frontiers in Big Data

HYPOTHESIS AND THEORY
published: 19 May 2020
doi: 10.3389/fdata.2020.00018

Theory In, Theory Out: The Uses of Social Theory in Machine Learning for Social Science

Jason Radford[1]* *and Kenneth Joseph*[2]*

[1] Department of Political Science, Northeastern University, Boston, MA, United States, [2] Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, United States

Research at the intersection of machine learning and the social sciences has provided critical new insights into social behavior. At the same time, a variety of issues have been identified with the machine learning models used to analyze social data. These issues range from technical problems with the data used and features constructed, to problematic modeling assumptions, to limited interpretability, to the models' contributions to bias and inequality. Computational researchers have sought out technical solutions to these problems. The primary contribution of the present work is to argue that there is a limit to these technical solutions. At this limit, we must instead turn to social theory. We show how social theory can be used to answer basic methodological and interpretive questions that technical solutions cannot when building machine learning models, and when assessing, comparing, and using those models. In both cases, we draw on related existing critiques, provide examples of how social theory has already been used constructively in existing work, and discuss where other existing work may have benefited from the use of specific social theories. We believe this paper can act as a guide for computer and social scientists alike to navigate the substantive questions involved in applying the tools of machine learning to social data.

Keywords: machine learning, computational social science, machine learning and social science, bias, fairness

OPEN ACCESS

Edited by:
Katja Mayer,
University of Vienna, Austria

Reviewed by:
Chao Lan,
University of Wyoming, United States
Luca Maria Aiello,
Nokia, United Kingdom

***Correspondence:**
Jason Radford
jasonscottradford@gmail.com
Kenneth Joseph
kjoseph@buffalo.edu

Specialty section:
This article was submitted to Data Mining and Management, a section of the journal Frontiers in Big Data

Received: *05 February 2020*
Accepted: *21 April 2020*
Published: *19 May 2020*

Citation:
Radford J and Joseph K (2020) Theory In, Theory Out: The Uses of Social Theory in Machine Learning for Social Science. Front. Big Data 3:18. doi: 10.3389/fdata.2020.00018

1. INTRODUCTION

Machine learning is increasingly being applied to vast quantities of social data generated from and about people (Lazer et al., 2009). Much of this work has been fruitful. For example, research using machine learning approaches on large social datasets has allowed us to provide accurate forecasts of state-level polls in U.S. elections (Beauchamp, 2017), study character development in novels (Bamman et al., 2014), and to better understand the structure and demographics of city neighborhoods (Cranshaw et al., 2012; Hipp et al., 2012). The increasing application of machine learning to social data has thus seen important success stories advancing our understanding of the social world.

At the same time, many (computational) social scientists have noted fundamental problems with a range of research that uses machine learning on social data (Lazer and Radford, 2017; Crawford et al., 2019; Jacobs and Wallach, 2019). For example, scholars have argued that machine learning models applied to social data often do not account for myriad biases that arise during the analysis pipeline that can undercut the validity of study claims (Olteanu et al., 2016). Attempts to identify criminality (Wu and Zhang, 2016) and sexuality (Wang and Kosinski, 2018) from people's faces and predicting recidivism using criminal justice records (Larson and Angwin, 2016) have led to

critiques that current attempts to apply machine learning to social data represent a new form of physiognomy (Aguera y Arcas et al., 2017). Physiognomy was the attempt to explain human behavior through body types and was characterized by poor theory and sloppy measurement (Gould, 1996). It ultimately served to merely re-enforce the racial, gender, and class privileges of scientists and other elites. Today it is considered pseudoscience.

Acknowledging these misappropriations of machine learning on social data, researchers have sought out technical solutions to address them. For example, in response to claims that algorithms embedded in policy decisions often provide unfair advantages and disadvantages across social groups, some scholars in the Fairness, Accountability and Transparency (FAccT) community have proposed new algorithms to make decisions more fair. Similarly, researchers in natural language processing have proposed several new methods to "de-bias" word embeddings' representation of gender, race, and other social identities and statuses (Bolukbasi et al., 2016).

The primary contribution of this paper is to put these challenges, criticisms, and searches for a solution into a single framework. Specifically, we argue and show that *at each step of the machine learning pipeline, problems arise which cannot be solved using a technical solution alone*. Instead, we explain how *social theory* helps us solve problems that arise throughout the process of building and evaluating machine learning models for social data. The steps in this process and an overview of how social theory can help us to perform the given step more effectively are outlined in **Figure 1**.

We define social theory broadly, as the set of scientifically-defined constructs like race, gender, social class, inequality, family, and institution, and their causes and consequences for one another. Using social theory in machine learning means engaging these constructs as they are defined and described scientifically and accounting for the established mechanisms and patterns of behavior engendered by these constructs. For example, Omi and Winant's (2014) *racial formation theory* argues that race is a social identity that is constantly being constructed by political, economic, and social forces. What makes someone "Black" or "White" in the United States and the opportunities and inequities associated with this distinction have changed dramatically throughout history and continues to change today. While there are other scientific definitions of race and active debates about its causes and consequences, engaging with them at each stage in the machine learning pipeline allows us to answer critical questions about what data we should use, features we should engineer, and what counts as fair (Benthall and Haynes, 2019; Hanna et al., 2019).

The paper is structured into two broad sections. In the *Theory In* section, we discuss how social theory can help us as we work through the model building pipeline. In the *Theory Out* section, we talk about a checklist of desiderata that we have for the models and results we produce, like generalizability, and discuss how social theory can help to improve these outputs of our work. Each subsection within Theory In and Theory Out focuses on a particular research problem or task and addresses a series of five questions:

1. What is the goal, or problem to be solved?
2. How have we tried to solve this problem computationally?
3. What are the limits to these technical solutions?
4. What solutions does social theory offer?
5. How can use social theories to solve these problems in our work?

In each subsection, we answer each question and use examples to illustrate our claims. Although each subsection addresses a unique element of Theory In or Theory Out, solutions identified in one step often enable us to address problems in others. For example, a model that is parsimonious is often more interpretable. A lack of a solution to a problem in one step can also prohibit a solution to issues that might arise downstream. At the highest level, a lack of social theory going into the model is likely to stymie drawing theory out. These overlaps are a strength, rather than a weakness, of the structure of this article. Like Olteanu et al. (2016), we believe that by emphasizing both the uniqueness and the critical relationships between different pieces of the pipeline, we can understand how the failure to address problems can propagate from one step to the next and ultimately impact what conclusions we draw from a study.

2. RELATED WORK

Social scientists have long established that theory can solve methodological and analytic issues that new techniques cannot. For example, Small (2017) has argued that theory alone can address questions of how best to measure what it means for one to have a close social tie. In the present work, we regularly draw on this literature, seeing many parallels between prior methodological innovations like linear models and sequence analysis (Abbott, 1988, 1995).

Other scholars working at the intersection of machine learning and social science have also proposed important critiques which we draw upon throughout the paper. These critiques fall into four broad categories. First, scholars have argued that many machine learning papers focus too heavily on prediction relative to explanation (Wallach, 2018) or measurement (Jacobs and Wallach, 2019). The prioritization of prediction over explanation leads to models that perform well for unknown reasons, leading to *ad hoc* justifications for model decisions and performance. Prioritizing prediction over measurement leads to a failure to acknowledge the always imperfect association between what we are trying to measure and what we are able to quantify.

