

Public Scrutiny of Automated Decisions:

Early Lessons and Emerging Methods

An Upturn and Omidyar Network Report



Acknowledgments

We would like to acknowledge the many people who gave us helpful insight and feedback during the course of our research, including Salmana Ahmed at Omidyar Network, Gemma Galdon Clavell at Eticas Research & Consulting, Luis Fernando Garcia at R3D, Luciano Floridi at the University of Oxford, John Havens at IEEE, Gus Hosein at Privacy International, Estelle Massé at Access Now, Valeria Milanés at ADC Digital, Vivian Ng at the Human Rights, Big Data and Technology Project at the Essex Human Rights Centre, Cathy O’Neil at O’Neil Risk Consulting & Algorithmic Auditing, Matthew Sheret at IF, Matthias Spielkamp at AlgorithmWatch, Martin Tisné at Omidyar Network, Frank Pasquale at the University of Maryland, Pablo Viollier at Derechos Digitales and Adrian Weller at the Leverhulme Centre for the Future of Intelligence.

Authors

Aaron Rieke is a Managing Director at Upturn. He holds a JD from Berkeley Law, with a Certificate of Law and Technology, and a BA in Philosophy from Pacific Lutheran University.

Miranda Bogen is a Policy Analyst at Upturn. She holds a Master’s degree in Law and Diplomacy with a focus on international technology policy from The Fletcher School of Law and Diplomacy at Tufts, and bachelor’s degrees in Political Science and Middle Eastern & North African Studies from UCLA.

David G. Robinson is a Managing Director and co-founder at Upturn. He holds a JD from Yale Law School, and bachelor’s degrees in philosophy from Princeton and Oxford, where he was a Rhodes Scholar. David also serves as an Adjunct Professor of Law at Georgetown University Law Center, where he teaches a seminar on Governing Automated Decisions.

Contributor

Martin Tisné is an Investment Partner at Omidyar Network where he leads policy, advocacy strategy, and related investments for the firm’s global Governance & Citizen Engagement initiative. He holds a B.A. from the University of Oxford and an M.Sc. from the London School of Economics.

Foreword

Over recent years, we have witnessed an exponential growth in the use of automation to power decisions that impact our lives and societies. While these bring many potential benefits (e.g. reducing human bias, speeding up decision-making) they are also fraught with risk (e.g. increasing machine bias, rendering the decision-making more opaque and remote to people). A consideration for the fairness, accountability, and transparency of these processes is therefore fundamental.

Omidyar Network's Governance & Citizen Engagement initiative started funding and advocating for open data in 2012. Following the Snowden revelations in the summer of 2013, we included privacy and surveillance reform as part of those efforts, realizing that openness and privacy were closely related. We focused on helping develop a framework governing the collection of, access to, storage and usage of, and rights over data.

More recently, we started looking at the impact of these data releases under a new light. As both processing power and access to data increased, decisions made by both government and corporate sector actors were becoming increasingly automated. If governments use algorithms to screen immigrants and allocate social services, it is vital that we know how to interrogate and hold these systems accountable.

Our main aim is to increase individuals' control and agency over the decisions that impact their lives and ultimately societies. We believe that the accountability and transparency agenda would benefit from better understanding, and ultimately scrutinizing, automated decision-making.

To explore some of these questions we partnered with Upturn to map out the landscape of public scrutiny of automated decision-making, both in terms of what civil society was or was not doing in this nascent sector and what laws/regulations were or were not in place to help regulate it. As the study developed it became clear that it could pave the way for a more textured, as well as practical, understanding of algorithmic transparency. Omidyar Network will explore some of these issues further in a companion paper to be released later this year.

Our hope is that this report will help civil society actors consider how much they have to gain in empowering the public and their audiences to effectively scrutinize, understand, and help govern automated decisions. We also hope that it will start laying a policy framework for this governance, adding to the growing and rich literature on the social and economic impact of such decisions. Finally, we hope that the report's findings and analysis will help inform other funders' decisions in this important and growing field, as it will our own.

Martin Tisné

*Investment Partner, Governance & Citizen Engagement
Omidyar Network*

Our hope is that this report will help civil society actors consider how much they have to gain in empowering the public and their audiences to effectively scrutinize, understand, and help govern automated decisions.

