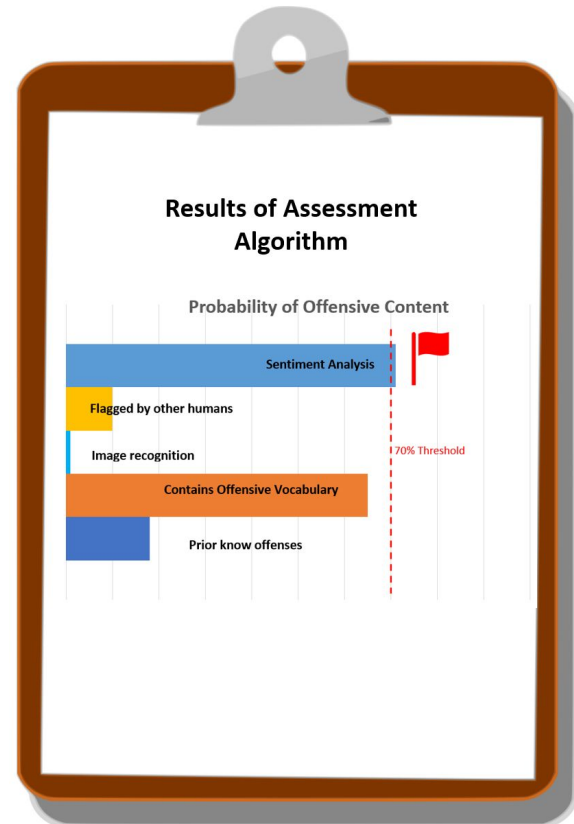


Figure 5: Treatment Explanation

Respondents were then asked a series of question to assess the subjects' reaction to the treatment and understand the difference in levels of acceptance of the decisions presented. We hypothesized that an explanation can increase trust and acceptance of decisions and that explanations have greater potential impact in contexts with decisions of greater personal significance.

The results of this study support our hypothesis that the presence of an explanation increases confidence in the algorithmic decision.

Methods

Survey Design

The pilot survey was created in Qualtrics. For the pilot, there were 4 treatment options, listed below in Table 1 and 2 contexts, Twitter and recidivism risk.

<u>Control</u> “An algorithm made this decision”	<u>Detailed Explanation</u> “An algorithm made this decision because of [factors]”
<u>Similarity to Humans</u> “An algorithm made this decision because the case is similar to decisions made by humans”	<u>Graphical Depiction of Factors</u> Graphical depiction of detailed explanation

Table 1: Pilot Study Explanations

For each treatment condition the subjects were asked:

For the Twitter Context

- How fair do you find this moderation flag?
- How accurately do you think the algorithm made this decision?
- How satisfied are you with the transparency of this decision?
- How useful do you find this explanation?

For the Recidivism Risk Context

- How fair do you find this risk assessment?
- How accurately do you think the algorithm made this decision?
- How satisfied are you with the transparency of this decision?
- How useful do you find this explanation?

- How clear do you find this explanation?
- How meaningful do you find this explanation?
- What information would make you more satisfied with the decision explanation?
- How clear do you find this explanation?
- How meaningful do you find this explanation?
- What information would make you more satisfied with the decision explanation?

Because we found that the graphical depiction of the detailed explanation produced a statistically significant improvement in multiple metrics for the recidivism context, this explanation was chosen for the full study.

Mechanical Turk

Study participants were recruited by Amazon Mechanical Turk (<https://www.mturk.com>). Amazon Mechanical Turk is a crowdsourcing site where ‘workers’ are compensated to complete tasks. While this process returned high numbers of responses in very short order, we understand that this population does not represent a true random selection of the general population. Typical MTurk workers reside in the United States with demographics similar to the overall internet population in the US. All of the respondents to this study were computer literate, and various sociodemographic questions revealed strong majorities as young (25-34, 48%), college educated (4 year degree, 42%) , white (66%), male (65%), and who use social media daily(80%).

The description of the task was *“This study is looking at the role that decisions made by computers play in different contexts. (~5 min)”*, with the instructions to *“Answer a survey about your opinions on decisions made by computers. (WARNING: This HIT may contain adult content. Worker discretion advised.)”*, with a link to a Qualtrics survey described below. The

task was restricted to unique respondents over 18 years of age. A warning was also posted due to possible offensive language in the survey. Each respondent was compensated \$0.60 for a completed survey in the pilot study and \$0.50 for a completed survey in the full study.

The pilot was conducted on July 13, 2017 and resulted in 80 compensated responses and the full study was completed on August 4, 2017 with 616 compensated responses. Each batch of surveys returned results very quickly. The full set of 616 responses was collected in under 2 hours.

Qualtrics

The survey itself was created and managed by software created by Qualtrics (www.qualtrics.com) and site licensed to the University of California, Berkeley. Qualtrics allows for very flexible survey creation and data collection.

Once the respondent from Mechanical Turk clicked on the survey link, they were directed to the Qualtrics site. After reading a brief introduction to the survey, the user was randomly assigned either to answer questions in the Twitter context or the criminal recidivism context first. After completing this section, the respondent was asked the same questions in the other context. Thus, each respondent answered questions in both contexts. Additionally, within each context the subject was randomly assigned treatment (the explanation) or control (no explanation). See the experimental design below in table 2.

Context Order	Context 1	Treatment 1 Assignment	Context 1 Treatment	Context 1 Response	Treatment 2 Assignment	Context 2 Treatment	Context 2 Response
R	Twitter	R	X	O	R	X	O
R	Twitter	R	X	O	R		O
R	Twitter	R		O	R	X	O
R	Twitter	R		O	R		O
R	Recidivism	R	X	O	R	X	O
R	Recidivism	R	X	O	R		O
R	Recidivism	R		O	R	X	O
R	Recidivism	R		O	R		O

Table 2: Research Design

Survey Questions

For the Twitter context, we showed the respondent the following description:

Assume for the sake of this study, that you see the following comment in your Twitter feed and it is flagged as inappropriate according to their rules of content and behavior.

