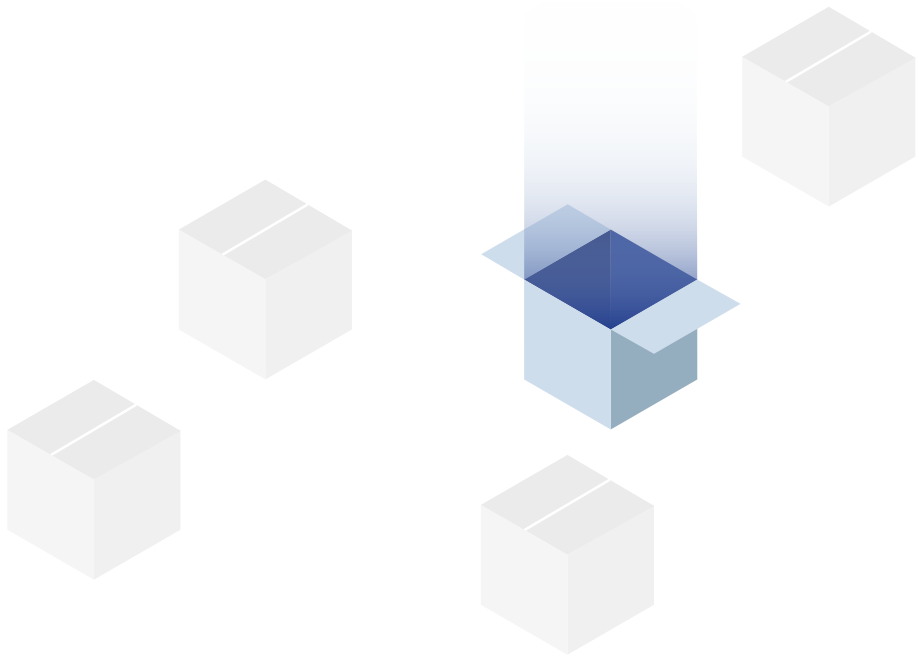
AINOW

TOOLKIT 01
OCTOBER 2018

Algorithmic Accountability Policy Toolkit



CONTENTS

INTRODUCTION

Algorithms are widely used in society to make decisions that affect most aspects of our lives, including which school a child can attend, whether a person will be offered credit from a bank, what products are advertised to consumers, and whether someone will receive an interview for a job. Federal, state and local governments are increasingly using algorithms to conduct government services. Algorithmic systems are used to make decisions about government resource allocation (e.g. where fire stations are built or where police are dispatched), expedite government procedures (e.g. public benefits eligibility and compliance), and aid government officials in making important decisions like whether a person will receive bail or a family will receive a follow up visit from a child welfare agency.

Despite the importance of these uses and decisions, government agencies frequently procure, develop, and implement algorithmic systems with minimal to no transparency, public notice, community input, oversight, or accountability measures. Procurement officers and agency staff often lack technical expertise to evaluate algorithmic systems, their capabilities, and potential consequences. This creates a knowledge imbalance in contracting, particularly because many algorithmic systems vendors almost exclusively sell to government agencies. Consequently, vendors are able to oversell the utility and value of a system or offer the system at reduced costs, which is difficult for resource constrained agencies to turn down.

Procurement officers and agency staff often lack technical expertise to evaluate algorithmic systems, their capabilities, and potential consequences.

Algorithms are fallible human creations, so they are embedded with errors and bias like human processes. When algorithmic tools are adopted by government agencies without adequate transparency, accountability, and oversight, their use can threaten civil liberties and exacerbate existing issues within government agencies (e.g. bias, inefficiencies, opacity regarding decision making). We know that federal, state and local governments are increasingly implementing algorithmic systems in their daily practices, but we still do not know how widespread and integrated such algorithmic systems are used at any level of government.

The following toolkit is intended to provide legal and policy advocates with a basic understanding of government use of algorithms including, a breakdown of key concepts and questions that may come up when engaging with this issue, an overview of existing research, and summaries of algorithmic systems currently used in government. This toolkit also includes resources for advocates interested in or currently engaged in work to uncover where algorithms are being used and to create transparency and accountability mechanisms.

Resource: Frequently Asked Questions

General Background

What are Algorithms? What are Automated Decision Systems?

An Algorithm is generally regarded as the mathematical logic behind any type of system that performs tasks or makes decisions. For example, how Facebook sorts what posts a user sees in their Facebook feed is an "algorithm." The logic used in a software program to assign criminal defendants a public safety risk score is also an "algorithm." "Algorithms" do not have to be based in software on computers. However, in the case of many types of risk assessments used in courts or human services agencies, the "algorithm" can be represented by a piece of paper that outlines the steps a human should take to evaluate a particular case.

What are Automated Decision Systems?

An Automated Decision[-making/-support] System is a system that uses automated reasoning to aid or replace a decision-making process that would otherwise be performed by humans. Oftentimes an automated decision system refers to a particular piece of software: an example would be a computer program that takes as its input the school choice preferences of students and outputs school placements. All automated decision systems are designed by humans and involve some degree of human involvement in their operation. Humans are ultimately responsible for how a system receives its inputs (e.g. who collects the data that feeds into a system), how the system is used, and how a system's outputs are interpreted and acted on.

When talking about automated systems used in government, you might hear people refer to "algorithms," "automated decision systems," or "algorithmic systems" loosely and interchangeably. "Automated decision system" was the phrase used in New York City for its algorithmic accountability task force, so we stick with that when talking about a complete end-to-end system used in government, from design, testing, and actual use, including the human operators.

What exactly does "Artificial Intelligence" (AI) mean?

Artificial Intelligence (AI) has many definitions, and can include a wide range of methods and tools, including machine learning, facial recognition, and natural language processing. But more importantly, AI should be understood as more than just technical approaches. It is also developed out of the dominant social practices of engineers and computer scientists who design the systems, and the industrial infrastructure and companies that run those systems. Thus, a more complete definition of AI includes technical approaches, social practices and industrial power.

What is "Machine Learning" (ML)? Is it the same thing as AI?

In current use, machine learning (ML) is the field most commonly associated with the current explosion of AI. Machine learning is a set of techniques and algorithms that can be used to "train" a computer program to automatically recognize patterns in a set of data. Many different tools fall under the umbrella of "machine learning." Though there are exceptions, ML generally uses "features" or "variables" (e.g. the location of fire departments in a city, data from surveillance cameras, attributes of criminal defendants) taken from a set of "training data" to learn these patterns without explicitly being told what those patterns are by humans. Machine learning has come to include things that have historically been more simply called "statistics." Machine learning is the technique at the heart of new automated decision systems, making it difficult for humans to understand the logic behind those systems.

Questions You May Have

How do algorithms relate to my work in immigration, criminal justice, education, housing, racial justice, national security, etc.?

Many different types of government agencies are increasingly using automated decision systems and other forms of predictive analytics. Automated decision systems can exist in any context where government bodies or agencies evaluate people or cases, allocate scarce resources, focus scrutiny or surveillance on communities, or make nearly any sort of decision. For example, in criminal justice we have seen algorithms used to assign recidivism risk scores to defendants or target policing activities through "predictive policing" systems. In education, we have seen algorithms used to evaluate teachers and match students to high school placements. Biases in the algorithms used by any of type of agency, including how they are formulated and what data they rely on, can lead to biased and harmful results for the people and communities most affected by the agencies. And even apart from the question of bias, there may be contexts in which it is always inappropriate or unlawful to have an automated decision system making consequential decisions

Where does the data used to build and train automated decision systems come from?

Sometimes the data used to train automated decision systems will come from the agency's own databases. Researchers studying automated decision systems have critiqued their use because of this limitation, since existing bias in an agency's decisions will be carried over in systems trained on biased agency data.[1]

Some government agencies, however, have access to data from other agencies for automated decision systems. This may become more common in the future as governments use more sophisticated technical platforms to manage and share their data.[2]

Some systems that government agencies use might be trained on data from third-parties that the government itself does not have, or private commercial data. For example, a vendor of automated decision systems might train a pre-trial risk assessment system on data from one jurisdiction, and then sell the trained risk assessment model to other jurisdictions.[3]

Has anyone litigated the use of algorithmic systems in government?

