# TABLE OF CONTENTS

Summary...........................................................................................................................................2

Introduction......................................................................................................................................2
   Changes in Computer Technology..................................................................................................2
   Fairness Challenges..........................................................................................................................3

A Framework of Responsible Use.................................................................................................4
   Data Analytics to Promote Fairness.................................................................................................6
      Consumer Credit.........................................................................................................................6
      Detecting and Remediying Discrimination..................................................................................7
   Legal Compliance..........................................................................................................................8
   Assessments Using Disparate Impact Analysis.............................................................................9
   Limitations of Source Code Disclosure........................................................................................12
   Issues for Discussion.....................................................................................................................13
      Stakeholder Roles......................................................................................................................13
      Metrics.......................................................................................................................................13
      Scope.........................................................................................................................................14
      Standards of Fairness................................................................................................................14

Conclusion......................................................................................................................................16

Endnotes.........................................................................................................................................17

---

The Software & Information Industry Association (SIIA) is an umbrella association representing 800+ technology, data, and media companies globally. Industry leaders work through SIIA’s divisions to address issues and challenges that impact their industry segments with the goal of driving innovation and growth for the industry and each member company. This is accomplished through in-person and online business development opportunities, peer networking, corporate education, intellectual property protection, and government relations.

For more information, visit siia.net.

Copyright © November 4, 2016. All rights reserved.

SIIA occasionally releases Issue Briefs on the range of issues that might be raised by particular developments in technology or in organizational practice. They are intended to scope the issues that might need to be addressed by policymakers but are not intended to take particular policy positions.
Summary

Big data analytics, including machine learning and artificial intelligence, are natural outgrowths of recent developments in computer technology including the availability of massive data sets, vast increases in computing power, and breakthroughs in analytical techniques. These techniques promise unprecedented benefits for consumers, workers and society at large. While they can also pose challenges for fairness, they are not exempt from existing anti-discrimination and consumer protection laws. Regulatory agencies and courts should enforce these laws against any abuses accomplished through big data analysis. Disclosure of source code, however, is not an effective way to respond to the challenges of designing and using unbiased algorithms. Instead, enterprises should develop and implement a framework for responsible use of data analytics that will provide for fairness by design and internal assessments of algorithms in use. Such a framework will need to develop appropriate standards of fairness and responses for findings of disparate impact. This will require moving beyond technical matters to address sensitive normative issues where the interests of different groups collide and moral intuitions diverge. A collaborative effort of businesses, government, academics, civil rights and public interest groups might sharpen the issues and allow sharing of information and best practices in a way that would benefit all.

Introduction

Discussions of algorithmic fairness have increased in the last several years, even though the underlying issues of disparate impact on protected classes have been around for decades. They appear primed to move into the mainstream of policy discussions in the new Congress and Administration. This renewed discussion derives in part from changes in computer technology that have driven increased use of powerful data analytics tools in more and more areas of social and economic life. This introduction provides some context for what is likely to be an increasing focus of policy discussion.

Changes in Computer Technology

Today analytics firms, data scientists and technology companies have valuable new tools at their disposal, derived from three interlocking developments in computer technology. Data sets have dramatically increased in volume, variety and velocity. Processing capacity and storage capacity has increased, accommodating and reinforcing these changes. And new analytic techniques that were not effective at lower scale and slower processing speeds have had spectacular successes.

Data is available to far more actors than previously and actors throughout the economy have access to the new analytic techniques. This increase in data availability and access to algorithms creates its own virtuous cycle of improvement in still better analytic techniques and even greater availability of data. Data volume is important, but the application of analytic techniques has created high quality data, smart data that reveals important relationships.¹

New analytic techniques can discover connections and patterns in data that were often invisible with smaller data sets and with older techniques. Earlier researchers would approach a defined data set with a well-formulated hypothesis and proceed to used standard statistical techniques such as multi-variate regression analysis to test the hypothesis. Researchers brought background
knowledge, theoretical understanding and intuitions into the process of hypothesis creation, and hoped to find a pattern in the data that would verify this hypothesis. But the data itself were silent and would tell them nothing. In contrast, new analytic techniques based on machine learning discover connections in the data that the researcher had not even dreamed of. The data speak for themselves, leading to completely novel and unexpected connections between factors that had previously been thought of as unrelated.

Machine learning programs get better as they are exposed to more data, which the spectacular growth of the Internet was able to provide in ever increasing amounts. These programs adjust themselves as they are exposed to new data, evolving not only from the original design of the programs but also from the weights developed from their exposure to earlier training data.

Because these new machine-learning techniques are not pre-programmed with humanly created rules, their operation can sometimes resist human comprehension. Often, it is “…impossible for the creators of machine learning programs to peer into their intricate, evolving structure to understand or explain what they know or how they solve a problem.”² In addition, they rely on correlations found in data, rather than empirically or theoretically comprehensible causal connections: “In a big-data world...we won’t have to be fixated on causality; instead we can discover patterns and correlations in the data that offer us novel and invaluable insights.”³

Despite the difficulty in discerning the logical or causal structure uncovered by these machine-learning algorithms, they are increasingly moving out of computer science departments and providing substantial benefits in real world applications.

