SIIA ISSUE BRIEF

Ethical Principles for Artificial Intelligence and Data Analytics

DEVELOPED BY THE PUBLIC POLICY DIVISION OF THE
SOFTWARE & INFORMATION INDUSTRY ASSOCIATION

Table of Contents

Introduction……………………………………….…………………..………………………2

General Ethical Principles………………………………………………..………………..4

Traditional Ethical Frameworks……………………………………………..4

General Principles for Data Practices…………..…………………………4

Principles of Human Experimentation……………………………..……6

Core Values for a Shared Ethical Framework……………..……………7

United Nations Principles on Business and Human Rights………..8

Ethical Principles for Disparate Impact Assessments……………………….8

General Principles for Disparate Impact Assessment……..…………8

Model and Data Governance Programs …………..………....9

Scope of Application of Disparate Impact Principles……….10

Responsibility to Produce Social Benefits………..……………11

Transparency and Explanations………………….……………12

Specific Principles for Disparate Impact Assessment……………...13

Conduct Disparate Impact Assessments…………..…………13

Collect and Use Adequate Data …………………..………..….14

Take Steps to Mitigate Disparate Impacts……………………..15

Going Beyond Disparate Impact ……………………………………15

Conclusion………………………………………………………………..…….………..…16

Reference Material and Further Reading……………………………….…17

This paper was originally published in draft form in September 2017. This is the final version.

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.

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.

Introduction

The application of big data analytics has already improved lives in innumerable ways. It has improved the way teachers instruct students, doctors diagnose and treat patients, lenders find creditworthy customers, financial service companies control money laundering and terrorist financing, and governments deliver services. It promises even more transformative benefits with self-driving cars and smart cities, and a host of other applications will drive fundamental improvements throughout society and the economy. Government policymakers have worked with developers and users of these advanced analytic techniques to promote and protect these publicly beneficial innovations, and they should continue to do so.

But the use of big data analytics also poses ethical risks involving fairness, accountability, and transparency. Assessments of their eligibility for credit, housing, employment, parole, school admission are enormously consequential for people. People can flourish with the right instructional material or flounder with a poor choice. Choosing the right medical treatment is often a matter of life or death. The bounty of big data analytics is not always widely available and too often its distribution fails to be equitable. When institutions use big data analytics for decision making in critical areas of life, the people affected deserve to be confident that their interests are being served and their rights are protected. Policymakers rightly want assurances that the benefits of data use are broadly and equitably shared and that people are being treated fairly and protected from harm.

Civil rights and civil liberties groups, privacy advocates, consumer activists, scholars and policymakers have called attention to these challenges and urged changes to current law to ensure fairness and to protect vulnerable groups. The AI Now Institute at New York University recently explored these concerns and in a new report concluded, "new ethical frameworks for AI need to move beyond individual responsibility to hold powerful industrial, governmental and military interests accountable as they design and employ AI." Policymakers have drawn attention to these issues as well, conducting workshops, issuing reports, proposing new regulations and legislation.

In many circumstances, current law and regulation provide an adequate framework for strong public protection. Most of the legal concerns that animate public discussions can be resolved through strong and vigorous enforcement of rules that apply to advanced and traditional analytical techniques alike. Using artificial intelligence and machine learning to discriminate against vulnerable groups is a violation of existing law, even if new technological means are used.

Moreover, new policy measures could be counterproductive, creating a substantial risk of overly-broad or overly specific rules, a lingering uncertainty as the ramifications of new rules are clarified, a prospect of significantly increased validation costs, a delay in the introduction of socially beneficial new services, and an organizational culture of passive compliance with external mandates rather than an openness to agile responses to rapidly changing circumstances and new challenges.

But organizations cannot rely on compliance with existing law alone as an adequate response to the ethical challenges of big data analytics. They must go beyond current law to respond fully to the ethical challenges that drive public concerns.

Organizations can meet their ethical obligations and persuade policymakers and the public that they are responsible users of data analytics only if they have policies and procedures in place for ethical review, publicly available ethical principles that they adhere to and a transparent communication program that allows them to describe in an accountable way their policies, procedures and principles.

The public oversight system in place in many jurisdictions throughout the world works through an active and vigorous advocacy community, scholarly research and media investigations to unearth and focus attention on problems. It relies on alert, informed regulators and policymakers who work with existing law to respond to problems and recommend changes when the limits of their competence and jurisdiction are reached. This oversight system functions effectively when organizations anticipate challenges and are prepared to react to these public pressures.

This issue brief focuses on a part of a framework of responsible data use, namely, ethical principles that institutions could use to assess the data and models they use and to make modifications when they are needed. It seeks to further international and national public discussion among policymakers, organizations developing and using data and models, activists, scholars, ethicists, and civil society.

To that end, it proposes general ethical principles to guide an organization's assessment of its data practices. As background and to structure the discussion of these principles, it outlines:

- Traditional Ethical Frameworks
- General Principles for Data Practices
- Principles of Human Experimentation
- Core Values for a Shared Ethical Framework
- United Nations Principles on Business and Human Rights

In addition, this issue brief focuses on the specific issue of data and analytical models that might have a disparate impact, that is, a disproportionate adverse effect, on vulnerable populations. It proposes particular principles and procedures intended to guide organizational assessment and mitigation of disparate impact.

A note on terminology in the rest of the issue brief. The phrase "data analytics system" refers to any method of processing or analyzing data, traditional or advanced, that reveals insights relevant for decision making. The phrase "advanced analytical model" refers to any of the newer techniques of data analysis including artificial intelligence and machine learning that rely of large quantities of data to train algorithms to produce a desired output, such as speech or facial recognition. The phrase "data practices" refers collectively to the process of data collection, analysis and use of information for decision making purposes.

General Ethical Principles

Traditional frameworks

Traditional frameworks of ethical theory approach questions of ethics and political philosophy from different vantage points, emphasizing either fundamental human rights, human welfare, or the cultivation of virtue. Understanding these perspectives can put the ethical challenges of data use into a form that will enable more effective resolution.