Table of Contents

Executive Summary	5
Introduction	7
Unpacking “Automated Decisions”	9
Insights	10
Existence	10
Purpose	10
Constitution	10
Impact	11
Artifacts	11
Policies	11
Inputs and Outputs	11
Training Data	12
Source Code	13
Scrutiny in Practice	15
Journalism	15
Qualitative Research	16
Legal Process	16
Black Box Testing	17
Examination of Training Data	18
Code Review	19
Designing for Accountability	20
Applying Metrics for Fairness	20
Interpretability	21
Procedural Regularity and Audit Trails	22
Perspectives on the Role of Law	23
Restrictions on Data Collection	23
Transparency	24
Explanations	24
Antidiscrimination	25
Data Access, Accuracy, and Redress	26
Opt-outs and Forbearance	26
Validation and Certification	27
Auditability	28
Competition	28
Product Liability	29
Conclusion and Paths Forward	30
Appendix: Examples of Public Scrutiny	31

Executive Summary

Automated decisions increasingly mediate civic life. Governments use algorithms to screen immigrants and allocate social services. Corporations rely on software to help make decisions in vital areas like hiring, credit, and political discourse. Trends like these have made algorithms objects of concern beyond the confines of computer science.

Advocates, policymakers, and technologists have begun demanding that these automated decisions be explained, justified, and audited. There is a growing desire to “open the black box” of complex algorithms and hold the institutions using them accountable. But across the globe, civil society faces a range of challenges as they pursue these goals.

This paper considers how the public can effectively scrutinize, understand, and govern automated decisions, based on an extensive review of computer and social science literature as well as dozens of semi-structured interviews and conversations with global digital rights advocates, regulators, technologists, and industry representatives. We also surveyed a broad array of real-world attempts to scrutinize automated systems, documenting the purpose of each inquiry, its methods, and its findings.

To encourage clear discussion, we offer a conceptual framework to describe how different elements of automated systems work together to produce a result. In addition to its source code, knowledge of a system’s existence, purpose, impact, constitution, policies, inputs and outputs, and training data can be helpful to both scrutinize and govern these systems. This framework highlights how non-technical insights about an automated system can be just as important, and often more important, than its technical, tangible artifacts.

We catalog a variety of both analog and technical approaches to scrutiny that we encountered during our research, and analyze each: journalism, qualitative research, legal process, black box testing, examination of training data, and code review. We also discuss emerging methods to design systems for accountability, including applying fairness metrics, designing for interpretability, cryptographically ensuring procedural regularity, and producing audit trails.

We consider these methods alongside common legal and regulatory approaches. Based on this appraisal, we highlight promising areas for progress and opportunities for further research, policy discussion, and advocacy.

Our key findings are:

- **Today’s automated decisions are socio-technical in nature: They emerge from a mix of human judgment, conventional software, and statistical models.** The non-technical properties of these systems — for example, their purpose and constraining policies — are just as important, and often more important, than their technical particulars. Automated systems vary in their goals and design, and demand different kinds of inquiry.
- **Scrutiny doesn’t have to be sophisticated to be successful.** Many of the most notable case studies we identified involved investigative reporting and basic observation of a system’s purpose, policies, inputs and outputs. Such approaches have led to productive public attention. However, more technically sophisticated types of scrutiny are beginning to bear fruit, especially in the realm of “black box testing.”
- **There are promising new methods for designing more accountable systems, but these remain largely theoretical.** Researchers are working hard on new ways to detect bias in datasets, to design predictive models

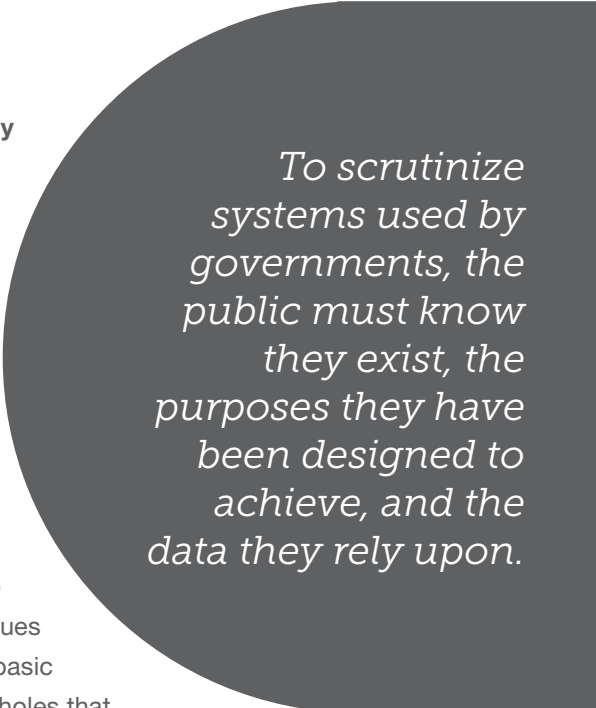
How can the public effectively scrutinize, understand, and govern automated decisions?