Recent studies document the domains in which these techniques are being used and where their impact is likely to be greatest in the coming years, including a 2014 Big Data Report from the Obama administration⁴ and a 2016 report by the AI Study Group, a panel of industry and academic experts.⁵ SIIA has also documented the benefits of these new technologies in a series of studies and presentations.⁶

**Fairness Challenges**

Although some uses of algorithms can provide benefits directly to historically disadvantaged groups and can be used to detect and remedy discrimination, the increased use of big data analytics also raises concerns about fairness. Are the algorithms accurate? Do they utilize characteristics like race and gender in ways that raise issues of discrimination? How do we know? Can people have redress if an algorithm gets it wrong or has a disparate impact on protected classes?⁷

Policymakers have issued reports and held workshops on these questions over the last two years. The Obama Administration’s 2016 report on big data and civil rights highlighted concerns about possible discriminatory use of big data in credit, employment, education, and criminal justice.⁸ The Federal Trade Commission held a workshop and issued a report on the possible use of big data as a
tool for exclusion.9 The Obama Administration’s series of workshops and final report on preparing for the future of AI also raised these issues.10

Even when companies do not intend to discriminate and deliberately avoid the use of suspect classifications like race and gender, the output of an analytical process can have a disparate impact on a protected class when a variable or combination of variables is correlated both with the suspect classification and the output variable. These correlations might be the result of historical discrimination that put vulnerable people at a disadvantage. The end result is that analytics relying on existing data could reinforce and worsen past discriminatory practices.11

Civil rights groups are concerned that the new techniques of data analysis will create additional challenges for minority groups. They have developed principles aimed at protecting civil liberties in an age of big data12 and have focused on the possibility that new techniques of analysis will be used to target minorities for discriminatory surveillance.13

In particular, the use of data analytics in sentencing has attracted substantial attention. One study from ProPublica of the recidivism models used for sentencing indicated that they had a disparate impact.14 In 2014, then-Attorney General Eric Holder questioned the legitimacy of using these models for sentencing, since they rely on variables correlated with race like education level, family circumstances and employment history that do not have anything to do with the crime, citing that these factors may deepen adverse impacts on already disadvantaged groups.15 Other studies suggest some scores commonly used in sentencing are not simply proxies for race.16 A recent Wisconsin case determined that courts could take risk assessments into account in sentencing, provided that these “risk scores may not be considered as the determinative factor in deciding whether the offender can be supervised safely and effectively in the community.”17

A Framework of Responsible Use

A framework of responsible use would help ensure algorithmic fairness. FTC Commissioner Terrell McSweeney has called for something like this in urging a framework of “responsibility by design” that would test algorithms – at the development stage - for potential bias. Fairness by design should be supplemented by internal assessment of algorithms in use to ensure that properly designed algorithms continue to operate properly.18 The Obama Administration called for a similar practice of “equal opportunity by design.”19

Individual enterprises could address the issues involved in the construction of a framework of responsible use, and in the end, it might be a matter of balancing business needs, legal risk and
social responsibility in ways that best fits the context of the individual company. For instance, care in the use of statistical models that rely on commute time can be a matter of a company's own individual determination of how to manage legal risk and public perceptions, and some companies choose not to use that information in their hiring decisions.20

Nevertheless, a collaborative effort involving a range of stakeholders might sharpen the issues and allow sharing of information and best practices in a way that would benefit all. In such a collaborative effort, businesses, government, academics and civil rights and public interest groups would come together to establish a clear operational framework for responsible use of big data analytics. The tech industry has begun to organize itself for this task with the formation of a Partnership on AI “to advance public understanding of artificial intelligence technologies (AI) and formulate best practices on the challenges and opportunities within the field.”21 A collaborative effort might build on industry efforts to formulate and implement essential elements of accountability to ensure that data is “created and used in a legal, fair and just way.”22

There are challenges in building a single framework for responsible use of big data analytics. It is certainly true that the risks and considerations involved in the use of a technique such as machine learning depends on the domain of use.23 The legal standards differ as well.24 Still a stakeholder group might assess whether there are actionable general principles that could be applied successfully in many fields.

It is important to begin to develop this framework now, and to get it right. The public needs to be confident in the fairness of algorithms, or a backlash will threaten their very real and substantial benefits. Stakeholders need to do more to ensure the uses of the new technology are, and are perceived to be, fair to all.

This section describes the key elements that might be a part of a framework of responsible use. First, companies might want to consider seeking out affirmative ways to use data analytics to ensure fairness, as illustrated by the use of alternative credit scores to improve the credit granting process for historically underserved groups, and the ways in which data analytics can detect and help remedy bias.

Second, the current framework of non-discrimination and consumer protection law applies to the use of big data analytics. Companies need to have adequate resources devoted to compliance with current law.

Third, proposals for disclosure of source code are unnecessary and counterproductive. However, companies might want to consider the extent to which they can provide an indication of the kinds of factors that go into the decision-making algorithms they use.

Fourth, when algorithms are used in ways that affect fundamental interests and rights of people, companies should conduct internal assessments both at the design stage and as they are actually employed in practice. The methods developed for disparate impact analysis used under current law provides a guide for doing this. A number of issues related to assessments need to be addressed
including the metrics used, the circumstances under which it makes sense to conduct them, and the role of different stakeholders, including outside researchers, who could provide valuable guidance on the methodologies for internal assessments. It is important to focus these efforts on uses that create consequential impacts on people’s lives and where there is a significant risk for individual or societal harm.