Others, like Nelson (2017), argue that machine learning applied to social data must come hand-in-hand with the development of new social theory. From this perspective, we do not necessarily know what model will work for what data, nor do we typically have theory to tell us what to expect. Consequently, we need to create new theory as we develop and run methods. This approach helps us to understand why machine learning models might present results that at first seem unintuitive, but that do reflect genuine patterns that should force us to reconsider our understanding of the social

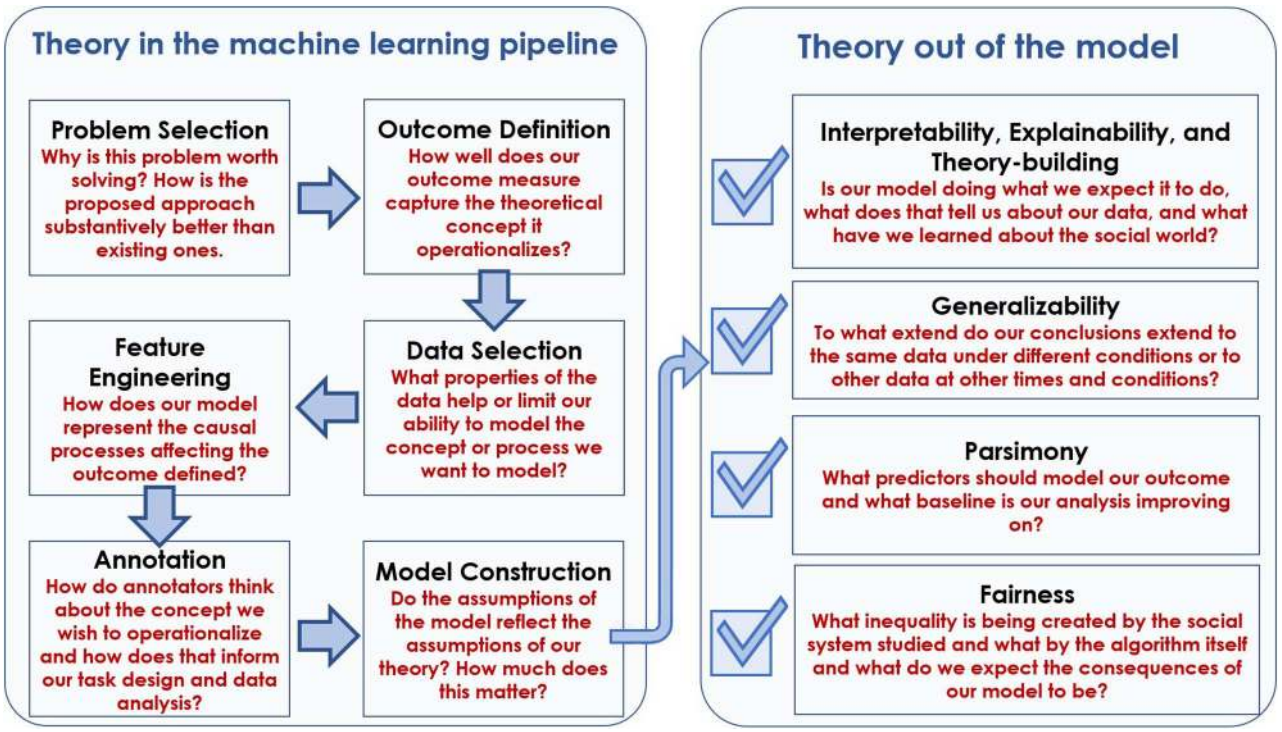

FIGURE 1 | An overview of the Theory In and Theory Out sections of the paper. Each box within Theory In and Theory Out is a subsection. The box gives the name of the subsection and the claim we make about how social theory can be used to improve our technical solutions for addressing that problem. We see Theory In (model building) as a pipeline, and Theory Out (model use) as a checklist of things we might potentially want from the model and/or it's outputs.

world. However, it also requires an a priori understanding of the potential theories that could apply, and seeks to adapt this existing theory, rather than create new theory entirely ex post-facto.

Still others have taken specific studies or sets of studies to task, arguing that they fail to understand the socio-technical context in which their data are produced (Lazer et al., 2009; Tufekci, 2014; Olteanu et al., 2016). For example, Tufekci (2014) argues that, despite generic claims above universal social behavior in many papers, research using Twitter data is unlikely to tell us much about social interaction on Facebook because the two have very different rules, data, and norms. Similarly, Olteanu et al. (2016) provide a dizzying array of potential pitfalls in analyzing social data, emphasizing the need for time-tested ideas from the social sciences, like matching and analyses of sample bias, to address these issues. These critiques point to the fact that machine learning is often done with one eye closed to the peculiarities of the data.

Finally, with the advent of algorithms that are making increasingly high-impact decisions, the FAccT community[1] has arisen to study how these algorithms can serve to reflect social biases embedded in the complex sociotechnical systems within which they are embedded, and how we might be able to address these issues. However, recent critiques of the fairness literature argue that far too much emphasis has been placed on technical "solutions" to unfair and/or "biased" algorithms, relative to the structural causes and consequences of those algorithms (Green, 2018; Crawford et al., 2019; Hoffmann, 2019). Such scholarship has argued that social science disciplines need be at the forefront of our understanding of how to address these root causes.

Each of these critiques—that prediction does not equal understanding, that we must be ready to construct new theory to interpret our results, that myriad biases lurk in the data and methods we use, and that these methods can result in discriminatory outcomes with systemic causes—is critical in pushing us toward a better, more responsible application of machine learning toward social data. Further, the many works reviewed below that apply machine learning to social data with these issues in mind provide a further glimpse into the potential for this version of the science.

In the present work, we seek to unify these critiques, arguing that each of them are levied at different pieces of the same underlying problem—attempts to use technology, or *ad hoc*, post-ex facto reasoning, to address problems only social theory can solve. We argue below that theory alone can lead us to the valid explanatory models sought by Wallach (2018), to ensure we draw correct conclusions from initially unintuitive results, to help us characterize dangerous assumptions in our data collection processes, and to help us understand and address discriminatory, biased, or unfair model behavior.

[1]http://fatconference.org

3. THEORY IN

In this section, we discuss the pipeline for studies that use machine learning to analyze social data, from problem conception through model selection.

Throughout the section, two broad themes arise. First, given the breadth of data available to us, we sometimes act opportunistically; we use what we have to jump quickly on new problems. This push to rapidly solve pressing social problems by using widely available data and methods leads us to, for example, use a dataset to answer a research question that the dataset is ill-suited for. Problems can arise from these decisions that entirely undermine the utility of the research—for example, selecting a bad sample of data can undermine the external validity of a study.

Second, we often rely on intuition to make decisions about research design. For example, when constructing annotation tasks, intuition can lead to overly simplified designs, when many other potential approaches could also be equally, or more, valid (Joseph et al., 2017a). Often, these intuitions are good enough for the task at hand. However, when our intuitions are wrong, the results can be problematic. For example, following misguided intuitions about sexuality and its causes can lead to incorrect claims, made on top of poor research design decisions, about the biological nature of sexuality (Wang and Kosinski, 2018).

This combination of opportunism and intuition can be particularly pernicious when combined with a lack of theory. While social scientists also often rely on intuition (Tavory and Timmermans, 2014), they can rely on the scaffolding provided by previous theoretical work, guiding them toward a better research design and/or understanding of their data. In section 3, we discuss how we can use social theory to help us constrain our opportunism and intuitions to existing knowledge of the social world provided to us by social theory. This increases our chance at producing new, lasting science that helps move forward our understanding of society.

3.1. Problem Selection and Framing

As researchers, we are constantly asking ourselves, "what problem should we be studying?"[2]

Unfortunately, while technical approaches can sometimes help identify oddities in social data worth investigating, there is no technical solution to identifying good social science research questions. These insights require an understanding of what is known already about the social world, and where gaps in this knowledge lie. However, with the onslaught of big data, we all too often optimize for convenience, using the data we have on hand to study problems just because they seem solvable, and because they seem to have real-world relevance. For example, we use publicly available Twitter data to predict people's movements within cities (Bauer et al., 2012) or aggregated search data from Google Trends to predict the prevalence of the flu (Lazer et al., 2014).

This convenience approach to problem selection and framing leads to two problems. First, it can lead us to formulate and tackle problems that seem important but in reality serve chiefly as an exercise in prediction, providing little new insight into the social world. Second, it can lead us to address problems that our intuitions accurately assume are important, but leave us struggling to frame the reasons *why* the problem is important. Social theory can help to alleviate these issues.

First, theory tells us which problems are worth solving. For example, election prediction is an essential research tool because it provides a model for understanding political processes (Beauchamp, 2017; Kennedy et al., 2017). However, theory tells us that because of polarization, gerrymandering, and campaign finance laws, most American elections today are very predictable with only one piece of data—knowing who the incumbent is. Theory also tells us, however, that in nominally competitive races, *polling* provides the next best predictor, because politics is driven by opinion. However, polling is expensive, and s only available for the most high-profile races. Theory thus suggests that within the domain of elections, the correct problem to study is modeling opinion in competitive and under-polled elections.

Second, theory can help us to motivate and frame problems that seem intuitively important. It may be apparent that predicting the prevalence of the flu can help save lives. However, less obvious is what problem is being solved when predicting, for example, a person's political affiliation based on their social media behavior (e.g., based on their tweets) (Cohen and Ruths, 2013). However, recent work on political polarization urges us to study affiliation as a function of partisan identity (Levendusky, 2009), and shows that such identities are rapidly undermining social and cultural stability in the United States (Doherty, 2017). Social theory therefore explains why predicting political affiliation is important—in order to study its association with cultural polarization (DellaPosta et al., 2015).