- Rights
  - Evaluate practices according their consistency with universal human rights
- Welfare
  - Evaluate practices according to their tendency to promote human welfare
- Virtue
  - Evaluate practices according to their tendency to promote ideal character traits

These traditional frameworks map well onto different privacy frameworks. The Kantian human rights perspective is embodied in various international human rights instruments, including the Universal Declaration of Human Rights, which contains a privacy principle. The European Data Protection Directive and its successor the General Data Protection Regulation are both designed to implement this fundamental right to privacy and data protection. The utilitarian tradition of Bentham and Mill seeks public policies that will maximize human welfare and relies on economic cost-benefit analysis as a touchstone for policy choice. It conceptualizes privacy as a legal right protecting people as consumers in their marketplace transactions with companies, an approach favored at the U.S. Federal Trade Commission. The virtue ethics approach derived from Aristotle focuses on the character traits and ethical norms needed for a good society to form and support people in their everyday efforts to live a life worth living. This social conception of ethics has been developed in modern times by Alastair MacIntyre and Michael Walzer and is reflected in Helen Nissenbaum's idea of privacy as contextual integrity. See the appendix on additional material for references and further reading related to this section on ethical frameworks

General Principles for Data Practices

Even though policy has focused on the privacy issues in the collection and use of data, data practices raise ethical challenges that go well beyond privacy. The contrasting ethical traditions can also help to illuminate these other normative challenges, allowing organizations to think more clearly about issues of rights, justice, welfare, and virtue as they conduct assessments of their own data practices. The following very general principles for data practices emerge from these traditions.

- **Rights -** Engage in data practices that respect internationally recognized principles of human rights.

The framework of human rights requires organizations to respect the equal dignity and autonomy of individuals through data practices that conform to the fundamental rights and freedoms that all people are entitled to regardless of nationality, sex, national or ethnic origin, race, religion, or language. These rights include the right to life, privacy, religion, property, freedom of thought, and due process before the law. These rights do not protect people in their social or economic role, for example, as a patient or client or customer. They do not derive from custom, laws, or traditions but from the status of people as free autonomous rational agents seeking to pursue their own interests and purposes. Organizations should validate these universal aspects of human nature by engaging only in data practices that respect fundamental human rights.

- **Justice -** Aim for an equitable distribution of the benefits of data practices and avoid data practices that disproportionately disadvantage vulnerable groups.

  Individuals have rights based on justice to a fair share of the benefits and burdens of social life. Extreme inequality in the distribution of the benefits of advanced analytical services can be unfair, especially in health care, education, job opportunities. The benefits of advanced analytical services should be equitably distributed and not arbitrarily made available only to restricted social groups or classes. Organizations should arrange their data practices so are as is practicable to ensure that opportunities for exercising talent and skill and earning a living through productive activity are available to all and not restricted based on arbitrary and irrelevant characteristics such as race, ethnicity, gender, or religion. Organizations should not be indifferent to how the models they develop are used and by whom and how the benefits of their new analytical services are distributed. They should aim for justice in the distribution of the services they make possible.

- **Welfare -** Aim to create the greatest possible benefit from the use of data and advanced modeling techniques

  Data and advanced modeling techniques are and should be used to increase human welfare through improvements in the provision of health care, workplace opportunities, insurance, credit granting, authentication, fraud prevention, marketing, personalized learning, recommendation engines, online advertising, to name just a few. Companies have a responsibility to use data and advanced modeling techniques especially in the growing area of AI and machine learning to bring these extraordinary benefits to people around the world.

- **Virtue** - Engage in data practices that encourage the practice of virtues that contribute to human flourishing.

  Data and advanced modeling techniques should be designed and implemented to enable people, individually and collectively, to become virtuous and to further their efforts to become people capable of living genuinely good lives. Data practices should be configured or re-configured to allow affected people to develop and maintain the moral virtues, that is, stable character traits such as honesty, courage, moderation, self-control, humility, empathy, civility, care, and patience that allow people to regularly and reliably engage in excellent, praiseworthy conduct.

Organizations need not choose one of these principles to the exclusion of the others. They should use them jointly as general guides to the development of ethical data practices and standards for the assessment and review of data practices already in use. These general ethical principles need to be supplemented with more detailed normative notions that can be used in the context of organizational evaluation of data analytics. The following sections examine some examples of these more detailed conceptions - the principles of human experimentation, core values for a shared ethical framework, and the United Nations principles on business and human rights.

Principles of Human Experimentation

The following notions are drawn from the law and principles governing experimentation on human subjects developed in the Belmont Report, the Menlo Report, and the Common Rule that guides human subject experimentation in U.S. law. See the appendix on additional material for references and further reading.

- **Respect for persons**
    - Give weight to the considered judgments of people affected by data practices.
    - Do not withhold information about data practices, when there are no compelling reasons to do so.
- **Beneficence**
    - Secure the well-being of people affected by data practices
    - Do not harm them
    - Maximize possible benefits and minimize possible harms
- **Justice**
    - Distribute the benefits and burdens of data practices equitably
    - Do not arbitrarily target groups based on attributes such as race, gender, religion, or ethnicity
- **Respect for Law**
    - Engage in due diligence to identify laws and regulations applicable to data practices.
    - Design and implement data practices that respects these restrictions.
- **Transparency**
    - Disclose the purposes of data practices, why data collection is required to fulfill those purposes, and how data will be used.
    - Clearly communication risk assessment and harm minimization related to data practices.
- **Accountability**
    - Document ethical evaluations and make them available in accordance with balancing risks and benefits.

The U.S. statute that governs human subject experimentation requires institutions to set up an Independent Review Board (IRB) and requires the board to determine that all of the following requirements are satisfied before authorizing the experiment:

- Risks to subjects are minimized;
- Risks to subjects are reasonable in relation to anticipated benefits, if any, to subjects, and the importance of the knowledge that may reasonably be expected to result;
- Selection of subjects is equitable;
- Informed consent will be sought from each prospective subject and documented (with some exceptions);
- Provisions are made to monitor the data collected to ensure the safety of subjects; and
- When appropriate, there are adequate provisions to protect the privacy of subjects and to maintain the confidentiality of data.

An IRB may grant exceptions from the requirement for informed consent when:

- The research involves no more than minimal risk to the subjects;
- The waiver or alteration will not adversely affect the rights and welfare of the subjects;
- The research could not practicably be carried out without the waiver or alteration; and
- Whenever appropriate, the subjects will be provided with additional pertinent information after participation.

These requirements can serve as a checklist for ethical data practices more generally and provide guidance for organizations conducting ethical assessments.

Core Values for a Shared Ethical Framework

Marty Abrams and his colleagues at the Information Accountability Project have assembled the following list of values that can be used to inform ethical assessments in a data governance program.

- **Beneficial**
    - Define the benefits that will be created by a data analytics project and identify the parties that gain tangible value from the effort.
- **Progressive**
    - Favor data analytic projects where the value created is materially better than not engaging in that project.
- **Sustainable**
    - Favor data analytic projects that effectively predict future behavior and generate beneficial insights over a reasonable period of time.
- **Respectful**
    - Provide notice in the context in which the data originated and disclose any contractual restrictions on how the data might be applied
- **Fair**
    - Avoid discrimination based on characteristics such as gender, race, genetics or age
    - Avoid data practices that cause substantial injury that cannot be avoided and have no countervailing benefits.
    - Favor data projects that share the benefits of technology and broader opportunities related to employment, health and safety.

