Justice Rising - The Growing Ethical Importance of Big Data, Survey Data, Models and AI

Richard Timpone, Ipsos
Yongwei Yang, Google

October 26, 2018

Abstract

In past work, the criteria of *Truth, Beauty*, and *Justice* have been leveraged to evaluate models (Lave and March 1993, Taber and Timpone 1996). Earlier, while relevant, *Justice* was seen as the least important of modeling considerations, but that is no longer the case. As the nature of data and computing power have opened new opportunities for the application of data and algorithms from public policy decision-making to technological advances like self-driving cars, the ethical considerations have become far more important in the work that researchers are doing.

While a growing literature has been highlighting ethical concerns of Big Data, algorithms and artificial intelligence, we take a practical approach of reviewing how decisions throughout the research process can result in unintended consequences in practice. Building off Gawande's (2009) approach of using checklists to reduce risks, we have developed an initial framework and set of checklist questions for researchers to consider the ethical implications of their analytic endeavors explicitly. While many aspects are considered those tied to *Truth* and accuracy, through our examples it will be seen that considering research design through the lens of *Justice* may lead to different research choices.

These checklists include questions on the collection of data (Big and Survey; including sources and measurement), how it is modeled and finally issues of transparency. These issues are of growing importance for practitioners from academia to industry to government and will allow us to advance the intended goals of our scientific and practical endeavors while avoiding potential risks and pitfalls.

As advances continue to progress in the volume and types of data being collected and the computing power to leverage them fundamentally change the face of research and its practical applications, related ethical issues are growing in importance. The numbers of examples and authors calling out issues around data and its usage continue to grow (some recent examples include- Eubanks 2017, Noble 2018, O'Neil 2016, Wachter-Boettcher 2017). It is our contention that the bulk of these issues are unintended consequences of research, data and modeling efforts.

While we do not feel the intentions of the social scientists, data scientists and technologists are generally in question, it is the unintended consequences of research decisions

Paper prepared for presentation at the BigSurv 2018 Conference of the European Survey Research Association; Barcelona, Spain.
Rich Timpone is Senior Vice President and Head of the Ipsos Science Center at Ipsos (rich.timpone@ipsos.com). Yongwei Yang is a Survey Research Scientist at Google, Inc. (yongwei@google.com).

that are creating the bulk of the ethical issues for their tools that are having more and more direct impact on individuals lives. The purpose of this paper is to help identify and classify the sources of some of these issues and provide some practical guidance on elements to consider and steps to take to help avoid them in the future.

Most of the elements that we discuss are ones that will be familiar to researchers. Here though, we summarize some of these and their consequences through the lens of their consequences to assist in avoiding potentially serious unintended consequences (Yang 2013). This builds on the approach of Atul Gawande (2009) of creating checklists for experts, from medical doctors to airline pilots, to ensure specific steps are remembered and not missed to reduce risks.

It must be highlighted that *Justice* and ethics in our data and models, like *Truth* and accuracy are "ideal[s] rather than a state of existence. We do not achieve it—we pursue it" (Lave and March 1993, p 74). Because each is aspirational, we feel that considering the same research questions through each of the different lenses can result in different research and design decisions. As the work of social scientists, technologists and applied data scientists work continues to grow in practical impact, we feel that continuing to reiterate these fundamental issues will be key to always advancing in each of our research pursuits.

Overview

Different types of models are being developed today that go beyond growing foundational theoretic understanding by social and other scientists to practical applications that are becoming more and more ubiquitous. The use of models to leverage multiple sources of data to influence policy decisions, from criminal sentencing to public services has grown with the

intent to be more objective and fact-based. Likewise, algorithms play more and more of a role in our lives from recommendations for books and movies we may like, to the news we may want to see, to our online searches, and looking forward how we will drive in the future.

The consequences of problems created by the sources and types of data leveraged to the algorithms and the analyses used, can have far ranging negative implications from creating emotional distress, undermining the dignity of individuals and groups to adversely affecting individuals lives from child welfare to prison sentences or even risk to life. Given the growing direct effect that these sources have on individual's lives, the importance to move from theoretic to practical considerations, to avoid negative and unintended becomes more important.

While there have been some questions of allowing algorithms to do things they may not have the intelligence to perform yet, practical questions exist about the appropriate balance for their growing roles (Dormehl 2017). While some argue that moving to more automation is a way to create more fairness and efficiency, as (Mayer-Schönberger and Cukier 2013, p 176) state, "[i]n the big-data era we... have to expand our understanding of justice, and require that it include safeguards for human agency." Friedman (2016, p 345) quotes Dov Seidman, CEO of LRN, who feared that "[w]e are letting technology do the work that human beings should never abdicate." Regardless of what we feel is the appropriate balance on the *Human + Machine* spectrum, automated decisions will continue to play a growing role in our lives. In addition to debating what is appropriate, as researchers we bear the burden of also ensuring the tools we use and develop are not creating unintended consequences that negatively impact people's lives.

While we will include a number of examples throughout the paper, a recent one can be seen when Indonesians posting about the hopes of people surviving and being unhurt in the

Lombok earthquake in August of 2018 had balloons added to their posts by an automated algorithm (Bach 2018, see Figure 1 from her article). This occurred due to the language parsing algorithm not recognizing that the term 'selamat' in Indonesian means congratulations as well as safe and unhurt. The framework of this paper will hopefully help researchers and teams avoid a number of unintended issues such as this as well, as those with even greater life affecting effects.

Figure 1: Facebook Balloons Added to Earthquake Posts

This first section of the paper considers a number of theoretic considerations from how we evaluate models and data at a high level to the shift in the relative importance of the criteria over time. While addressing ethics, we distinguish normative considerations from models that deviate from their intent or produce unintended consequences. We generally leave the former debates for political and philosophical consideration while the latter is where we focus the practical considerations.

The second part of the paper emphasizes these practical considerations focusing on a handful of specific dimensions: who and what we are collecting data on, what we are specifically

collecting, how we build the models themselves, how learning algorithms are trained (where applicable), how algorithms behave in action and how we know what is driving their recommendations and actions. This paper is a first phase of an on-going iterative development of how we can consider traditional research questions through the lens of *Justice* and ethics to help our data and models to serve us better in the future.

Dimensions of Evaluation

Previously, we had leveraged Lave and March's (1993) criteria for evaluating models in the social sciences to our evaluations of the family of computational models (Taber and Timpone 1996). These are the dimensions of *Truth*, *Beauty* and *Justice*. Two decades ago we noted that "[t]he criterion of Justice is less universally accepted than either Truth or Beauty" (ibid, p 81). Given the growing practical implications, and potential for causing harm, we feel that it now must be explicitly considered for those leveraging multiple types of data and building models for scientific discovery as well as practical applications. Many of our positions are evaluating traditional research elements as we agree with Strong (2015) that even in the era of Big Data many of the old rules still apply.

Truth has been viewed as the most important of the criteria, and as we will see it has significant implications for each of the other two dimensions, although passing thresholds for success in one is not sufficient to avoid issues in the others which is why considering them as separate dimensions of evaluation is necessary. In short, here we ask if data and models are accurate in what they intend to measure. We must consider both data and models, since not only problems in our models and algorithms can have negative effects but even models that are accurate that are fed flawed data will be just as problematic (Gilbert and Troitzsch 2005).

Beauty is more about aesthetic and utilitarian aspects of the data and models. In particular, we consider whether these purposeful representations are appropriately parsimonious and generate many testable hypotheses and surprising results (Gilbert and Troitzsch 2005, Lave and March 1993, Starfield et al. 1990). The criteria of simplicity and balancing accuracy with parsimony is a key question as we do not want to abstract away elements that are important in the interest of simplicity. In fact, this balance is one reason that we believe addressing the same research questions while considering the evaluation for *Truth* and *Justice* separately can lead to different conclusions.

"*Justice* takes into account the real-world implications of a model, especially those that affect quality of life" (ibid p. 71). It makes sense that going from advising social scientists in their research to the current ubiquity of data, model and algorithm based decisions in life over the past two decades, that the importance of *Justice* and ethical considerations would be rising in importance and require researchers to consider this far more actively and thoroughly. The more practical the application, the greater the risk if we fail to address the implications adequately.

As with *Beauty*, the evaluation of *Justice* is strongly tied to *Truth*. As we will see in the second section of the paper, many of the areas that are highlighted for researchers to avoid unintended ethical issues focus on accurately reflecting what we intend to measure and the processes modeled. While we still emphasize the importance of *Truth* for scientific discovery, we no longer place greater importance on its considerations than the ethical ones when we examine the practical implications of the use of data and models on people's lives with both of these dimensions being critically important today.

As we have stated, while many aspects we will discuss seem to emphasize the dimension of *Truth* and accuracy, the fact that *Truth* itself is a goal rather than something actually attained

emphasizes the tradeoffs made in our decision-making on measurement, models etc. The issue is that evaluating these same research design questions through the lens of *Justice* and the impact on people's lives, can lead to different considerations when we are balancing the desire for parsimony and where we may feel that our measures and models are 'accurate enough.' In short, we believe going through our checklists through a lens of *Justice* may lead to some different design decisions than if we were to continue considering this to be a less relevant dimension.