Thus, while there may be situations in which the problem to be addressed can be motivated solely by the need for increased accuracy (e.g., correctly identifying a consenting individual's location information from WiFi signals), many machine learning problems can be made more interesting and relevant when grounded in underlying theory about the social behavior under study. There are many examples where scholars using machine learning on social data have used theory to identify and frame important problems. For example, several scholars have addressed precisely the problem of opinion polling in competitive and under-polled elections using big data (Wang et al., 2015; Beauchamp, 2017). And Cranshaw et al. (2012) take an intuitively interesting task—clustering venues on foursquare by the patrons they have in common—and ground it in theory from urban studies on the meaning of the term "neighborhood" to motivate and frame their work as addressing an unsolved problem of how neighborhood boundaries should be defined.

3.2. Outcome Definition

Having established a problem of interest, we turn to the task of defining and measuring our outcome. Our outcome measure is ideally meant to be the ground truth for our phenomenon we're modeling, i.e., an observation of the phenomenon itself.

[2]Importantly, not all researchers ask this question. Others, for example, may seek to understand the consequences of the practices they are studying. Thank you to Stef Shuster for pointing this out.

For example, if we are interested in studying partisanship, we can establish ground truth through a variety of means—whether someone votes for a single party (Poole and Rosenthal, 1991), who they donate money to (Bonica, 2014), or what topics they tweet about (Tsur et al., 2015). Unfortunately, this data is often not easily available. The easiest "technical solution" to this is simply to use a variable available in our data as ground truth, or, as we discuss in section 3.5, to construct the variable through a rapid crowdsourced annotation task.

However, this technical solution fails to help us fully characterize the link between the variable we select as an outcome and the concept we are interested in studying. For example, no technical solution can determine whether voting behavior or political sentiment in tweets is a more valid measure of partisanship (Cohen and Ruths, 2013). Answering these kinds of questions requires social theory. In this case, theory is needed to help identify what we mean by partisanship, or more specifically, by liberal vs. conservative. In turn, we must therefore approach ground truth as something being theorized by researchers. It therefore makes sense to do so in a way that existing social theory tells us is valid in capturing the construct we seek to measure (Hacking, 1986).

Returning to liberalism and conservatism, for example, political theories of partisan sorting and ideological alignment shows that people and sociotechnical systems shape the "ground truth." Only recently have liberal and conservative labels for partisanship aligned with the Democratic and Republican parties in the United States—what is called partisan sorting (Mason, 2015). For example, Gentzkow et al. (2016) show that partisan ideology has only become distinguishable in Congressional floor speeches since 1980. That is, language has only become partisan in the past 40 years.

These theoretical insights, in turn, help us create a valid outcome. Instead of predicting liberalism/conservatism, a measure that has only recently come to align with partisanship, partisan identity theory (Van Bavel and Pereira, 2018) suggests we should instead focus on if someone is a Democrat or a Republican. Theory can further explain how to identify this outcome of interest in social media data. Specifically, partisan identity theory claims that party membership is driven by party identification. What makes someone a Democrat is not that they support public health care or market regulation but that they identify with Democrats. Thus, if we want to infer someone's political party identification from their tweets, we should look to whose side they take in a debate, rather than the specific issues they support. In his campaign, Donald Trump famously supported liberal policies like public health care and criticized the war in Iraq. These stances did not make him a moderate conservative. They made him a populist Republican as opposed to an establishment Republican.

3.3. Data Selection

The process of data selection is defined as the identification of one or more datasets that can be used to address the problem under study. Data selection is typically carried out using either precedent (i.e., using existing data) or convenience (i.e., using easily collectable data) as a heuristic.

This use of precedence and convenience stems from our interest not only in answering questions about the social world, but in desiring to do so via novel methodologies. For example, when constructing novel solutions to existing problems, we tend to reach for established datasets for which prior results exist for comparison. And for novel data collection, our methods often require large datasets, and so convenient collection of this data is almost a prerequisite.

But relying on either convenience or precedent can cause issues for social science questions, because all data contain both inclusions and exclusions that manifest in varying forms of bias (Olteanu et al., 2016). By taking shortcuts with data selection, we often choose to ignore or brush over these inclusions and exclusions. For example, Blodgett et al. (2016) show that language identification tools shown to perform well on a wide array of text corpora (Lui and Baldwin, 2012) suffer significantly at distinguishing African-American English as English in social media texts. Because scholars have often used this tool to filter out non-English tweets, the result is a set of studies on social media data where the voices of African Americans are diminished relative to white Americans.

As Blodgett et al. (2016) suggest, socio-linguistic theory could have helped us to anticipate the potential issues in using the convenient language classifier they studied to identify English vs. non-English content. Specifically, theoretical models of how dialects form emphasize that variations of written English may not readily align in terms of the primary features used by the language classification model, n-grams of characters (Rickford and Labov, 1999). Further, socio-linguistic theory emphasizing the importance of African American English and its distinctions from other English dialects in the presentation of the self online for Americans (Florini, 2014) would have emphasized the need for social media scholars to reconsider the notion that there is a single definition of English that they wish to study.

More broadly, then, social theory helps us to understand the implications of how we make our decisions on what data to include or not include in our sample. This is especially critical when we expect others will reuse our data or the models we construct from them. For example, a significant amount of research has used pre-trained word vectors from an unreleased Google news corpus, meaning the biases in the data are both unclear and unknown. On the contrary, Lundberg et al. (2019) use the statistical sampling and survey measurement theories baked into the Fragile Families Well-being Study to create the Fragile Families Challenge—a common data set computational social scientists can use to develop models predicting critical social factors like income, health, and housing. The use of theory to identify and explain important inclusion and exclusion variables have allowed research conducted during the challenge to contribute successfully to social scientific knowledge on child development (Salganik et al., 2019).

3.4. Feature Engineering

Feature engineering encompasses the process of converting our raw data to quantities that can be input into a machine learning model. The key question is, of course, how do we know that we have engineered the right features?

A technical solution to this question has typically privileged model performance. The problem with this approach to feature engineering is that the features we select might boost our performance but may not help us distinguish genuine from spurious signal. Overfitting to noise is one way in which injudicious feature selection can inflate performance. Another is to include features that are spuriously related with our outcome of interest or exclude features that are directly related.

Take the case of recidivism prediction as an example. To predict who will return to prison, not only do we need features that signal a person's propensity to commit a crime, but also features that capture police and judicial processes around who is likely to be arrested and convicted of a crime. For example, Sudnow's concept of "Normal Crimes" captures how the daily work of prosecution routinizes how certain kinds of cases from certain kinds of defendants are processed, in particularly who gets what plea agreements and whether jail time is recommended (Sudnow, 1965). Omitting features capturing both crime commission and criminal conviction yields a poorly-specified model that performs well.

Automated causal inference at scale is an as-yet unattained holy grail in machine learning. Thus, without theory, we cannot enumerate which features we should include and which we should exclude. Specifying the theoretical model in advance is the only way to enumerate what features we should generate (Rohrer, 2018; Pearl, 2019). Building theoretical models allow us to identify which features should be included, and if they are deemed important by a model, what they might mean. More concretely, Wallach (2018) argues that we should always be informing our selection of features and understanding of the problem with theory-based causal models in mind.

This argument is, of course, at odds with claims of "featureless" models, as many claim deep learning models to be. For example, where before we may have needed to provide a model for Named Entity Recognition with part-of-speech tags for each input word, modern deep learning architectures do not require this feature engineering step (Goldberg, 2016). However, even with such models, we are still making implicit decisions about our features, for example, by deciding whether to use words or characters as input to the model (Devlin et al., 2018). Further, the causal processes of interest often lay beyond decisions on whether or not to use words or characters. For example, regardless of what deep NLP model we choose to model an individual's language, word choices are often driven by more difficult-to-capture variables, like age and gender (Schwartz et al., 2013).

3.5. Annotation

Oftentimes we cannot identify some critical feature we want to model from our data. For example, Twitter does not provide data on the gender or religious affiliation of their users. In such cases, we often ask humans, be it ourselves or others, to go through the data and identify the feature of interest by hand.