The source of these values is the Universal Declaration of Human Rights and they are meant to define an ethical framework for advanced analytics and big data that can be used both by organizations for their own internal assessments and by external oversight agencies, including regulators.

United Nations Guiding Principles on Business and Human Rights

The United Nations developed foundational and operational principles that can be used by business to respect human rights. The following ethical principles for data practices are based on their foundational principles.

- Human Rights
    - Engage in only those data practices that respect internationally recognized human rights in the International Bill of Human Rights
- Adverse Human Rights Impacts
    - Avoid data practices that cause or contribute to adverse human rights impacts
    - Seek to prevent or mitigate adverse human rights impacts that are directly linked to data practices through business relationships.
- Policy Commitment
    - Establish and maintain in place a policy commitment to engage only in data practices that respect human rights.
- Due Diligence
    - Have in place processes to identify, prevent and mitigate any impacts on human rights resulting from data practices.
- Remediation
    - Establish processes to provide for effective remediation of any adverse human rights impacts caused by data practices
- Communication
    - Communicate to outside parties how risks of human rights impacts are addressed

These are a selection from a more detailed set of guiding principles that might be useful for organizations to assess as they are setting up a data governance program that provides for ethical assessments.

Ethical Principles for Disparate Impact Assessments

The different strands of ethical reflection surveyed in the previous section suggest ethical principles relating to an organization’s responsibilities to avoid disparate impact on vulnerable populations. Organizations have an ethical duty to avoid such disparate impacts in their development, implementation, and use of advanced data analytics. The normative principles that follow are meant to guide organizations in fulfilling this obligation. While they derive motivating force from the above ethical principles, they reflect non-discrimination law and policy in the United States and the widespread international norm that high-stakes decisions about people should not disadvantage vulnerable populations based on characteristics such as their race, gender, ethnicity, or religion.

All ethical traditions condemn practices that impose disproportionate adverse effects on vulnerable groups. Such practices violate norms of fairness and justice and principles of equal treatment; they perpetuate unjust distributions of the burdens and benefits of collective life; they subordinate the welfare and rights of vulnerable populations to the interests of dominant groups. They rely on arbitrary and irrational bases for decision making and so impose needless costs and burdens on vulnerable groups and reduce the overall welfare of society. They conflict with widespread, entrenched ethical norms and interfere with the ability of targeted groups to develop the character traits needed to live a good life as equals in their own communities.

For these reasons, organizations have an ethical responsibility to take steps to determine whether their analytics systems have discriminatory effects and take steps to mitigate these effects.

Discriminatory harms could be intentional, but it is assumed that organizations have in place policies and procedures to prevent intentional infliction of harm on vulnerable populations. The more complex issues arise with unintentional disparate impacts. Disparate impacts are adverse consequences on vulnerable groups that are not intended; they arise indirectly and inadvertently through attempts to achieve legitimate organizational objectives. This section deals specifically with principles for organizations seeking to address these unintentional, inadvertent discriminatory harms.

General Principles for Disparate Impact Assessment

We start with several background principles: model governance, application, benefits and transparency. Organizations need to have a general model and data governance program into which they can insert disparate impact assessments (model governance); they need to make a threshold determination when to apply disparate impact principles (application); they need to keep in mind an overriding obligation to provide data analytics services that allows for the greatest possible benefit to society overall (benefits); and they need to disclose operations to ensure public trust (transparency). These are general principles that should govern all aspects of data practices and are relevant to other ethical obligations beyond disparate impact, such as those relating to privacy. But they are crucial for an adequate organizational response to the challenges of disparate impacts.

Model and Data Governance Programs

All organizations that collect and process data using traditional or advanced analytical systems should have policies and procedures in place to guide the development and use of their systems. The following two principles express these responsibilities.

- Establish and maintain effective and comprehensive data and model governance programs, including policies and procedures to assess the ethical implications of data analysis.

- Build governance, controls, and ethical assessment into the process of developing, revising, and updating models, rather than conducting ad-hoc after-the-fact reviews.

Advanced analytical systems will produce organizational and societal benefits only if the model development, validation, implementation, and use is subject to rigorous controls to ensure accuracy and fairness. Organizations should adopt and maintain appropriate governance and control mechanisms including proper management oversight and appropriate incentive and organizational structure. Ethical principles and values need to be considered as intrinsic elements of an adequate data model governance program. The details of such a model governance programs will vary by industry, use and context, but policies and principles developed for the model management programs in the financial services might provide useful guidelines for companies using data and modeling techniques in other areas of economic life These programs are more general than, but can include elements from, privacy risk assessments, consumer subject review boards and research-focused independent review boards.

Separate principles should govern both data governance and model governance programs. Data governance programs should include:

- An assessment and documentation of all data assets
- How is the data collected, and from whom? Directly from the consumer or from third party sources about the consumer?
- Is the data collected in an authorized manner? If from third parties, can they attest to the authorized collection of the data?
- Has consumer been informed of data collection, and where appropriate, provided permission about its use?
- How transparent is the notice provided to consumers about collection and use?
- Is the data held securely from unauthorized access, consistent with the sensitivity of the data?

In addition, a model governance program should include an assessment of all analytical techniques, traditional or advanced, that are used for consequential decision making about people. It includes:

- An inventory and documentation of all models used for automated decisions
- What data is used in the model
- Is the data accurate for the purposes of the model?
- Where did the data come from? If from a third party, what collection and permission mechanisms were in place?
- Define what the model is intended to predict
- Test the model to ensure it is predictive of the purpose intended
- Retest and validate the model on a regular basis to ensure it is working pursuant to its purpose.
- Exclude data that is not relevant to predicting the outcome

Scope of Application for Disparate Impact Principles

Organizations need to determine when to expend resources to assess the impact of their data practices on vulnerable groups prior to engaging in these assessments. When they regularly develop, implement, or use data analytic systems that might have a discriminatory effect on vulnerable groups, they should have policies and procedures in place to determine when to conduct further assessments. If a consideration of relevant factors suggests that a data analytic system could have a significant potential for harmful impacts on these groups, this triggers an obligation to assess and address potential discriminatory harms. The following principles of scope of application are meant to aid organizations in making this determination. They are intended to help organizations determine whether they should have procedures and standards in place to evaluate when in a particular case it should conduct a full disparate impact assessment.

- In determining the need for applying these disparate impact principles, companies should focus on the groups being protected, the context of use, the nature of the application of the data analytic system and the potential for substantial harm.
- When the people affected by data practices 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, organizations should evaluate whether they should conduct disparate impact assessments and address any significant disproportionate adverse impact on the members of that class in accordance with these principles.