Yes. Cases across the country have challenged the use of automated decision systems on different grounds, including in criminal sentencing and in public benefits.

In State v. Loomis (2016) in front of the Wisconsin Supreme Court, the defendant claimed that the court's use of a risk assessment when determining his sentence was a due process violation. Other cases have tried to force agencies to divulge information about algorithmic systems (including their source code) in criminal proceedings.

---

1 One study on predictive policing, for example, demonstrated that predictive policing systems that use police records on drug crimes to predict drug use will come to more biased decisions than if they were trained on drug use data from different sources. More broadly, predictive policing algorithms frequently attempt to predict future crime based on past arrest locations, which is problematic because arrest locations reflect human decisions about where to focus policing efforts. Unchecked, this propagates past biased policing decisions to focus on communities of color and/or low income communities into predictive policing systems.

2 For example, the Allegheny County (PA) Department of Human Services uses an automated decision system for evaluating the risk of child abuse or neglect, which augments its own data with data from local police departments, mental health services, and public benefits agencies. Research from sociologist Virginia Eubanks has explored how this use of data focuses scrutiny on poor people since they disproportionately receive attention from government agencies.

3 US Immigration and Customs Enforcement, for example has proposed using social media data to evaluate immigration applications.

Questions You May Have

What about the use of automated decision systems in private companies?

Private companies use algorithms every day in the course of regular business, some of which urgently need public scrutiny. The effort to regulate those sorts of algorithms can benefit from what we learn about holding public algorithms accountable, but the methods may have to look quite different.

Work by ACLU's Racial Justice Project has highlighted how Facebook's ad-targeting platform can allow advertisers to illegally discriminate against people of color by limiting their audience for housing advertisements by "ethnic affinity."

Who studies algorithmic systems from a technical viewpoint?

A growing interdisciplinary community of academic researchers studies fairness, accountability, and transparency in algorithmic systems, with contributions from computer science, the social sciences, and the law.

The annual Fairness, Accountability, and Transparency Conference (FAT*) has been a gathering point for this community.

How does fairness/bias get defined in the technical world?

The technical community has made many attempts to define "fairness" mathematically, so that machine learning systems can be made to meet some provable standard of "fair." This effort is the subject of ongoing research in machine learning.[4] It should be noted that this conversation often ignores notions of "justice:" in seeking to address these systems, justice is what we're after, not merely fairness. They are not the same thing, but the academic machine learning research focuses on quantifying the latter.[5]

For example, consider an ML-based pretrial risk assessment that attempts to rate a defendant's risk of rearrest as either low risk or high risk. Under one definition of "fairness" centered on "calibration," the risk assessment should be deemed fair if "high-risk" and "low-risk" mean the same thing for Black and White defendants; that is, if "high-risk" for a Black defendant means there is a 70% probability they will be rearrested, then "high-risk" for a White defendant should also mean there is a 70% probability they will be rearrested.

Under a different definition of "fairness" based on error rate balance, then the risk assessment should be considered fair if it mis-scores Black and White defendants at similar rates (i.e. a Black defendant who ultimately does not get rearrested is just as likely to be given a high risk score as a White defendant who does not ultimately get rearrested).

These sorts of definitions sometimes directly contradict each other. Research has found that in the above example, as long as the base rates of rearrest among Black and White defendants differ and as long as perfect prediction is impossible, a risk assessment cannot meet these two definitions of fairness ("calibration" and "error rate balance") at the same time. There are many more definitions of fairness that go beyond these two relatively simple ones, and their differences can have meaningful differences in how systems using machine learning treat people.

---

4 A recent talk by computer science professor Arvind Narayanan highlights the work being done in defining fairness, the trade-offs and ethical considerations reflected in the different definitions, and the limitations of the entire technical field of study. A guide on "quantitative fairness" by statisticians also offers a view into how the field of statistics and machine learning views fairness.

5 Though there are exceptions, and the field is rapidly changing: research by computer scientist Reuben Binns has looked at understanding fairness and justice in machine learning with "lessons from political philosophy."

Questions You May Have

Can you understand an automated decision system simply by looking at the source code (that is, the programming written by a human)?

In most cases, no. The source code of a system that uses machine learning will not reveal the "rules" the machine learning model uses to make decisions. Instead of source code, it is helpful to have access to the system's training data or the "model" or "weights" that the ML algorithm learned.

Source code might be helpful in the case of "expert systems" or other simpler automated decision systems where a human explicitly writes decision making rules into code. But in the case of the growing use of AI and machine learning in government, the source code is neither sufficient for understanding the system nor is it often necessary.

Aren't automated decision systems infallible? After all, they're computers.

We have to remember that humans are a necessary part of automated decision systems. Automated decision systems are not built and used in a vacuum: humans classify what data should be collected to be used in automated decision systems, collect the data, determine the goals and uses of the systems, decide how to train and evaluate the performance of the systems, and ultimately act on the decisions and assessments made by the systems. So, like humans, they are not infallible.

Why do we need a separate effort to regulate the use of algorithms in government? Aren't existing laws that address discrimination and harm in education/criminal justice/employment/etc. sufficient?

The opacity and inscrutability of algorithms present a new threat to our ability to understand how government agencies. We need a new approach to identify and address bias and discrimination in automated decision systems. New laws and practices are needed to encourage the safe development of these systems if they are ever to be used and to enable new forms of oversight. At the same time, we must also find ways to enforce existing laws and standards even in the face of algorithms that might muddy the picture.

Automated decision systems are not built and used in a vacuum: humans classify what data should be collected to be used in automated decision systems, collect the data, determine the goals and uses of the systems, decide how to train and evaluate the performance of the systems, and ultimately act on the decisions and assessments made by the systems.

Questions You May Receive

If they're based off of data from the real world, how can algorithmic systems be biased? Don't they just learn from reality?

Humans are responsible for defining what data should be collected, how it will be collected, and how it will be used. Automated decision systems are largely built on finding patterns in that data. Because the collection and use of data is such a human process, we should not take for granted that the data is "correct," or representative of a reality that we want to perpetuate in future.[6] Automated systems are not inherently scientifically objective.

Why should we be concerned with bias in automated decision systems? Aren't they replacing humans who are already biased?

It is true that individual humans who make decisions already have biases, but automated decision systems can amplify bias on an unprecedented scale, and give that bias the appearance of scientific objectivity.[7] Automated decision systems also raise due process concerns that we do not yet know how to address, especially when people cannot access or understand the technical systems that produced the decision. Responsibility matters, but it's hard to hold algorithms accountable for their decisions.

What makes an automated decision system different than some types of excel spreadsheets or other simple tools already used by government?

In some cases, if a public agency uses simple tools like Microsoft Excel to automate non-trivial decision processes, we might want to consider those uses of the tool their own kind of automated decision system. We would not necessarily want to say that Microsoft Excel is an automated decision system, but we would want to clarify that it can be used to implement more complex logic that should be scrutinized. It will be difficult to draw a boundary around these uses of seemingly simple tools.

Haven't we had these types of systems for years? Why do we care now?

Automated decision systems are not an entirely new phenomenon, but their deployment is rapidly expanding into new areas of government. This is in large part thanks to the availability of more data and increased interest in "smart cities" and "smart government." The stakes have also changed — automated decision systems are increasingly being used in more and more sensitive decisions that impact human welfare.

6 As referenced earlier, research has shown that the choice of what data to use to solve a particular problem can lead to different outcomes in the decisions made.

7 See, Danielle Keats Citron's foundational paper "Technological Due Process" for a discussion of "automation bias:" the tendency to trust automated processes over human reasoning, resulting in undue deference to algorithms.

Resource: Types of of Algorithmic Systems used in Government and Legal Concerns

Humans are responsible for defining what data should be collected, how it will be collected, and how it will be used. Automated decision systems are largely built on finding patterns in that data. Because the collection and use of data is such a human process, we should not take for granted that the data is "correct," or representative of a reality that we want to perpetuate in future. Automated systems are not inherently scientifically objective.