As a result Twitter may lock or suspend the commenter's account.

Then, we displayed an image of a tweet shown in Figure 6.



Figure 6: Offensive Tweet

This example was chosen to maintain slight ambiguity in whether this tweet should be flagged. In our pilot study, stronger language in the tweet resulted in respondents seemingly not questioning the algorithm. The response to the fairness metric was very high under the control, leaving the explanation unable to significantly affect the metric.

For the control group, the subject was told: ***“The comment has been flagged by an algorithm.”***

For the treatment group, the subject was told: ***“The comment was flagged by an algorithm based on the strength of the factors shown below.”***, with the image in Figure 7.

Both the treatment and control groups were asked 3 questions regarding the decision.

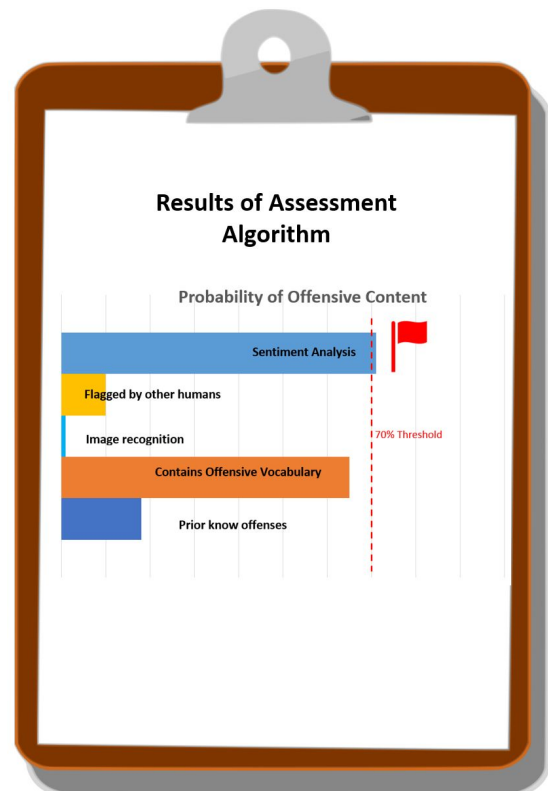


Figure 7: Twitter Treatment

1. How fair do you find this moderation flag?

2. How accurately do you think the algorithm made this decision?
3. How satisfied are you with the transparency of this decision?

Then 3 questions regarding the explanation were asked.

1. How useful do you find this explanation?
2. How clear do you find this explanation?
3. How meaningful do you find this explanation?

The response choices were based on a 5 level likert-type scale, ranging from “Extremely” to “Not at all”. Finally, we asked, “What information would make you more satisfied with the decision explanation?” The respondents were presented with the choices of 1) Examples of other levels of decision output; 2) Relative importance of the characteristics that led to the decision; 3) Detailed description of how the algorithm works; or 4) Other.

For the context of criminal recidivism, the subject was told:

For the purposes of this study, please consider yourself to be a criminal defense attorney. Your client is charged with armed robbery.

The court uses a software that assigns your client a risk assessment of high risk of committing future crimes. The judge is going to take this risk assessment into account to set bail.

For the control group, the subject was told: “**The defendant is assessed as high risk by the algorithm.**”

For the treatment group, the subject was told: “**The defendant was assessed as high risk of committing another crime by the algorithm on the strength of the factors shown below.**”, with the image in Figure 8.

Again, both the treatment and control groups were asked similar questions about the decision and the explanation.

We also collected demographic information. The respondents' age group, race or ethnic category, gender category, social media usage and highest level of education were optionally requested.

Once the survey was completed, a randomly generated six digit code was issued to the respondent and was to be submitted in the Mechanical Turk task to verify completion. The random codes assigned by Qualtrics were correlated with the completed tasks codes in Mechanical Turk, and the respondents were compensated.

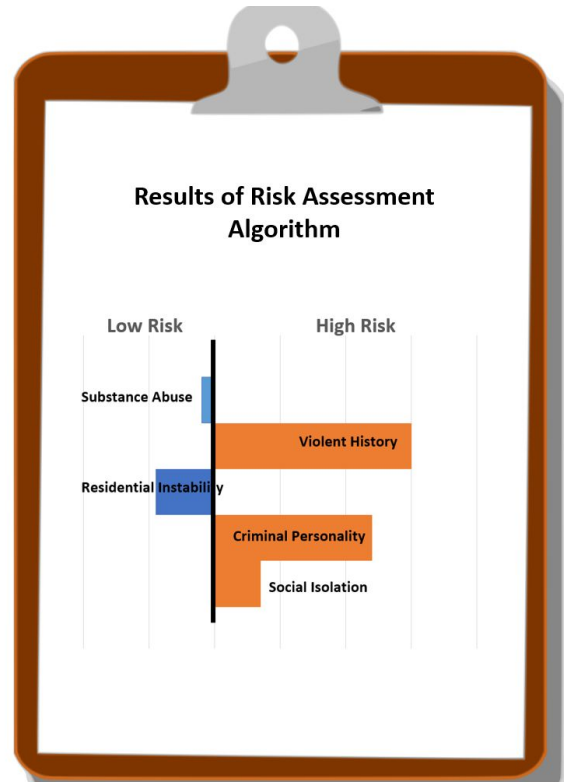


Figure 8: Recidivism Treatment

The total number of responses collected from Qualtrics was 641, with very little attrition (19). Six respondents failed to submit the code to Mechanical Turk and were not compensated. Randomization checks confirmed balanced random assignment. See Figure 9.