Organizations should put themselves in a position to conduct disparate impact assessments in a particular case when they regularly develop, implement, or use data analytic systems that might have a discriminatory effect on vulnerable groups. Organizations that are in the business of designing, implementing or using data analytics systems should be thoughtful about the potential for discriminatory effects in any field, regardless of whether there is a legal obligation pertaining to vulnerable groups. They should be especially careful about this potential for discriminatory harm when their use of data and modeling techniques has consequential impacts on the lives of people deserving special protection because of their vulnerable status or because of historical or continuing practices aimed at subordination. These vulnerable populations include but are not limited to groups protected by law in many jurisdictions around the world such as those defined by age, gender, race, and ethnicity.

Responsibility to Produce Social Benefits

Organizations devote time, energy, human resources and capital to develop data analytic systems that provide insight into important areas of human decision making. They have a responsibility to use the resources at their disposal to provide products and services that fulfill important social purposes and do so in a more effective and efficient fashion than comparable use of resources. Part of this duty is the obligation to avoid unnecessary expenditures on activities that do not benefit the public. The following principles capture these responsibilities

- Organizations should aim to maximize the social benefits of their use of data and advanced modeling techniques
- Organizational assessments of possible discriminatory harms should be proportional and necessary considering the costs involved.

Organizations currently use data and advanced modeling techniques to improve outcomes and lives in health care, employment, insurance, credit granting, authentication, fraud prevention, marketing, personalized learning, recommendation engines, and online advertising, to name just a few. Organizations have a responsibility to continue to innovate in the use of data and advanced modeling techniques especially in the growing area of AI and machine learning to bring these extraordinary, transformative benefits to people around the world. Often the benefits of these new techniques are improvements in advancing public policy objectives or delivering urgent government services or meeting the needs of unserved or underserved populations. In these cases, the need to demonstrate social benefit from the use of data analytics systems is easily satisfied. But this requirement to demonstrate social benefits can also be satisfied by a showing of widespread consumer demand for products and services made possible or improved by big data analytics. To protect the societal interest in bringing these social benefits to the public, companies should be prudent in their expenditure of resources for assessing potential discriminatory harm. The steps they take to assess and ensure fairness need to be proportional and necessary considering the costs involved.

Transparency and Explanations

A key aspect of ethical use of data is an organization's willingness to be accountable to outside oversight about the processes and outcomes of data analytic systems. Accountability cannot be effective without transparency to the outside world and a commitment to conveying clearly and comprehensively how an organization's processes and standards address the ethical issues raised by data use, including how an organization assesses and remedies disparate impacts. Several U.S. and European regulations, described in the appendix on additional material, call for disclosures of explanations. The following principles regulate how an organization should approach these transparency questions.

- Organizations should disclose what data they collect, the purposes for which it is used, and which analytic techniques and models are used to process data and produce an outcome.
- Organizations should provide explanations of how advanced modeling techniques produce their results, including disclosing, where available and appropriate, the key factors that contribute to the outcome of an analytic process.
- Organizations should publicly describe the model governance programs they have in place to detect and remedy any possible discriminatory effects of the data and models they use, including the standards they use to determine whether and how to modify algorithms to be fairer.

Trust in the fairness of a data analytic system relies on public awareness of data and the analytical systems used as well as the basis for organizational steps to detect and mitigate disparate impacts. Transparency about the process and standards used is especially important for disparate impact assessments, where ethical intuitions differ and social consensus on the right course of action might not be possible. The need to consult with public officials and the affected communities is especially strong in the cases, discussed below, of using sensitive variable in data analytic systems and determining how to navigate the tradeoff between accuracy and fairness when a data analytics system might not be able to fully satisfy both values.

Organizations do not need to disclose source code of proprietary algorithms for several reasons. Disclosure is not useful for accountability purposes, especially in the case of advanced analytical techniques that improve themselves in use. Source code disclosure would likely produce counterproductive efforts to game analytical systems in ways that defeat their purpose. Disclosure would allow anyone to use or benefit from systems that require extensive development resources, thereby weakening the economic incentive in creating these systems. For these reasons, disclosure has not been required for heavily regulated traditional scoring systems such as credit scores that have been in use for decades.

If organizations do not reveal their source code, they must take other steps to provide for transparency and accountability. Organizations should be prepared to communicate to outside parties the key factors that go into their scores, and to provide evidence on a regular basis of the continuing validity and reliability of the predictive models they use. Public trust in the fairness of algorithms requires sufficient disclosure so that people feel able to comprehend and assess the process used to produce insights that might have important effects on their lives.

Specific Principles for Disparate Impact Assessment

In addition to these general principles that provide a framework for conducting disparate impact assessments, organizations should be guided by more specific principles. They specify what steps to take in conducting a disparate impact assessment (disparate impact assessment); what data needs to be available to conduct them (data adequacy); what is a disparate impact test and what steps to take to ensure that their modeling techniques pass such a test; (mitigation steps); and what additional steps might need to be taken when model passes an assessment (beyond disparate impact).

Conduct Disparate Impact Assessment

Since disparate impact occurs inadvertently, the only way an organization will discover on its own that its data practices have a disparate impact is to look for it. As noted above in the scope principle, organizations should put in place procedures and standards to determine when to conduct a full disparate impact assessment when they regularly develop, implement or use data analytic systems that might have a discriminatory effect on vulnerable groups. The following principles specify when a data analytic system should be subjected to a full disparate impact and what the elements of a disparate impact assessment are.

- Organizations should evaluate a data analytic system for disparate impact when the design, implementation or use of that data analytic system has a significant potential for substantial and consequential discriminatory effects on vulnerable groups.
- A disparate impact assessment determines whether a data analytic system has a substantial disproportionate adverse impact on a vulnerable group, examines whether the use of the system advances legitimate organizational objectives and compares it to alternative systems that might have a lesser disparate impact.

Organizations regularly operating in areas that have consequential impacts on people’s lives should evaluate data analytic system techniques for disparate impact when the design, implementation or use of data analytic systems has a significant potential for discriminatory effects.

A disparate impact assessment has three steps. The first is to determine whether the data analytics system under review has a disproportionate adverse impact on a vulnerable group. This can be measured by standard statistical characteristics of the data analytic system such as departures from statistical parity or equal group error rates. Organizations should devise or adopt – in collaboration with academics, advocates, and independent technical experts – accurate and reliable guidelines and methodologies for detecting disparate impacts.

The second step is examination of how the data system in question serves organizational objectives. Notwithstanding any disproportionate adverse effect on vulnerable groups, a data analytic system can pass a disparate impact assessment if it furthers a legitimate organizational interest. Avoiding disparate impact cannot be a requirement to abandon the values and goals that constitute an organizations mission. But furthering a legitimate objective is not sufficient to pass a disparate impact assessment, because there might be an alternative system that also furthers organizational objectives, but does so with a smaller impact on the vulnerable group.