When annotating data, a primary goal is to ensure that annotators agree. Due to both noise and intrinsic variation amongst individuals, different people look at the same data and come up with different labels. Our interest, particularly when searching for some objective ground truth, is to ensure that despite these differences, we can identify some annotated value on which most annotators roughly agree. Scholars in the social sciences have long established statistical measures of agreement in annotation (Krippendorff, 2004), which are readily used in the machine learning pipeline. However, machine learning researchers have also sought to increase agreement in various ways (Snow et al., 2008). These technical efforts to increase agreement largely rely on either trying to find the best annotators [i.e., those that tend to agree most frequently with others (Ipeirotis et al., 2010)], finding better aggregation schemes (Raykar et al., 2010; Passonneau and Carpenter, 2014), or simply by increasing the amount of data labeled (Snow et al., 2008).

At the core of many disagreements between annotators, however, is that the constructs we are seeking to annotate are often difficult to define. For example, Joseph et al. (2016) built a classifier to identify social identities in tweets, a concept that is notoriously varied in its meaning in the social sciences. Thus, even experts disagree on exactly what a social identity constitutes. Unsurprisingly, then, Joseph et al. found that non-expert annotators provided unreliable annotations, even after a discussion period. Annotations of hate speech have seen similar struggles, with limited agreement across annotators (Davidson et al., 2017) and with significant differences across annotators with different demographics (Waseem, 2016).

In such cases where the construct is difficult to define, technical solutions like adding more annotators or performing different aggregation schemes are unlikely to increase agreement. This is because, as with outcome definition, technical solutions cannot address the fundamental issue—defining the construct itself. In other words, technical solutions cannot be used to answer the questions, "what is a social identity?" Or, "what is hate speech?" Instead, we must rely on theory to provide a definition. For example, Affect Control Theory in sociology focuses not on the general idea of social identity, but rather on "cultural identity labels," defined as "(1) the role-identities indicating positions in the social structure, (2) the social identities indicating membership in groups, and (3) the category memberships that come from identification with some characteristic, trait, or attribute" (Smith-Lovin, 2007, p. 110). Upon using this definition, and annotations from Affect Control theorists, Joseph et al. (2016) noted a significant increase in annotation quality.

Annotation, particularly with complex phenomena like identity, hate speech, or fake news (Grinberg et al., 2019), therefore requires starting with a theory of the construct we wish to measure and its intersection with the subjective processes of our annotators. One additional tool worth noting for this task that social scientists have developed is *cognitive interviewing* (Beatty and Willis, 2007). Cognitive interviewing involves talking to potential annotators about how they think of the construct, its potential labels, how they would identify those labels, and then having them actually try to apply our task to some test data. While similar to the idea of a pilot annotation task that machine learning researchers are likely familiar with, cognitive interviewing outlines specific ways in which theory can be applied before, during, and after the pilot to help shape the definition of the construct. Finally, although beyond the scope of the

present work, it is also critical that annotation follows best methodological practices for structured content analysis in the social sciences (Geiger et al., 2019).

3.6. Model Construction

In building a machine learning model for social data, our goal is to predict, describe, and/or explain some social phenomenon. Our job, then, is to identify the model that best accomplishes this goal, under some definition of best. Our challenge is to determine which of the many modeling approaches (e.g., a deep neural network vs. a Random Forest) we can take, and which specific model(s) (e.g., which model architecture with what hyperparameters) within this broad array we will use for analysis.

It can be, and often is, overwhelming to select which model to use for a given analysis. Consider, for example, the goal of understanding the topics in a corpora of text. The early topic modeling work of Blei et al. (2003), has been cited over 28,000 times. Many of these citations are from extensions of the original model. For example, there are topic models for incorporating author characteristics (Rosen-Zvi et al., 2004), author characteristics and sentiment (Mukherjee, 2014), author community (Liu et al., 2009), that deal specifically with short text (Yan et al., 2013), that incorporate neural embeddings of words (Card et al., 2017), and that emphasize sparsity (Eisenstein et al., 2011). How do we construct a model that is best, or right, for our analysis?

O'Connor et al. (2011) describe this kind of modeling choice as occurring along two axes—computational complexity and domain assumptions. Computational complexity is used loosely to represent complexity in computational time and "horsepower." Domain assumptions vary from few assumptions, essentially assuming "the model will learn everything," to cases where we explicitly model theory. However, O'Connor et al. leave open the question of where in this space the "right" model for a particular problem is likely to fall, or how to define the right domain assumptions.

This is where theory comes in. By defining the goal of the model—prediction, explanation, description, and so on; and providing clear expectations for what our domain assumptions are, theory helps us navigate the computation/domain space. In the context of topic modeling, the Structural Topic Model (STM) (Roberts et al., 2013, 2014) provides a generic framework for defining our domain assumptions based on the factors we expect to be important for shaping the topics that appear in a document. By incorporating covariates into the modeling process that we theorize to be relevant, we can leverage theory both to create a model that "fits the data better," and get outputs of the model that we can use to directly test extensions to our theory. The right model, then, is defined by theory. For example, Farrell (2016) uses theories of polarization through "contrarian campaigns" that originate in well-funded organizations to determine a particular instantiation of the Structural Topic Model that they use to study how polarization has emerged on the topic of climate change.

The STM is therefore useful in that, given an established set of generic modeling assumptions and a defined level of computational complexity, we can use theory to define the specific model we construct. Similar efforts have been made in other areas of text analysis as well. For example, Hovy and Fornaciari (2018) use the concept of homophily, that people with similar social statuses use similar language, to retrofit their word embedding model. This theory-driven change allowed the model to leverage new information, yielding a more performant model. As such, the use of theory to guide natural language processing models can serve as a blueprint for the application of theory in other domains of social data.

4. THEORY OUT

Machine learning has traditionally concerned itself with maximizing predictive performance. This means that the first results reported in machine learning papers, those in "Table 1," are often a report on the model's predictive performance relative to some baselines. However, scholars are increasingly interested in other aspects of model output, like interpretability and fairness. In applied research, it is important for scholars to demonstrate that their model helps us understand the data and explains why particular predictions are made. These new demands for the output of machine learning models create problems for which technical solutions have been proposed. In this section, we argue that this technical innovation is insufficient on its own. We must engage with relevant social theories if we are to use our models to learn about social world.

4.1. Interpretability, Explainability, and Theory-Building

Few criticisms have been leveled against machine learning models more than the charge that they are uninterpretable. While a concrete definition of interpretability has been elusive (Lipton, 2016), the general critique has been that machine learning models are often"black boxes," performing complex and unobservable procedures that produce outputs we are expected to trust and use. In trying to open the black box and account for our models, three distinct questions are often treated interchangeably:

- *What did the model learn, and how well did it learn it?* Meaning, given a particular input, how does the model translate this to an output and how accurately does this output match what we expect? We refer to this as the question of **interpretability**.
- *Why did the model learn this?* What is it about the (social) world that led to the model learning these particular relationships between inputs and outputs? We will refer to this as the question of **explainability**.
- *What did we learn about the world from this model?* What new knowledge about the social world can be gleaned from the results of our model? We refer to this as the question of **theory-building**.

Interpretability, explainability, and theory-building get lumped together in the technical solutions that have been developed to open the black box. For example, sparsity-inducing mechanisms like regularization (Friedman et al., 2009) and attention (in neural networks; Vaswani et al., 2017) increase interpretability by minimizing the number of parameters to inspect. In turn, these

technical solutions are used help us explain how the parameters relate to the data generating process (Zagoruyko and Komodakis, 2016). We also use model-based simulations, tweaking inputs to show how they produce different outputs (Ribeiro et al., 2016) and adversarial examples that fool the model to explore its performance (interpretability) the limits of its understanding about the world (theory-building) (Wallace et al., 2019).

However, while there are many methodological overlaps; interpretation, explanation, and theory-building are distinct research questions requiring different uses for social theories.

When interpreting models, social theory enables us to go beyond the technical question of *how* to look at the model to *what to look at*. In order to choose what parts of the model to visualize, we need to have pre-defined expectations about how the model is *supposed* to work and what it is *supposed* to do based on theories about how the phenomenon we're studying is represented in our data. For example, social theory, like Sen and Wasow's model of race as a "bundle of sticks" (Sen and Wasow, 2016), tells us that race is constituted by many different dimensions beyond just skin color. For example, racial bias driven by skin tone, called "colorism," is different from racial bias driven by cultural codes like accent and hair style (Todorov et al., 2008). Consequently, if we want to understand how race is represented in a computer vision model, we should look to the different dimensions along which race is constructed. This can help differentiate, for example, whether biases in the model derive from cultural norms embedded in the population that make up the training data or from under-representation of individuals with certain skin tones in the training data (or both) (Benthall and Haynes, 2019; Hanna et al., 2019).