Data and Modeling

The risks of unintended consequences from different sources of data are relevant for traditional data sources, such as survey data, as well as the breadth of recent advances in Big Data. While the "Three V's", of *Volume, Velocity* and *Variety*, from the META Group (now Gartner) are commonly used as a describe Big Data (Beyer and Laney 2012), we like Cielen, Meysman and Mohamed's (2016) simple definition that it can be considered any collection of data or datasets so complex or large that traditional data management approaches become unsuitable. This issue of complexity in data and algorithms is the source of unintended risks in their deployment.

This definition of Big Data recognizes the dramatic expansion in the collection and storage of many types of data from internet data (online text, video, sound data), social media data, website data (logs, cookies, transactions, website analytics), Internet of Things, behavioral data, transaction data, administrative data, commercially available databases, satellite imagery, machine generated sensor data, and even the digital exhaust that is produced and leftover from transactions and other activities performed online, on mobile devices and other areas where data is collected (Callegaro and Yang 2018). The dramatic increase in what is being captured and

stored allows for many exciting possibilities, however it also increases challenges and concerns as we are often leveraging data for purposes distinct from the intended uses they were collected for (Biemer 2017, Callegaro and Yang 2018, Japec et al. 2015, Tolich 2018).

These criteria for evaluation and the research questions we leverage in our checklists are ones that are very familiar to social scientists and survey researchers. In fact, a number of them build from the ideas of Total Survey Error that have been extended to Big Data Total Error (Biemer 2010, 2017; Callegaro and Yang 2018; Japec et al. 2015). This framework notes the risks of specification errors, measurement issues, sample frame, non-response, data processing and other issues. While our checklists group these and extend some, a key distinction is the shift in focus from their focus on accuracy and *Truth* to considering them through the lens of *Justice*. This is even more important as Biemer's (2010) piece on Total Survey Error emphasizes the tradeoffsof optimizing survey quality within constraints of budgetary and time constraints. These are the exact tradeoffs that effect all practical algorithms and considering the tradeoffs through the question of unintended consequences on people's lives can produce different research choices than focusing on veracity alone.

While the measures and sources of our data are relevant for considerations of the quality and implications of their implementation, the bulk of the scientific discovery and practical applications come from progressively complex models. This could be from any type of model from natural language to mathematical (including statistics and logic) to apps and computational models (Ostrom 1988, Taber and Timpone 1996).

As the lines between these general languages of models becomes blurred with many modern applications bridging mathematical and computational models, we will focus on a few specific aspects of different classes of models and algorithms as they affect some of the questions

we pose for research teams. These are specifically relevant for the nature of their construction (researcher driven, emergent from data, or a combination), the level of transparency in the models themselves, and whether the implementation includes dynamic elements on top of the core algorithms.

These distinctions are relevant as they affect the confidence in the accuracy or *Truth* of the models, our ability to understand what is occurring in them, and to recognize how they can perform and vary in practice. As we have stated, *Truth* is an aspiration rather than the goal for the simplified representations that are our models, but we know there are decisions that can be made for the inputs, combinations and other aspects to maximize the verisimilitude of our models and minimize unintended consequences. These may entail researcher decisions or machine generated ones.

Depending on the model choices we use, how much we can understand the outcomes will vary. At the clearest end, the logic of models that leverage rule-based approaches or simple statistical models can be readily understood. With a rule-based approach we can readily see what is driving the outcome of a model, whether an Expert System or other sequential process framework (Taber 1992, Taber and Timpone 1994, 1996). Statistical models, from simple linear regression to more complex ones that take into account aspects like selection bias and multicollinearity with clear parameters, likewise, are totally transparent on what is going into determining the outcomes and how each of them is weighted.

At the other end of the spectrum are Deep Learning algorithms. While their inputs will be known, which ones are actually utilized in determining the outcome, how they are weighted and combined (linear or not, additive or interactive, etc.) is generally unknown. As O'Neil (2016, p 173) states "oceans of behavioral data, in coming years, will feed straight into artificial

intelligence systems. And these will remain, to human eyes, black boxes." There are significant efforts going on to provide insight into these boxes and this is an increasingly important question being explored by policymakers and practitioners alike.

These aspects of data and modeling will feed into our framework and questions for the research checklists to be evaluated through the lens of *Justice*. Before moving into the practical research questions, it is worth highlighting the distinction between unintended consequences and those that behave as the developers intended but that observers may disagree with from a normative perspective.

Normative Issues and Justice of Data and Models

Stepping back for a moment, it is worth thinking more broadly on questions of *Justice* and ethics. The domain of ethics builds from the division of philosophy concerned with what is considered to be morally correct or good (Boone 2017). From a philosophical perspective, the questions of ethics go far beyond what we are touching on here. While we do not argue for fundamental primacy of consequentialist theories of ethics, we do focus on those aspects of data and models' actions and effects on people, groups and society as we are making our recommendations on evaluating models in this dimension.

From social science theoretic studies to practical applications of data and models, the normative question of what is "the proper course of action for human behavior" (ibid, p 59) is sometimes an unstated but driving aspect of domains of study and models developed. In practical applications, especially those in policy domains, this is far more explicitly important.

While each of the authors of this paper have strong views on normative issues of policy and human nature, those are outside the scope of this paper. Some critics of modern use of Big

Data, algorithms and AI in practice often blur the distinction between the effects that are unintended and those that may be driven by normative views that are at odds with the authors worldview. While Eubanks (2017) raises a number of concerns about the effects of various programs, some of the issues she raises are driven by a specific normative view of poverty, and notes that political debates "are more than informational; they are about values, group membership and balancing conflicting interests. (p 124)"

While there is some blending of normative and objective issues, sometimes the normative debates are explicit and public. This may be the case with companies' direct positions or those that create internal debate among its employees. This was seen recently with employees at Facebook expressing outrage when one of the company's top global policy executives publicly supported the nomination of U.S. Supreme Court Justice Brett Kavanaugh (Seetharaman 2018).

This is reflecting the fact that ideological divisions across the globe have become more polarized and divisive (Young 2018). The impact of this divide is growing in prevalence and recently some politicians in the United States recently have been challenging whether there is an anti-conservative bias in search results at Google. Sundar Pichai, CEO of Google, for instance, met with members of the U.S. Congress in September to discuss among other things "how the company sets up its teams and codes its algorithms to prevent bias" (Romm 2018). The question is what we can do as researchers and data scientists to avoid those deviations from what we are intending from our algorithms.

From a normative perspective, if members of a group have fundamental differences in their preferences this can create issues for their performance. Unlike diversity being beneficial for problem solving and prediction, if teams disagree on what they are trying to accomplish effectiveness will be compromised (Page 2007). Again, these differences in fundamental world

views should not be ignored but are most effectively addressed outside of the practical efforts in developing solutions to problems.

This distinction is not always as clear as we just laid out though as the lack of transparency in some work makes it impossible to identify what was intended from what may be unintended. While transparency is one of the aspects we cover in practical checklists, we still focus here on where outcomes and effects are negative, separate from what each of us believes is a better world. Other than ideological distinctions, we believe most social scientists, data scientists and technologists are attempting to build tools for what they view as 'good', thus we are focused more on the unintended consequences of data and modeling efforts and leave discussion of underlying motivations for elsewhere.

While we have focused on normative questions in this section, it should be clear that most of the types of data and models we are discussing are intended to serve objective rather than ideological purposes. While some have elements of policy with different normative views, most are seeking objectivity. Thus, whether an image recognition program for instance demonstrates bias is not a normative issue but rather one where tools are not behaving as intended and could still result in negative implications.

What Can We Do?

Many argue that the objective nature of algorithms and use of Big Data will create more efficiency and fairness. As we will see, while this may be the case, we must evaluate how well this is reflected in how our projects are designed and behave. We firmly believe in the *Human + Machine* idea of people working together with technology for the realization of our goals.

As Daugherty and Wilson (2018) discuss, *Human + Machine* partnerships actually reflect a spectrum from the dominance of one to the other. For us, at the least, people's roles are fundamental in the training and development parts of the process. Beyond this though, they also argue for the area of humans and technology complementing each other in the middle of the spectrum. Here people play more of a role in explaining models (which we will focus on in our transparency discussion) as well as a sustaining role which sets limits on the technology and evaluates it for ethical issues.

While we feel the placement of Big Data and algorithms on the *Human + Machine* spectrum will vary, it is fundamental to explicitly consider the consequences of our choices. The more we believe these issues will be resolved with 'good' data and 'good' models without considering the broader issues is where problems are most likely to creep into our processes. As Silver (2012, p 9) said, "[i]t is when we deny our role in the process that the odds of failure rise." We believe that having teams explicitly consider their decisions throughout the development, design and implementation phases of research through the lens of *Justice* considerations that we are most likely to obtain our research goals and avoid unintended consequences.

The issues that have created challenges in other domains, such as medicine, show a parallel in the development of our modern Big Data and algorithmic research programs. Like them, we are dealing with areas with high complexity where we often divide up the work. In fact, complexity in our modern data sources is one of the hallmarks of Big Data as we defined earlier. Thus, Gawande's (2009) focus is extremely relevant here as is his argument that leveraging the simple tool of a checklist can help overcome failures.