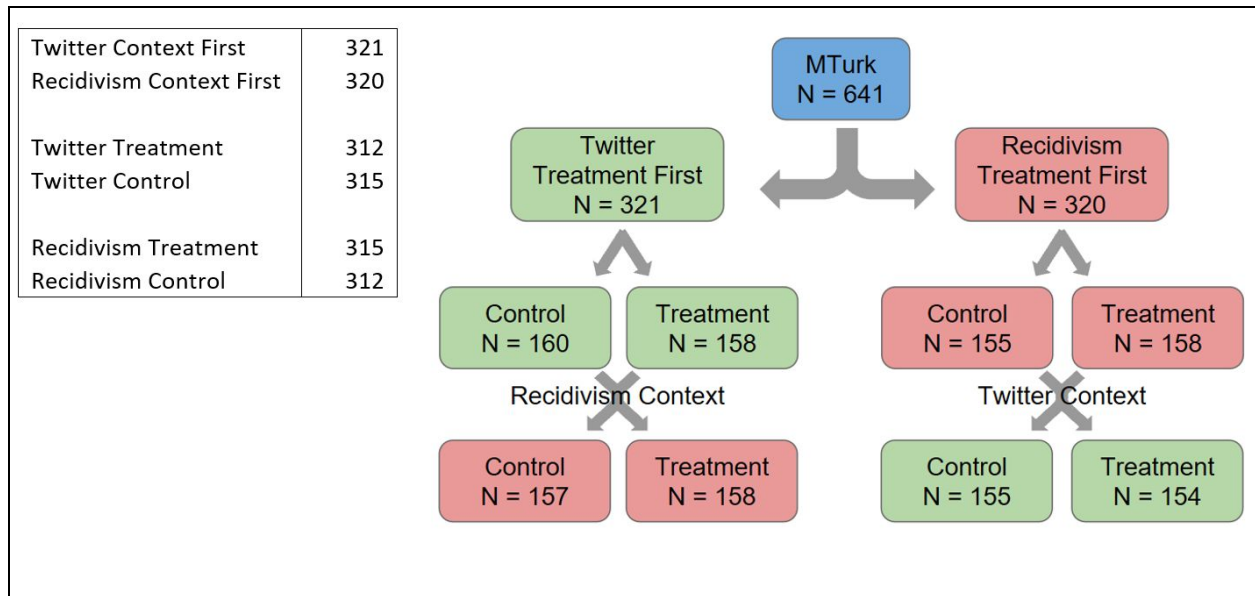


Figure 9: Survey Random Assignment

Modeling and Analysis

Data was exported from Qualtrics in a comma separated values (csv) format. Data was analyzed using the R statistical package. A full analysis can be found attached as Appendix A. Before performing any modeling of results, several data checks were performed.

Data Preparation and Checks

The questions were based on a 5 point Likert scale. For each metric, the answers varied from "Extremely" to "Not at All." Most Qualtrics questions were set so the extreme positive value was the first choice (1). In order to show an increase in acceptance or trust as positive, we rescaled these values and flipped them around the median value (3). Because the Qualtrics format required each question to be different, we had to consolidate the responses for each metric into a single value per context.

We had no way of measuring compliance, which in this case would entail respondents answering questions without reading the treatment.

The first check we performed was to ensure that randomization occurred as designed. As shown in Figure 9 above, the order of context ended up fairly equal, as did the number of respondents who received treatment or control within each context.

Our second check was to review the attrition rates within the survey. As mentioned previously, 19 of the 641 respondents did not complete the survey. We did notice discrepancies between the number of completed surveys in Qualtrics and Mechanical Turk. Not all respondents who were assigned a random number code after completing the survey in Qualtrics submitted that code in Mechanical Turk. We found some duplicate IP addresses, but the worker codes in Mechanical Turk were all unique, meaning they came from different Mechanical Turk accounts in the same physical location. Given that Mechanical Turk approves accounts based on unique social security numbers, it may be that these are valid responses from multiple people using the same computer.

Also, some of the incomplete surveys appear from the same IP address as completed surveys, indicating that someone got part of the way through our survey without finishing it but opened the link again and completed it. Without too much detail about the individual instances, it would be more conservative to leave the results in the analysis, which is what we have done. These are roughly evenly distributed and represent a very small fraction of the overall number of respondents.

When analyzing the potential patterns of attrition in greater detail, we also found that these instances of attrition occurred at various stages in the survey. Most of the attrition (10 incidences) occurred before respondents were even assigned to one of the contexts. These respondents followed the link to Qualtrics and saw the description of the survey, but did not

proceed further. In the other instances, we saw an even distribution of attrition occurring within both contexts, and between the first and second contexts. Because of this even distribution, attrition does not seem to be indicative of any specific problem with our survey design, and it does not affect one context more than the other. We proceeded with our analysis by using any record that was assigned treatment or control for a given context. This means that our measured treatment effects are actually intent to treat effects. When comparing contexts, we only used respondents who were assigned to treatment or control in both contexts. This would allow for inclusion of records where respondents completed the first context and started the second, even if they did not complete it. But it filters out records where respondents attrited during their first context.

Tests Employed for Statistical Significance

Because the output for all questions was measured in a 5 point Likert scale, our first test for significance within each of the two contexts was a Wilcoxon rank sum test. This test is appropriate to determine if ordinal data shows a significant difference in means. This was applied to each metric in each context to see if there was a significant difference between the control group mean and the treatment group mean.

When comparing contexts, we used a Wilcox signed rank test. This test was appropriate because it allowed us to compare individuals who received either treatment or control in both contexts as dependent samples.

Linear models are not the most appropriate test for statistical significance of ordinal data. However, we did perform some linear regression to demonstrate the magnitude of the shifts that we saw across metrics. These results are not intended to be interpreted as tied to the difference in steps between possible responses, but they are informative when comparing treatment

effects across the different metrics we used because they were consistent with each other. However, these numbers cannot be used to compare to other studies.