A good example of how theory can be used to guide interpretation is the work from Bamman et al. (2014), who identify tropes of literary characters. They validate their model by testing specific parameters against a slate of theory-based hypotheses. These hypotheses, derived from theory on the writing styles of authors during the time period of study, took the form of "character X is more similar to character Y than either X or Y is to a distractor character Z." Good models were those that accurately predicted these theorized character similarities.

Oftentimes, we take our interpretation of model behavior and develop an account of why the model did what it did based on that interpretation. This often serves as an explanation of what the model did and an effort to build new theory. However, when we build explanations based only on model behavior, we are at risk of developing *folk theory* (d'Andrade, 1995). Folk theory involves leaning on common understanding to "read the tea leaves" (Chang et al., 2009) characterizing human behavior as simply "making sense." This is dangerous, however (Kerr, 1998). Models will always output *something* and some model will always outperform others on some metrics. Building theory only from model output often serves to reinforce myths and biases.

For our explanations to contribute to a broader understanding of the social world, we need to not only find the right explanation for each model, but to also integrate many models and explanations into a coherent account of the world. Nelson's work on the development of second wave feminism is a prime example. She used social network and feminist theory to build different machine-learning based models for the structure of feminist communities in New York and Chicago. She then compared the structures of the social and idea networks to show that the ideas central to feminist community in New York were more aligned with what we understand today to be "second wave" feminism and that their community was more densely connected than that in Chicago. She argues this dense connectivity enabled feminists in New York to set the agenda for feminism in the 1960s and 70s.

Nelson and Bamman et al.'s work also provide a blueprint on how machine learning can help us to revise old or build new theory given empirical results from a machine learning model. To do so, their work tells us, one must first acknowledge the existing theoretical frames that have been used to characterize the problem. Following their work, one way to do so is to use these theories to generate hypotheses about what empirical results might look like, and to provide alternative hypotheses for what results might look like of a new or revised theory was instead true.

Another way to do so is to build a machine learning model that matches the theoretical model, and to then show that adding additional components to the model, inspired by new or revised theory, improve the performance of that model. For example, Joseph et al. (2017b) show that Affect Control Theory's (Heise, 2007) model of stereotyping may be insufficient by incorporating additional model components based on cognitive theories of stereotyping based on parallel constraint satisfaction models (Kunda and Thagard, 1996).

4.2. Generalizability

Generalizability refers to the goal of understanding how well results apply to the cases that were not tested. For example, if we develop a model to predict unemployment using mobile phone data in Europe (Toole et al., 2015), an analysis of generalizability might involve assessing whether the same approach would work in Algeria, Canada, or Mexico or on other kinds of data like internet searches or transit data.

In machine learning, generalizability is often addressed technically by reapplying the same methodology to other data to see whether it performs similarly to the original. For example, the generalizability of a topic model might be tested by applying a fitted model to different kinds of data. We also test the generalizability of a particular analytic approach by reapplying it in different domains. For example, Lucas et al. (2015) use machine translation across multiple languages to study whether politics in different countries were constituted by the same issues being discussed in the same ways. Finally, recent efforts have been made to train a model that learns representations of some universal input which can then be fine-tuned to apply to a variety of problems. For example, ResNet (Szegedy et al., 2017) and BERT (Devlin et al., 2018) learn generic representations for images and sentences, respectively, and can then be fine-tuned for various classification tasks.

While these technical solutions can make individual models more generalizable, they cannot help us establish why a result on one dataset can be generalized (or not) to others. For this, we need theories that tell us what similarities and differences are salient. Tufekci (2014) makes this point when arguing that we cannot treat one online platform (i.e., Twitter) as a stand-in for all others—as a *model organism* for society. The platform

rules, social dynamics, and population that make Twitter worth engaging in for its users also distinguish it fundamentally from services like Facebook, Instagram, and WhatsApp. For example, theories of homophily suggests that, on any platform, people will associate with others like them. Yet, the commonalities on which we build connections depend on the platform itself. Our friends, colleagues, and public figures are on Twitter and our family is on Facebook. Following Goffman's theory of presentation of self, these differences in audiences drive people to behave differently on different platforms (Goffman, 1959).

Of course, there is no such thing as the perfect dataset, and science must be able to proceed in spite of this. Social theory can be used moving forward not as a way to find perfect data, but rather as a way to develop paradigms for understanding the particular strengths and weaknesses of different kinds of data, like data from different social media platforms, and for how models might be tweaked to generalize beyond the imperfect data they were trained on.

4.3. Parsimony

Parsimony refers to the goal of building a model with as few parameters as possible while still maximizing performance. Machine learning models benefit from parsimony because it decreases complexity and cost, limits the danger of overfitting, and makes it easier to visualize (Hastie et al., 2009).

A variety of technical approaches for constructing parsimonious models exist. For example, we can use regularization, topics, or factoring to reduce feature dimensionality. In the case of neural networks, we also use techniques like drop-out (Gal and Ghahramani, 2016) or batch normalization (Ioffe and Szegedy, 2015).

There are, however, three common flaws with these technical approaches. First, because many features are correlated with one another and the outcome, these approaches often arbitrarily select certain correlated features and not others. This arbitrary selection can make it difficult to differentiate between truly irrelevant features and those that are simply correlated with other relevant features. Second, decisions on when the model is "parsimonious enough" rely largely on heuristic comparisons between model performance on training and validation data [e.g., the "1-Standard Error rule" used in the popular `glmnet` package in `R` (Friedman et al., 2009)]. Finally, the standard machine learning assumption that we need many features can be incorrect even at relatively low values. It is often the case in social science problems that a small set of variables can easily explain a large part of the variance. A regularizer may select 1,000 features out of 10,000 while the best model may only need 50.

Social theory provides a solution to these issues by helping us define small sets, or "buckets," of variables that we expect to explain a large portion of the variance in the outcome. Theories point us in the direction of the most important variables. Instead of starting with many features and trying to weed out irrelevant ones, we can use theory to create a baseline level of predictability from which we can assess whether additional features provide additional performance. Similarly, because theory provides us with the features we expect to be important, we may be able to identify cases in which regularization removes important, stable predictors due to correlation amongst variables.

The idea of identifying parsimonious, theoretically-informed baseline models for comparison has been shown to work well in practice. Theories of network centrality and homophily have proven to be robust predictors on a variety of tasks. For example, in their study of Twitter cascades, Goel et al. (2015) show that a simple model which accounts only for popularity of the user is an extremely strong baseline for predicting the size of a retweet cascade. These ideas align with theories of source credibility (Hovland and Weiss, 1951) and information spreading (Marsden and Friedkin, 1993).

Efforts to push the limits of predictability have informed the development of more formal social theory on the limits of predictability in social systems (Hofman et al., 2017), which may further extend our ability to estimate the degree of parsimony expected for particular problems. For example, in the Fragile Families challenge, the best submissions using thousands of variables and various models were not very good at predicting life outcomes like GPA, material hardship, and grit and were only marginally better than baseline models using only four variables (Salganik et al., 2020). In considering parsimony moving forward, we need to better understand the cases when the tools of machine learning add substantively to our model of the world beyond existing theory.

4.4. Fairness

In both popular media (Li, 2019) and academic literature (Mitchell et al., 2018), significant attention has turned to the question of how machine learning models may lead to increased discrimination against, or *unfairness* toward, certain social groups. The bulk of the work to ensure fairness has focused on making the input data more representative or modifying existing models to ensure fair outcomes (Kamishima et al., 2011; Kearns et al., 2017). Scholars have also recently focused on developing measures that account for sociologically relevant phenomena like intersectionality[3] (Foulds and Pan, 2018), on the tradeoffs between existing measures (Kleinberg, 2018), and on a better understanding of the causal assumptions of different measures (Glymour and Herington, 2019) amongst other tasks.

However, as argued by a rash of recent work, there are important complications to defining fairness technically (Crawford, 2016; Barocas et al., 2017; Green, 2018; Selbst et al., 2018; Hoffmann, 2019; Mitchell et al., 2020). First, different people have different views on what is fair. Second, the views of those in power are the views that are most likely to be used. Third, models emerge from a vast and complex sociotechnical landscape where discrimination emerges from many other places beyond the models themselves. Finally, "Fairness" may not be the correct metric along which the harms of algorithms should be quantified. One conclusion has been that a fair algorithm cannot fix a discriminatory process. For example, recidivism prediction algorithms will almost certainly be used in a discriminatory fashion, regardless of whether or not the models themselves are fair (Green, 2018). We need social theory, e.g., critical race theory (Hanna et al., 2019), to better understand the social processes in which these algorithms and the data they are based on, are embedded. As this prior work has argued, social theory enables us

[3] Although see the critique from Hoffmann (2019).

to distinguish discrimination caused by the algorithm from that originating in the social system itself.