Fifth, companies should develop standards of fairness to guide their actions in response to their internal assessments. In particular, they should understand the extent to which they are aiming at accuracy in prediction and the extent to which they are seeking to prevent disproportionate adverse impacts on protected groups. These standards will help to guide their actions in response to any findings of disparate impact.

Data Analytics to Ensure Fairness

Two areas where affirmative use of data analytics can ensure fairness are consumer credit and detecting discrimination. These examples illustrate the potential for these new technologies to vastly improve the economic and social conditions for historically disadvantaged groups.

Consumer Credit

Credit scoring models have been used for decades to increase the accuracy and efficiency of credit granting. They help as many people as possible to receive offers of credit on terms they can afford; and they allow lenders to efficiently manage credit risk. They improve upon the older judgmental systems that relied excessively on subjective assessments by loan officers.

The traditional credit scores are built from information in credit bureau reports and typically use variables relating to credit history. But these traditional credit scores are not able to score approximately 70 million individuals who lack credit reports or have “thin” credit reports without enough data to generate a credit score.

This inability to score no-file or thin-file individuals differentially affects historically disadvantaged minorities. A recent Lexis-Nexis study found that 41% of historically underserved minority populations of Hispanics and African-Americans could not be scored using traditional methods, while the unscorable rate for the general population was only 24%. Minorities face an unscorable rate that is 1.7 times the rate – almost twice – the rate for the general population.²⁵

To remedy this limitation, companies are looking beyond the information contained in credit reports to alternative data sources and they are building credit scores based on this additional data. For instance, an alternative credit score, called RiskView, built by Lexis-Nexis relies on public and institutional data such as educational history and professional licensing, property asset and ownership data such as home ownership, and court-sourced items such as foreclosures, evictions, bankruptcies, and tax liens.

The Lexis-Nexis report demonstrated the extent to which credit risk scores built from alternative data can help to extend credit to unscorable consumers, finding that fully 81% of the unscorable minorities received a RiskView score. A major benefit of alternative credit scores is the improvement in the availability of credit for historically underserved minority groups.
Not every new model based on alternative data will be as predictive as the standard credit scoring models and there is an important distinction between mainstream alternative data and less proven data sources. But there are a number of checks in place to prevent abuse. New scoring models based on alternative data are subject to the same regulatory scrutiny as traditional scores. Some social media platforms do not use information about their users for eligibility decisions and do not allow other companies to use this information for eligibility decisions. Moreover, the market will not support inaccurate models. For instance, some rethinking is already taking place on the appropriate role of social media information in determining creditworthiness.

**Detecting and Remediary Discrimination**

Human biases are notorious and often unconscious. Classical music orchestras were almost entirely male for generations, despite the denials of bias by conductors who exhibited no gender bias in any other aspect of their lives. But arranging auditions to be held behind a screen that hid the gender of the aspiring musician produced a dramatic change toward gender neutrality. Eliminating information that biased human judgment led to fairer outcomes.

More elaborate data analysis can also detect totally unconscious biases. Judges are trained to conscientiously make good faith efforts to be impartial. Still one study in Israel found that at the beginning of the workday, judges granted around two-thirds of parole requests, but that approvals fell steadily until food breaks, after which the judges again granted most of the parole requests.

Statistical techniques can be used to assess whether employment hiring and promotion practices are fair and provide the basis for taking remedial steps. Google publishes its diversity report regularly. Other tech companies including Amazon, Apple, Facebook, LinkedIn, Microsoft, Twitter and Yahoo now do so as well. Google has pioneered efforts to diversify its workplace through workshops to its employees on detected and dealing with unconscious bias. Software recruiting tools can also be used to help employers diversify their workforce to correct the underrepresentation of certain groups in their workforces.

Data analysis can detect whether a statistical model has disproportionate adverse effects on protected classes. For instance, non-mortgage financial institutions do not have information about the race and ethnicity of their applicants and customers. To assess whether their statistical models comply with fair lending rules they can use publicly available information on surnames and geolocation as reliable predictors of these characteristics, and advanced statistical techniques can improve the predictive accuracy of these factors.

Finally, it is often important to collect accurate and current data about underrepresented groups and protected classes in order to correct historical biases and incompleteness in data sets and to
ensure that they receive services that are generally available to the public and to detect and remedy discriminatory practices.\textsuperscript{35}

**Legal Compliance**

A key element of a framework for responsible use of big data analytics is compliance with the current legal framework. Any technology, new or old, can further illegal or harmful activities, and big data analysis is no exception. But neither are the latest computational tools an exception from existing laws that protect consumers and citizens from harm and discrimination.