Results

Responses

Twitter Moderation Context

Within the Twitter moderation context, the histograms of survey responses show a notable shift to the right for each metric. In Figure 10 below, the control responses are shown in blue and the treatment responses are shown in red. As you can see, in the control the accuracy of Twitter's algorithmic decision is already rated fairly high. This is likely due to the fact that respondents can see the tweet and judge the decision themselves and deemed the tweet offensive.

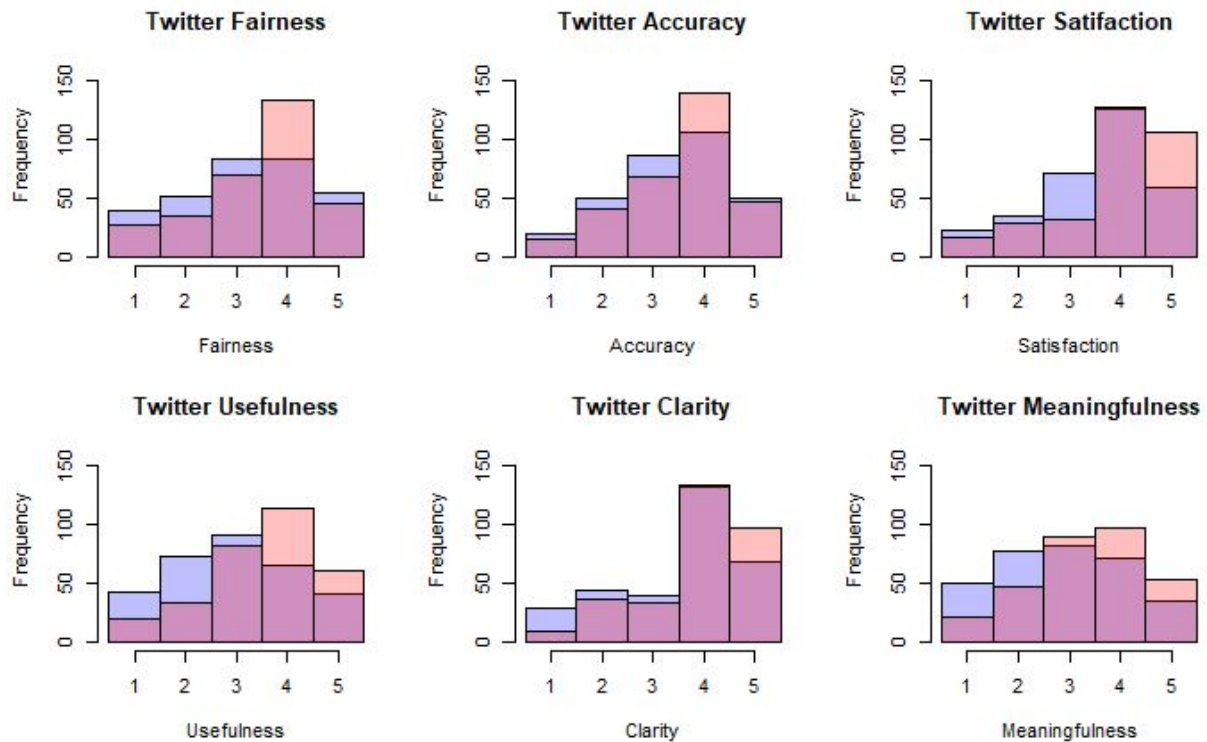


Figure 10: Histograms of Twitter Responses

Criminal Recidivism Risk Assessment Context

Within the criminal recidivism risk assessment context, the histograms of survey responses also show a notable shift to the right for each metric. In Figure 11 below, the control responses are shown in blue and the treatment responses are shown in red. The control in this context rates much lower than in the Twitter context. This could relate to something about the survey design or about the context itself.