that are interpretable, and to verify the behavior of important software. However, these techniques — many of which require proactive cooperation from institutions of interest — are likely to remain in the lab until civil society makes a clearer case for their adoption.

- **Many existing legal and regulatory frameworks remain relevant to automated systems' behavior and output, but their applicability is often unclear or untested.** Some laws have been recently updated to specifically address automated decisions, but they remain largely untested. Others may require updating to remain effective in the era of widespread automation.

In conclusion, we highlight the following promising paths forward:

- **Increased investment in what we term exploratory scrutiny, especially by journalists and advocacy organizations.** Many of today's highest-impact efforts have provided just enough insight to spark important debates in domains like policing and credit. To engage a wider audience in debates about how automated systems should function, the field needs more work to find evidence about and clearly explain how important systems work in practice. This work can help build the case for new policies and technical requirements.
- **Strategic evaluation of right to information laws' ability to facilitate effective transparency for today's automated decisions.** **To scrutinize systems used by governments, the public must know they exist, the purposes they have been designed to achieve, and the data they rely upon.** Once those basic details are available, other techniques for scrutiny and ensuring integrity will be useful — but all too often, these basic facts remain obscured, and most transparency laws have concerning loopholes that prevent this and other relevant information from being shared with the public.
- **Consideration of policy mandates that certain automated systems be auditable and interpretable.** New technical methods are making it more feasible to design automated systems that are more amenable to scrutiny, but these methods may not be adopted without outside pressures. For public-sector automated systems, this could mean requirements that systems be designed to meet auditability requirements. For private-sector systems, this could mean "interpretability" requirements for important automated systems, such as those used in the context of employment and credit.
- **Further advancement of normative dialogues.** Many new policies and technical proposals presuppose standards and benchmarks that do not yet exist. Policymakers and the public must think more concretely about what "fairness" and "accountability" ought to mean in particular social contexts.



To scrutinize systems used by governments, the public must know they exist, the purposes they have been designed to achieve, and the data they rely upon.

Introduction

Automated decisions increasingly mediate civic life. Governments use algorithms to screen immigrants and allocate social services. Corporations rely on software to help make decisions in vital areas like hiring, credit, and political discourse. Algorithms also influence the production and dissemination of news, social and professional interactions among people, the delivery of social services, and the stability of financial markets. These trends have made algorithms objects of concern beyond the confines of computer science.

In some cases, automating such decisions can promote efficiency, consistency, and fairness. But in others, doing so can reinforce historical discrimination or obscure undesirable behavior.

Government bodies in the US and the EU have recently expressed concern about automated decisions, asking questions, issuing reports, and even updating relevant laws. The Obama White House published several reports on automated systems and social issues,¹ the UK Parliament recently asked for input on use and regulatory oversight of decision-making algorithms,² and the United Nations has engaged in ongoing conversations about autonomous weapons.³ The EU has updated its data protection laws, aiming in part to ensure that certain automated systems be “explainable” to data subjects.⁴ France has even moved to classify software source code used by government agencies as a public record subject to transparency laws.

At the same time, a striking number of research groups, standards bodies, and private companies have announced proactive efforts to ensure that automated decisions are “accountable.” An international, interdisciplinary collection of technologists formed a research community called Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) and published a normative set of “Principles for Accountable Algorithms.”⁵ The US branch of the Association for Computing Machinery, a major technical professional organization, released slightly more applied guidance for system developers called “Principles for Algorithmic Transparency and Accountability,”⁶ and a collection of major technology companies including Facebook, Google, IBM, and Amazon formed the Partnership on AI, grounded in a set of tenets focused on ensuring automated systems serve to benefit people.⁷ Formal industry groups have released policy principles on artificial intelligence,⁸ standard-setting bodies like the IEEE have begun the process of defining ethical concerns and technical standards related to autonomous systems,⁹ and individual countries like Japan are pushing for international conversations around shared guidelines for AI research and development.¹⁰

However, despite all these efforts, the use of automated decisions is far outpacing the evolution of frameworks to understand and govern them.