Human Resources/Public Benefits

Child Risk and Safety Assessment	An instrument that assesses the risk of current and future harm to a child. The tool can be used at different stages in the decision making process at a child welfare agency. Common uses include workers assessing whether a family should receive a secondary visit by a social service worker, or whether a family should receive intervention services.
Genogram and Ecomap Software	An assessment tool that allows child welfare caseworkers to map family trees, identify gaps in family history, organize information amassed from family, and assess interventions.
Homelessness Prioritization	A tool used to prioritize individuals in need of temporary or permanent housing, so the most vulnerable can be helped first. Depending on the housing options available locally, it can be used for placement in housing as well as eligibility for housing subsidies.
Medicaid eligibility assessment	A tool that determines eligibility and compliance for Medicaid. Similar tools are used to assess eligibility and compliance for other public benefits. Known Vendor: **IBM, APS Healthcare**

Public Health

Disease treatment	An algorithm used to identify individual with chronic hepatitis C for treatment and cure. The systems also analyzing health surveillance data to monitor treatment and cure rates within a municipality to assess progress towards treatment goals.
Prescription drug monitoring databases	Some states are using proprietary algorithms applied to prescription drug monitoring databases to identify possible doctor shopping or improper prescribing. Known Vendor: **Appriss**

Criminal Justice

Surveillance Technologies

Many surveillance technologies used by local and state law enforcement use algorithms including but not limited to, facial recognition (including on body cameras), automatic license plate readers, and visual or data analytics systems. Law enforcement agencies also data-mining software that processes large quantities of data from commercial and government sources to identify relationships or connections between people, places, and things.

Known Vendors: **Palantir, Vigilant Solutions, Cognitec, Amazon, Microsoft, Motorola, IBM, Axon**

Predictive Policing

Any system that analyzes available data to predict either (A) where a crime may happen in a given time window (place-based) or (B) who will be involved in a crime as either victim or perpetrator (person-based). A predictive policing system then must convey that information to police officers or other social service providers so that they can take some course of action.

Known Vendors: **Predpol; Azavea (Hunchlab); Palantir, Starlight, Bair Analytics, IBM, RTMDx**

DNA Analysis

Also known as probabilistic genotyping, these systems interpret forensic DNA samples by performing statistical analysis on a mixture of DNA from different people to determine the probability that a sample derives from a potential suspect.

Known Vendors: **Strmix, TrueAllele, Cybergenetics**

Pretrial Risk Assessment

A system that analyzes information collected during interviews with an arrested person to assess the person's likelihood of nonappearance, rearrest, and rearrest for a violent crime. Most pretrial risk assessments use a simple algorithm that is reliant on a small number of input variables, which are usually determined by state law. When predicting what defendants might not appear in court, such risk assessments are sometimes called "failure to appear" tools.

Known Vendors: **Northpointe (COMPAS); University of Chicago Crime Lab**

Sentencing Risk Assessment

A system designed to reduce recidivism by targeting defendants that are considered "high risk" and reduce prison populations by diverting "low risk" defendants from prison.

Known Examples: **MHS Assessments (Juvenile Sentencing tool); PA Sentence Risk Assessment Instrument**

Inmate Housing Classification

Any system that analyzes a variety of criminal justice data and outcomes to determine the conditions of confinement and overall housing arrangements of inmates in a jail or prison.

Parole

A system analyzing a variety of criminal justice data and determinations to assist a decision regarding whether an inmate should receive parole and terms of parole.

Known Vendors: **Northpointe**

Education

Teacher Evaluation	The most commonly used version of this tool is the value added model evaluation, which aims to measure how each teacher contributes to student educational achievement. Typically, the tool compares student test scores over time, but the actual variables used for evaluation are often unknown because of proprietary claims made by vendors.
School Assignment	Many school assignments are determined using a simple match algorithm that evaluates school choices selected by parents and a school districts admission preferences and seat availability. Knowns Vendors: **Institute for Innovation in Public School Choice**[8]
Controlled Choice	A student assignment algorithm that is designed to achieve school diversity and optimize choice/distribution balance, particularly in school districts experiencing racial and/or socioeconomic segregation. Parents rank-order schools by preference, and students are assigned to school based on constraints set by the school district to achieve a balanced student distribution goal. Known Vendors: **Michael Alves (education consultant)**
School Violence Risk Assessment	A tool designed to identify students who are at a high risk for school related violence (e.g. homicide, suicide). A recent study used the BRACHA (Brief Rating of Aggression by Children and Adolescents) scale measures aggressive behavior, and the School Safety Scale, which measures behavioral changes that may indicate violence, and manual annotation of student interview to make predictions about likelihood of violence.

Automated decision systems can exist in any context where government bodies or agencies evaluate people or cases, allocate scarce resources, focus scrutiny or surveillance on communities, or make nearly any sort of decision.

8 Benjamin Herold, "Custom Software Helps Cities Manage School Choice," Education Week, December 4, 2013.

Resource: Relevant Literature

The following document is a review of relevant books, academic articles, reports and other publications that can help provide advocates understand the current concerns, debates and legal or technical analyses related to algorithmic accountability. Though, the publications are organized by the issue area headings, many of the recommended publications cover several topics so advocates should explore the full list and read the descriptive summaries to determine which publications are most useful for specific advocacy objectives. This literature review was derived from a larger AI Now Law and Policy Reading List, which includes more robust list of publications as well as relevant cases, statutes, and legislation.

Basic Background

Public Scrutiny of Automated Decisions: Early Lessons and Emerging Methods

Upturn
Omidyar Network

A report studying how journalists, researchers, and lawyers have so far kept the use of automated decision systems in government accountable. This report systematically documents specific cases where scrutiny has been applied to automated decision systems, the specific technical, legal, and journalistic techniques and tools of accountability used in those cases, and potential directions for future policymakers to take to keep public automated decision systems accountable. This resource can provide a more in-depth treatment of how algorithms used in government work, accessible to policymakers.

Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy

Cathy O'Neil

The author of this book, a mathematician and data scientist, discusses the high-level problems with using models and big data to drive significant decisions. After providing a primer on models and how they may fail, the book focuses on higher education, online advertising, criminal justice, employment, credit, insurance, and civic life, positing that the systems are discriminatory in part because the algorithms backing them are unregulated and difficult to challenge.

Public Records Requests

Algorithmic Transparency for the Smart City

Robert Brauneis
Ellen P. Goodman

In this article, researchers attempt to use public records requests to investigate the use of algorithms in public agencies and evaluate their usefulness for holding algorithms accountable. They describe useful information that public records requests should contain. This article is a useful resource for making effective public records requests for algorithms with real-world examples from criminal justice and child welfare contexts.

Limitations of Transparency

Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability

Mike Ananny
Kate Crawford

This article explores the inadequacy of transparency as a type accountability in algorithmic systems. It argues that transparency ideals fall short when attempting to understand or govern algorithmic systems and proposes alternative approaches to creating algorithmic accountability.

Due Process

The Scored Society: Due Process for Automated Predictions

Danielle Keats Citron
Frank Pasquale

This article details how algorithmic outputs can be inherently biased, and thus discriminatory. Using the case study of credit scoring, the authors highlight three main areas of failure for such algorithms: opacity, arbitrary assessment, and disparate impact. They then describe safeguards such as regulatory oversight, transparency, and notice. This article is useful in understanding objections to the scoring of individuals, with a particular focus on credit scores.

Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms

Jason Schultz
Kate Crawford

This article explores the limitations of existing privacy laws in addressing the new and novel risks and harms presented by big data and predictive analytics. It then explores the history of procedural due process to argue that procedural due process principles can appropriately address these risks and harms.

Courtroom Algorithms

The Big Data Jury

Andrew G. Ferguson

This article examines the use of big data with regard to juries, specifically representative jury pool selection and providing litigants with personal information about the jurors themselves. It explores the multitude of ways that jury selection can be impacted by algorithmic technologies, as well as the potential constitutional implications of using such technologies.