Again, we feel the research domain fits his framework as well as "[w]e have accumulated stupendous know-how. We have put it in the hands of some of the most highly trained, highly

skilled, and hardworking people in our society. And, with it, they have indeed accomplished extraordinary things (ibid., p 13)." But like the other domains, the failures are too common and demoralizing.

The research enterprise is not a singular event like those that he advocates for the process improvement. Unlike pre-surgery or pre-flight checklists, the research timescale we are discussing could extend into weeks, months and years. Even acknowledging this, we still feel that having an explicit list of research considerations to review through the lens of impact on people, groups and society can, like those more time constrained events, help reduce the failures in consequences that can occur. Given the scope of research and algorithmic tool development, issues will never be totally eliminated, but we do believe that consciously considering these issues by leveraging a tool like our research justice checklists will help hopefully catch some that would have slipped through the process otherwise.

Where Do We Go From Here?

As discussed in the previous section, our intent is to focus on those areas where teams can make additional considerations to avoid unintended consequences that have negative ethical implications. Along with not addressing intended normative issues, we also are not focusing on issues of data privacy. Like normative debates, this is not because we do not consider that to be important, but rather it is worthy of its own area of discussion and is also tangential to the content and consequences of the application of data and models on which we are focused. We are certain that discussions of data privacy and ownership will continue to grow, building on the European Union's implementation of the General Data Protection Regulation (GDPR) in May of 2018 (Pardes 2018) and current debates in the U.S. government.

In the future, we hope there are more open discussions of the intent of algorithms and data so that the different perspectives of worldview are transparent and discussed including normative and privacy issues. We also expect there will be more discussion on the consequences of technology on societal changes on the differential welfare of groups and countries as change affects the levels of economic inequality and other societal dynamics (Brynjolfsson and McAfee 2014, Dormehl 2017, Milanovic 2016).

We also hope that more explicit discussions of the ethical considerations of measurement and development of tools and algorithms ensure people's quality of life is not negatively impacted due to inadequate consideration of potential issues. We can leverage the experience from actual examples of where problems have occurred as well as recognize how the deep research knowledge gained from survey and other research in the past can help identify potential sources of problems (i.e. Egner and Timpone 2015, Strong 2015). From these experiences we can identify a number of specific areas where special vigilance can help avoid unintended ethical issues.

While the next section provides an overview of key topic areas and approaches researchers and research teams can employ to avoid some issues, this paper represents an initial set of questions that we are engaging other practitioners and researchers to help build from. As a reader of this paper, we invite you to engage with our exercise and help us further expand on topics and specific steps that can be leveraged to avoid such issues in the future.

In this initial set of considerations, we summarize some opportunities for avoiding unintended consequences by examining aspects of the sample, measurement, model development, behavior in action and transparency. We believe that a dialogue around agreement and

disagreement with these questions is healthy to bring the implications more to the forefront and help all of us be more effective in our research and applications.

Specific Data and Model Considerations

Thus far we have set the stage that ethical considerations are more important for our data and modeling considerations today than ever before, and we have provided a high-level overview of types of data and models we are referencing. Figure 2 presents the initial framework for the research domains we focus on as we explore practical challenges that can result in negative unintended consequences.

In short, we have developed four separate checklists for different phases and dimensions of the research process to be considered through the lens of *Justice*. As we have stated earlier many of these are considerations that research teams already are addressing as they go through the tradeoffs of data availability, model complexity and other decisions usually considered from the perspective of accuracy. However, considering many of these same questions through the lens of the impact and use of the data and models on peoples' lives may lead to different considerations and hopefully avoid some of the unintended consequences that have occurred.

The first phase of research that we delve into focuses on who we are including in the data sources themselves (considering how representativeness could affect implications of the results). Following this is what we are examining and including in our analysis and model in terms of the quality of the measures for the tasks we are putting them to. Next are broad considerations of model development, training, testing and quality. While we only touch on a few aspects of this, we still believe the points we highlight in our checklist will help think about consequences of the choices in a fuller way. Finally, we deal with the issue of transparency and the fact that we don't always, and in some cases can't, know exactly what is happening in our models.

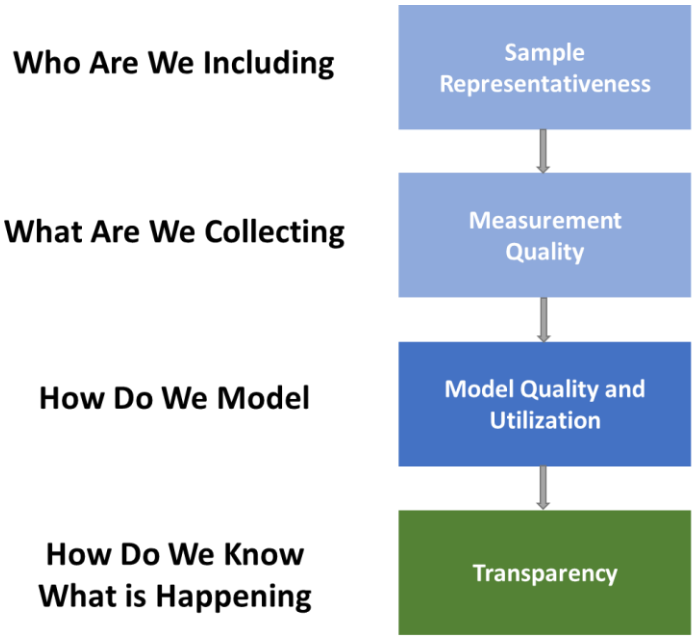

Figure 2: Dimensions of Practical Consideration

Again, these are fundamental research design questions explored from a research quality perspective. Considering each of the questions through the lens of their consequence and impact on people, groups and society allows us to identify implications that could be missed otherwise.

Who We Are Including

For each of these sections we begin with our list of questions for reference and then walk through each one with some of the specific considerations we are calling out and examples of some of the challenges that have been seen where failing to adequately consider the specific question has led to issues in practice.

Some have argued that because we have less error from sampling with Big Data that we can accept more error elsewhere in the process (Mayer-Schönberger and Cukier 2013). The reality is that who we have data for and coverage bias are still fundamental considerations and those who ignore this with Big Data can run risks on the quality of their research and the consequences of their algorithms. It turns out that many of the traditional survey considerations of sampling concerns apply to newer sources of Big Data as well (Biemer 2010, 2017; Callegaro and Yang 2018; Japec et al. 2015; Strong 2015).

Figure 3 provides the checklist for the considerations around our sample and who we are including in the study.

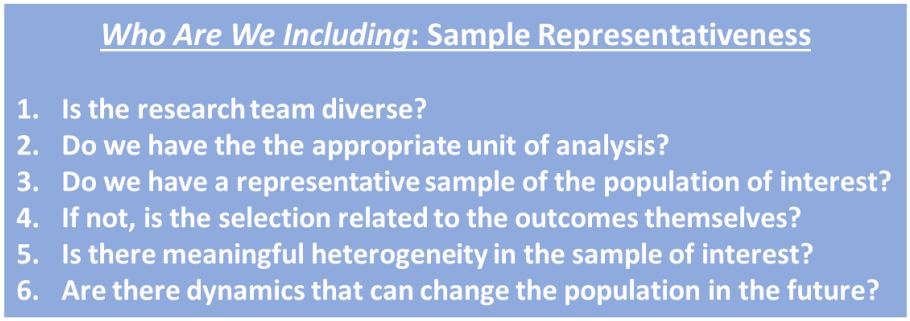

Figure 3: *Who Are We Including* Questions

Our first question which is relevant for each of the practical domains and included in each list is *Is the Research Team Diverse*? While some consider this to be the fundamentally important question (c.f. Wachter-Boettcher 2017), and we do consider this important, it alone will not be sufficient to address many of the issues discussed. We must consider different

perspectives of the issue at hand and not having a homogenous group from demographic, experience, worldview and other dimensions is important to avoid falling into some of the traps we are discussing. As Page (2007) found, groups that are diverse in how they see, categorize, understand and go about improving the world are better at problem solving and prediction efforts. While we do advocate breadth on teams, all relevant perspectives may never be represented, so considering each of the questions that follow in the sample, measurement, testing and other research aspects is important.

Stepping back to revisit the questions we have summarized from research design considerations recommends a level of research humility to avoid the mindset that is too common "where someone assumes they have all the answers about a product and leaves out anyone with a different perspective (Wachter-Boettcher 2017, p 16)."

In terms of who we are studying, a diverse team, and consideration of diverse perspectives, will allow teams to consider what the right unit of analysis is, whether there are issues with how information was collected and stored, potential heterogeneity in the population of interest and all of the questions regarding sample quality and representativeness. We will see that the value of considering diverse experiences, backgrounds, motivations and behaviors will flow throughout the questions in each domain. Again, adding diversity to the team is not sufficient to avoid problems, but will increase the likelihood of anticipating potential implications of research decisions.