Perhaps equally important, theory can also help us to understand the *consequences* of unfair and/or biased algorithms. Take, for example, recent work showing that search algorithms return gender and race stereotypical images for various occupations (Kay et al., 2015). Social psychological theories, e.g., the Brilliance Hypothesis (Bian et al., 2017), focusing on representation emphasize that from a young age, we internalize representations of occupations and skills that cause us to shift toward those that are stereotypical of our own perceived gender. Thus, while technical solutions may help us to identify such problems, they cannot explain the impacts of these biases and thus why they should be addressed and how.

Finally, social theory helps to identify how unfair machine learning impacts our knowledge about the world. Biased algorithms, such as those that detect gender and race for demographic comparisons (Jung et al., 2017), can bias the science we produce. Standpoint theory and other critical epistemological theories have shown how who does science and whose data are used for what analysis affects what we know about the social world (Haraway, 1988; Harding, 2004). We do not want to replicate the patterns of exclusion and stigmatization found in the history of medicine (Martin, 1991), psychology (Foucault, 1990), and sociology (Zuberi and Bonilla-Silva, 2008) by throwing out data from marginalized people, only studying marginalized people as the Other, or not allowing marginalized people speak for themselves about their data.

Recently, similar critiques have been made by Jacobs and Wallach (2019). They argue that measurement theory, a particular domain of social theory engaging in the validity and reliability of different ways of measuring social constructs, can provide a concrete and useful language with which different definitions of fairness, and the impacts of algorithms, can be assessed. Their work provides an important example of how social theory can be used to bring old, socio-theoretic perspectives to bear in an area of current research in machine learning on social data.

5. CONCLUSION

The combination of machine learning methods and big social data offers us an exciting array of scientific possibilities. However, work in this area too often privileges machine learning models that perform well over models that are founded in a deeper understanding of the society under study. At best, this trade-off puts us in danger of advancing only computer science rather than both computer science and social science. At worst, these efforts push the use of machine learning for social data toward pseudoscience, where misappropriated algorithms are deployed to make discriminatory decisions and baseless social scientific claims are made.

However, as the many positive examples we have highlighted here show, machine learning and big social data can be used to produce important, ground-breaking research. To do so, the examples we highlight have baked social theory into each step of the machine learning pipeline. These works do not cherry-pick one theory, ex post-facto, to support their claims. Instead, they use multiple, potentially competing theories, at every step of the pipeline, to justify their inputs and help validate their outputs. In using, or at least acknowledging, competing theories, we can elucidate where disagreements exist and therefore which technical trade-offs are most important.

The positive examples we highlight, our review of negative examples, and the related work we draw on pave the way forward for the scientifically-grounded, ethical application of machine learning to social data. But our efforts must move beyond the way we produce research to the ways we review it, consume it, and encourage it as a research community. As reviewers, for example, we must ask ourselves if the work we are looking at is justified not only by statistical theory, but by social theory as well. And as a community, we must find ways to feature and promote papers that may not have the flashiest "Table 1," but that provide a careful and well-grounded social scientific study.

Machine learning can and should become a critical piece of social science. The solution does not necessarily require a computer scientist to "go find a social scientist," or vice versa. There is already a wealth of knowledge to draw from, and we should not allow ourselves or others to avoid delving into it simply because it is "out of our field." For those who do not know where to start, we hope this paper is a guide to anyone for how to use that knowledge to address specific questions in the research. Similarly, social science should become an increasingly important part of machine learning. To be sure, certain problems faced by machine learning are computational issues (e.g., how to efficiently sample from a complex distribution) for which social theory will be of little use. But in incorporating social theory into their work, machine learning researchers need not reliquish model performance as the ultimate goal; we have argued here that, instead, theory can help guide the path to even better models and predictive performance.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

KJ was supported by an NSF grant, IIS-1939579.

ACKNOWLEDGMENTS

We greatly appreciate the assistance of Laura Nelson, Celeste Campos-Casillo, Stef Shuster, Atri Rudra, Emiel van Miltenburg, and David Lazer, who all provided invaluable feedback on earlier versions of this work. That said, these individuals of course bear no responsibility for the current content, all issues, errors, and omissions are the fault of the authors alone.

REFERENCES

Abbott, A. (1988). Transcending general linear reality. *Sociol. Theory* 6:169. doi: 10.2307/202114

Abbott, A. (1995). Sequence analysis: new methods for old ideas. *Annu. Rev. Sociol.* 21, 93–113. doi: 10.1146/annurev.so.21.080195.000521

Aguera y Arcas, B., Mitchell, M., and Todorov, A. (2017). *Physiognomy's New Clothes*. Available online at: https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a (accessed May 6, 2017).

Bamman, D., Underwood, T., and Smith, N. A. (2014). "A bayesian mixed effects model of literary character," in *Proceedings of the 52st Annual Meeting of the Association for Computational Linguistics (ACL'14)* (Baltimore, MD). doi: 10.3115/v1/P14-1035

Barocas, S., Boyd, D., Friedler, S., and Wallach, H. (2017). Social and technical trade-offs in data science. *Big Data* 5, 71–72. doi: 10.1089/big.2017.29020.stt

Bauer, S., Noulas, A., Seaghdha, D., Clark, S., and Mascolo, C. (2012). "Talking places: modelling and analysing linguistic content in foursquare," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)* (Amsterdam), 348–357. doi: 10.1109/SocialCom-PASSAT.2012.107

Beatty, P. C., and Willis, G. B. (2007). Research synthesis: the practice of cognitive interviewing. *Public Opin. Q.* 71, 287–311. doi: 10.1093/poq/nfm006

Beauchamp, N. (2017). Predicting and interpolating state-level polls using twitter textual data. *Am. J. Polit. Sci.* 61, 490–503. doi: 10.1111/ajps.12274

Benthall, S., and Haynes, B. D. (2019). "Racial categories in machine learning," in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (New York, NY: ACM), 289–298. doi: 10.1145/3287560.3287575

Bian, L., Leslie, S.-J., and Cimpian, A. (2017). Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science* 355, 389–391. doi: 10.1126/science.aah6524

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022. doi: 10.1162/jmlr.2003.3.4-5.993

Blodgett, S. L., Green, L., and O'Connor, B. (2016). "Demographic dialectal variation in social media: a case study of African-American English," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (Copenhagen). doi: 10.18653/v1/D16-1120

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Advances in Neural Information Processing Systems* (Barcelona), 4349–4357.

Bonica, A. (2014). Mapping the ideological marketplace. *Am. J. Polit. Sci.* 58, 367–386. doi: 10.1111/ajps.12062

Card, D., Tan, C., and Smith, N. A. (2017). A neural framework for generalized topic models. *arXiv* 1705.09296.

Chang, J., Boyd-Graber, J. L., Gerrish, S., Wang, C., and Blei, D. M. (2009). Reading tea leaves: how humans interpret topic models. *NIPS Proc.* 22, 288–296. Available online at: http://dl.acm.org/citation.cfm?id=2984093.2984126

Cohen, R., and Ruths, D. (2013). "Classifying political orientation on Twitter: it's not easy!" in *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media* (Cambridge, MA).

Cranshaw, J., Schwartz, R., Hong, J. I., and Sadeh, N. (2012). "The livehoods project: utilizing social media to understand the dynamics of a city," in *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, ICWSM '12* (Dublin: AAAI).

Crawford, K. (2016). Can an algorithm be agonistic? Ten scenes from life in calculated publics. *Sci. Technol. Hum. Values* 41, 77–92. doi: 10.1177/0162243915589635

Crawford, K., Dobbe, R., Theodora, D., Fried, G., Green, B., Kaziunas, E., et al. (2019). *AI Now 2019 Report*, AI Now Institute.

d'Andrade, R. G. (1995). *The Development of Cognitive Anthropology*. Boston, MA: Cambridge University Press. doi: 10.1017/CBO9781139166645

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). "Automated hate speech detection and the problem of offensive language," in *Eleventh International AAAI Conference on Web and Social Media* (Montreal, QC).

DellaPosta, D., Shi, Y., and Macy, M. (2015). Why do liberals drink lattes? *Am. J. Sociol.* 120, 1473–1511. doi: 10.1086/681254

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv* 1810.04805.