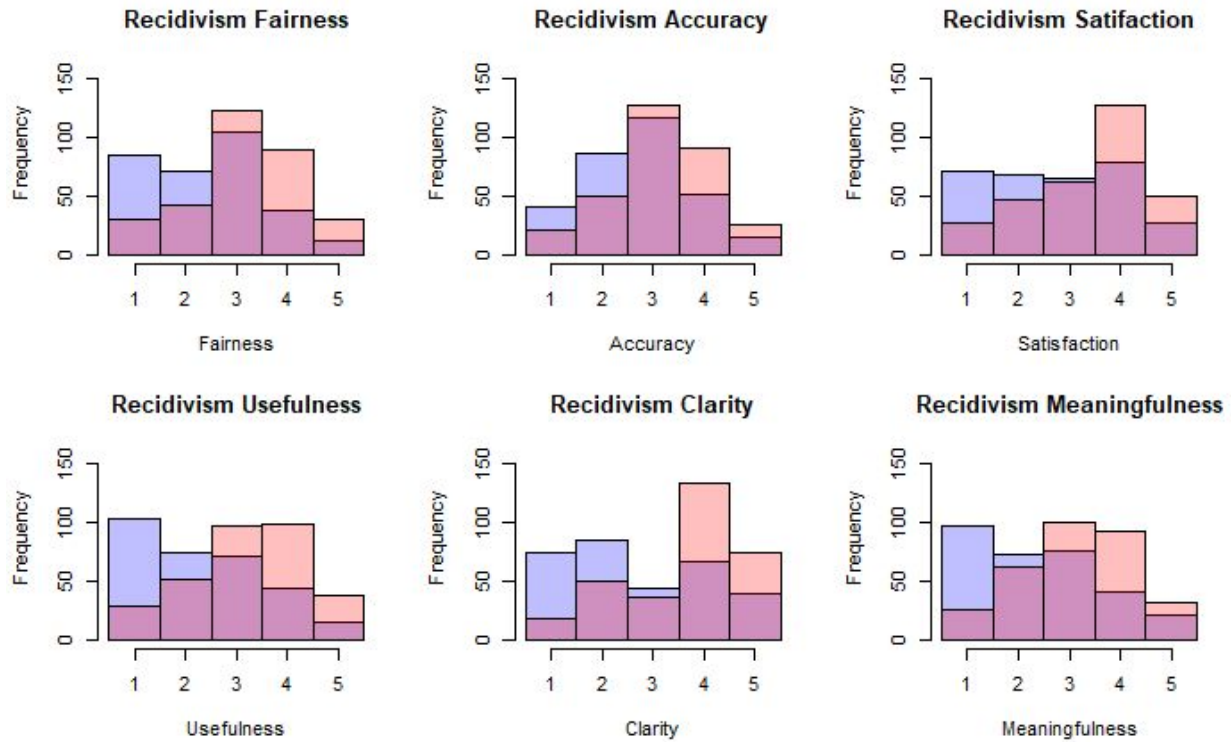


Figure 11: Histograms of Recidivism Responses

Statistical Significance

Using the Wilcoxon rank sum test, we see that each metric shows statistical significance in each of the two contexts. These p-values are quite small, although some are larger than others. In particular, we see the Twitter fairness and accuracy metrics are many orders of magnitude larger than the other metrics, though still well below a 0.01 statistical significance level.

Wilcoxon Rank Sum Significance						
	Fairness	Accuracy	Satisfaction	Usefulness	Clarity	Meaningfulness
Twitter	6.90e-04	5.27e-03	4.30e-08	6.80e-11	1.80e-05	2.50e-08
Criminal Recidivism	8.60e-18	8.90e-10	2.20e-12	1.40e-20	1.60e-18	2.50e-16

Table 3: Treatment Effect Significance

We also compared contexts using the Wilcoxon signed rank test, which is the appropriate test of significance for the ordinal data when viewing differences within subjects. We filtered to only include subjects who were assigned the same treatment in both contexts (control or explanation). We again see statistical significance to the difference in contexts across all metrics. Interestingly, this significance is stronger with the decision metrics than it is for the explanation metrics. This supports what we saw earlier where the Twitter context showed a lower degree of significance for treatment effects regarding the decision.

Wilcoxon Signed Rank Significance						
	Fairness	Accuracy	Satisfaction	Usefulness	Clarity	Meaningfulness
Contexts	1.39e-07	1.15e-09	1.24e-10	8.30e-07	2.50e-08	1.62e-04

Table 4: Significance of Difference in Contexts

A basic view of the data showed there was a change in the distributions. The Wilcoxon tests showed the significance of the treatment effect of the explanation within each context as well as the difference between contexts. While linear modelling of Likert scale data can be misleading, we will use regression models to allow us to more fully gauge the significance of these changes. We have created linear models for each question for each context (Twitter and recidivism). The models subset the data to look only at those respondents that were assigned to either treatment or control for that context. In this way, someone who attrited in the first context will not count against the second context, but someone who dropped out midway through a response to context will count in the effects.

Twitter Moderation

Twitter Moderation						
	Fairness (1)	Accuracy (2)	Satisfaction (3)	Usefulness (4)	Clarity (5)	Meaningfulness (6)
Explanation	0.242** (0.096)	0.157* (0.087)	0.367*** (0.091)	0.545*** (0.095)	0.332*** (0.094)	0.467*** (0.096)
Constant	3.197*** (0.071)	3.371*** (0.063)	3.524*** (0.064)	2.965*** (0.069)	3.530*** (0.070)	2.886*** (0.070)
Observations	627	627	627	627	627	627
R2	0.010	0.005	0.025	0.050	0.020	0.036
Adjusted R2	0.008	0.004	0.024	0.049	0.018	0.035
Residual Std. Error (df = 625)	1.203	1.091	1.139	1.183	1.170	1.202
F Statistic (df = 1; 625)	6.357**	3.266*	16.292***	33.215***	12.623***	23.643***

Note: *p<0.1; **p<0.05; ***p<0.01

Table 5: Twitter Moderation

The last three metrics show that the treatment explanation received significantly better responses than the control. The decision metrics (fairness, accuracy, satisfaction) also show significant increases but they are not as strong. The Accuracy metric was the only metric to not receive a significant effect at the 0.05 level. The Wilcox test performed previously showed this was a significant effect, but this view of the ordinal output shows that accuracy is the metric that is affected the least by the treatment.

Criminal Recidivism

Recidivism Risk Assessment						
	Fairness (1)	Accuracy (2)	Satisfaction (3)	Usefulness (4)	Clarity (5)	Meaningfulness (6)
Explanation	0.726*** (0.088)	0.440*** (0.082)	0.649*** (0.099)	0.872*** (0.095)	0.906*** (0.104)	0.736*** (0.095)
Constant	2.423*** (0.064)	2.718*** (0.059)	2.750*** (0.073)	2.324*** (0.070)	2.705*** (0.079)	2.388*** (0.071)
Observations	628	628	628	628	628	628
R2	0.098	0.044	0.064	0.118	0.109	0.088
Adjusted R2	0.097	0.042	0.063	0.117	0.108	0.086
Residual Std. Error (df = 626)	1.102	1.029	1.239	1.192	1.295	1.189
F Statistic (df = 1; 626)	68.079***	28.723***	43.073***	84.045***	76.767***	60.139***
Note: *p<0.1; **p<0.05; ***p<0.01						

Table 6: Recidivism Risk Assessment

In the recidivism context, we see statistical significance at the 0.01 level for treatment in all metrics. These are also higher in magnitude than all of the Twitter effects. Again, the effects are larger in magnitude for the explanation based metrics than for the decision metrics. Also as we saw in the Twitter context, satisfaction is the most affected out of the decision metrics. This shows that even if people don't agree with the decision, they are more willing to accept it.