Predictive Policing

Stuck in a Pattern: Early evidence on "predictive policing" and civil rights

David G. Robinson
Logan Koepke
(An Upturn report)

This resource explains what predictive policing is, how it works, and what it may mean for civil rights. It comprises prior literature, analysis of predictive policing systems, and surveys of major police departments, and it provides a list of critiques of predictive policing methods. This study is designed as a starting point for questioning, evaluating, and challenging predictive policing systems.

Big Data Surveillance: The Case of Policing

Sarah Brayne

This articles summarizes the observations and findings from an empirical study of the Los Angeles Police Department's adoption of big data analytics. The study found that big data analytics have amplified prior surveillance practices that create greater social inequalities and consequences. Based on these findings and other observations, Brayne developed a theoretical model of big data surveillance that can be applied in several institutional domains.

Illuminating Black Data Policing

Policing Predictive Policing

Andrew G. Ferguson

These articles contend that predictive policing is marred by inadequate, "black" data that is opaque, racially biased, and insufficient for big data methods. They also review constitutional doctrine and upcoming questions regarding the use of predictive policing and other police technologies. Policing Predictive Policing in particular provides a list of the different aspects of predictive policing systems that make them problematic.

To Predict and Serve?

Kristian Lum
William Isaac

This paper demonstrates how the origin of data used in predictive policing systems can lead to biased outcomes. For example, predicting the location of drug crimes using historic policing data leads to more bias against communities of color versus prediction using public health data on drug use. This paper is useful when discussing athe correctness of data used in AI should not be taken for granted.

Predictable Policing: Predictive Crime Mapping and Geographies of Policing and Race

Brian Jordan Jefferson

This article examines Chicago's predictive crime mapping system using geographic information systems as a tool. It suggests, after this map-based analysis, that predictive policing exacerbates racial disparities in policing and perpetuates geographically-based racism.

Criminal Justice Risk Assessment

"Fair" Risk Assessment: A Precarious Approach for Criminal Justice Reform

Ben Green

This articles interrogates how risk assessment tools that are considered fair by technical definitions, can actually be unfair and hinder or dilute existing criminal justice reform efforts. It suggests that the machine learning field should expand the considerations and questions used to evaluate the opportunities and challenges presented by the use of risk assessments

Assessing Risk Assessment in Action

Megan Stevenson

This article assesses the results of a pretrial risk assessment used in Kentucky, concluding that pretrial risk assessment did not significantly increase the rate of pretrial releases, and that judges generally return to their old habits even when using the risk assessment. In addition to the case study, the article provides background on pretrial risk assessment programs and suggests takeaways from Kentucky's experiment for other jurisdictions seeking to implement a similar program. This resource is most useful when evaluating claims that risk assessments are inherently decarceral in nature, a claim this article provides evidence against.

Danger Ahead: Risk Assessment and the Future of Bail Reform

Logan Koepke
David G. Robinson

This resource considers the impact of pretrial risk assessment on pretrial detention, specifically noting that current tools rely on faulty data and overestimate risk, escape public scrutiny, and provide a veneer of scientific objectivity for social concepts like "dangerousness." The authors also explain how systems should be developed to avoid harms, specifically ensuring quality inputs and good governance. The resource provides a history of bail and an explanation of pretrial risk assessment, and is most useful for its explanations of how to address the challenges presented by pretrial risk assessment systems.

The Use of Risk Assessment at Sentencing: Implications for Research and Policy

Jordan Hyatt
Steven L. Chanenson

This report provides background on risk assessment for sentencing, surveys judicial attitudes toward sentencing, compiles select court opinions on the topic, and provides case studies from Virginia, Pennsylvania, and Utah. It serves as a means for understanding the judicial lay of the land when it comes to risk assessment, and suggests benefits and drawbacks to relying on algorithmic decision-making in this arena.

The Accuracy, Fairness, and Limits of Predicting Recidivism

Julia Dressel
Hany Farid

This paper is a quantitative examination of how a particular pretrial risk assessment tool — Northpointe's COMPAS system — functions. It concludes the software is no more accurate or fair than predictions made by people, and provides an alternate, more simplified model that would generate the same results. It is most valuable as a guide for how to similarly evaluate other pretrial risk assessment tools, providing a model and methodology that could be ported to other kinds of algorithmic assessment as well.

Human Resources/Public Benefits

Automating Inequality

Virginia Eubanks

This book provides extensive case studies on automated welfare benefits determination in Indiana, a homeless registry in California, and predictive models for child welfare in Pennsylvania. The author describes how automated systems have failed to provide services in a fair and efficient manner, and especially focuses on how they impact poor, minority, and otherwise vulnerable populations. The resource is most helpful in approaching the problems of algorithms from an advocate's perspective.

The Empirical Turn in Family Law

Claire Huntington

This article details the history of empirical analysis in family law (covering issues such as abortion, marriage inequality, domestic violence, juvenile sentencing, and child custody), where the empirical tools have been beneficial, and when they have failed. The author considers the use of empirics overall as contributing to good governance, but proposes frameworks that prevent the misuse of empirics and protect families. The article is targeted at legal scholars and puts algorithms in the context of court systems.

Foretelling the Future: A Critical Perspective on the Use of Predictive Analytics in Child Welfare

Kelly Capatosto (Kirwan Institute)

This white paper details potential problems with using predictive analytics in the field of child welfare. It first provides background on how data is used to identify risk for youth, discusses cognitive and structural reasons why the data or predictive methods can contribute to bias or other suboptimal outcomes, and suggests remedies. This resource can serve as a primer for how to assess and critique algorithms used in child welfare systems.

Technological Due Process

Danielle Keats Citron

This article describes how automated decision systems can fail to provide adequate due process for people subject to administrative decisions. It explains how the creation of automated decision systems blurs the line between adjudication and rulemaking in how agencies make decisions, and argues for the need for enhanced due process protections for new algorithmic systems.

"What happens when an algorithm cuts your health care"

Colin Lecher

This article describes the use of algorithms in home health care assessments, focusing on the deployment of a new algorithmic system in Arkansas but also highlighting cases in Colorado, California, and Idaho. It provides anecdotal descriptions of how individuals have challenged home care algorithms in court, and describes the importance of explainability, transparency, and notice. This article provides a useful case study of automated decision making in the provision of public benefits and its failures in the specific instance of determining home care hours.

Healthcare

Regulating Black-Box Medicine

W. Nicholson Price

This resource provides a short history and explanation of the use of algorithms in medicine that help in guiding care. It also discusses the FDA's current regulatory scheme for such algorithms and suggests potential pitfalls and possible improvements. This article is particularly useful when considering the role of algorithms in high-risk domains and new ways to regulate algorithms.

Education

"The Broken Promises of Choice in New York City Schools"

Elizabeth A. Harris
Ford Fessenden

This article describes the process of selecting a high school under former Mayor Bloomberg's school choice system, describing its structural inequities and focusing in particular on the opaqueness of the matching program. It provides useful background on how discrimination and segregation can persist despite the policy goals of an automated decision system.

Immigration

Algorithmic Jim Crow

Margaret Hu

This resource examines how algorithm-based biometric systems can be leveraged to create security systems that unjustly discriminate against individuals of different races, national origins, and religions at the border. The author explains how classification and screening methods, when coupled with big data, may facilitate discrimination against minorities. Also provided is a section explaining how to litigate against these algorithmic tools, proposing strategies and possible shifts in doctrine that may occur as a result.

Algorithmic Accountability Frameworks

Algorithmic Impact Assessments: A practical framework for public agency accountability

AI Now Institute

This report proposes "Algorithmic Impact Assessments" (AIAs) as a framework to assess automated decision systems and ensure public accountability. It lays out the process and components of an AIA and can serve as a resource for government agencies and members of the public to understand what tools are needed to keep automated decision systems accountable.

MAPPING SYSTEMS, DATA AND VENDORS

Public Records Request Guidance

Public records requests have been a useful tool in identifying where and how government agencies are using algorithmic tools. They can also provide useful information to help challenge government use of algorithmic tools. Yet, it is not always apparent what to ask for, what information is subject to public records law, and what information is useful for civil liberties analysis. The following guidance will help affiliate staffs determine what to look and ask for to assess and potentially challenge government use of algorithmic systems.