From the composition of who is working on the problem at hand, we move to substantive questions, and the first fundamental one any research project or practical application is *Do we have the appropriate unit of analysis* for the study. While this is often left unstated and assumed, that is not always the best path and while we focus on individuals and individual level data in

this paper, this is not the only unit of analysis we may consider. In our previous work, we created a typology of models in the Social Sciences, reproduced as Figure 4, that is relevant for all of the practical applications as well as theoretic ones developed for scientific discovery (Taber and Timpone 1996, p 9).

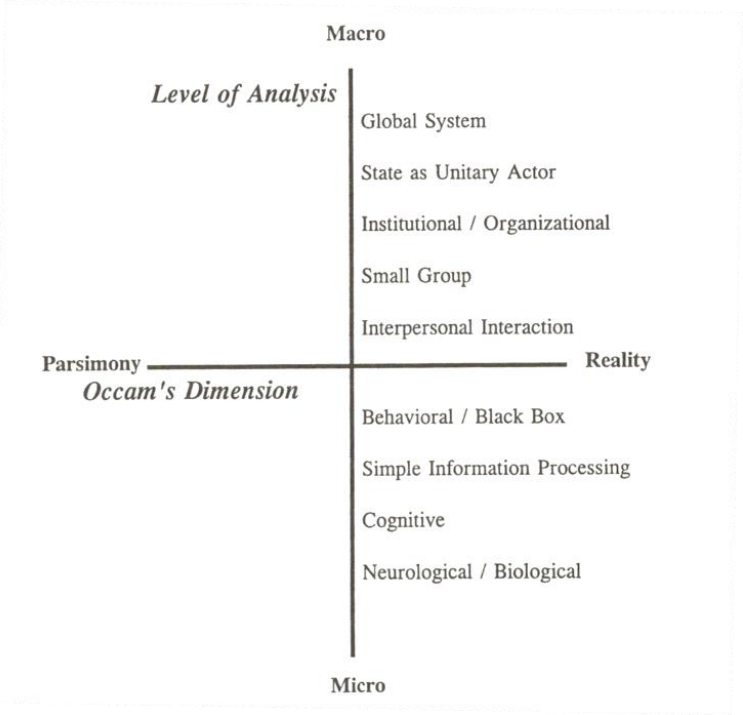

Figure 4: Taber and Timpone (1996) Typology of Models

This taxonomy has two dimensions. The first, we dub Occam's Dimension, focuses on the level of complexity of our models and ties to both the aspects of *Beauty* and *Truth* for model evaluation. Considering how much we can simplify the models to accurately fulfill the purpose we are interested in will be revisited in the *How We Model* section with potentially different choices needed from an accuracy and consequence perspective. Here though we focus on the other aspect on whether we are ranging from the individual level to organizational and beyond. While practical implications of what information is available at each level sometimes affect research choices, we must consider if these practical decisions may be impacting the quality if the opportunity for action is at a different level.

In our summary of this section we focus on the question of sample representativeness and so the next question is the basic one of *Do we have a representative sample of the population of interest.* Paired with this question is a second one that *If not, is the selection related to the outcomes themselves?*

While we will delve further into the questions of representation and selection bias, it is worth highlighting this more generally as with the advances in Big Data, this is sometimes not considered an issue at all. While collecting vast amounts of data has great value, it must be noted that large does not mean unbiased. The lessons learned by survey researchers over 80 years ago when the *Literary Digest* dramatically mis-predicted the outcome of the 1936 U.S. presidential election even with an unprecedented survey sample of 2 million respondents are often missed as we move to the domain of big data (Egner and Timpone 2015, Strong 2015). Even as we move to petabytes of data and beyond, if there are systematic differences in who data exists for there can be dire consequences for model performance.

The issue of selection bias is an old one that is well documented analytically (Heckman 1979), and with practical implications for policy areas where administrative or other processes can lead to non-random subsets of a population of interest (Timpone 1998a, 2002). While these projects demonstrated how addressing administrative issues like voter registration could improve the understanding of voting behavior and issues of group representation, areas of contemporary algorithmic policy assistance and decision-making has far greater consequences if the sample is not representative.

There are multiple challenges that may commonly come up that have systematic ethical implications for the application of algorithms. Even when designed to be more objective, efficient and fair, sample issues can create and reinforce existing biases. In the case study that

Eubanks (2017) shares on the Allegheny County Office of Children, Youth and Families, she notes a number of issues that will come up in our checklists of consideration. The first is that the data that they collect comes disproportionately from individuals in their system for public services.

In this case, the sample for whom data on many of the measures being leveraged for the model is available is a non-random selection to start with. By entering the system, even for programs designed to improve parents and their children's lives, they become further entrenched in the data collection policies and flagged as risks in the evaluative criteria. This creates a cyclical reinforcement of the initial biases as such in-depth information would often be fought by those not in need of government services. This type of cycle also is seen in criminal justice which can create a "pernicious feedback loop. The policing itself spawns new data, which justifies more policing" (O'Neil 2016, p 87).

As Wachter-Boettcher (2017, p 139) notes, "if a system… is fed data that reflects historical biases, then those biases will be reflected in the result[s]." While the Allegheny example highlights how data collection based on the convenience sampling of policies, systematic biases historically can become coded into the problems directly. This is a concern with the increasing use of algorithmic aids in criminal justice decisions including sentencing. In the United States, the criminal justice system has an "overrepresentation of people of color,… disproportionate sentencing of racial minorities,… targeted prosecution of drug crimes in poor communities,… [and] criminalization of new immigrants and undocumented people" (Stevenson 2014, p 301). Thus, even when a cross section of individuals is available, systematic patterns will be reinforced in a way that steady improvement in civil rights would move away from.

In the policy cases above, systematic non-randomness in the sample data comes from the administrative processes themselves. Those are not the only sources of difference though, which is why we include the question *Is there meaningful heterogeneity in the sample of interest.* While there will always be heterogeneity in groups, here we are posing the question for research teams on whether who data exists for is non-random, again in ways that are related to the outcomes of interest.

Many statistical methods explicitly focus on averages and this has been extended to many of our algorithmic processes as well (Rose 2015). While we do not think this is as fundamental a problem as Rose (2015), we must consider the nature of heterogeneity in a system. As Milanovic (2016, p 235) notes, with research we no longer want to cover up meaningful differences by averaging, but now rather try to uncover dissimilarities. This is true not only for our data source samples, but also for the measurement and model considerations as well.

Clear examples of this exist with online consumer generated data. While the volume of information transferred across the internet exceeds a Zettabyte of data each year, again vast volumes does not mean the same thing as representative. Individuals who share views on politics, products and other topics are often not a representative sample as they are generally more passionate and not necessarily distributed as in the population. While that does not mean there is not great value in such information, it does mean that how it is leveraged needs to consider the implications of this for the research and algorithms at hand.

Finally, we end this section with the question of change and *Are there dynamics that can change the population in the future*. As in the question of heterogeneity, we know the populations are always evolving, but the question here, like in the previous one is whether it is expected to change in ways that are meaningful for the quality of the sample and inferences that

will be made in the models. While "Big Data processes codify the past" this becomes an issue when the past does not reflect current situations and is more critical if they are left unadjusted to changing dynamic situations (O'Neil 2016, p 204).

The level of concern varies by the nature of the data and models used and how large the dynamic shifts are, which will be addressed in the next sections. Technological changes can be critically important for sample dynamics as well. In many ways, we are still early in the expansion of Big Data in many domains. For governmental and corporate purposes, who information is collected for is steadily growing. However, that can create challenges when models are built for specific purposes with selected samples and then the sample frame expands to a more representative, or different expanded selected, sample. In each case, the models developed on the limited base may no longer be appropriate reflections. Considering who information is collected from over time may lead us to explore more dramatic changes in our data and models than simply the dynamics of the system may lead us to believe. Thus, once again, who we are collecting data for is foundational for many of our research questions and the risks of not just accuracy but also impact on people's lives and policy begins there and which lens is used can affect the degree of acceptance with different sample decisions.

What Are We Collecting

In the first section, we focused on the foundational issues of who is included. While the issues of sampling and its quality have been fundamental to survey design, it often receives less attention than it needs with Big Data sources. In this section, we move from who we are collecting information for to specific questions of what we are collecting. This is another area less emphasized with Big Data applications. In our summary here, we are focusing on the

quality of our measures that are going into our scientific explorations and practical applications. Figure 5 includes the questions for our research quality ethical consideration checklist that we will be exploring in more detail in this section.