So, the third step in a disparate impact assessment is a comparison of the data system to alternatives. This step should involve an active search for alternatives to or modifications of the system being reviewed. It should not be restricted to an assessment of obvious or readily available alternatives. Organizations should develop and assess alternatives to algorithms with a disparate impact to ascertain the extent to which they achieve organizational objectives.

A data analytic system passes a disparate impact test, despite having a disproportionate adverse impact on a vulnerable group, when after an appropriate search for alternatives, an organization finds there is no alternative algorithm that furthers institutional objectives with a lesser impact.

Disparate impact assessments should be conducted at the same frequency as other reviews needed to ensure the validity and reliability of models. Especially in the case of advanced analytic systems that improve in use, impact assessments need to be conducted frequently.

Collect and Use Adequate Data

Organizations should have in place procedures to conduct disparate impact assessments when they are likely to need to examine data analytic systems for disparate impact in the regular course of their business. Once they have made that determination, they also need to have available the data they need to conduct disparate impact assessments. After-the-fact aggregation of data needed to conduct assessments should not be the regular practice; rather organizations need to plan to have the needed data available. Moreover, the use of sensitive data, while subject to substantial legal restrictions, might need to be included in data analytic system themselves to provide for both fairness and accuracy.

- Adopt and maintain policies and procedures reasonably designed to collect information sufficient to conduct assessments that would detect any significant disparate impacts, including, if necessary, collecting sensitive information such as race, gender, ethnicity, and religion or constructing accurate proxies for such sensitive information.
- Where permitted by law, organizations should consider whether and the extent to which they should use sensitive data in data analytic system.

Organizations do not always collect sensitive information such as race, gender, ethnicity, and religion.

Sometimes collecting sensitive information is prohibited by national law to protect vulnerable populations from explicit decisions to discriminate against them. In other cases, organizations are permitted to collect this information only with explicit affirmative consent of the data subject and only for conducting disparate impact assessments to ensure that their decisions are free from unintentional discrimination. Organizations will often need sensitive information to identify potential discriminatory outcomes of data analytic systems. However, the absence of such information is not always fatal. It is often possible to construct accurate proxies that could be used in assessments instead of the sensitive variables themselves. Organizations should obtain sensitive information or construct proxies that will allow them to conduct disparate impact assessments.

In addition, organizations need to assess carefully and consciously the need to use sensitive variables in data analytic systems. In certain cases, relevant attributes and characteristics are distributed differently in vulnerable groups than in the general population. Including variables for vulnerable groups in data analytic systems can improve both fairness and accuracy and can increase social welfare through more effective achievement of institutional objectives. These issues touch on controversial and unsettled legal and normative issues. Ideally, the use of sensitive variables in data analytic systems should be undertaken only in consultation with public officials and the affected communities.

Take Steps to Mitigate Disparate Impacts

A disparate impact assessment does not automatically lead to action. A further question is when the results of an assessment should lead an organization to modify an algorithm to make it fairer or adopt an alternative that has less of an impact. The following principle prescribes action when a data analytic system fails a disparate impact test.

- Organizations should not develop, implement, use or continue to use a data analytic system if an alternative is available that achieves the same legitimate organizational objective with a lesser disparate impact on vulnerable populations.

One purpose of assessing advanced analytic models for disparate impact is to compare them against alternatives that might be equally effective but have a less harmful impact on vulnerable groups. After conducting this assessment of alternatives, organizations have an obligation to avoid using or continuing to use a model with a disparate impact if an equally effective but less impactful alternative is available. Using or continuing to use the more impactful model will impose unnecessary harms on vulnerable groups without any compensating benefits to the organization or to society. In U.S. law regulating discrimination in housing, education, employment, and the granting of credit, a failure to follow this principle can expose an organization to legal liability. However, organizations should follow this principle even when it is not required by law in cases where the harmful effects on vulnerable groups are substantial and consequential.

Going Beyond Disparate Impact

What should an organization do when a full disparate impact assessment, including a thorough search for alternatives, reveals that the use of a data analytic system to further organizational objectives has a substantial and consequential adverse effect on a vulnerable population but there is no equally effective alternative? The answer is not obvious or easy. Organizations cannot simply avoid the question or jump to a quick unreflective response. Instead, they should consider the alternatives carefully and develop standards and principles that will enable them to respond to this possibility.

- When there is no effective but less impactful data analytic system available, organizations have a further ethical obligation to consider whether and how-to re-design their system to have less of a disparate impact on vulnerable groups, even if this sacrifices some organizational effectiveness, taking into account the benefits and costs of taking this step.

Data analytic systems that have a disparate impact on protected classes might be the most accurate available systems. Any system that produces a smaller disparate impact will also be less accurate and so less capable of achieving organizational objectives. This can happen when relevant characteristics are distributed differently in a vulnerable group. Even though these uses pass a disparate impact test, they raise questions of fairness because they can perpetuate historical discrimination. In such circumstances, organizations should consider whether to sacrifice some accuracy for more fairness. This might be the only way to more towards greater equality in organizational outcomes. Organizations need to consider what steps to take in this circumstance and have explicit policies and procedures designed to address it.

Organizations should approach with great care the decision to go beyond the requirement that a data analytic system pass a disparate impact assessment. Such steps toward greater fairness can have adverse effects of their own. The resulting algorithms or decision rules might be less predictive and make it harder to attain legitimate objectives such as public safety or financial risk management. They might also run counter to the principle of treating similarly situated people the same. Organizations should assess the extent of these costs in considering whether to adjust algorithms for fairness, and be prepared to communicate the results of these considerations and the standards and values embodied in them.

It is likely that the most appropriate answer to this ethical dilemma varies by context. In a setting such as criminal sentencing or parole, the demands of racial fairness might be predominant; while in other contexts such as granting credit or insurance accuracy might prevail. As with the use of sensitive variables in data analytic systems, an organization should ideally decide what to do in these cases in open and transparent consultations with public officials and the affected communities.

Conclusion

Organizations are not always comfortable taking into consideration or talking about the ethical aspects of their institutional activity. The crisper language of legal requirements and achieving organizational objectives seems more suited to a public conversation. But ethical aspects of organizational decision making are unavoidable; ignoring them or translating them in more comfortable formulas will not make them go away. Justice, respect for rights, human welfare, and the moral virtues that enable people and communities to live good lives are real aspects of our lives. And they are affected by organizational activities and decisions.

Organizations should rely increasingly on data analytic systems to achieve their institutional objectives. The advantages, the transformative benefits of the new analytical systems especially, are too great to be ignored. Indeed, as we have seen, one aspect of the ethical responsibilities of organizations is to embrace the new technological advances in data science to bring innovative new products and services to the betterment of local and global communities.