Comparison of Contexts for Heterogeneous Context Effects

	Context Comparison					
	Fairness (1)	Accuracy (2)	Satisfaction (3)	Usefulness (4)	Clarity (5)	Meaningfulness (6)
Recidivism Context	-0.795*** (0.096)	-0.671*** (0.087)	-0.798*** (0.097)	-0.830*** (0.099)	-0.649*** (0.105)	-0.504*** (0.100)
Treatment	0.233** (0.096)	0.147* (0.087)	0.354*** (0.091)	0.335*** (0.094)	0.546*** (0.093)	0.469*** (0.096)
Recidivism Treatment	0.510*** (0.131)	0.306** (0.120)	0.315** (0.135)	0.577** (0.134)	0.332*** (0.139)	0.275** (0.135)
Constant	3.204*** (0.072)	3.380*** (0.063)	3.534*** (0.064)	3.540*** (0.069)	2.974*** (0.070)	2.895*** (0.070)
Observations	1,248	1,248	1,248	1,248	1,248	1,248
R2	0.101	0.078	0.110	0.113	0.121	0.084
Adjusted R2	0.098	0.076	0.108	0.111	0.119	0.082
Residual Std. Error (df = 1244)	1.151	1.059	1.186	1.223	1.179	1.187
F Statistic (df = 3; 1244)	46.403***	35.233***	51.221***	52.900***	57.212***	38.257***
Note: *p<0.1; **p<0.05; ***p<0.01						

Table 7: Comparison of Contexts

Our second research question asked if there was a difference between how respondents evaluated the explanation in two different contexts of varying importance or personal significance. When comparing treatment effects across contexts, we see statistical significance in the baseline constant for all metrics in the fourth row. This represents the constant in the Twitter control group. In the first row of the regression, we see the change to recidivism has a significant negative effect in all metrics. This means that respondents were less accepting of the algorithm's decision with the control explanation than they were in the Twitter context. This may be partly attributable to the design of the survey. Criminal recidivism is a complicated problem with more inputs than a 140 character tweet. Because we included the full tweet, people were able to judge the appropriateness of the decision by themselves. In the recidivism context,

respondents were only given a brief description of the case, with just the offense the defendant was being charged with. Including some information about criminal history or other factors may have made this a more appropriate comparison.

In the second row, we see what is effectively the same effects with the same significance of the Twitter treatment that we saw in the Twitter only models. The numbers are slightly different here because this analysis looks only at individuals who were assigned to treatment or control in both contexts, whereas previous models included records that may not have made it to a second context. So a few instances are missing in this regression where an individual did not make it to the second half of their survey. In the third row, we see the effect of the recidivism treatment compared to the effect of the Twitter treatment. Again, we have high statistical significance in all metrics as we did in the recidivism only models. However, the significance of the differences is less than of the treatment itself, dropping in 4 of the 6 cases below the 0.01 level, although still significant at a level of 0.05.

When combining the results from the first and third rows, this suggests that the recidivism treatment started from a lower baseline due to the control explanation and therefore had more room to improve. This also shows that the explanation in this context did in fact provide greater effects than the explanation in the Twitter context. In the decision metrics, the sum of the treatment and recidivism treatment effects are approximately equal in magnitude but opposite in direction to the effect of the recidivism context. That means that the Twitter and recidivism contexts ended up at approximately equal ratings for those three metrics. In the explanations metrics, the treatment effects outweigh the negative effect of the recidivism control compared to the Twitter control. This shows that the recidivism explanation ratings ended up higher than the Twitter context, though not by too much.

Difference in Order

We also discussed looking at the difference in responses depending on the order of contexts. Significant effects here would show whether answering one context first created a bias in the response to the second context.

Effects of Seeing Twitter Moderation After Recidivism Treatment						
	Fairness (1)	Accuracy (2)	Satisfaction (3)	Usefulness (4)	Clarity (5)	Meaningfulness (6)
Twitter Treatment	0.138 (0.204)	-0.097 (0.181)	0.129 (0.194)	0.279 (0.191)	0.202 (0.189)	0.190 (0.192)
Constant	3.052*** (0.142)	3.351*** (0.126)	3.416*** (0.127)	2.987*** (0.134)	3.506*** (0.134)	3.013*** (0.132)
Observations	156	156	156	156	156	156
R2	0.003	0.002	0.003	0.014	0.007	0.006
Adjusted R2	-0.003	-0.005	-0.004	0.007	0.001	-0.0001
Residual Std. Error (df = 154)	1.264	1.124	1.207	1.185	1.172	1.195
F Statistic (df = 1; 154)	0.464	0.293	0.444	2.157	1.162	0.982
Note: *p<0.1; **p<0.05; ***p<0.01						

Table 8: Effects on Twitter Moderation After Recidivism Treatment

In the case of Twitter moderation, there does seem to be a difference based on order of context. There is no significant effect to receiving the Twitter treatment if the respondent had already seen the recidivism treatment. This is different from our overall Twitter treatment effect, which implies that the effect is stronger in the other groups.

Effects of Seeing Recidivism Context After Twitter Treatment						
	Fairness (1)	Accuracy (2)	Satisfaction (3)	Usefulness (4)	Clarity (5)	Meaningfulness (6)
Treatment	0.954*** (0.171)	0.535*** (0.171)	0.766*** (0.194)	0.978*** (0.191)	1.136*** (0.201)	1.041*** (0.191)
Constant	2.418*** (0.129)	2.734*** (0.130)	2.785*** (0.146)	2.304*** (0.147)	2.582*** (0.154)	2.215*** (0.151)
Observations	157	157	157	157	157	157
R2	0.169	0.060	0.092	0.147	0.172	0.162
Adjusted R2	0.164	0.054	0.086	0.141	0.167	0.157
Residual Std. Error (df = 155)	1.064	1.068	1.209	1.187	1.252	1.192
F Statistic (df = 1; 155)	31.565***	9.853***	15.769***	26.644***	32.287***	29.950***
Note:				*p<0.1; **p<0.05; ***p<0.01		

Table 9: Effects on Recidivism After Seeing Twitter Treatment

Interestingly, people who saw Twitter treatment first still showed a significant effect for the recidivism treatment. This is opposite what we saw from respondents who saw recidivism treatment before Twitter treatment. The magnitude of effect for all metrics is also larger than the corresponding treatment effects we saw in the overall comparison.