The Fair Credit Reporting Act sets out requirements for credit reporting agencies, including access and correction and notification of adverse action.\textsuperscript{36} FCRA was put in place to deal with computerized credit reporting agencies in the 1970s, but it applies to decisions made using big data and the latest machine learning algorithms, including to third party companies that combined social media data with other information to create profiles of people applying for jobs.\textsuperscript{37}

A second element of the current legal framework is the prohibition on discrimination against protected groups for particular activities. These statutory constraints on discrimination include:

- Title VII of the Civil Rights Act of 1964 makes it unlawful for employers and employment agencies to discriminate against an applicant or employee because of such individual’s “race, color, religion, sex, or national origin.”\textsuperscript{38}

- The Equal Credit Opportunity Act makes it unlawful for any creditor to discriminate against any applicant for credit on the basis of “race, color, religion, national origin, sex or marital status, or age.”\textsuperscript{39}

- Title VIII of the Civil Rights Act of 1968, the Fair Housing Act, prohibits discrimination in the sale, rental or financing of housing “because of race, color, religion, sex, familial status, or national origin.”\textsuperscript{40} The act also protects people with disabilities and families with children.

- The Age Discrimination in Employment Act of 1967 (ADEA) makes it unlawful for an employer to refuse to hire or to discharge or to otherwise discriminate against any individual or, because of the individual’s age\textsuperscript{41}

- The Genetic Information Nondiscrimination Act of 2008 prohibits U.S. health insurance companies and employers from discriminating on the basis of information derived from genetic tests.\textsuperscript{42}

- Section 1557 of the Affordable Care Act of 2010 prohibits discrimination in health care and health insurance based on race, color, national origin, age, disability, or sex.\textsuperscript{43}
These laws apply to the use of any statistical techniques, including big data analytics, as the Obama Administration recognized when they recommended that regulatory agencies “should expand their technical expertise to be able to identify practices and outcomes facilitated by big data analytics that have a discriminatory impact on protected classes, and develop a plan for investigating and resolving violations of law in such cases.”

The U.S. constitution also protects citizens against discriminatory practices. The equal protection clause of the Fourteenth Amendment of the U.S. constitution provides that no state shall deny to any citizen within its jurisdiction the “equal protection of the laws.” The due process clause of the Fourteenth Amendment provides that no State shall “deprive any person of life, liberty, or property, without due process of law.” The due process clause of the Fifth Amendment extends this due process protection providing that no person shall “be deprived of life, liberty, or property, without due process of law.”

The Supreme Court invoked the equal protection clause in its Brown v Board of Education decision that outlawed racial segregation in public schools. In the jurisprudence that followed, the Court applied tiered scrutiny, focusing on whether the government action involved fundamental rights and whether it involved a suspect classification.

Assessments Using Disparate Impact Analysis

It is true that big data analytics might have discriminatory effects, even when companies do not intend to discriminate and do not use sensitive classifiers like race and gender. But social scientists and policymakers have long known that statistical techniques and inferences can have discriminatory effects. When discrimination arises indirectly through the use of statistical techniques, regulatory agencies and courts often use a disparate impact assessment to determine whether the practice is prohibited discrimination.

Disparate impact analysis is controversial because it focuses on the effects of a policy, practice or procedure rather than on its motivation or intent. Yet courts have upheld the use of disparate impact analysis to assess discrimination in a wide variety of circumstances.

Title VII of the Civil Rights Act of 1964 forbids any employment practice that causes a disparate impact on a prohibited basis if the practice is not “job related for the position in question and consistent with business necessity” or if there exists an “alternative employment practice” that could meet the employer or employment agency’s needs without causing the disparate impact.

On June 25, 2015, the Supreme Court, by a five-to-four margin, upheld the application of disparate impact under the Fair Housing Act.

While there has been no separate Supreme Court ruling on disparate impact under the fair lending laws, the Consumer Financial Protection Board has held that disparate impact claims are cognizable
under ECOA.\textsuperscript{50} It reaffirmed this view following the decision upholding disparate impact analysis under the housing laws.\textsuperscript{51}

The Supreme Court has recognized that disparate impact analysis is permitted under the Age Discrimination in Employment Act of 1967 (ADEA). The plaintiff must show that employers used a specific policy practice or procedure that resulted in the disparate impact, and employers can defend themselves by showing that this policy or practice was based on reasonable factors other than age and they do not have to prove that this reasonable business practice had less impact on older workers than other possible alternatives.\textsuperscript{52} Standard best practice for employers, however, recommends that companies examine alternative policies to see if they would have a smaller disparate impact.\textsuperscript{53}

In some areas of non-discrimination law, disparate impact analysis is limited. The Genetic Information Nondiscrimination Act of 2008 explicitly prohibits disparate impact analysis.\textsuperscript{54} But some commentators have suggested adding a clause to GINA that allows for disparate impact cases.\textsuperscript{55}

Supreme Court cases have ruled that disparate impact is not covered under Title VI of the Civil Rights Act of 1964 and regulations prohibiting Federal agency practices that have a discriminatory effect are not enforceable by private litigants.\textsuperscript{56} Nevertheless, most Federal agencies have promulgated regulations that forbid agency practices that have “the effect of subjecting individuals to discrimination because of their race, color or national origin.”\textsuperscript{57}

In 1977, the Court ruled that a disproportionate impact on a protected class was not enough to show a violation of the equal protection clause: “Proof of racially discriminatory intent or purpose is required to show a violation of the Equal Protection Clause...Official action will not be held unconstitutional solely because it results in a racially disproportionate impact.... Such impact...is not the sole touchstone of an invidious racial discrimination...A racially discriminatory intent, as evidenced by such factors as disproportionate impact, the historical background of the challenged decision, the specific antecedent events, departures from normal procedures, and contemporary statements of the decision makers, must be shown.”\textsuperscript{58}

Assessing discriminatory effects is a key part of assessing compliance with statutory and constitutional prohibitions on discrimination. It will also be key to fairness in the design of decision-making algorithms and also to assessments of algorithms after they have been designed to ensure that they maintain fairness in practice and as they evolve and adjust themselves as they are used.