As organizations step up to that responsibility to make the world a better place through better analytics, they need to keep in mind the demands of justice, fairness, and virtue. This issue brief is a contribution to the needed discussion of these additional ethical demands on organizations. Its aim is not to arrive at definitive solutions to the thorny ethical dilemmas that organizations will confront, but to stimulate the needed collaborative work that will help to develop and improve practical ways to address them in the real world.

This issue brief has focused on the ethical issues raised by the fact that data analytics systems can and often do have unintended disparate impact on vulnerable groups. Organizations will often need to develop policies and procedures to manage the adverse effects on these vulnerable groups and to mitigate them where possible. It directs attention to the ethical issues involved – to the value of equal treatment of all regardless of race, religion, gender, and other sensitive characteristics and to the demands of justice to end the subordination of disadvantaged groups. It urges organizations to confront these ethical dilemmas in an open and transparent way with full consultation with public officials and the affected communities.

Reference Material and Further Reading

This material provides background for the discussion in the text and can be used as a guide to further reading in the topics covered in each section of the issue brief.

Introduction

Alex Campolo, Madelyn Sanfilippo, Meredith Whittaker, Kate Crawford, "AI Now 2017 Report," New York University and AI Now, 2017, available at https://assets.contentful.com/8wprhhvnpfc0/1A9c3ZTCZa2KEYM64Wsc2a/8636557c5fb14f2b74b2be64c3ce0c78/_AI_Now_Institute_2017_Report_.pdf. A joint report by New York University and AI Now that provides policy recommendations and research on artificial intelligence and ethics.

Cathy O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Crown/Archetype, 2016. A thorough review of the ethical challenges that face developers and users of data analytic systems in employment, health care, transportation, education and credit granting.

Frank Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information, Harvard University Press, 2015. A comprehensive study of the ways in which secrecy regarding data analytic systems can perpetuate injustice and a call for transparency as one of the remedies.

Erik Brynjolfsson and Andrew McAfee, "The Business of Artificial Intelligence," Harvard Business Review, July 18, 2017. Available at https://hbr.org/cover-story/2017/07/the-business-of-artificial-intelligence. Two MIT business professors explain what machine learning means for business, asserting that "the most important general-purpose technology of our era is artificial intelligence, particularly machine learning (ML) — that is, the machine's ability to keep improving its performance without humans having to explain exactly how to accomplish all the tasks it's given."

The Federal Trade Commission. “Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues.” January 2016. Available at https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf. The FTC held a hearing on the ways in which big data analytics could help underserved populations and how it could exacerbate existing inequalities.

Executive Office of the President, Preparing for the Future of Artificial Intelligence, October 2016, available at https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf. This White House report outlines the history of AI and its application to social good, and to industry and especially the automobile industry.

Committee on Legal Affairs, European Parliament, Report with Recommendations to the Commission on Civil Law Rules on Robotics, January 27, 2017, available at http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+REPORT+A8-2017-0005+0+DOC+XML+V0//EN This sweeping report recommends codes of conduct for engineers working on autonomous AI-systems and other oversight mechanisms to ensure that they are used for social good.

General Ethical Principles

Traditional Frameworks

John Mizzoni, Ethics: The Basics, 2nd Edition, Wiley, 2016. The following chapters provide the basic three ethical frameworks: Introduction, Chapter 2: Virtue Ethics, Chapter 5: Utilitarian Ethics, and Chapter 6: Deontological Ethics.

Michael Sandel, Justice, Farrar, Strauss and Giroux. This is an excellent summary of the different classical traditions with applications to contemporary ethical issues.

Immanuel Kant, Groundwork for the Metaphysic of Morals, available at http://hj.tpnicdn.net/lecture-11-mind-your-motive/#1477505675034-57204795-9b61 The classic exposition of the view that morality is a matter of doing the right thing, regardless of utility.

Jeremy Bentham, Principles of Morals and Legislation, available at http://hj.tpnicdn.net/lecture-2-the-case-for-cannibalism/#1477504398770-429c5ad9-ee6c. What could morality be about except increasing human welfare?

John Stuart Mill, Utilitarianism, available at http://hj.tpnicdn.net/lecture-4-how-to-measure-pleasure/#1477504398584-5d76ef29-aab0. The canonical arguments for basing ethical and social decisions on utility.

Aristotle, Nicomachean Ethics, available at http://classics.mit.edu/Aristotle/nicomachaen.html The classic Western version of virtue ethics; Politics available at http://classics.mit.edu/Aristotle/politics.1.one.html. Book I contains the definitive conception of people as intrinsically social and political beings.

Alasdair MacIntyre, After Virtue, 3rd ed. University of Notre Dame Press, 2007. The best contemporary version of virtue ethics.

Michael Walzer, Spheres Of Justice: A Defense Of Pluralism And Equality, Basic Books, 1983. This bases our notions of justice and equality in the norms and meanings of our social world and was an inspiration for Helen Nissenbaum's work on privacy as contextual integrity.

Shannon Vallor, Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting, Oxford University Press, 2016. An extraordinarily clear exposition of virtue ethics, including both Chinese and Buddhist versions in addition to the classic Western version from Aristotle.

Principles for Human Experimentation

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, Belmont Report: Ethical Principles And Guidelines For The Protection Of Human Subjects Of Research (1979) *available at* http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html or https://videocast.nih.gov/pdf/ohrp_belmont_report.pdf This is the seminal Belmont Report that proposed the principles of respect for persons, beneficence and justice to govern experiments on human subjects and was the basis for the IRB reviews mandated for research receiving federal funding.

David Dittrich and Erin Kenneally, U.S. Department Of Homeland Security, The Menlo Report: Ethical Principles Guiding Information And Communication Technology Research, August 2012, *available at* https://www.caida.org/publications/papers/2012/menlo_report_actual_formatted/menlo_report_actual_formatted.pdf This report updates the Belmont principles and applies them to information and communications technologies, adding the principle of respect for law and public interest to the Belmont Report's original three.

Department of Health and Human Services, Federal Policy for the Protection of Human Subjects ('Common Rule'), available at https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html Regulations for human subject experimentation based on the Belmont Report, including the establishment of Independent Review Boards. The general regulations are at 45 CFR 46, https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html The finding IRBs must make to approve experiments are at 45 CFR §46.111. The exceptions from informed consent are at 45 CFR §46.116(d).

Polonetsky J, Tene O and Jerome J, Beyond the Common Rule: Ethical Structures for Data Research in Non-Academic Settings. *Colorado Technology Law Journal* 13 (2015) at https://fpf.org/wp-content/uploads/Polonetsky-Tene-final.pdf This article adapts the IRB standards (the Common Rule) for online research.

Core Values for a Shared Ethical Framework

Martin Abrams, et al, "Unified Ethical Frame for Big Data Analysis," Information Accountability Foundation, March 2015, available at http://informationaccountability.org/wp-content/uploads/IAF-Unified-Ethical-Frame.pdf This framework for ethical review of algorithms and their use builds on privacy assessments and would look to whether the use is beneficial, progressive, sustainable, respectful and fair.