Influence of Other Factors

We checked for influence of other factors on responses, including gender, race, social media use, education, and level of education. It became difficult to draw many conclusions on this data because of the lack of diversity in our sample population. As described previously, we had large majorities in each of these categories. For instance, over two-thirds of our subjects were white. This made it difficult to get samples of the various minority groups that we could draw reasonable conclusions from. As a results, we tried to identify if there significant difference in effects of treatment in white and minority respondents. This showed no significant differences, but the individual minorities were not uniform in their responses. With a larger or more diverse

sample, we may see significant differences. More discussion of these various factors can be found in the appendix.

Other Explanations

After rating the algorithm and explanation, respondents were asked what else they wanted from the explanation. Our results show that relative importance of factors used in the decision was the most requested, followed closely by a detailed description of the algorithm. Finally, representative examples of other levels of output had only about half as many responses as the other two categories. The relative importance of factors is what we attempted to include in our explanation, so that seems to reinforce the results of our pilot study as well as the effectiveness of our treatment in our main experiment.

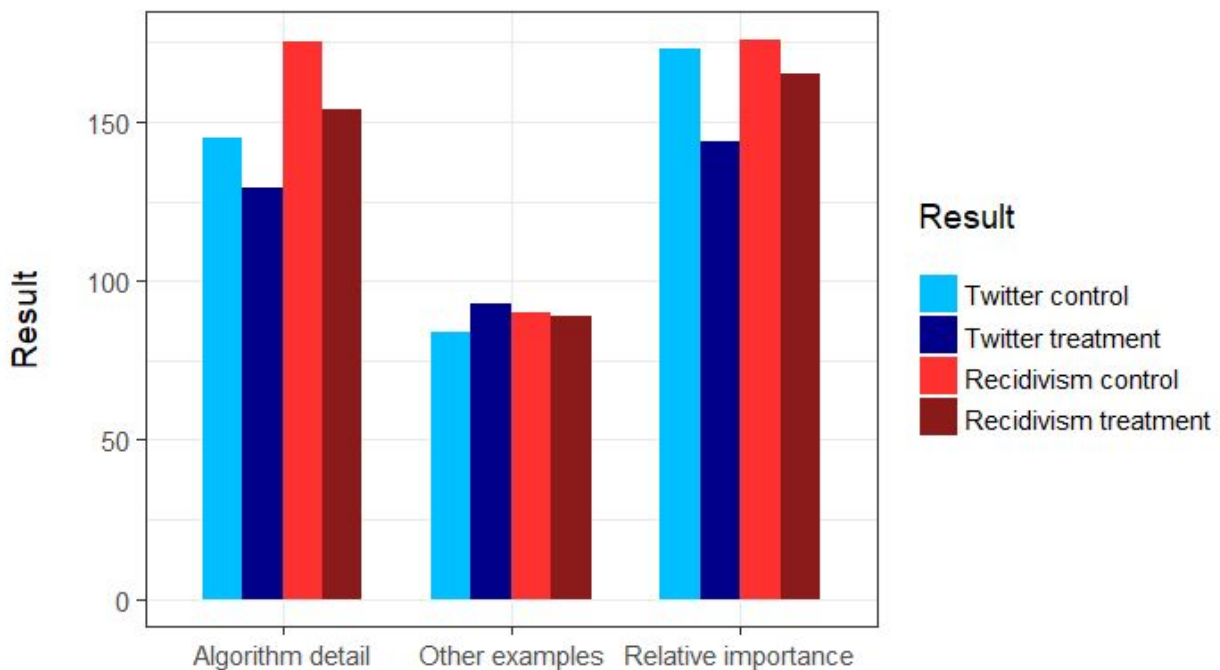


Figure 12: Additional Explanations Requested

We also included a free text input so respondents could give further feedback. Some respondents did not think that algorithms should be used at all and were not aware that such

practices were going on. Since part of our objective was education, this was a useful outcome. One respondent specifically was bothered by the lack of oversight in the decision and lack of appeals process.

In the moderation context, people thought that algorithmic moderation was a violation of free speech. Some suggested that individuals block an offensive tweeter rather than Twitter blocking his comments. Other respondents wanted to know about patterns of behavior of the Twitter user and prior offenses. We also saw feedback that respondents were confused by “sentiment analysis,” which in retrospect we should have clarified or used simpler language for the benefit of non-technical audiences. Some respondents were worried about the arbitrary nature of the 70% threshold. Future experiments could include a justification for the threshold.

In the recidivism context, people were more agitated by the use of algorithms. Some respondents suggested that a psychologist or social worker weigh in on the algorithm's output. Other respondents did not like that “criminal personality” was used in decisionmaking since it seemed arbitrary. The categories we used were examples from existing algorithms. One respondent asked for the source code and regression test results.

Overall, there was a wide variety of responses, including some people being completely satisfied with our explanation. The strongest themes were questions on why algorithms are used in these contexts at all as well as the seemingly arbitrary nature of some parts of the explanation.

Discussion

Before running the experiment, we felt good about our design but had some uncertainty about how some elements would be received. In our pilot, the tweet we showed was much more offensive, and the baseline fairness and accuracy measures were so high that the treatment

could not produce a much higher score. Seeing the results that we ended up with, we have several more questions about how the setup might be affecting the different contexts.

Of the explanations we looked at, the weighting of factors contributing to the decision shown in a clear manner performed the best. This was backed up by written responses to another question about what they would like to see in an explanation.

We would recommend changing the question format to a numeric scale (e.g. “How fair do you find this decision, on a scale of 1-10?”) rather than text answers. This would provide for better statistical analysis and a more easily interpretable output.