Interpretation of discrimination law on disparate impact analysis is complex and controversial. It often involves shifts in the burden of proof that have important legal ramifications, but might not be crucial in the conduct of a non-legal assessment. But in general terms, disparate impact assessment analysis has three stages: evidence of a disproportionate impact caused by a policy or procedure, an
assessment of whether and the extent to which the policy or procedure serves a valid purpose, and
an assessment of whether there are alternative policies or procedures that would achieve the
legitimate objective with less of a disparate impact.

The first stage is an understanding of the effect of a law, policy or procedure on members of the
protected class compared to other people who are not members of this class. The assessment has to
find an adverse effect on a protected class. A finding of a difference that is not adverse might not be
a signal that there is an issue. For instance, a product or service designed to meet the special
interests of a racial, ethnic or religious community might be targeted marketed to these groups. An
assessment of this marketing effort that shows it disproportionately reached these groups would be
a sign of success, not discriminatory action harming the protected group.

The law is generally clear that there must be a nexus or causal connection between some element of
institutional practices and the disparate outcome in order for there to be a finding of illegal
discrimination. A disparity all by itself does not establish a discriminatory issue. In addition, a
plaintiff must be able to point to a policy or procedure causing the disparity.59 Often, however, the
discovery of the presence of a disproportionate adverse impact can lead an organization to take
steps to change its practices to avoid this outcome, even in the absence of legal liability. Google
took the step after an internal assessment showed that hiring and promotion were less favorable for
women and minorities.

The second step is the evaluation of the rationale for the policy or procedure. Generally,
institutional policies are not contrary to the discrimination laws unless they are “artificial, arbitrary,
and unnecessary barriers.”60 The discrimination laws are not intended to prevent institutions from
achieving legitimate objectives and valid purposes. But if a policy or procedure creates is a
disproportionate adverse impact, that policy or procedure must satisfy a legitimate need. The policy
must have a reasonable basis. The second stage examines the institutional rationale for its practice
or procedure, evaluates the legitimacy of this goal and assesses the extent to which it meets the
objective.

The last step is to assess whether there is an alternative policy that would achieve this legitimate
interest with a smaller disparate impact. This is often a crucial step in assessing responses to a
finding of disparate impact. It is the step that explains why the practice of redlining is discriminatory.
Refusing to lend to all people in a heavily minority neighborhood has a disproportionate adverse
impact on a protected class, but it might have a legitimate purpose, namely, to avoid making loans
that could not be repaid. On average, it might be argued, people in those neighborhoods are more
likely to be bad credit risks. But there are more effective ways of achieving the same purpose such
as relying on credit scores that assess people in the neighborhood on the basis of their own
individual risks rather than relying on the neighborhood average.

A 2007 study of credit insurance scores by the FTC illustrates how a statistical technique can be
assessed for disparate impact. The study found that credit insurance scores predict automobile
insurance risk; protected classes had lower scores and so paid higher insurance premiums; the
predictive power of the score did not derive from correlation with a protected class; no alternative
model had equivalent predictive power but less adverse impact on protected classes.61
The study seems to indicate that the scores would pass a disparate impact test. There was a disproportionate adverse impact on a protected class. But the score was not just a proxy for race, and it satisfied the legitimate business need of controlling auto insurance risk. And no alternative model satisfied the business need with less impact on the protected class.

Credit scoring models are routinely tested for compliance with fair lending laws and methodologies have been developed to assess the risk of failing a disparate impact test. Reviews of some of these assessments have been made public. Studies of disparate impact in the financial world include the Federal Trade Commission study on insurance credit scores, a Payment Card Center study of credit cards, and a Federal Reserve Board study of credit scores and the availability of credit.

Existing rules for due diligence apply when newer techniques of big data analysis such as machine learning algorithms are used. When these techniques are used in the regulated contexts of housing, credit granting, employment and insurance, they are subject to the same regulatory controls and validation requirements that apply to any statistical methodology used in these contexts.

Limitations of Source Code Disclosure

To address the questions of fairness in the use of new computational techniques, some commentators have called for the disclosure of the source code embodied in a decision-making or classificatory algorithm. In this view, one of the major changes of rendering decision making more computational under big data analysis is that the standards and criteria for making decisions have become more opaque to public scrutiny and understanding. Disclosure would allow outsiders an effective way to evaluate the basis for the decisions made by the programs. Along with a right to appeal a decision before an independent body, it would provide due process protection for people when algorithms are used to make decisions about them.