You should be aware that public records requests of algorithmic tools can become resource intense. Some of the information you will request may be held by a third party vendor, so a government agency may suggest that they do not have responsive records to your request because they are not technically retained by the agency. In these circumstances, you will need to be prepared to challenge the government's response through an administrative appeal. You can also consider advocating for legislative reform of your state's open records law to address this problem.

Because the collection and use of data is such a human process, we should not take for granted that the data is "correct," or representative of a reality that we want to perpetuate in future.

Steps to a Public Records Request

1 **Review your State's Open Records Law**

While many state public records laws are very similar to the federal law, several vary drastically so it is important follow this guidance in context of your state law.

2 **Compile Sources or References to the System**

In addition to tailoring your request to your state law, you should try to tailor your request in accordance with known or speculative information about the system. Potential sources of references to systems may include news articles, public statements or press releases by public officials, agency or legislative budgets, agency or legislative public hearing notes or minutes, or relevant databases (e.g. MuckRock Project Public Records Request Archive) . Compiling these reference sources in advance may make your request more specific and help if you have to appeal or challenge the response to your public records request.

3 **Draft Public Records Request**

You can draft your request using the annotated model public records request as a template but use your sources or reference documents to narrow your request. In addition to being specific, you should try to include examples of what you are looking for so it is clear to the public records officer or other government officials who are making determinations about what is responsive to your request.

Annotated Model Public Records Request:

Most of the provisions in this model request are in reference to a software based algorithmic-system. Be mindful that some systems may function as a hybrid, where government officials perform certain tasks or functions of the algorithm. Therefore, you must modify the language to capture these distinctions so your request is not rejected (in part or whole) because of semantics or misinterpretations.

Re: **STATE'S PUBLIC RECORDS LAW** Request
ACTUAL OR DESCRIPTIVE NAME OF SYSTEM

Dear **AGENCY PUBLIC RECORD OFFICER OR DESIGNEE**:

The American Civil Liberties Union of _____ is filing this request for records pursuant to **STATE LAW NAME AND STATUTORY CITATION**, to seek information on the **NAME OF AGENCY**'s **ACTUAL OR DESCRIPTIVE NAME OF SYSTEM** used to **SUCCINCTLY DESCRIBE TO THE BEST OF YOUR ABILITY WHAT THE SYSTEM DOES OR HOW IT IS ALLEGEDLY USED**.

The American Civil Liberties Union of _____ request the following records:

This provision seeks the actual algorithm and other relevant technical information. It is important to specific to the best of your knowledge what the system does. Questions to consider when describing the system:

- Does it predict behavior or actions?
- Is it used to determine what resources a person will receive or how they may be treated?
- Is it classifying an individual or group of people?

1. All records including information relating to the algorithm that **DESCRIBE WHAT IT DOES** within **NAME OF AGENCY, DIVISION OR FACILITY**, including but not limited to its source code, models, developer documentation, and operator manuals.

These provisions seek information about the variables a system may use to produce an output (e.g. a prediction or determination).

2. All records relating to the training data used to develop, or train, the algorithm.
3. All records, including but not limited to documentation or internal communications, about the traits, characteristic, or factors used to develop the data fields in the System.

Use your source references to try to identify specific outputs of a system to specify here because if this provision is too broad the public records officer may reject the request because they do not know what may be response.

4. All records showing the full list of the data fields in the System.
5. All records of de-identified input data in the System.
6. All de-identified records of algorithm outputs, including but not limited to **ADD A SPECIFIC EXAMPLE OF SOMETHING YOU ARE SEEKING**.

This provision seeks information to help you determine what level of deference agency staff may give to the system's outputs. Typically, the lack of any information, directives, and/or guidances suggests a fairly subjective process

7. All records showing how **NAME OF AGENCY** staff use algorithm outputs to determine **INSERT ANY KNOWN PREDICTIONS OR DETERMINATIONS MADE BY THE SYSTEM**.

8. All records of, including communications regarding, audits, internal reviews, or validation studies of the System.
9. Any internal policies, practices, procedures, memoranda and training materials for using the System, and for storing, accessing, and sharing data inputs and analysis created by the System.
10. Any internal policies, practices, procedures, memoranda and training materials for sharing data inputs and outputs created by the System with entities outside of **NAME OF AGENCY**, including the **SPECIFY AN OUTSIDE ENTITY OF CONCERN**.
11. Any records showing which entities outside of **NAME OF AGENCY** have accessed, used or requested to use, the System.
12. Any records reflecting any agreements for or permission to develop, use, test, or evaluate an algorithmic system used to **SPECIFY FUNCTIONS AND/OR OUTPUTS** and services with any third-party vendor or consultants, including the **SPECIFY THIRD-PARTY THAT MAY HAVE COLLABORATED OR REFERENCED IN SOURCE DOCUMENTS**.
13. Any Records referencing the public process preceding the procurement or acquisition of the System, including public meeting agendas or minutes, public notice, analyses, or communications between **NAME OF AGENCY** and elected officials or other public servants.

If possible, please provide requested records in electronic format. Please contact us before retrieving the records so that we can ensure that the retrieved records are in a usable and readable format.

Upon locating the requested documents, please contact us before any reproduction or photocopying and advise us of the actual costs of duplication so that we may decide whether it is necessary to narrow our request.

We would appreciate a response as soon as possible and look forward to hearing from you shortly. Please furnish the requested records to:

NAME
ADDRESS
PHONE NUMBER
EMAIL

If any portion of this request is denied for any reason, please inform us of the reason for the denial in writing and provide the name and address of the person or body to whom an appeal should be directed.

This provisions seeks information on whether the agency or vendor has audited or tested the system for bias and errors.

This provision seeks information on how the system is used, as well as how information may be shared. Ideally, this information will help identify security features of the system and agencies procedures to prevent or address misuse, abuse, etc.

This provision seeks information about policies or procedures for external sharing. You should specify external entities of concern like law enforcement or federal agencies.

This provision seeks information about whether information has ACTUALLY been shared externally, regardless of whether there are policies or procedures and whether they have been followed.

This provision is seeking information about contracts with consultants or other third-parties.

This provision seeks to identify if there was any public communications about the system. This information can be useful for any advocacy regarding public procurement loopholes or problems.

Little Sis Tracking

Little Sis (https://littlesis.org/) is a free and open database detailing the connections between people and organizations. Little Sis can be used to track the relationships between government officials, vendors, lobbyist, business leaders, philanthropic organizations, and independent donors. The database is populated by its users so while it can be resource intensive to use this tool, it can help illuminate relationships and transactions that can be important in the algorithmic accountability contexts.

The database consist of:

- **Entities** (people of organizations)
- **Relationships** between entities in various categories (Position, Education, Membership, Donation/ Grant, Service/Transaction, Ownershivp, Hierarchy, or Generic)
- **Lists** that allow for the groups of Entities
- **Maps** that visualize relationships between entities in the form of Diagrams

Little Sis is currently being used to identify and map relationships between government agencies, vendors, and the recently appointed automated decision systems task force in New York City. Below are screenshots of how the tool is being used to map relationships.

Two lists were created to identify New York City Agencies and Vendors.