Doherty, C. (2017). *Key Takeaways on Americans' Growing Partisan Divide Over Political Values*. Washington, DC: Pew Research Center.

Eisenstein, J., Ahmed, A., and Xing, E. P. (2011). "Sparse additive generative models of text," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (Bellevue, WA), 1041–1048.

Farrell, J. (2016). Corporate funding and ideological polarization about climate change. *Proc. Natl. Acad. Sci. U.S.A.* 113, 92–97. doi: 10.1073/pnas.1509433112

Florini, S. (2014). Tweets, tweeps, and signifyin' communication and cultural performance on "black twitter". *Televis. New Media* 15, 223–237. doi: 10.1177/1527476413480247

Foucault, M. (1990). *The History of Sexuality: An Introduction*. New York, NY: Vintage.

Foulds, J., and Pan, S. (2018). An intersectional definition of fairness. *arXiv* 1807.08362.

Friedman, J., Hastie, T., and Tibshirani, R. (2009). *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*. R package version. Available online at: http://crcui.r-project.org/web/pack-ages/glmnet/index.htm

Gal, Y., and Ghahramani, Z. (2016). "Dropout as a bayesian approximation: representing model uncertainty in deep learning," in *International Conference on Machine Learning* (New York, NY), 1050–1059.

Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., et al. (2019). Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? *arXiv* 1912.08320.

Gentzkow, M., Shapiro, J., and Taddy, M. (2016). *Measuring Polarization in High-Dimensional Data: Method and Application to Congressional Speech*. Technical report, eSocialSciences. doi: 10.3386/w22423

Glymour, B., and Herington, J. (2019). "Measuring the biases that matter: the ethical and casual foundations for measures of fairness in algorithms," in *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19* (New York, NY: ACM), 269–278. doi: 10.1145/3287560.3287573

Goel, S., Anderson, A., Hofman, J., and Watts, D. J. (2015). The structural virality of online diffusion. *Manag. Sci.* 62, 180–196. doi: 10.1287/mnsc.2015.2158

Goffman, E. (1959). *The Presentation of Self in Everyday Life*. Garden City, NY: Anchor.

Goldberg, Y. (2016). A primer on neural network models for natural language processing. *J. Artif. Intell. Res.* 57, 345–420. doi: 10.1613/jair.4992

Gould, S. J. (1996). *The Mismeasure of Man*. New York, NY: WW Norton & Company.

Green, B. (2018). ""Fair" risk assessments: a precarious approach for criminal justice reform," in *5th Workshop on Fairness, Accountability, and Transparency in Machine Learning* (Stockholm).

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., and Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science* 363, 374–378. doi: 10.1126/science.aau2706

Hacking, I. (1986). *Making Up People*. Palo Alto, CA: Stanford University Press.

Hanna, A., Denton, E., Smart, A., and Smith-Loud, J. (2019). Towards a critical race methodology in algorithmic fairness. *arXiv* 1912.03593.

Haraway, D. (1988). Situated knowledges: the science question in feminism and the privilege of partial perspective. *Femin. Stud.* 14, 575–599. doi: 10.2307/3178066

Harding, S. G. (2004). *The Feminist Standpoint Theory Reader: Intellectual and Political Controversies*. London, UK: Psychology Press.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Berlin: Springer Science & Business Media.

Heise, D. R. (2007). *Expressive Order*. New York, NY: Springer.

Hipp, J. R., Faris, R. W., and Boessen, A. (2012). Measuring 'neighborhood': constructing network neighborhoods. *Soc. Netw.* 34, 128–140. doi: 10.1016/j.socnet.2011.05.002

Hoffmann, A. L. (2019). Where fairness fails: on data, algorithms, and the limits of antidiscrimination discourse. *Inform. Commun. Soc.* 22, 900–915. doi: 10.1080/1369118X.2019.1573912

Hofman, J. M., Sharma, A., and Watts, D. J. (2017). Prediction and explanation in social systems. *Science* 355, 486–488. doi: 10.1126/science.aal3856

Hovland, C. I., and Weiss, W. (1951). The influence of source credibility on communication effectiveness. *Public Opin. Q.* 15, 635–650. doi: 10.1086/266350

Hovy, D., and Fornaciari, T. (2018). "Increasing in-class similarity by retrofitting embeddings with demographic information," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels: Association for Computational Linguistics), 671–677. doi: 10.18653/v1/D18-1070

Ioffe, S., and Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. *arXiv* 1502.03167.

Ipeirotis, P. G., Provost, F., Sheng, V., and Wang, J. (2010). *Repeated Labeling Using Multiple Noisy Labelers*. New York, NY: SSRN eLibrary.

Jacobs, A. Z., and Wallach, H. (2019). *Measurement and Fairness*.

Joseph, K., Friedland, L., Hobbs, W., Lazer, D., and Tsur, O. (2017a). "ConStance: modeling annotation contexts to improve stance classification," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (Copenhagen: Association for Computational Linguistics), 1115–1124. doi: 10.18653/v1/D17-1116

Joseph, K., Wei, W., and Carley, K. M. (2016). "Exploring patterns of identity usage in tweets: a new problem, solution and case study," in *Proceedings of the 25th International Conference on World Wide Web* (Montreal, QC: International World Wide Web Conferences Steering Committee), 401–412. doi: 10.1145/2872427.2883027

Joseph, K., Wei, W., and Carley, K. M. (2017b). "Girls rule, boys drool: extracting semantic and affective stereotypes from Twitter," in *2017 ACM Conference on Computer Supported Cooperative Work (CSCW)* (Seattle, WA). doi: 10.1145/2998181.2998187

Jung, S.-G., An, J., Kwak, H., Salminen, J., and Jansen, B. J. (2017). "Inferring social media users demographics from profile pictures: a face++ analysis on twitter users," in *Proceedings of 17th International Conference on Electronic Business* (Dubai).

Kamishima, T., Akaho, S., and Sakuma, J. (2011). "Fairness-aware learning through regularization approach," in *2011 IEEE 11th International Conference on Data Mining Workshops* (Washington, DC: IEEE), 643–650. doi: 10.1109/ICDMW.2011.83

Kay, M., Matuszek, C., and Munson, S. A. (2015). "Unequal representation and gender stereotypes in image search results for occupations," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul: ACM), 3819–3828. doi: 10.1145/2702123.2702520

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. (2017). Preventing fairness gerrymandering: auditing and learning for subgroup fairness. *arXiv* 1711.05144.

Kennedy, R., Wojcik, S., and Lazer, D. (2017). Improving election prediction internationally. *Science* 355, 515–520. doi: 10.1126/science.aal2887

Kerr, N. L. (1998). HARKing: hypothesizing after the results are known. *Pers. Soc. Psychol. Rev*. 2, 196–217. doi: 10.1207/s15327957pspr0203_4

Kleinberg, J. (2018). Inherent trade-offs in algorithmic fairness. *ACM SIGMETRICS Perform. Eval. Rev*. 46:40. doi: 10.1145/3219617.3219634

Krippendorff, K. (2004). Reliability in content analysis. *Hum. Commun. Res*. 30, 411–433. doi: 10.1111/j.1468-2958.2004.tb00738.x

Kunda, Z., and Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: a parallel-constraint-satisfaction theory. *Psychol. Rev*. 103, 284–308. doi: 10.1037/0033-295X.103.2.284

Larson, J., and Angwin, J. (2016). *How We Analyzed the COMPAS Recidivism Algorithm*. New York, NY: ProPublica.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of Google flu: traps in big data analysis. *Science* 343, 1203–1205. doi: 10.1126/science.1248506

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., et al. (2009). Computational social science. *Science* 323, 721. doi: 10.1126/science.1167742

Lazer, D., and Radford, J. (2017). Data ex machina: introduction to big data. *Annu. Rev. Sociol*. 43, 19–39. doi: 10.1146/annurev-soc-060116-053457

Levendusky, M. (2009). *The Partisan Sort: How Liberals Became Democrats and Conservatives Became Republicans*. Chicago, IL: University of Chicago Press. doi: 10.7208/chicago/9780226473673.001.0001

Li, D. (2019). *AOC Is Right: Algorithms Will Always Be Biased as Long as There's Systemic Racism in This Country*. Available online at: https://slate.com/news-and-politics/2019/02/aoc-algorithms-racist-bias.html

Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv* 1606.03490.