Public disclosure of source code would be a mistake. Commercial algorithms are often proprietary and are deliberately kept as trade secrets in order to provide companies with a competitive advantage. In addition, revealing enough about the algorithm so that outside parties can predict its outcomes can defeat the goal of using the formula. For instance, the process and criteria for deciding who to audit for tax purposes or who to select for terrorist screening must be opaque to prevent people from gaming the system. Revealing source code could also allow hackers to game the system, thereby creating security risks.

In addition, disclosure of code will not really address the problem of bias in decision-making. Source code is only understandable by experts. And even for them it is hard to understand what a program will do based solely on the source code. In the case of machine learning algorithms, the decision rule is not imposed from outside, but emerges from the data under analysis. Even experts have little understanding of why the decisional output is what it is. In addition, the weights associated with each of the factors in a machine learning system change as new data is fed into the system and the
program updates itself to improve accuracy. Knowing what the code is at any one time will not provide an understanding of how the system evolves in use.68

Still some understanding of the “narrative” behind algorithms might accomplish the goals of algorithmic transparency. Traditional credit scoring companies such as FICO routinely release the general factors that power their models and the rough importance of these factors. For instance, payment history contributes 35% to the overall score and amounts owed contributes 30%.69 Designers and users of newer statistical techniques might consider the extent to which they could provide the public with a story to accompany the output of their statistical models. For instance, Carnegie Mellon researchers have developed a method for determining why an AI system makes particular decisions without having to divulge the underlying workings of the system or code.70

**Issues for Discussion**

*Stakeholder Roles*

The framework would need to address the proper roles of the public, developers and users of algorithms, regulators, independent researchers, and subject matter experts, including ethics experts. How much does the public need to know about the inner workings of algorithms? What are the different responsibilities of the developers of analytics, the furnishers of data and the users? In what circumstance, if at all, should regulators be involved in the assessment of fairness? In areas where there are no legal responsibilities is there a role for government to act as a convener? In what ways can companies, leverage the expertise and work of independent researchers in order to ensure the fairness of their practices? 71

*Metrics*

Fairness involves protecting certain classes of people against disproportionate adverse impacts in certain areas. But how do we measure this? Statisticians, economists, and computer scientists, among others, are working in the growing field of metrics designed to measure disparate impact.72 Deviation from statistical parity is one measure. But there are others.73 Sometimes analysis of training data can reveal the possibility of disparate impact in the use of algorithms.74

An “80 percent” rule is used as a rule of thumb in employment law, for instance. Under that rule, an employment policy or practice has a disproportionate adverse impact on a protected class when the members of that class are selected at a rate less than 80% of the rate at which non-protected members are selected.75

It might be important to develop similar thresholds of disproportionate burden that suggest possible illegal discrimination in other fields. It is also important to measure how much loss of accuracy would result from using alternative statistical models. To assess the central normative questions,
we need good measurement of the size of the disparate impact and the loss of accuracy, if any, that might be involved in remedial action. Many of these questions may ultimately be a matter of judgment but effective metrics can structure these judgments so that the issues can be faced squarely.76

Individual enterprises or stakeholders will need to survey these methods, develop a process for keeping up with developments in this fast-moving field and integrating the most effective methodologies into the process of internal assessment and testing.

**Scope**

Current law does not protect people in suspect classifications in every area of economic and social life. Nevertheless, those who design, implement and use data analytics systems should be thoughtful about the potential for discriminatory effects in any field. Two key elements in assessing fairness in decision-making are the group of people who are being assessed and the context of assessment. When the people involved belong to a class that deserves special protection and the decision-making is in a context that affects their fundamental rights and creates a significant potential for individual or societal harm, any significant deviation from statistical parity must be examined carefully. In determining the need for internal assessments, it is of paramount importance to focus on the context of use, the nature of the application of the data analytic system and the potential for substantial harm.

In many cases, businesses would want to know whether their use of these tools has disproportionate adverse impacts on protected classes. But a norm calling for assessments of all statistical models regardless of context, application or potential for harm is too broad, since it would extend to areas where there is no consensus that making distinctions is an issue. In the stakeholder approach to developing a common framework of responsible use, interested parties would need to discuss which areas of social and economic life to include in a responsible use model.

**Standards of Fairness**

As mentioned earlier, the FTC found that credit insurance scores have a disparate impact, and are also predictive of automobile insurance risk, not simply proxies for race and ethnicity.77 The Federal Reserve Board had a similar finding that traditional credit scores have a disparate impact and are also accurate predictors of creditworthiness. These findings were viewed as confirmation that the scores were not discriminatory.78 Still many think these uses are unfair and most states limit the use of credit insurance scores.79

Assessments of algorithms appear to be about data, statistics and analytics. But in reality, they are often disputes about contested standards of fairness. Is fairness a matter of reducing the subordination of disadvantaged groups or the avoiding the arbitrary misclassification of individuals.80 Should analytics aim only at accurate predictions or should it also aim at statistical parity of protected groups? 81 Is fairness simply accuracy in classification? Or does fairness call for
some sacrifice of accuracy in order to protect vulnerable groups? Are there some factors other than suspect classifications that should not be taken into account even if they are predictive because to do so would be unfair? Should we be judged solely by what we do and never by who we are? Should we allow new evidence from statistical models to change our pre-existing notions of what is relevant to a certain decision?