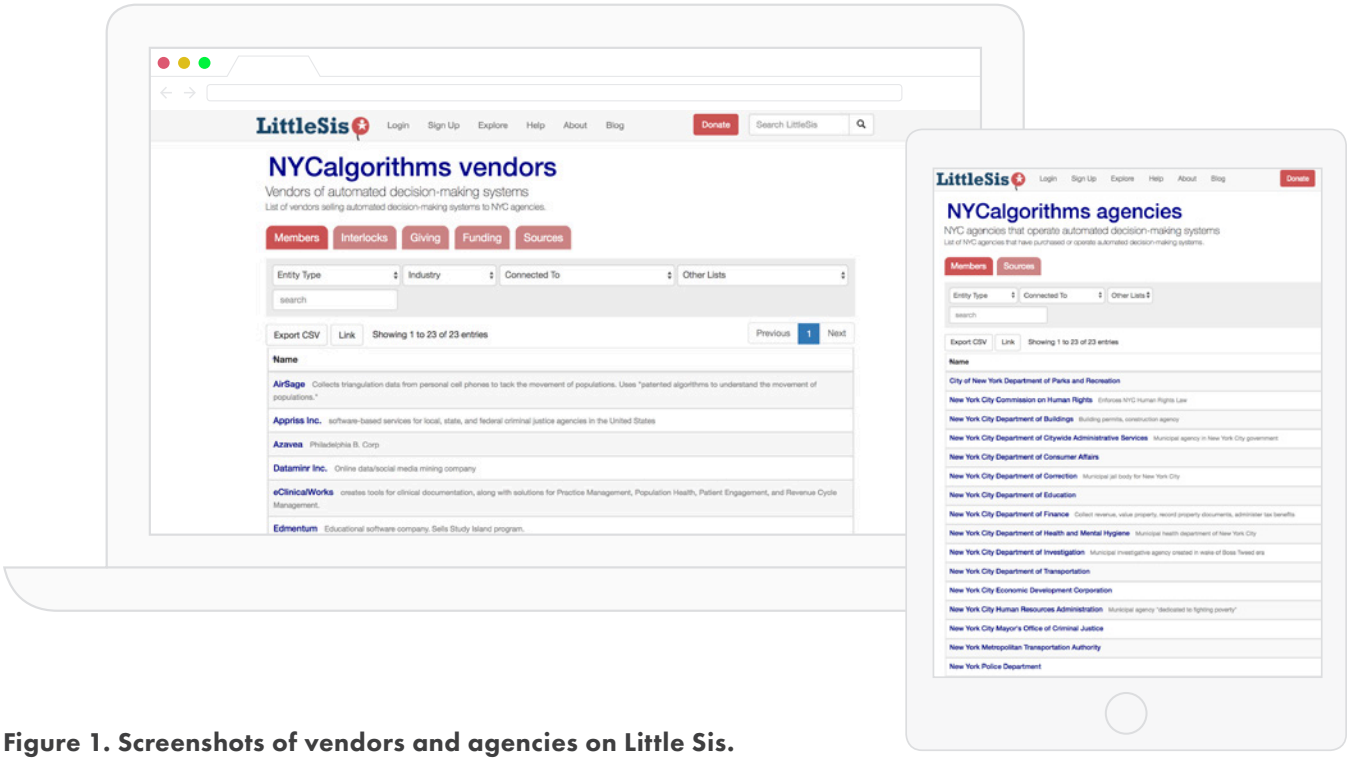


Figure 1. Screenshots of vendors and agencies on Little Sis.

Using public records and media reports, relationships and transactions between vendors and agencies were entered into Little Sis to develop a map that illustrates relationships.

Figure 2. Screenshot of entities and relationships on Little Sis.

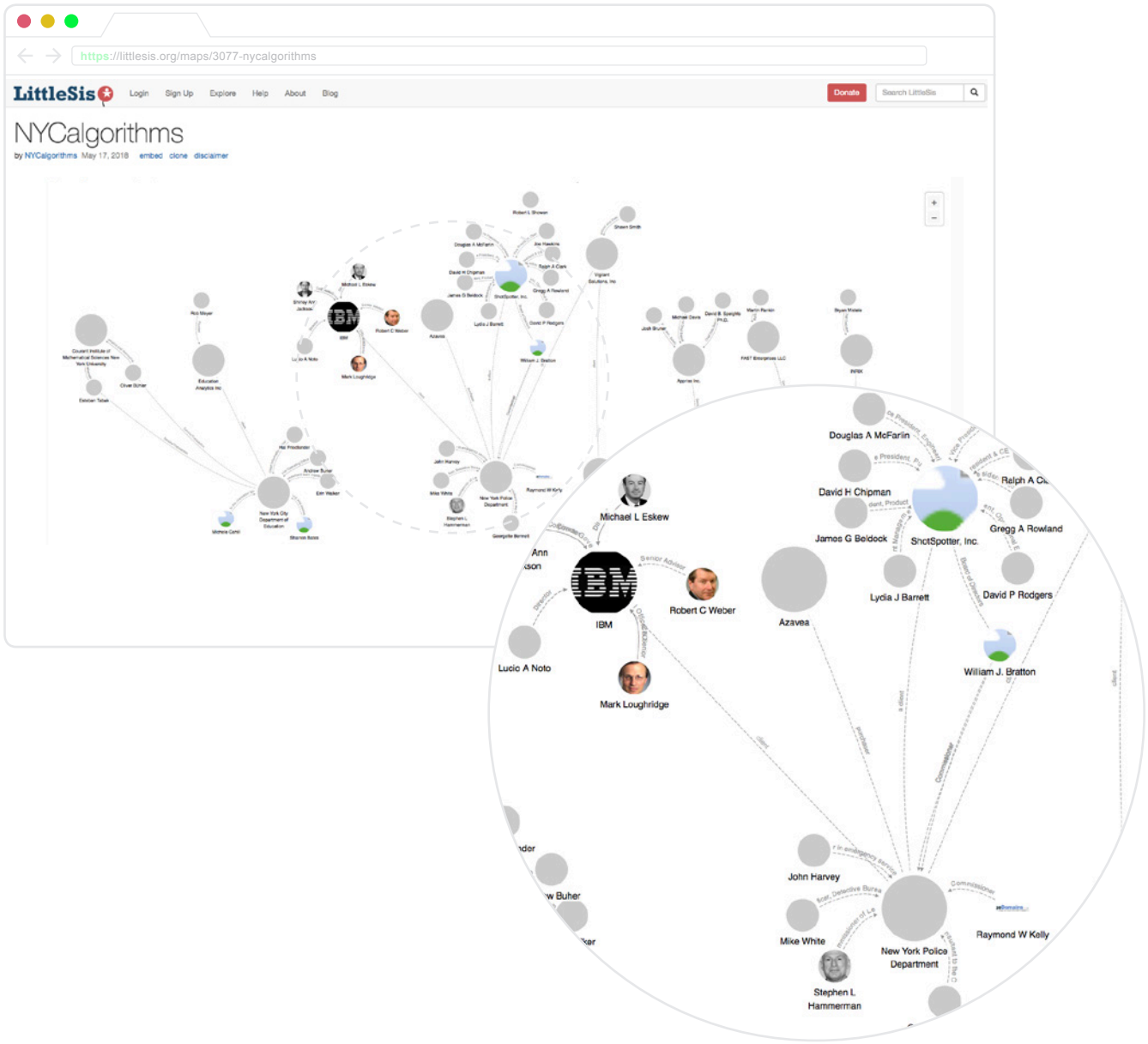


Figure 3. Little Sis screenshot detail.

ADVOCACY RESOURCES

Government Contract Provisions/Requirements

Many, if not most, automated systems come into government via procurement processes.

Figure 4. A model procurement process (NYC Mayor's Office).

These processes can serve as powerful moments to raise and address accountability concerns through the procurement contract between the government entity and automated system vendors. Below we provide recommended language for inclusion in such contracts by category of concern, focusing on contracts for automated decision systems or algorithmic systems that involve statistical modeling.

These processes can serve as powerful moments to raise and address accountability concerns through the procurement contract between the government entity and automated system vendors.