Liu, Y., Niculescu-Mizil, A., and Gryc, W. (2009). "Topic-link LDA: joint models of topic and author community," in *Proceedings of the 26th Annual International Conference on Machine Learning* (Montreal, QC: ACM), 665–672. doi: 10.1145/1553374.1553460

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., and Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Polit. Anal*. 23, 254–277. doi: 10.1093/pan/mpu019

Lui, M., and Baldwin, T. (2012). "langid.py: An off-the-shelf language identification tool," in *Proceedings of the ACL 2012 System Demonstrations* (Jeju Island: Association for Computational Linguistics), 25–30.

Lundberg, I., Narayanan, A., Levy, K., and Salganik, M. J. (2019). Privacy, ethics, and data access: a case study of the fragile families challenge. *Socius* 5:2378023118813023. doi: 10.1177/2378023118813023

Marsden, P. V., and Friedkin, N. E. (1993). Network studies of social influence. *Sociol. Methods Res*. 22, 127–151. doi: 10.1177/0049124193022001006

Martin, E. (1991). The egg and the sperm: how science has constructed a romance based on stereotypical male-female roles. *Signs J. Women Cult. Soc*. 16, 485–501. doi: 10.1086/494680

Mason, L. (2015). "I disrespectfully agree": the differential effects of Partisan sorting on social and issue polarization. *Am. J. Polit. Sci*. 59, 128–145. doi: 10.1111/ajps.12089

Mitchell, M., Baker, D., Moorosi, N., Denton, E., Hutchinson, B., Hanna, A., et al. (2020). "Diversity and inclusion metrics in subset selection," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (None), 117–123. doi: 10.1145/3375627.3375832

Mitchell, S., Potash, E., and Barocas, S. (2018). Prediction-based decisions and fairness: a catalogue of choices, assumptions, and definitions. *arXiv* 1811.07867.

Mukherjee, S. (2014). "Joint author sentiment topic model," in *SIAM International Conference in Data Mining (SDM 2014)* (Pennsylvania, PA). doi: 10.1137/1.9781611973440.43

Nelson, L. K. (2017). Computational grounded theory: a methodological framework. *Sociol. Methods Res*. 49:0049124117729703. doi: 10.1177/0049124117729703

O'Connor, B., Bamman, D., and Smith, N. A. (2011). "Computational text analysis for social science: model assumptions and complexity," in *NIPS Workshop on Computational Social Science and the Wisdom of Crowds*.

Olteanu, A., Castillo, C., Diaz, F., and Kiciman, E. (2016). *Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries*. SSRN Scholarly Paper ID 2886526. Rochester, NY: Social Science Research Network. doi: 10.2139/ssrn.2886526

Omi, M., and Winant, H. (2014). *Racial Formation in the United States*. New York, NY: Routledge. doi: 10.4324/9780203076804

Passonneau, R. J., and Carpenter, B. (2014). The benefits of a model of annotation. *Trans. Assoc. Comput. Linguist*. 2, 311–326. doi: 10.1162/tacl_a_00185

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Commun. ACM* 62, 54–60. doi: 10.1145/3241036

Poole, K. T., and Rosenthal, H. (1991). Patterns of congressional voting. *American Journal of Political Science*. 35, 228–278. doi: 10.2307/2111445

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., et al. (2010). Learning from crowds. *J. Mach. Learn. Res*. 11, 1297–1322. Available online at: http://dl.acm.org/citation.cfm?id=1756006.1859894

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should i trust you?": explaining the predictions of any classifier. *arXiv* 1602.04938. doi: 10.18653/v1/N16-3020

Rickford, J. R., and Labov, W. (1999). *African American Vernacular English: Features, Evolution, Educational Implications*. Malden, MA: Blackwell.

Roberts, M. E., Stewart, B. M., Tingley, D., and Airoldi, E. M. (2013). "The structural topic model and applied social science," in *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation* (Lake Tahoe, UT), 1–20.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., et al. (2014). Structural topic models for open-ended survey responses. *Am. J. Polit. Sci*. 58, 1064–1082. doi: 10.1111/ajps.12103

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: graphical causal models for observational data. *Adv. Methods Pract. Psychol. Sci*. 1, 27–42. doi: 10.1177/2515245917745629

Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. (2004). "The author-topic model for authors and documents," in *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (Arlington, VA: AUAI Press), 487–494.

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., et al. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proc. Natl. Acad. Sci.* 117, 8398–8403. doi: 10.1073/pnas.1915006117

Salganik, M. J., Lundberg, I., Kindel, A. T., and McLanahan, S. (2019). Introduction to the special collection on the fragile families challenge. *Socius* 5:2378023119871580. doi: 10.1177/2378023119871580

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., et al. (2013). Personality, gender, and age in the language of social media: the open-vocabulary approach. *PLoS ONE* 8:e73791. doi: 10.1371/journal.pone.0073791

Selbst, A. D., Boyd, D., Friedler, S., Venkatasubramanian, S., and Vertesi, J. (2018). *Fairness and Abstraction in Sociotechnical Systems*. SSRN Scholarly Paper ID 3265913. Rochester, NY: Social Science Research Network. doi: 10.1145/3287560.3287598

Sen, M., and Wasow, O. (2016). Race as a bundle of sticks: designs that estimate effects of seemingly immutable characteristics. *Annu. Rev. Polit. Sci.* 19, 499–522. doi: 10.1146/annurev-polisci-032015-010015

Small, M. L. (2017). *Someone to Talk To*. Oxford, UK: Oxford University Press. doi: 10.1093/oso/9780190661427.001.0001

Smith-Lovin, L. (2007). The strength of weak identities: social structural sources of self, situation and emotional experience. *Soc. Psychol. Q.* 70, 106–124. doi: 10.1177/019027250707000203

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). "Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Honolulu, HI: Association for Computational Linguistics), 254–263. doi: 10.3115/1613715.1613751

Sudnow, D. (1965). Normal crimes: sociological features of the penal code in a public defender office. *Soc. Probl.* 12, 255–276. doi: 10.2307/798932

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. (2017). "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence* (San Francisco, CA).

Tavory, I., and Timmermans, S. (2014). *Abductive Analysis: Theorizing Qualitative Research*. Chicago, IL: University of Chicago Press. doi: 10.7208/chicago/9780226180458.001.0001

Todorov, A., Said, C. P., Engell, A. D., and Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends Cogn. Sci.* 12, 455–460. doi: 10.1016/j.tics.2008.10.001

Toole, J. L., Lin, Y.-R., Muehlegger, E., Shoag, D., González, M. C., and Lazer, D. (2015). Tracking employment shocks using mobile phone data. *J. R. Soc. Interface* 12:20150185. doi: 10.1098/rsif.2015.0185

Tsur, O., Calacci, D., and Lazer, D. (2015). "A frame of mind: using statistical models for detection of framing and agenda setting campaigns," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Lisbon), 1629–1638. doi: 10.3115/v1/P15-1157

Tufekci, Z. (2014). "Big questions for social media big data: representativeness, validity and other methodological pitfalls," in *ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media* (Ann Arbor, MI).

Van Bavel, J. J., and Pereira, A. (2018). The Partisan brain: an identity-based model of political belief. *Trends Cogn. Sci.* 22, 213–224. doi: 10.1016/j.tics.2018.01.004

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems* (Long Beach, CA), 5998–6008.

Wallace, E., Feng, S., Kandpal, N., Gardner, M., and Singh, S. (2019). Universal adversarial triggers for nlp. *arXiv* 1908.07125. doi: 10.18653/v1/D19-1221

Wallach, H. (2018). Computational social science ≠ computer science + social data. *Commun. ACM* 61, 42–44. doi: 10.1145/3132698

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015). Forecasting elections with non-representative polls. *Int. J. Forecast.* 31, 980–991. doi: 10.1016/j.ijforecast.2014.06.001

Wang, Y., and Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *J. Pers. Soc. Psychol.* 114:246. doi: 10.1037/pspa0000098

Waseem, Z. (2016). "Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter," in NLP+ CSS 2016, 138. Austin, TX. doi: 10.18653/v1/W16-5618

Wu, X., and Zhang, X. (2016). Automated inference on criminality using face images. *arXiv* 1611.04135.

Yan, X., Guo, J., Lan, Y., and Cheng, X. (2013). "A biterm topic model for short texts," in *Proceedings of the 22nd International Conference on World Wide Web* (Rio de Janeiro: ACM), 1445–1456. doi: 10.1145/2488388.2488514

Zagoruyko, S., and Komodakis, N. (2016). Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. *arXiv* 1612.03928.

Zuberi, T., and Bonilla-Silva, E. (2008). *White Logic, White Methods: Racism and Methodology*. Lanham, MD: Rowman & Littlefield Publishers.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.