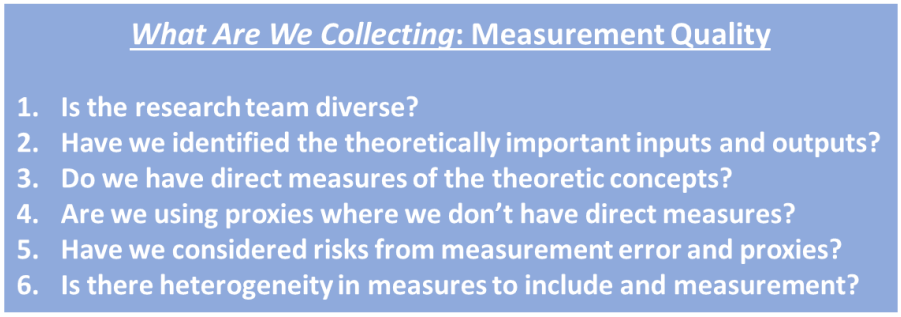

Figure 5: *What Are We Collecting* Questions

Just as in the first section, we again ask *Is the research team diverse*? In this section, this is important because the focus is on what to measure, how we measure it and the quality of those measures. After deciding who to study, researchers must decide what is important for the process and thus what should be included. Not only do algorithms "reflect the values of their creators" (Wachter-Boettcher 2017, p 121), but so does the even more foundational question of what is important to include. Diverse perspectives, including those with different experiences in the domain in question, as well as experts may identify things that may be missed that are important in general or for specific subsets of individuals.

Before considering measurement questions, we must first determine what needs measuring and consideration for inclusion. Here we ask *Have we identified the theoretically important inputs and outputs*? This step is too often skipped in considerations with Big Data for two reasons. The first is that with the wealth of data we often now have access to, the idea is that this is adequate for the tasks we are applying the data (Mayer-Schönberger and Cukier 2013). The second part is that often we have so many measures that atheoretic data mining will allow us to identify the measures of interest from vast data lakes. While it is true, as Japec et. al. (2015)

note that new forms of Big Data can change how we think about behavior more holistically, is that the problem with both of these issues is that very frequently, with survey data, and even more often with Big Data, we are generally leveraging data that was collected for another purpose. Thus, we could have significant gaps in the measures collected from what we may think is theoretically important for decisions. As the data and our algorithms impact people's lives, these gaps can have significant effects if the core drivers are left out of consideration. Thus, taking time to consider what should be included needs to be a first step before moving onto explorations of our data sources more directly. In this way, the idea that we need to marry theoretic considerations from social science with our utilization of data will help to highlight the justice implications of what we are collecting for our models and how we do so (Strong 2015, Yang 2013).

Once we have a sense of what we believe should be at least considered for inclusion, we can ask *Do we have direct measures of the theoretic concept*. As Lazer et. al (2014, p. 1203) state in discussing some traps in Big Data analysis, "quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability and dependencies among data." And we still believe that "[a]pplying measurement theories to facilitate the design of big data driven assessments and evaluate the validity of the outcomes they produce" can assist with and facilitate best practices (Yang 2013, p 125).

Bailey (1988) developed a three-level framework for validity which moves from the theoretic concept to its empirical occurrence or 'true score', and the actual empirical measures captured and used in analysis. Timpone (1998b) extended this to acknowledge that we often do not have the direct measures, and this leads to the question *Are we using proxies where we don't have direct measures.* Figure 6 is the framework from previous work on connections in models of measurement.

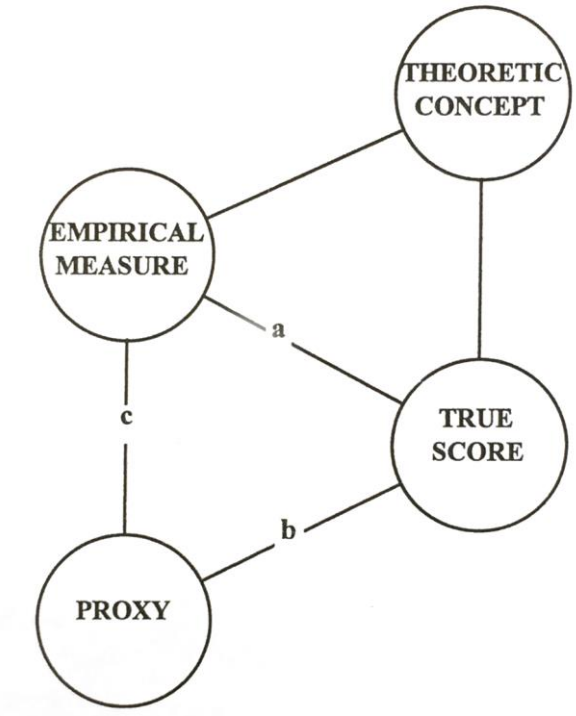

Figure 6: Timpone (1998b) Levels of Measurement

While these questions regarding levels of measurement are generally focused on in the accuracy and *Truth* evaluation of models, here we focus on how failures in these more theoretic elements of our results can have serious consequences for the practical application and their effects in action. O'Neil (2016) notes that in the development of apps, teams routinely lack data for the concepts they are interested in that moves them to use to proxies that can be inexact, unfair and prone to being gamed. In the Allegheny County example mentioned earlier, Eubanks

(2017) notes the challenges for measuring child neglect and maltreatment directly results in using items that are more indicators of poverty and/or subjective. This results in elevating algorithmic evaluations of risk in ways that could have significant consequences such as the removal of children from their families for factors not directly related to neglect or treatment.

Considering the fundamental concepts we believe need to be included goes beyond the question of whether we have quality measures or proxies of them. As many exercises start with the data available, this creates a risk of totally missing critical components. Silver (2012, p 92) highlights the issue of many researchers' biases, where "[o]ne of the most pernicious is to assume that if something cannot easily be quantified, it does not matter." Thus again, starting with the fundamental theoretic questions can help us avoid falling into the traps of starting with the data we have easily available.

While our practical applications and problems usually focus less on the theoretic framework of measurement reflected in Figure 6, as these examples show, the more our models impact people's lives the more we need to explicitly include these considerations in our checklist. While researchers often consider link b that proxies are quality direct indicators of the True Score, the question of this link as well as the risk of measurement error and bias from link a (direct empirical measures) need to be considered. Thus, our checklist question *Have we considered the risks from measurement error and proxies*?

As Achen (1992) made a strong case for not including demographics and proxies in social science research, their consequences for applications directly touching people's lives makes this recommendation even stronger. While we had taken a balanced view previously (Timpone 1998b), given the risks to *Justice* from such decisions, today we side closer to Achen in practice. That said, we do acknowledge that direct measures do not always exist. The question is whether

they can be developed, or how we at least take this issue seriously and be more systematic in establishing the quality of link c and the compounded measurement error that exists through links a and c in Figure 6. Thus, the earlier work highlighting the importance of evaluating these questions is even more important as these decisions (often not consciously considered) impact lives.

Advances in Big Data create additional opportunities whose implications also tie into this issue of measurement error. Given the level of analysis chosen in the first set of questions, there are not only challenges with identifying and capturing appropriate measures, but they may not all be contained in the same databases. The challenge is whether the data across datasets can be linked at the level of analysis chosen. While this is easier at more aggregate levels, such as the nation state, at the individual level, this may not be possible because of who is included in each data source paired with data privacy issues.

If the data can be linked at the level of analysis, the questions above on the quality of the measures is the primary focus. However, if not, the question is whether moving to an aggregated level where the data is still valid (or at least as valid as at the lower level based on measurement and representativeness) will meet the research needs. The additional issue comes up when this is not viewed to be the case. Here data fusion methods are being employed more often to combine Big Data sources. Unfortunately, this is often not done with the full consideration of accuracy at the unit of analysis and the additional measurement error that is introduced.

Again, since the data is often captured for other purposes, the question of how adequate the measures linking sources are in actually predicting the variability in the measures being fused needs to be considered. Like many other specific items on our lists, a full discussion goes beyond the scope of this paper but acknowledging the impact for our data is important.

Before moving to the next question, it is worth building on our discussions of measurement theory. The discussions in this section have focused on questions of accuracy that underlie questions of criterion, construct and predictive validity that are most commonly used in the social sciences (Carmines and Zeller 1979). Thus, our general comments of considering these questions through the lens of *Justice* make sense against that backdrop. It is worth noting that there are domains that view measurement theory more broadly, in particular Education Research. In Educational measurement, the evaluation of validity of measures explicitly includes the consideration of their proposed uses with the idea that the evaluations can vary with different value perspectives on a consequentialist basis (Kane 2013, Messick 1998, Sireci 2013). In this case the precious considerations of measurement theory would include the lens we are recommending, but for most social scientists, as well as practitioners, this is likely a shift in perspective.

We end this section by asking *Is there heterogeneity in the measures to be included and their measurement*. The potential issues of heterogeneity can be relevant for a number of the questions in this section. One of the reasons we focused on diversity is to ensure that we are not missing relevant factors that are important for meaningful subsegments of the population of interest. Beyond this though, how information is measured, especially if different groups of people come from different sources (where stacking creates potential issues different from the fusing of variables) deserves revisiting. While we expect distinct groups to score differently on measures, the question here is whether there are differences in the *quality* of the data that may differ by group. If so, the team should at least explicitly consider how this may affect the performance and implications of the research endeavor, models and practical applications.

How Do We Model

Like Starfield et al. (1990), we consider models to be the purposeful representations of the research question at hand and it is because of the practical applications that more and more models are being leveraged for that their ethical implications and quality must be considered. While Lave and March's (1993) overview of Social Science models as speculations of human existence is still useful for consideration to, it is the transition from speculation and understanding of human behavior (Strong 2015) to action and consequence that leads to raising the importance of the evaluative dimension of *Justice*. While a full consideration of modeling is well beyond the scope of this paper, Figure 7 provides an initial high-level checklist of questions for teams to explicitly answer to avoid further negative consequences of their work.