1. REQUIREMENTS AND RESTRICTIONS FOR SYSTEM DESIGN, PRODUCTION, AND CONFIGURATION:

a. Input, Training, and Testing Data

i. Before **VENDOR** uses any input or training data in the design, production, or configuration of the **SYSTEM**, **VENDOR** must provide **AGENCY** with a comprehensive list of all proposed data sets that **VENDOR** intends to use for each of these purposes.

ii. For each data set on the list, **VENDOR** shall also provide (1) a **DATASHEET**, identical or substantially similar to the one appended to this agreement as Exhibit **[TK]**; (2) an assessment of the quality of the initial data in the data set; (3) an explanation of any proposed manipulation of said data as part of the design, production, or configuration processes; and (4) records, documents, or other evidence of possible sources of bias and a proposed plan for **VENDOR** to take into account possible sources of bias in data collection, including but not limited to bias on the basis of **[race, gender, religion, neighborhood etc]**.

iii. Upon submission of the list, **AGENCY** must then review and approve each listed data set for each use before **VENDOR** may proceed with the proposed use of said data set.

iv. If, after the submission of the initial list above in 1.a.i., **VENDOR** identifies other data sets that it proposes to use for design, production, configuration, or testing purposes, it shall provide the materials specified above in 1.a.ii., and obtain the approval above in 1.a.ii., before proceeding with the proposed uses of said other data sets.

b. Methodologies

i. Before **VENDOR** applies any methodologies in the design, production, or configuration of the **SYSTEM**, **VENDOR** must provide **AGENCY** with a comprehensive list of all proposed methodologies that **VENDOR** intends to use for each of these purposes for **AGENCY** to review and approve. If, after submission of the initial list, **VENDOR** identifies other methodologies it proposes to use, it must present that methodology to and obtain approval of **AGENCY**.

c. Documentation

i. **VENDOR** shall establish and maintain records, documents, or other evidence of design decisions made, methodologies employed, and analyses conducted in the production of the **SYSTEM**.

1. REQUIREMENTS AND RESTRICTIONS FOR SYSTEM DESIGN, PRODUCTION, AND CONFIGURATION (CONTINUED)

d. Requirements for Test Version (or examination of pre-existing technology):

i. Before deployment, **VENDOR** must produce a test version of the **SYSTEM**.

ii. After production of test version, **VENDOR** must:

The purpose of the feasibility analysis is to determine whether the task of the SYSTEM is even possible: given a set of data from the AGENCY, can the VENDOR accomplish the SYSTEM's intended purpose. If the task is to accomplish some purpose without being biased, this analysis might also overlap with the validation study.

1. Conduct a feasibility analysis of said test version. Feasibility analyses will include, but are not limited to: evaluation of the predictive power of **SYSTEM**'s models, the choice of particular target variables over alternatives, the cost of archiving data used in the creation of **SYSTEM** models, the cost and feasibility of updating **SYSTEM** models with new data, **[other requirements for feasibility]**.

The purpose of this analysis is to establish that prediction uncertainty should be reflected in how the SYSTEM operates.

2. Conduct a performance analysis of said test version, including determining how to establish "confidence intervals" on **SYSTEM** prediction and how **SYSTEM** should represent confidence intervals and prediction uncertainty to **SYSTEM** operators. Performance analysis will also study potential operator use of the **SYSTEM** to understand patterns of use and how operators interpret and act on **SYSTEM** outputs.

The purpose of the validation study is to analyze the SYSTEM's use on different comparison groups, to detail predictive accuracy/error/bias.

"identifiable populations" could be enumerated to include groups that differ by race, neighborhood, gender, age, or any other protected attribute that should be analyzed within the validation study

3. Conduct a validation study of accuracy and performance of said test version on a data from one or more approved data sets that was not used in the construction of **SYSTEM** models, in which **VENDOR** shall determine predictive accuracy and the likelihood of error in reference to different identifiable populations subject to the **SYSTEM**. Different groups for analysis should include (but are not limited to) groups that differ by: **[race, neighborhood, etc]**

1. REQUIREMENTS AND RESTRICTIONS FOR SYSTEM DESIGN, PRODUCTION, AND CONFIGURATION (CONTINUED)

e. Waiver of legal claims against activities intended to audit or assess **SYSTEM** Accountability

i. In the interest of promoting **SYSTEM** accountability, **VENDOR** hereby agrees not to assert any legal claims against **[AGENCY or any AGENCY officials or agents]** or any third party for conducting research on **SYSTEM** concerning any actions intended to test, audit, examine, or otherwise understand **SYSTEM**'s effects on an individual or group of individuals impacted by **SYSTEM**'s outputs, including but not limited to concerns over bias, due process, disparate impact, or fairness. Non-limiting examples of such legal claims include claims for infringement of patent, copyright, or trademark rights, trade secret misappropriation, breach of contract, interference with business relationships, or violations of state or federal anti-hacking laws.

f. Agency Review of System Before Acceptance/Launch:

i. **VENDOR** will prepare documentation to support **AGENCY** and **VENDOR**'s presentation of its methodology to **[relevant reviewing committees]**.

ii. **VENDOR** will prepare a white paper explaining the process and results form the analyses conducted above. Such paper shall include findings of the feasibility analyses and recommendations about whether, when, and how to proceed with the development of **TECHNOLOGY**.

iii. **VENDOR** will present the white paper to **AGENCY** for comment and review.

iv. **VENDOR** will modify the model and white paper based on **AGENCY** review and discussion.

v. With **AGENCY** approval of this analysis and white paper, **VENDOR** will complete all of the requirements **[reference to technical requirements documentation]**.

2. EXPLANATION OF SYSTEM:

a. **VENDOR** will produce, under **AGENCY** direction, materials, such as training materials, for appropriate audience, an explanatory presentation for each model, and sample reports with explanations that can be used to explain the system in general to all stakeholders including **[stakeholders' names]**.

b. **VENDOR** will provide a technical manual for **SYSTEM** that provides user, design, and code documentation **[as described in technical requirements documentation]**.

3. IMPACT ASSESSMENT:

a. **VENDOR** agrees to assist and support **AGENCY** in any reasonable manner to produce and provide an Algorithmic Impact Assessment as part of **AGENCY**'s mission of public accountability.

This could include findings from the validation study required above, which calls for an analysis of the SYSTEM on different communities.

b. As part of this assistance and support, **VENDOR** will use historical data to provide a report evaluating the impact of **TECHNOLOGY** on **[specify which communities, agency processes, financials, etc.]**

4. ONGOING MODEL ASSESSMENT AND SUPPORT:

a. General: **VENDOR** agrees to cooperate with the **AGENCY** with respect to challenges to the technical methodology and analytical analyses work performed by **VENDOR** under this contract for a period of **[X]** after the challenged model is first implemented.

   i. During such period, **VENDOR** shall make available appropriate employees and produce any existing evidence or other existing documentation to help refute a challenge to the validity and reliability of such technical methodology and analytical analyses that underlie the **SYSTEM** at no cost to the **AGENCY**.

   ii. During this same such period, **VENDOR** shall also pay the costs for (i) any external expert witness testimony desired by the **VENDOR** and (ii) any expert witness reasonably required by the **AGENCY**, if **VENDOR** cannot provide an employee capable of being qualified as an expert witness by the **[relevant decision-making body]**.

b. **VENDOR** shall provide **AGENCY** notice of any claims, suits or actions related to the **SYSTEM** that might impact **AGENCY**'s use of said system. **AGENCY** reserves the right to join such an action, at its sole expense, when it determines there is an issue involving a significant public interest.

c. For claims, suits or actions instituted against the **AGENCY** of policy decisions and/or any work under or arising from this contract which is unrelated to the technical methodology and analytical analyses of the **TECHNOLOGY**, **VENDOR** agrees to cooperate and assist the **AGENCY** in defense of any such claims, suits or actions at the **AGENCY**'s expense.

4. ONGOING MODEL ASSESSMENT AND SUPPORT (CONTINUED)

d. Upon request by **AGENCY**, **VENDOR** also agrees to revalidate **SYSTEM** if the **SYSTEM**'s logic is changed or modified or if there are significant changes to populations that are the intended target of the **SYSTEM** or whose data has been used to train the **SYSTEM**.

e. **VENDOR** also agrees to disclose to **AGENCY**, in a timely manner, any evidence, analysis, or reports of known or discovered flaws in **SYSTEM**'s logic, analytical model or any data sets listed per requirements above, including but not limited to issues of bias, discrimination, disparate impact, or fairness.

5. REGULATORY COMPLIANCE/ACCOUNTABILITY:

a. **VENDOR** shall provide services and meet the program objectives summarized in **[RFP Document]** in accordance with: provisions of this **AGREEMENT**; relevant laws, rules and regulations, administrative and fiscal guidelines; and where applicable, operating certificates for facilities or licenses for an activity or program.

b. If **VENDOR** enters into subcontracts for the performance of work pursuant to this **AGREEMENT**, the **VENDOR** shall take responsibility for the acts and omissions of its subcontractors. Nothing in the subcontract shall impair the rights of the **AGENCY** under this **AGREEMENT**. No contractual relationship shall be deemed to exist between the subcontractor and the **AGENCY**.