Some strong moral intuitions regarding fairness are widely shared, and can form a basis for responsible actions even when they are not legally required. But this is an area where the interests of different groups collide and moral intuitions diverge. One advantage of a collaborative approach to a responsible use framework would be to facilitate discussions to air these differences and seek commonality, most productively under the guidance of philosophers and legal scholars who have a wide understanding of different approaches to ethical questions. Coming to some consensus or even isolating key differences in standards of fairness are not technical questions. They are normative and need to be approached as such in developing a framework of responsible use.

Standards of fairness will also help enterprises to determine a course of action following a finding of disproportionate adverse impact, discovered either in the design stage or as an assessment in use. While companies need to make their own decisions, a stakeholder approach might help to develop and share common approaches.

Current non-discrimination law applies only to certain industries and contexts, and even in those contexts, it does not require designing away features of algorithms that pass a legal test for disparate impact. Still some designers and users of data analytics feel a need to do more than reflect the realities of what was, and in many respects, still is, a discriminatory society. They are examining the extent to which they should take steps to free their algorithms as much as possible of what they view as harmful biases.

Standards of fairness can help to address the question of what to do when steps to provide fair treatment for protected classes through adjustments in algorithms might disadvantage other citizens who feel that this treatment is itself a form of discrimination. These are contentious issues concerning the extent to which institutional practice should be race-conscious and the extent to which efforts to avoid disparate impact run afoul of the duty to avoid discriminatory treatment.

Detecting disparate impact and designing alternatives that are less impactful are technical questions. But the decision to modify an algorithm to be fair is not. It involves legal, ethical, business and social matters that go beyond technical expertise in system design. For this reason, people from many disciplines and with a broad array of knowledge, expertise and experience need to be involved in assessing what to do with analytical structures that have or could have a disparate impact. Since the risks, laws and other considerations vary from domain to domain, it is unlikely that there will be one response to the question of what to do with an algorithm that has a disparate impact.
Conclusion

The powerful new tools of data analysis making their way into our social and economic life are designed and used by people, acting in their institutional capacities. They can be designed and used in ways that preserve fairness for all, but this will not happen automatically. If we want these outcomes, we have to design these features into our algorithmic systems and use them in ways that preserve these values.
Endnotes

6. SIIA, Data-Driven Innovation Study, Testimony at FTC, and AI comments.
13. On April 8, 2016, Georgetown Law and the Center on Privacy & Technology held a conference this issue of minority surveillance, entitled The Color of Surveillance: Government Monitoring of the African American Community. See https://www.law.georgetown.edu/academics/centers-institutes/privacy-technology/events/index.cfm
20. Evolv Inc., a tech firm that helps companies hire and manage hourly workers, “…is cautious about exploiting some of the relationships it turns up for fear of violating equal opportunity laws. While it has found employees who live farther from call-center jobs are more likely to quit, it doesn’t use that information in its scoring in the U.S. because it could be linked to race.” Joseph Walker, “Meet the New Boss: Big Data. Companies Trade In Hunch-Based Hiring for Computer Modeling,” Wall Street Journal, September 2012, available at http://www.wsj.com/articles/SB10000872396390443890304578006252019616768
23. This dovetails with the idea that regulation of AI as such would be mistaken. See Study Group, p. 48: “…attempts to regulate “AI” in general would be misguided, since there is no clear definition of AI (it isn’t any one thing), and the risks and considerations are very different in different domains.”
24. For instance, disparate impact analysis in employment is subject to a business necessity test, but disparate impact for age discrimination is subject to a less stringent “reasonableness” standard. See Smith v City of Jackson, 544 U.S. 228 (2005), available at https://www.law.cornell.edu/supct/pdf/03-1160P.ZO.
27. Facebook, for instance, says that their users’ data should not be used to “make decisions about eligibility, including whether to approve or reject an application or how much interest to charge on a loan.” See section 3.15 of their platform policy, available at https://developers.facebook.com/policy/.
30. “I think it’s time we broke for lunch….Court rulings depend partly on when the judge last had a snack,” The Economist, Apr 14th 2011, available at http://www.economist.com/node/18557594
31. Google’s January 2016 report showed that its workforce was 59% white and 69% male. See Google Diversity Index, available at https://www.google.com/diversity/index.html
34. CFPB recently revealed the methodology it uses to assess disparate impact for fair lending compliance. Consumer Financial Protection Board, Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity, Summer 2014 at http://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf. (CFPB Methodology) It does not mandate that anyone use this methodology but companies seeking to assess fair lending compliance risk are now in a position to more reliably make these assessments.
NIST points out that information externalities can produce this kind of harm: “Inferential disclosure may result in group harms to an entire class of individuals, including individuals whose data do not appear in the dataset. For example, if a specific demographic group is well represented in a data set, and if that group has a high rate of a stigmatizing diagnosis in the data set, then all individuals in that demographic may be stigmatized, even though it may not be statistically appropriate to do so.” Simpson Garfinkel, De-Identification of Personal Information, National Institute of Standards and Technology, NIST IR 8053, October 2015, p. 12, available at http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf


50. Village of Arlington Heights v. Metropolitan Housing Development Corp, 429 U.S. 252 (1977), available at https://supreme.justia.com/cases/federal/us/429/252/case.html. The boundary between the effects versus purpose basis for a finding of constitutional discrimination is not entirely clear. The same set of facts that prove discriminatory effect may also demonstrate discriminatory purpose: “Necessarily, an invidious discriminatory purpose may often be inferred from the totality of the relevant facts, including the fact, if it is true, that the law bears more heavily on one race than another. It is also not infrequently true that the discriminatory impact -- in the jury cases, for example, the total or seriously disproportionate exclusion of Negroes from jury venires -- may for all practical purposes demonstrate unconstitutionality because, in various circumstances, the discrimination

59. See Texas v Inclusive Community, at 20


71. The American Civil Liberties Union has filed a court case alleging that the Computer Fraud and Abuse Act is preventing independent researchers from conducting audits of websites. See Esha Bhandari and Rachel Goodman, ACLU Challenges Computer Crimes Law That is Thwarting Research on Discrimination Online, June 29, 2016, available at https://www.aclu.org/blog/free-future/aclu-challenges-computer-crimes-law-thwarting-research-discrimination-online

72. Kroll et al. have a good discussion of the field in Accountable Algorithms; Sandvig et al. discuss different auditing techniques in Auditing Algorithms. A good general framework from a computer science perspective is set out in


75. “A selection rate for any race, sex, or ethnic group which is less than four-fifths (or 80%) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact.... " Uniform Guidelines on Employee Selection Procedures (1978), 29 C.F.R. § 1607.40 (1987), available at http://uniformguidelines.com/uniguideprint.html. However, "(t)his "4/5ths" or "80%" rule of thumb is not intended as a legal definition, but is a practical means of keeping the attention of the enforcement agencies on serious discrepancies in rates of hiring, promotion and other selection decisions." See Adoption of Questions and Answers To Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures, Federal Register / Vol. 44, No. 43 / Friday, March 2, 1979, available at https://www.eeoc.gov/policy/docs/qanda_clarify_procedures.html

76. For example, in their discussion of fair lending risk, Charles River Associates notes: “If a variable only marginally contributes to predictive power or business objectives, but produces a disproportionate adverse impact on a prohibited basis, then it is worth considering whether use of the variable presents an acceptable regulatory compliance risk.” Charles River, p. 5.


78. Commissioner Julie Brill said that the FTC and the companion Federal Reserve study, “...found that the scores they examined largely did not serve as proxies for race or ethnicity,” Remarks of FTC Commissioner Julie Brill at the FTC Workshop, “Big Data: A Tool for Inclusion or Exclusion?” September 15, 2014, available at http://www.ftc.gov/system/files/documents/public_statements/140915bigdataworkshop1.pdf

79. “As of June 2006, forty-eight states have taken some form of legislative or regulatory action addressing the use of consumer credit information in insurance underwriting and rating.” FTC Study, p. 17


81. See Jill Gaulding, Race Sex and Genetic Discrimination in Insurance: What’s Fair, 80 Cornell L. Rev. 1646 (1995), available at: http://scholarship.law.cornell.edu/clr/vol80/iss6/4: “From the efficient discrimination perspective, we have a right not to be classified for insurance purposes unless the classification corresponds to an accurate prediction of risk. From the anti-discrimination perspective, we have a right not to be classified for insurance purposes on the basis of unacceptable classifiers such as race, sex, or genetic factors.”

82. Dwork in Fairness through Awareness joins the accuracy group with her definition of individual fairness as closeness to an accurate assessment of ground truth about individuals and her rejection of statistical parity as an adequate notion of fairness. But her framework allows for a form of “affirmative action” by seeking ways to preserve statistical parity with the minimum sacrifice of accuracy.

83. Cathy O’Neil reflects widespread moral intuitions in saying that it is unjust to base sentencing on factors that could not be admitted in evidence such as the criminal record of a defendant’s friends and family. “These details should not be relevant to a criminal case or a sentencing,” she says because “(w)e are judged by what we do not by who we are.” See Weapons of Math Destruction, at Kindle location 423. Eric Holder holds this view as well.

84. In Fisher v. University of Texas at Austin, 579 U.S. (2016), available at https://www.supremecourt.gov/opinions/15pdf/14-981_4g15.pdf, the Supreme Court allowed the University of Texas to continue using a race-conscious admissions program, ruling that the program was permitted under the equal protection clause.

85. Under the Supreme Court decision in Ricci v. DeStefano, 557 U.S. 557 (2009), available at https://www.supremecourt.gov/opinions/08pdf/07-1428.pdf, an employer might not be permitted to respond to a finding that an employment test has a disparate impact by taking steps that would consciously disadvantage other groups without a “strong basis in evidence to believe that it will be subject to disparate impact liability” if it continues to use that employment test. See Kroll, Accountable Algorithms, for further discussion of the idea that this “strong-basis-evidence” test counsels for building fairness into algorithms in the design stage rather than revising them after discovering a disparate impact in use.

86. Rob Atkinson emphasizes this key point of human agency: “AI systems are not independent from their developers and, more importantly, from the organizations using them. See Rob Atkinson, “Will Smart Machines Be Less Biased than Humans?” Brink, August 15, 2016, available at http://www.brinknews.com/will-smart-machines-will-be-less-biased-than-humans/?mc_cid=feaec2cdf1&mc_eid=aa397779d1