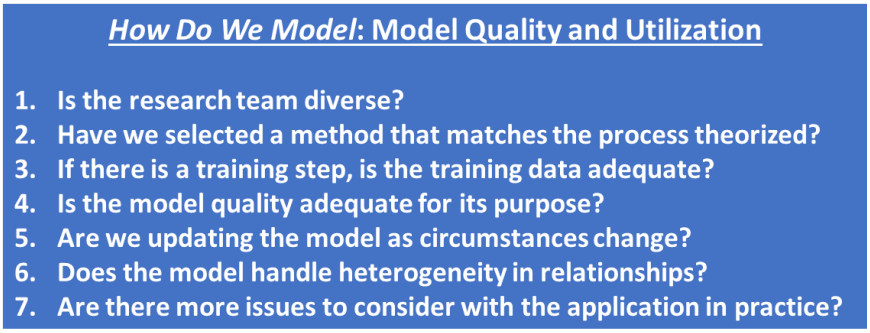

Figure 7: *How Do We Model* Questions

As in each of the previous sections, we again start by asking *Is the research team diverse*? As models are formalizations of the speculations about the processes in question, a more diverse group considering these, especially as they tie directly to human behavior, the better. Just as in the previous areas though, diversity is not a sufficient condition for avoiding negative implications. Every group should consider the breadth of the issue at hand to be better suited to consider the complexity of the situation before moving to the more parsimonious solutions that all research approaches entail.

The next question is one where we will only touch on a handful of elements as we consider *Have we selected a method that matches the process theorized.* Obviously, this is a huge question but considering what the intent is and how to approach it is of fundamental importance. We recommend considering the framework that Pearl and Mackenzie (2018) and determine if methods of association are adequate or if methods that predict the effects of intervention accurately are needed or if moving into the realm of out of sample counterfactual consideration is the appropriate level for the application in question. Sometimes knowing *what* is happening may be enough (Mayer-Schönberger and Cukier 2013), but given ethical issues and risks we often do need to go deeper. This is an area that has been noted could be a fruitful partnership between Big Data and survey research (Callegaro and Yang 2018, Japec et al. 2015).

This framework is relevant to consider many of the modeling assumptions from whether basic approaches are deterministic or probabilistic and how this is incorporated. Additionally, what types of processes reflect the system whether additive, structural, complex and adaptive etc. should be applied as appropriate. Moving to newer methods of Machine Learning and Deep Learning does not eliminate these aspects. In fact, it was shown that even neural network models may be inadequate for handling structural interrelationships while being robust to non-linear and non-additive relations (Egner, Porter and Hart 2011). How far one would need to extend a Deep Learning algorithm to accurately account for structural interrelationships is itself a question to consider if that is the theorized process. Regardless of these different considerations, acknowledging the uncertainty that exists in many algorithms is important to avoid misuse.

While many algorithms still build from regression type frameworks these are designed to maximize the fit to the outcome while the impacts, holding all else constant, can be substantively counterintuitive and in the opposite direction of relationships in a specific set of data when there

are structural interrelationships and multicollinearity among the items. While many applications are moving to Machine and Deep Learning applications, some of the ones noted through this paper for policy recommendations are still built off the regression family of methods (including logit, probit, etc.).

As we move to Machine Learning and Deep Learning methods, the quality of the training data becomes more critical. Thus, the paired question depending on the methods being leveraged *If there is a training step, is the training data adequate?* The earlier example in Figure 1 of the Facebook balloon identification algorithm was an issue where the text parsing and language rules did not adequately include the diverse meanings of the word 'selamat' in Indonesian. This is far from an isolated issue and Wachter-Boettcher (2017, p 145) quotes Computer Science Professor Sorelle Friedler and co-chair of Fairness, Accountability and Transparency in Machine Learning that "one of the biggest concerns is the training data" used to build the models.

An example that was immediately apologized for and addressed was when the Google Photos image tagging was being developed and rolled out. Issues with the adequacy of the method used and the training data when the app first rolled out identified black people as gorillas (Barr 2015, Guynn 2015). While Google immediately acknowledged this was not acceptable and appalling, and immediately addressed the issue, this example further highlights the importance of considering the quality of our models more generally. While they are often probabilistic and will never be deterministically perfect (hence *Truth* as an aspiration and goal not to actually be obtained), the question is *Is the model quality adequate for its task*. If they fall short, are there fixes (even if it means removing some features in deployment, like in the prior example until it can be corrected) to avoid negative consequences for users or individuals affected by the algorithms.

There are many aspects of model quality that are relevant for our different methods. This includes approaches to using out of sample data to avoid overfitting, conducting sensitivity analysis and basic error checking (Gilbert and Troitzsch 2005, Lazer et. al. 2014, Silver 2012, Starfield et al. 1990, Taber and Timpone 1996). The issue of error checking and debugging cannot be overstated. We previously discussed an early example with implications for *Justice* of the Club of Rome report from the early 1970s (Taber and Timpone 1996). This report influenced policy debates for decades and yet some of its most dire predictions were driven by a clerical error in the programming that was not caught in the debugging process. Thus, fundamental issues that are part of model development and testing cannot be overstated given their potential to avoid introducing fundamental flaws and weaknesses with practical implications.

While a number of measures of model quality are necessary but not sufficient, they are useful in at least identifying that there is not any reason for fundamental concern. Eubanks (2017) shares a different model to evaluate individual children at risk of abuse or neglect that highlights this issue. In a model developed in New Zealand, while recommended for deployment, it was discovered that it mis-predicted 70% of the children it identified as highest risk in the historical data itself.

Teams must objectively evaluate whether there are concerns to consider the recommendations from our models. The first author has worked with one corporate team that was leveraging a classification algorithm from an academic in their work. In that case they correctly classified 65% of the cases, and this was used as evidence for the quality of the model. However, it was discovered that the outcome was dichotomous and over 60% were in one category in the data. In this case, the model would not predict worse than this and the evidence was questionable at best. Again, this paper is targeted at research teams and we hope we all are

focused on providing insights, tools and guidance that we truly believe is appropriate and ask these questions of ourselves to avoid negative unintended consequences on people and society.

Even once we are confident in a model that has been developed the next question is *Are we updating the model as circumstances change*? This question is tied to the earlier question of what method was selected for the model and application itself. With true learning models in the Machine Learning and Deep Learning family, information may be updated and used to allow the models to recalibrate themselves. Even in these cases we must consider whether dynamics may go beyond how the elements in the model are reweighted and integrated and potentially need to add new elements that had not been relevant and included in the process previously. Moving to more static rule-based and statistical model implementations, the need for re-evaluating the elements, framework and weighting by the research team takes even more active engagement. Given the impact that these tools will have on people's lives, ensuring they change with the domain to continue to be accurate and fair evaluators is important.

The issue of addressing dynamics is important as Tenner (2018) in his summaries on limitations of Big Data discusses how algorithmic systems are often developed that are less responsive to changing circumstances than they need to be. In this way, algorithms may have an inability to break out of their established patterns even if it would be appropriate for them to do so.

While we have incorporated the question of heterogeneity in each of the previous question lists, it is very important depending on the type of method selected for the modeling. Thus, we again include it in this level and ask *Does the model handle heterogeneity in relationships*? For many of the predictive and classification models used in practice, aggregate outcomes are predicted based on the set of inputs. In these cases, pooling heterogenous groups

can radically alter the outcomes. The issue of Simpson's paradox is well known in theory, but we have seen applications of survey and behavioral data where multiple groups that are evaluated on an outcome fundamentally differently have seen the pooled relationship reversed for one or more groups in practice. This is not simply a theoretic concern and the common thought that if groups differ we would get a weighted average is sometimes simply wishful thinking. If there are different groups, it is useful to examine them separately to determine the robustness of any fuller model that pools them together.

Even if we have developed objectively accurate models and applications that do not produce unintended consequences because of their implementation, certain models can still create issues if they adjust to information that may come from other sources dynamically. Thus, we end this section by asking for those tools, *Are there more issues to consider with the application in practice?*

Obviously for most methods this is not an issue. From rule-based to statistical to even Deep Learning models, in most cases when a set of inputs are put in, a prediction comes out. Even if the outcome has probabilistic elements, if the earlier considerations of sample, measurement and model were answered affirmatively, systematic issues would not be expected here. Again, this may arise if even a truly objective algorithm leverages additional, external dynamic information.

Noble (2018) emphasizes this type of unintended consequence with the auto-fill recommendations in search engines. She contended how the biased views of other users can create offensive lists and recommendations in that "search is a mirror of users' beliefs and that society still holds a variety of sexist ideas about women." (ibid, p 15). Her concern, then, is that the decisions to objectively leverage other users' behavior may lead to negative consequences

and this, itself, is a choice. Such arguments reinforce that active consideration and coding may be necessary if the input of individuals outside of the coding team could create negative unintended consequences outside of their efforts even if theoretically objective and dynamic.