Ultimately, we saw the results that we hypothesized, so we feel confident in saying that a well formed explanation can help improve the trust and acceptance of algorithmic decisions, provided the decisions are accurate and justifiable in the first place. We also feel confident in showing that explanations become more important as the impact of a decision increases.

However, there are many pertinent questions left to be answered. It would be interesting to explore heterogeneous treatment effects of different explanations. This would be an expansion of our pilot. As stated earlier, we were interested to see what would happen if we changed the control from an algorithmic decision to a decision made by humans. It would also be interesting to explore explanations in additional contexts like recommender systems.

We had some additional questions for the contexts that we implemented. The recidivism context was difficult to operationalize. How would our results have changed if we had a different crime, if the defendant was rated as medium risk, or if we gave additional details of the case? How would our results have changed if we had a different tweet that instead of offensive against women was offensive to another group or if we changed how offensive our tweet was?

Overall, we think we chose a manageable research question and hypotheses, but would be interested in continuing to explore these avenues of research further.

Conclusion

We set out to test whether providing an explanation for an algorithmic decision influenced people's trust of the algorithmic decision and if this effect changed between Twitter moderation and recidivism risk assessments in courts. Using a Qualtrics survey and recruiting subjects from Mechanical Turk, we used linear regression models to see that providing an explanation improved both the perception of fairness, accuracy, and satisfaction with the algorithmic decision and the usefulness, clarity, and meaningfulness of the explanation. As we hypothesized, we saw a larger effect in the recidivism context than in the Twitter context.

Algorithm creators want people to trust their algorithms and algorithm consumers want to understand how algorithmic decisions are made. Providing algorithmic explanations is a net win both for the designers of algorithms and the consumers. This benefits the creators because people are more likely to perceive the algorithmic decisions as fair, accurate, and satisfactory. Explanations benefit the consumers of algorithms because they can better understand how and why a decision was made, as shown by their rating of explanations as more useful, clear and meaningful. As the GDPR roll out quickly approaches, we will see how automated decision making ends up in practice--what explanations are offered by tech companies and how people react to the explanations of these decisions. We look forward to seeing future research that explores new contexts, different levels of decisions (high/medium/low), or different modes of explanations.

As algorithmic decision become more commonplace, explanations for algorithms will need to evolve to inspire confidence in their decisions and provide explanations that allow humans to determine if they should trust the algorithmic decision.

Acknowledgements

Thank you to D. Alex Hughes and David Reiley for helping us with experimental design and to Nathan Good for helping us think through our research question.

References

- Administrative Office of the United States Courts Office of Probation and Pretrial Services (2011, September). An Overview of the Federal Post Conviction Risk Assessment). Retrieved from http://www.uscourts.gov/sites/default/files/pcra_sep_2011_0.pdf
- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23) Machine Bias <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- Burt, A. (2017, June 1). Is there a 'right to explanation' for machine learning in the GDPR? Retrieved July 31, 2017, from <https://iapp.org/news/a/is-there-a-right-to-explanation-for-machine-learning-in-the-gdpr/>
- Bolukbasi, T., Chang, K., Zou, J., Saligrama, V., & Kalai, A. (2016, July 21). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. Retrieved August 18, 2017, from <https://arxiv.org/abs/1607.06520>
- Chan, D. S., van Hemmen, S., Sharma, A., Shen, J., Todeschini, A., & Hughes, D. A. (n.d.). Empathization. Retrieved August 22, 2017, from <http://empathization.info/>
- Danker, J., Glenn, P., Mahar, M., Witt Advisor, E., & Mulligan, D. (2017). The Moderation Machine. University of California, Berkeley. Retrieved from https://www.ischool.berkeley.edu/sites/default/files/sproject_attachments/moderationmachine.pdf
- Datta, A., Sen, S., & Zick, Y. (2016, May). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In Security and Privacy (SP), 2016 IEEE Symposium on (pp. 598-617). IEEE.
- Duggan, M. (2014). Online Harassment | Pew Research Center. Retrieved from <http://www.pewinternet.org/2014/10/22/online-harassment/>
- Finley, K. (2017, June 29). The Real Impact of Google's Big EU Fine. Retrieved August 19, 2017, from <https://www.wired.com/story/google-big-eu-fine/>

- Geiger, R. S. (2016). Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society*, 19(6), 787–803. <https://doi.org/10.1080/1369118X.2016.1153700>
- Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., & Newell, C. (2012). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5), 441-504.
- Lipton, Z. C. (2016). The mythos of model interpretability. arXiv preprint arXiv:1606.03490.
- Loomis v. Wisconsin, No. 16-6387 Brief for the United States as Amicus Curiae (2017, May 23). Retrieved from <http://www.scotusblog.com/wp-content/uploads/2017/05/16-6387-CVSG-Loomis-AC-Pet.pdf>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135-1144). ACM.
- Skeem, J. & Lowenkamp, C. (2016, June 14). Risk, Race, & Recidivism: Predictive Bias and Disparate Impact. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2687339
- Stuart, Guy, Databases, Felons, and Voting: Errors and Bias in the Florida Felons Exclusion List in the 2000 Presidential Elections (September 2002). KSG Working Paper Series RWP 02-041. Available at SSRN: <https://ssrn.com/abstract=336540> or <http://dx.doi.org/10.2139/ssrn.336540>
- Tamagnini, P., Krause, J., Dasgupta, A., & Bertini, E. (2017, May). Interpreting Black-Box Classifiers Using Instance-Level Visual Explanations. In *HILDA@ SIGMOD* (pp. 6-1).
- Wachter S., Mittelstadt B., Floridi L.; Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law* 2017; 7 (2): 76-99. doi: 10.1093/idpl/ix005

Appendices

- Appendix A: Statistical Analysis
- Appendix B: Pilot Study Summary
- Appendix C: Survey Questions