United Nations Guiding Principles on Business and Human Rights

United Nations, Human Rights Council, Report of the Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises, John Ruggie, Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework, 2011, available at http://www.ohchr.org/Documents/Issues/Business/A-HRC-17-31_AEV.pdf This report (Ruggie Report) establishes guiding principles for companies to use to assess and manage the human rights risks created by their activity or to which they contribute through their business relationships.

General Ethical Principles for Disparate Impact Assessment

Data and Model Governance Programs

Office of the Comptroller of the Currency, Fair Lending: Comptroller's Handbook for Consumer Compliance Examination, January 2010, https://www.occ.gov/publications/publications-by-type/comptrollers-handbook/fairlend.pdf This latest guideline outlines what regulators look for from national banks to demonstrate compliance with fair lending laws and regulations.

Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency, Supervisory Guidance on Model Risk Management, April 4, 2011, https://www.occ.treas.gov/news-issuances/bulletins/2011/bulletin-2011-12a.pdf This regulatory guidance provides a comprehensive roadmap for developing and maintaining model governance programs.

Michael Versace and Karen Massey, A Framework for Model Governance: Opinions and Insights from Bankers Across North America, IDC Financial, June 2013, http://www.fico.com/en/latest-thinking/white-papers/white-paper-framework-for-model-governance . This industry white paper sponsored by FICO provides insights into model governance programs that would be useful for companies throughout the economy.

Ryan Calo, Consumer Subject Review Boards: A Thought Experiment, 66 Stanford Law Review Online 97 (2013), at http://pacscenter.stanford.edu/wp-content/uploads/2016/05/Calo-Consumer-Subject-Review-Boards.pdf This article calls for companies to set up ethical review boards to review initiatives for consistency with a broad array of ethical standards.

Molly Jackman & Lauri Kanerva, Evolving the IRB: Building Robust Review for Industry Research, 72 Wash. & Lee L. Rev. Online 442 (2016), http://scholarlycommons.law.wlu.edu/wlulr-online/vol72/iss3/8 This law review article describes Facebook's procedure for online experimentation.

Scope of Application of Disparate Impact Principles

Marty Abrams, Accountability in an Observational, Analytics-Driven Digital Economy, Information Accountability Foundation, September 28, 2016, available at http://informationaccountability.org/accountability-in-an-observational-analytics-driven-digital-economy/ This provides some discussion of when companies need to have an accountability system in place.

Responsibility to Produce Social Benefits

Viktor Mayer-Schonberger and Kenneth Cukier, Big Data: A Revolution That Will Transform How We

Live, Work, and Think, Houghton Mifflin Harcourt, 2013. An excellent review of the developments in big data analytics and their transformative benefits.

The White House, “Big Data: Seizing Opportunities, Preserving Values,” May 2014, available at https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf This report welcomes the social benefits of big data analytics and urges companies to innovate wisely to achieve them.

Artificial Intelligence and Life In 2030, One Hundred Year Study On Artificial Intelligence, Report Of The 2015 Study Panel, Stanford University, September 2016, p. 7, available at https://ai100.stanford.edu/sites/default/files/ai_100_report_0831fnl.pdf. This report from industry and academic experts focuses on how artificial intelligence and machine learning will transform health care, transportation and other fields.

Transparency and Explanations

Danielle Keats Citron and Frank Pasquale. “The Scored Society: Due Process for Automated Predictions.” University of Maryland Francis King Carey School of Law, Legal Studies Research Paper, No. 2014 -8. (2014) 89 Wash. L. Rev 1] http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2376209. This seminal paper shows that concerns about scoring go well beyond privacy issues and implicate a range of normative concerns including fairness, discrimination, accountability and due process. Transparency is a key to maintaining trust. The authors call for source code disclosure.

Citron, Danielle Keats, Technological Due Process. U of Maryland Legal Studies Research Paper No. 2007-26; Washington University Law Review, Vol. 85, pp. 1249-1313, 2007. Available at SSRN: http://ssrn.com/abstract=1012360. This article argues persuasively that due process requires substantial openness.

Equal Credit Opportunity Act (ECOA). ECOA requires that “each applicant against whom adverse action is taken shall be entitled to a statement of reasons for such action from the creditor…A statement of reasons meets the requirements of this section only if it contains the specific reasons for the adverse action taken.” See 15 U.S.C. § 1691(d)(2)-(3), available at https://www.law.cornell.edu/uscode/text/15/1691

Fair Credit Reporting Act (FCRA). FCRA requires consumer reporting agencies to disclose “all of the key factors that adversely affected the credit score of the consumer in the model used, the total number of which shall not exceed 4. See 15 U.S. Code § 1681g(f)(1)(c), available at https://www.law.cornell.edu/uscode/text/15/1681g. In cases of an adverse action based on information in a credit report, FCRA requires disclosure of these key factors. See 15 U.S. Code § 1681m(a), available at https://www.law.cornell.edu/uscode/text/15/1681m
Regulation B. Both ECOA and FCRA require disclosures of explanations of specific outcomes of automated decisions. Regulation B (12 C.F.R § 1002), available at https://www.consumerfinance.gov/eregulations/1002) implements ECOA and provides guidance for FCRA disclosures as well. An appendix to Regulation B contains a Sample Notice of Action Taken and Statement of Reasons that contains a list of 24 factors that can be used to satisfy these disclosure obligations. See 12 C.F.R. § 1002, Appendix C —Sample Notification Forms, available at https://www.consumerfinance.gov/eregulations/1002-C/2011-31714#1002-C-5-h1-p5

General Data Protection Regulation (GDPR). The European Union's GDPR will go into effect in May 2018. See Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC, 2016 O.J. L 119/1, available at http://data.consilium.europa.eu/doc/document/ST-5419-2016-INIT/en/pdf. Article 15 of GDPR grants data subjects access to "meaningful information about the logic involved" in "automated decision-making." See GDPR, Article 15(1)(h). This goes beyond the text in Article 12 of the earlier Data Protection Directive which called only for a right of access to "knowledge of the logic involved" in automated decisions, suggesting that GDPR requires a higher level of explanation. This new level of explanation seems to require some detail on the general way in which decision-making algorithms work, rather than an explanation of an outcome in a specific case. The UK's Independent Commissioner's Office has issued guidance in this area that requiring data processors to "ensure processing is fair and transparent by providing meaningful information about the logic involved, as well as the significance and the envisaged consequences." See https://ico.org.uk/for-organisations/data-protection-reform/overview-of-the-gdpr/individuals-rights/rights-related-to-automated-decision-making-and-profiling/

Specific Principles for Disparate Impact Assessment

Overview Material

Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson & Harlan Yu *Accountable Algorithms*, 165 U. Pa. L. Rev. 633 (2017). Available at: http://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3. This comprehensive law review article discusses why new algorithms raise ethical issues and how to create accountability mechanisms to deal with them.