6. OWNERSHIP:

a. Any materials, processes, and products produced for **AGENCY** pursuant to this contract, including but not limited to methodologies, measures, software, analysis, **SYSTEM** code, analysis or outputs, documentation, white papers, reports, implementation guidance, training materials, evaluation forms, data complications, and reports shall be the sole and exclusive property of the **AGENCY**. Should vendor use the services of consultants or other organizations or individuals who are not regular employees of the **VENDOR**, the subcontract agreement shall provide that any work produced pursuant to the agreement shall be the sole and exclusive property of the **AGENCY**. **VENDOR** and any subcontractors will comply promptly with any **AGENCY** request to deliver to **AGENCY** any materials or products that are in the possession of **VENDOR** or subcontractors.

7. SECURITY/CONFIDENTIALITY:

a. **VENDOR** must adhere to **AGENCY**'s security protocols regarding the storage of secure materials. **VENDOR** must also adhere to **AGENCY**'s security protocols regarding the transmission of secure materials, including use of encryption. Electronic transfer via e-mail, Internet, or facsimile (FAX) of any data capable of identifying any specific individuals or groups or of any secure test data is prohibited unless authorized by the **AGENCY** on a case-by-case basis.

b. All materials supplied by **AGENCY** are to be held strictly confidential and must not be copied, duplicated, or disseminated in any manner or discussed with anyone other than persons authorized by the **AGENCY**.

GLOSSARY

This list presents the technical terms that can come up in conversations around the use of algorithms in government. Note that many of these terms are not well-defined, even in the world of AI research. Terms like "algorithm," "model," and even "artificial intelligence" itself are often used vaguely and imprecisely. We try to reflect that imprecision below, while also giving definitions that are most useful for policy conversations. The terms are presented in an order from more general concepts and ideas to more specific terms, so that the reader can understand the terms in context.

"SUPERVISED" MACHINE LEARNING

A broad category of machine learning ("ML") in which an algorithm uses input data to learn a pattern that it can then use to predict a particular outcome (a "target value" or "label") when it sees different data. Generally, automated decision systems built with ML use "supervised" techniques. One example would be a system that predicts whether a new loan recipient is likely to default on their loan by learning from data on past loan recipients and whether or not they defaulted on a loan.

"UNSUPERVISED" MACHINE LEARNING

Machine learning that does not try to predict a "target variable" but merely learn patterns from the training data (e.g. an algorithm that clusters people together based on their features to identify social groups).

MODELS

The "rules" and relationships a machine learning algorithm learns to perform a task from training data. Those patterns might include the "weights" the algorithm learned to assign to different variables ("features") in the training data to predict some output.

TRAINING DATA

The input data used by a machine learning algorithm to find patterns. For example, when creating an ML model for use in government, a developer might use records of past administrative decisions made by a government agency to "train" the system on how to make future decisions. For machine learning that examines images or videos, training data will include those images or videos themselves. Training data is generally made up of "features" (variables) and a "target variable" or "training label" that a machine learning model will attempt to predict.

- **Structured training data** is training data that follows an easily machine-readable format. For example, the excel spreadsheet of New York City's Stop and Frisk database is in a structured format.
- **Unstructured training data** is training data that machines cannot readily parse and understand. Images, videos, emails, and freeform text documents like emails are some of the most common examples of unstructured data.

FEATURES

The different variables machine learning algorithms use to predict outcomes. Features can be things like "a defendant's number of prior arrests," "age," "credit score," or anything that can be measured and put into a dataset.

LABELS OR TARGET VARIABLES

The outcomes that machine learning algorithms attempt to predict. These can either be numerical (e.g. predicting stock prices based on historic data) or categorical (e.g. predicting whether or not a defendant will be fail to appear in court). In the case of supervised machine learning, the training data will have "labeled data" so that the algorithm can learn the relationship between the data and its label.

MODEL PARAMETERS

What many machine learning algorithms learn so that they can make predictions. "**Weights**," a term that has seen use in policy contexts, can also be used to refer to the values learned by machine learning algorithms. In simple machine learning algorithms like "linear regression," they represent the relative importance of each feature in a dataset for deciding an outcome. For instance, in predicting recidivism, an ML algorithm might learn to assign the value "-3" to age and "5" to number of past arrests to assign the relative importance of those features in deciding "riskiness." In practice, the parameters learned by an algorithm can be so numerous and the relationships between them so complex that they make ML models hard to understand by a human: there will not be a direct relationship between the "weight" learned by the algorithm and the corresponding feature's relative importance. This is especially the case in "neural networks" and algorithms described as "deep learning."

DEEP LEARNING

A class of machine learning algorithms that perform their tasks by building abstractions of the input data it analyzes. The most popular deep learning methods use "neural networks," machine learning algorithm architectures that learn potentially hundreds of millions of parameters. These are most commonly applied to images, since neural networks can learn how to "understand" the different features of an image without a human explicitly telling the network how to extract meaning from the image. The models created through deep learning are notoriously difficult to understand by a human.

EXPLAINABILITY

A property of a machine learning model that makes it easy for a human to understand why an ML model makes one prediction and not another. There is an active debate among policymakers and researchers about what sort of explanations are most valuable, and whether relying on explanations as a form of transparency places too big of a burden on individual citizens to hold government systems accountable themselves. Some explanations might not get at the underlying cause of the pattern and thus may not have much value (e.g. a model that explains that it denied someone a loan because they were born on a Tuesday is not very useful). A relatively new subfield of computer science research on "**interpretability**" studies how to make models and their decisions more understandable and useful.

VALIDATION STUDY

Research that studies the accuracy, efficiency, or other properties of an automated decision system. Validation studies generally look at how well an automated decision system is at performing its task. A more comprehensive validation study can also analyze how well an automated decision system performs its task on different populations, studying, for example, if it is possible for the system to produce disparate impacts.

RISK ASSESSMENTS

Particular types of automated decision systems that evaluate people and assign them a risk score for some definition of "risk." For example, risk assessments might assign criminal defendants a risk score representing how likely they are to fail to appear in court, using a machine learning model trained on past defendants and whether or not they failed to appear. The most well-studied types of risk assessments are in criminal justice, but risk assessments are also used in child welfare, homelessness services, and other domains, and are constantly being deployed into more contexts. Risk assessments have not always used machine learning: actuarial risk assessments have existed since before the current explosion of machine learning, and are based on more traditional statistical means of assessing risk. Actuarial risk assessments raise similar concerns to risk assessments built on machine learning.

SOURCE CODE

The instructions written by a human to create a computer program.

- **Open source software** are computer programs for which the source code is publicly available.
- **Closed source software** are computer programs for which the source code is not public.

PREDICTIVE ANALYTICS

The use of statistics to make predictions about certain events or outcomes. Machine learning is a form of predictive analytics, but predictive analytics might also include analytical techniques that are not based on learning predictions from training data. "Predictive analytics" and "machine learning" are often conflated and used interchangeably.

EXPERT SYSTEMS

Automated decision systems that help make decisions using rules explicitly created by a human. Contrast this with a system built using machine learning: with ML, decision rules are automatically learned by a machine learning algorithm. An example of an expert system would be a medical diagnosis system in which the rules for diagnosing disease were explicitly decided by a doctor (e.g. "if a patient has a temperature over 100 degrees and a blood pressure over 140/90, the patient has a fever"). Expert systems are historically considered artificial intelligence.

ACKNOWLEDGEMENTS

Thank you to **Esha Bhandari**, **Rachel Goodman**, **Kade Crockford**, **Nicole Triplett**, **Robert Brauneis**, **Olivia Zhu**, **William Isaac**, **Timnit Gebru**, and **Cassie Deskus** for helpful feedback and contributions while developing this toolkit.

AINOWINSTITUTE.ORG