How Do We Know What's Happening

In our final set of ethical consideration model questions, we focus on the issues of transparency and how we even know what is driving the outcomes and recommendations that may be impacting our lives. Opaque applications can raise questions of bias, and as the consequences of these tools increase in importance, the ethical considerations become even more critical. Figure 8 shares the questions, we feel teams should be asking themselves to be able to understand for themselves and disclose to others as much as possible to ensure individuals are *and* feel they are being treated fairly by the increasingly automated tools in our lives. Thus, beyond the objective questions, we need to ensure that the perceptions regarding them are considered as well.

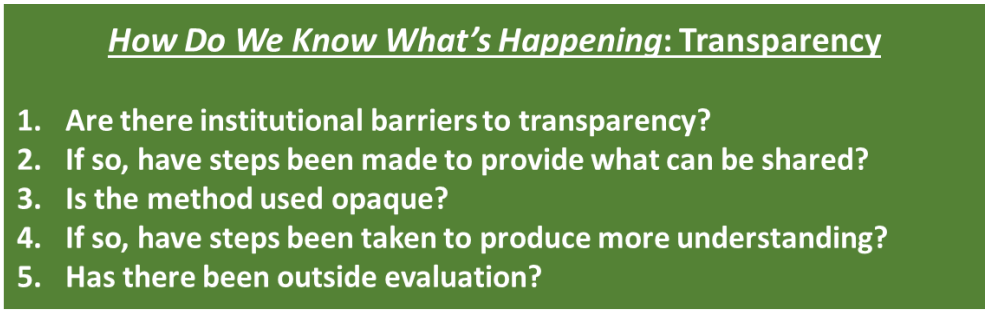

Figure 8: *How Do We Know What's Happening* Questions

This is relevant as practices that are barred by governments could resurface again when it is unclear what goes into and is influencing decisions. For instance, in the United States it is illegal to consider race in housing, lending and other activities, these included policies that were used in the past known as *redlining*. Noble (2018) raises the concern that digital tools may

provide new means of racial profiling which she has dubbed *technological redlining*. The possibility that factors like race, gender and age can feed into processes such as employment decisions as well as other areas where regulations and laws bars this is a risk especially when proxies are used and how decisions in algorithms are made are unclear (Yang 2013). While the previous steps help to avoid such issues occurring unintentionally, addressing the opaque nature of many of these tools will help to further identify where concerns may actually exist and where transparency can help rule out the concerns.

The questions in this section fall into 3 groups, institutional challenges, technical challenges, and outside evaluation. The first question is *Are there institutional barriers to transparency.* Given the value of digital tools, one of the biggest challenges is the protection of the intellectual property of those creating tools designed to make our lives better, more fair, easier etc. Protection of intellectual property is a valuable principle, but it is a double-edged sword for transparency. In some cases, even clients (including policymakers) may not have the information on how the tools work and when they do they often are not able to evaluate them.

It is because of this that we believe the second question here is critical, *If so, have steps been made to provide what can be shared*? This entails both the sharing of specifics but also sharing in a way that can be understood by those making decisions on the tools. As the intent of this paper is to help teams to ask questions to avoid unintended negative ramifications for individuals, groups and society, it must be highlighted that this is not about how to sell but considering how to communicate and disclose the information about tools built off the previous section, so that efforts to avoid unintended consequences are addressed. While we are writing for research teams, we will end the section with the idea of outside evaluation, but before that we

must address the fact that there are technical issues that limit transparency in a growing number of cases as well.

As noted, techniques differ in their fundamental transparency and so we ask the paired questions *Is the method used opaque*, and *If so, have steps been taken to produce more understanding.* The expanding area of prediction explanation to help clarify what is going on in algorithms will continue to grow in importance as more opaque methods like Deep Learning extend their reach. That said, many of the methods do not have fundamental issues of transparency for the research team. Rule-based algorithms and statistic models with clear parameters are clear on what is included in a model, how they are combined and weighting to determine the outcomes. In these cases, the models may not be disclosed for IP reasons, but what is happening in them is clear. This is one reason that these models are still preferred in some applications from personnel selection to determining learning progress placement of students or schools.

This clarity is not the case in all methods and at the extreme is with Deep Learning algorithms. This is one driver for Daugherty and Wilson's (2018) view that an important role for humans in the future will be to help explain what our algorithms and tools are doing. As Deep Learning algorithms continue to be leveraged for more opportunities, the challenges of being able to identify what they are actually including in their decision making and how it is combined will continue to raise questions. Significant effort on predictive explanation is being conducted by academics, companies and governments.

While much work continues to develop in this burgeoning space, we will mention one approach we have been working on that could aid teams in helping to answer the questions for their own algorithms. One domain that our team has been exploring is the idea of low level

vectorization which is an extension of Transfer Learning in Deep Learning algorithms of Goodfellow et al. (2016). This is often used to generalize Deep Learning models to other areas, however we see applications for this to help understand the algorithms themselves better.

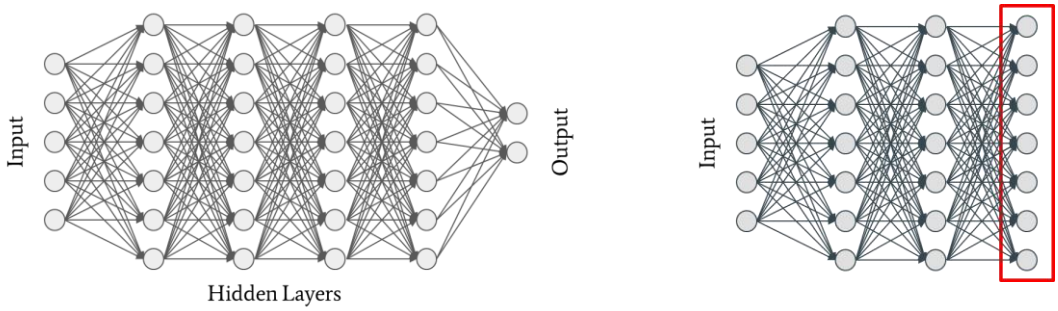

Figure 9: Low Level Vectorization for Deep Learning Insights

Figure 9 provides a visualization of how a deep learning neural net can be truncated for Transfer Learning and to obtain more insight into what is going on in an algorithm. The image on the left shows a full network model from input to final classification. If we truncate the process and obtain the values for the nodes at levels above the final classification, we can examine their attributes to gain some insight into the black box and what is feeding into the outcomes. While only a partial solution, it does allow addressing some of the transparency challenges that exist with more of the newer learning models. This and other techniques will become critical as regulation and clients demand more transparency in practice in the future.

Finally, we ask *Has outside evaluation taken place*. Again, the intent of this paper was to assist research teams. That said, for both the perception of fairness as well as its actualization, being able to produce confidence in the intents and actions of teams is often aided by having independent evaluation. While the need to protect intellectual property will affect what may be done in this space, having others evaluate has benefits for the research effort as well. Actually, there are two different types of outside evaluation that may be considered for the researchers who have made it this far in the paper.

The first level of 'outside' evaluation is *outside of the team.* Thus, in companies, government agencies, academic research teams, it may be useful to have others from within the organization provide an objective review with their fresh, even if not totally independent, perspective. Having this done through the process could avoid discovering fundamental issues far down the research process, but at any stage in the process having an unintended consequence review could be a valuable contribution for the research team itself. The second level of outside evaluation is *outside of the organization* itself. This can provide a more objective set of evaluations, but as noted balancing how this can be done while protecting IP is an issue that must be addressed.

Concluding Comments

We discussed the three standards of evaluating models that we have applied to the practical applications more and more directly affecting our lives- *Truth, Beauty* and *Justice*. Given the risks of harm and manipulation, we have taken the position that the importance of *Justice* and ethical concerns is more important now for these tools than ever before. It should be clear after going through the checklists we have created for research teams to consider though that many of them are also aspects of *Truth* and accuracy.

Now that we have reviewed our *Justice* consideration question lists in detail, one of their features is worth acknowledging. If one goes back and reviews the questions, it can be seen that we posed them each as yes-no items. Even though the timespan of the research process is fundamentally different than Gawande's (2009) other checklists, this was a conscious decision. In our experience, telling people to consider the many items we listed would lead to more superficial acceptance of their status quo assumptions and decisions. Asking more direct yes-no

items forces teams to address potential failures more directly. Thus, we moved from soft consideration that could be more easily ignored to a practical tool to hopefully aid teams in uncovering actual issues.

While many failings of accuracy in our models have ethical consequences, we feel that each of the questions need to be evaluated explicitly through the lens of the impact on different groups and the ethical issues and not simply accuracy. As we noted *Truth* is not something that can actually be obtained. All modelers, academic and applied practitioners, must make simplifying assumptions. It is our concern that when making these choices, the consequences of unintended effects and potentially seriously negative consequences can be missed. If the questions shared are evaluated through the lens of *Justice* as well as accuracy, our tools are more likely to obtain their intended purposes and we hope in the end play a role in making the world a better place.

References

Achen, C. (1992). “Social psychology, demographic variables, and linear regression: Breaking the iron triangle in voting research.” *Political Behavior* 14: 195-211.