Solon Barocas and Andrew D. Selbst, Big Data's Disparate Impact, 104 California Law Review 671 (2016), available at http://ssrn.com/abstract=2477899. A superb review of the ways big data can be discriminatory at work – even when that is not the intention.

Software & Information Industry Association, Algorithmic Fairness, September 2016, available at http://www.siia.net/Portals/0/pdf/Policy/Algorithmic%20Fairness%20Issue%20Brief.pdf . This issue brief surveys the legal landscape in the U.S. regarding disparate impact obligations and discusses a framework of responsible data use and recommends the use of disparate impact analysis generally.

Conduct Disparate Impact Assessments

Board of Governors of the Federal Reserve System, Report to Congress On Credit Scoring and Its Effects On The Availability and Affordability of Credit, August 2007, available at http://www.federalreserve.gov/boarddocs/rptcongress/creditscore/creditscore.pdf This study suggests that credit scores are accurate predictors of credit risk and their disparate impact reflects differences in the underlying credit-worthiness of different groups.

Robinson + Yu, Knowing the Score: New Data, Underwriting, and Marketing in the Consumer Credit Marketplace, October 2014, at https://www.teamupturn.com/static/files/Knowing_the_Score_Oct_2014_v1_1.pdf This white paper surveys the older data and methods used for credit scoring and compares them to the newer data sources and models.

FICO, Understanding FICO Scores, https://www.myfico.com/downloads/files/myfico_uyfs_booklet.pdf This short white paper explains how credit scores work and what the major factors determining a credit score.

Consumer Financial Protection Bureau, Request for Information Regarding Use of Alternative Data and Modeling Techniques in the Credit Process, 82 Federal Register 11183 (February 21, 2017) at https://www.gpo.gov/fdsys/pkg/FR-2017-02-21/pdf/2017-03361.pdf. This short notice outlines the issues raised by the development of new data sources and new analytic techniques. The comments filed in response can be found here: https://www.regulations.gov/docketBrowser?rpp=50&so=DESC&sb=postedDate&po=0&dct=PS&D=CFPB-2017-0005. SIIA's comment is here: https://www.regulations.gov/document?D=CFPB-2017-0005-0072

Federal Trade Commission, Credit-Based Insurance Scores: Impacts On Consumers Of Automobile Insurance, July 2007, available at http://www.ftc.gov/sites/default/files/documents/reports/credit-based-insurance-scores-impacts-consumers-automobile-insurance-report-congress-federal-trade/p044804facta_report_credit-based_insurance_scores.pdf. Credit scores can also be used to predict automobile insurance risk. This FTC study evaluates whether these scores have a disparate impact.

Robert Avery, et al. "Does Credit Scoring Produce a Disparate Impact?" Staff Working Paper, Finance and Economics Discussion Series, Divisions of Research & Statistics and Monetary Affairs, Federal Reserve Board, October 2010, available at https://www.federalreserve.gov/pubs/feds/2010/201058/201058pap.pdf This working paper from the Federal Reserve Board is a good summary of the disparate impact studies in the area of credit scoring.

Experian, FICO Reason Codes, at https://www.figfcu.com/documents/fico/FICO%C2%AE%20Score%20Factors%20Guide%20-%20Experian.pdf These codes are meant to explain the reason for a lower score and suggest actions to take that might make things better.

Center for Financial Services Innovation, The Predictive Value of Alternative Credit Scores, November 26, 2007 at http://www.cfsinnovation.com/node/330262?article_id=330262 Some data suggests that alternative data can improve credit scoring.

Charles River Associates, Evaluating the Fair Lending Risk of Credit Scoring Models, February 2014 available at http://www.crai.com/sites/default/files/publications/FE-Insights-Fair-lending-risk-credit-scoring-models-0214.pdf. This economic analysis shows how lenders can use analysis to reduce fair lending risk.

Collect and Use Adequate Data

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 recommends this methodology for proxy variables to use to assess disparate impact for fair lending compliance, when variables for race and ethnicity are not available.

Hamilton, Melissa, Risk-Needs Assessment: Constitutional and Ethical Challenges (January 26, 2015). American Criminal Law Review, Forthcoming; U of Houston Law Center No. 2014-W-2. Available at

SSRN: https://ssrn.com/abstract=2506397 or http://dx.doi.org/10.2139/ssrn.2506397 This comprehensive approach argues that a sensitive variable such as gender might need to be included in risk assessment tools in order to promote both accuracy and fairness.

Take Steps to Mitigate Disparate Impact

A. Romei and S. Ruggieri, "A multidisciplinary survey on discrimination analysis," The Knowledge Engineering Review, pages 1–57, April 3, 2013, available at http://pages.di.unipi.it/ruggieri/Papers/ker.pdf. This is a standard survey of the different ways organizations can use statistical modifications of algorithms to improve fairness.

Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, Suresh Venkatasubramanian, "Certifying and removing disparate impact," arXiv:1412.3756v3 [stat.ML], 2015, available at https://arxiv.org/abs/1412.3756. This article reviews several techniques for mitigating disparate impact by moving algorithms toward statistical parity.

Moritz Hardt, et al., Equality of opportunity in supervised learning, in Advances In Neural Information Processing Systems (2016) arXiv:1610.02413v1 This article describes a method of equalizing group error rates that preserves more accuracy that adjusting algorithms for statistical parity.

Beyond Disparate Impact

Alexandra Chouldechova, *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments* 1 (October 2016), available at https://arxiv.org/abs/1610.07524/ .This article illustrates the conflict between individual level fairness and equal group error rates in the context of recidivism predictions.

Sam Corbett-Davies, et al., Algorithmic decision making and the cost of fairness (June 10, 2017) available online at https://arxiv.org/abs/1701.08230v4 This article illustrates the sacrifice in accuracy and organizational objectives from steps moving toward equalizing group error rates

Julia Angwin, Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say, ProPublica (December 30, 2016), https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say In this follow up article, Angwin reports the conflict between equalizing group error rates and predictive parity and argues for equalizing group error rates.

Mark MacCarthy, Measures of Algorithmic Fairness Move Beyond Predictive Parity to Focus on Disparate Error Rates, SIIA, March 21, 2017, available at http://www.siia.net/blog/index/Post/71370/Measures-of-Algorithmic-Fairness-Move-Beyond-Predictive-Parity-to-Focus-on-Disparate-Error-Rates Choosing which measure of statistical fairness to use is inevitable and value-laden. Organizations need to be clear about what they are doing to mitigate disparate impact and why.