Bach, N. (2018). “Facebook Apologizes for Algorithm Mishap That Threw Balloons and Confetti on Indonesia Earthquake Posts.” 8 August, 2018. Fortune.com. (http://fortune.com/2018/08/08/facebook-indonesia-earthquake-selamat-balloons/)

Bailey, K.D. (1988). “The conceptualization of validity: Current Perspectives.” *Social Science Research* 170:117-136.

Barr, A. (2015). “Google Mistakenly Tags Black People as ‘Gorillas,’ Showing Limits of Algorithms.” 1 July 2015. wsj.com (https://blogs.wsj.com/digits/2015/07/01/google-mistakenly-tags-black-people-as-gorillas-showing-limits-of-algorithms/)

Beyer, M.A. and D. Laney. (2012). “The Importance of ‘Big Data’: A Definition.” G00235055. Stamford, CT: Gartner.

Biemer, P.P. (2010). “Total Survey Error: Design, Implementation, and Evaluation.” *Public Opinion Quarterly* 74(5): 817-848.

Biemer, P.P. (2017). “Errors and Inference.” Ch 10 in *Big Data and Social Science: A Practical Guide to Methods and Tools*, eds. I. Foster, R. Ghani, R.S. Jarmin, F. Kreuter, and J. Lane. Boca Raton, London, and New York: CRC Press.

Boone, B. (2017). *Ethics 101*. New York, London: Adams Media.

Brynjolfsson, E.B. and A. McAfee (2014). *The Second Machine Age*. New York and London: W.W. Norton & Co.

Callegaro, M., and Y. Yang (2018). “The Role of Surveys in the Era of ‘Big Data’” Chapter 23 in *The Palgrave Handbook of Survey Research*. Eds. D.L. Vannette and J.A. Krosnick. Palgrave Macmillan.

Carmines, E.G. and R.A. Zeller. (1979). *Reliability and Validity Assessment*, Quantitative Applications in the Social Sciences #17, Sage Publications, Thousand Oaks, London and New Delhi.

Cielen, D., A. Meysman, and A. Mohamed. (2016). *Introducing Data Science: Big Data, Machine Learning, and more, using Python Tools.* Manning Publications.

Daugherty, P.R. and H.J. Wilson. (2018). *Human + Machine: Reimagining Work in the Age of AI.* Boston: Harvard Business Review Press.

Doerr, J. (2018). *Measure What Matters: How Google, Bono, and the Gates Foundation Rock the World with OKRs.* New York: Portfolio/Penguin.

Dormehl, L. (2017) *Thinking Machines: The Quest for Artificial Intelligence—and Where It’s Taking Us Next.* New York: TarcherPerigree.

Egner, M., S. Porter, and R. Hart. (2011). “Key Drivers Methods in Market Research: A Comparative Analysis.” Paper presented at the annual ART Forum.

Egner, M. and R. Timpone. (2015). “Forecasting with Big Data” The Oracle, July 16(3).

Eubanks, V. (2017). *Automating Inequality: How High- Tech Tools Profile, Police, and Punish the Poor*, St. Martin’s Press, New York.

Friedman, T.L. (2016). *Thank You for Being Late: An Optimist’s Guide to Thriving in the Age of Accelerations.* New York: Farrar, Straus and Giroux.

Gawande, A. (2009). *The Checklist Manifesto: How to Get Things Right*. New York: Metropolitan Books, Holt and Co.

Gilbert, N. and K.G. Troitzsch. (2005). *Simulation for the Social Scientist, second edition.* Berkshire and New York: Open University Press.

Goodfellow, I., Y. Bengio, and A. Courville. (2016) *Deep Learning*. Cambridge: MIT Press.

Guynn, J. (2015). "Google Photos labeled black people 'gorillas.'" 1 July, 2015 USAToday.com (https://www.usatoday.com/story/tech/2015/07/01/google-apologizes-after-photos-identify-black-people-as-gorillas/29567465/)

Heckman, J.J. (1979). "Sample Selection Bias as a Specification Error." *Econometrica* 47(January): 153-61.

Japec, L., F. Kreuter, M. Berg, P. Biemer, P. Decker, C. Lampe, J. Lane, C. O'Neil, and A. Usher. (2015). "Big Data in Survey Research: AAPOR Task Force Report." *Public Opinion Quarterly* 79(4): 839-880.

Kane, M.T. (2013). "Validating the Interpretations and Uses of Test Scores." *Journal of Educational Measurement* 50(1): 1-73.

Lave, C.A., and J.G. March. (1993). *An Introduction to Models in the Social Sciences.* University Press of America, Lanham, New York, and London.

Lazer, D., R. Kennedy, G. King, and A Vespignani. (2014). "The Parable of Google Flu: Traps in Big Data Analysis." *Science* vol 343, 14 March: 1203-1205.

Mayer-Schönberger and K. Cukier. (2013). *Big Data: A Revolution that Will Transform How We Live, Work, and Think*. Boston, New York: Eamon Dolan: Houghton Mifflin Harcourt.

Messick, S. (1998). "Test Validity: A Matter of Consequence." *Social Indicators Research* 45: 35-44.

Milanovic, B. (2016). *Global Inequality: A New Approach for the Age of Globalization*. Cambridge, Mass. and London: Belknap Press of Harvard University Press.

Miller, J.H. and S.E. Page (2007). *Complex Adaptive Systems: An Introduction to Computational Models of Social Life.* Princeton and Oxford: Princeton University Press.

Noble, S.U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown New York.

Ostrom, T.M. (1988). "Computer Simulation: The Third Symbol System." *Journal of Experimental Social Psychology* 24:381-391.

Page, S.E. (2007). *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies.* Princeton and Oxford: Princeton University Press.

Pardes, A. (2018). "What is GDPR and Why Should You Care?" 24 May, 2018. Wired.com. (https://www.wired.com/story/how-gdpr-affects-you/)

Pearl, J. and D. Mackenzie. (2018) *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.

Romm, T. (2018). "Google CEO visits White House and Congress to combat charges of anti-conservative bias ahead of key hearing." 28 September, 2018. WashingtonPost.com (https://www.washingtonpost.com/technology/2018/09/28/google-ceo-visits-congress-combat-charges-conservative-bias-ahead-key-hearing)

Rose, T. (2015). *The End of Average: How We Succeed in a World That Values Sameness*. New York: Harper One.

Seetharaman, D. (2018). "Pro-Kavanaugh Executive Stirs Ire at Facebook." 5 October, 2018. The Wall Street Journal.

Silver, N. (2012). *The Signal and the Noise: Why So Many Predictions Fail- But Some Don't.* New York: Penguin.

Simonite, T. (2018). "When It Comes to Gorillas, Google Photos Remains Blind." 11 January, 2018. Wired.com (https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/)

Sireci, S.G. (2013). "Agreeing on Validity Arguments." *Journal of Educational Measurement* 50(1): 99-104.

Starfield, A.M., K.A. Smith, and A.L. Bleloch. (1990). *How to Model It: Problem Solving for the Computer Age.* New York: McGraw Hill.

Stevenson, B. (2014). *Just Mercy: A Story of Justice and Redemption*. New York: Spiegel & Grau.

Strong, C. (2015). *Humanizing Big Data: Marketing at the Meeting of Data, Social Science & Consumer Insight*. London, Philadelphia, New Delhi: Kogan Page.

Taber, C.S. (1992) "POLI: An expert system model of U.S. foreign policy belief systems." *American Political Science Review, 86:* 888-904.

Taber, C.S. and R.J. Timpone. (1994). "The Policy Arguer: The Architecture of an Expert System." *Social Science Computer Review* 12: 1-25.

Taber, C.S. and R.J. Timpone. (1996). *Computational Modeling*, Quantitative Applications in the Social Sciences #113, Sage Publications, Thousand Oaks, London and New Delhi.

Tenner, E. (2018). *The Efficiency Paradox: What Big Data Can't Do*. New York: Alfred A. Knopf.

Timpone, R.J. (1998a). "Structure, Behavior and Voter Turnout in the United States." *American Political Science Review* 92:145-158.

Timpone, R.J. (1998b). “Ties that Bind: Measurement, Demographics, and Social Connectedness.” *Political Behavior* 20(1): 53-77.

Timpone, R.J. (2002). “Estimating Aggregate Policy Reform Effects: New Baselines for Registration, Participation, and Representation.” *Political Analysis* 10(2): 154-177.

Timpone, R. (2016) Big Data: A Guided Tour. Ipsos Views #3.

Tolich, M. (2018). “Big Data's Front-Ended Ethical Considerations Ignore How Results Can Stigmatize Identifiable Groups: Examining Big Wastewater Data in New Zealand”. Paper prepared for the BigSurv 2018 Conference.

Wachter-Boettcher, S. (2017). *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech*. W.W. Norton and Co., New York and London.

Yang, Y. (2013). “’Big Data’ Technologies: Problem or Solution?” *The Industrial Organizational Psychologist* 51(2): 119-126.

Young, C. (2018) Our Age of Uncertainty. Ipsos Point of View.