This paper considers how the public can effectively scrutinize, understand, and govern automated decisions.

We leave to other projects the question of which automated decisions are most likely to help or harm, and which deserve closer scrutiny. Here, we ask instead: What practical approaches and legal tools are available to understand and shape how algorithms work in the real world, so that benefit can be maximized, and harm be identified and mitigated?

Automating decisions can promote efficiency, consistency, and fairness, but can also reinforce historical discrimination or obscure undesirable behavior.

To answer this question, we conducted an extensive review of computer and social science literature as well as dozens of semi-structured interviews and conversations with global digital rights advocates, regulators, technologists, and industry representatives. We also surveyed a broad array of real-world attempts to scrutinize automated systems, documenting the purpose of each inquiry, its methods, and its findings to uncover lessons for successful scrutiny of consequential automated systems.

The use of automated decisions is far outpacing the evolution of frameworks to understand and govern them.

This paper proceeds as follows:

Unpacking “Automated Decisions”

First, we offer a conceptual map for thinking about automated decisions to help the reader think about how their different elements work in tandem to produce important real-world outcomes. We emphasize that non-technical insights about an automated system can be just as important, and often more important, than its technical artifacts.

Scrutiny in Practice

Second, we describe ways public actors have successfully scrutinized existing systems, analyzing case studies from around the globe. We observe that much of the most successful scrutiny to date has been relatively non-technical in nature — e.g., relying on investigative reporting and simple observation of a system’s inputs and outputs. However, more technically sophisticated types of scrutiny are beginning to bear fruit.

Designing for Accountability

Third, we explore new ways that automated systems can be proactively designed to be fairer, more interpretable, and more auditable. These techniques are exciting, but remain largely theoretical today.

Perspectives on The Role of Law

Fourth, we review common approaches to regulating automated decisions, incorporating views from global digital rights advocates. We summarize different themes and emphases from major jurisdictions across the globe.

Conclusion and Paths Forward

In closing, we suggest some promising paths forward, including increased “exploratory scrutiny,” a close look at right to information laws’ ability to provide needed transparency, and tech-neutral policy mandates for socially important automated systems.

Unpacking “Automated Decisions”

Automated decisions are decisions made with the aid of systems that limit human judgment. They precede the advent of the digital computer: Complex and opaque rules have long guided the decisions of corporate and governmental institutions. However, modern technologies have the potential to make automated decisions even more important and less scrutable than ever before.¹¹

Today’s discourse on automated decisions is rife with technology buzzwords.

Algorithms — specific sequences of steps used to accomplish some task, especially those embedded in a computer — have captured the public imagination. **Machine learning** — a family of techniques that allow computers to learn directly from examples, data, and experience, finding rules or patterns that a human programmer did not explicitly specify — is allowing for new kinds of data-based automation.¹² And **artificial intelligence**, an umbrella term for a range of computer systems that act in seemingly intelligent ways, is gaining popularity in newspapers and corporate marketing pitches.

These technologies are transforming how important decisions are made. But today’s automated decisions are not defined by algorithms alone.¹³ Rather, they emerge from *automated systems* that mix human judgment, conventional software, and statistical models, all designed to serve human goals and purposes.¹⁴ Discerning and debating the social impact of these systems requires a holistic approach.

To encourage clear discussion, we offer the following conceptual framework. Each of the “elements” described below is a different way to evaluate an automated system. Some of these elements, which we call “insights,” can yield important information without reference to any technical specifics of the system. The remaining elements, which we call “artifacts,” refer to tangible parts of a system that are often more technical in nature. In a later section, we discuss how these different properties have been scrutinized by journalists, advocates, and policymakers.

Automated decisions are decisions made with the aid of systems that limit human judgment.

INSIGHTS



EXISTENCE



PURPOSE



CONSTITUTION



IMPACT

ARTIFACTS



POLICIES



INPUTS AND OUTPUTS



TRAINING DATA



SOURCE CODE

Insights

Existence

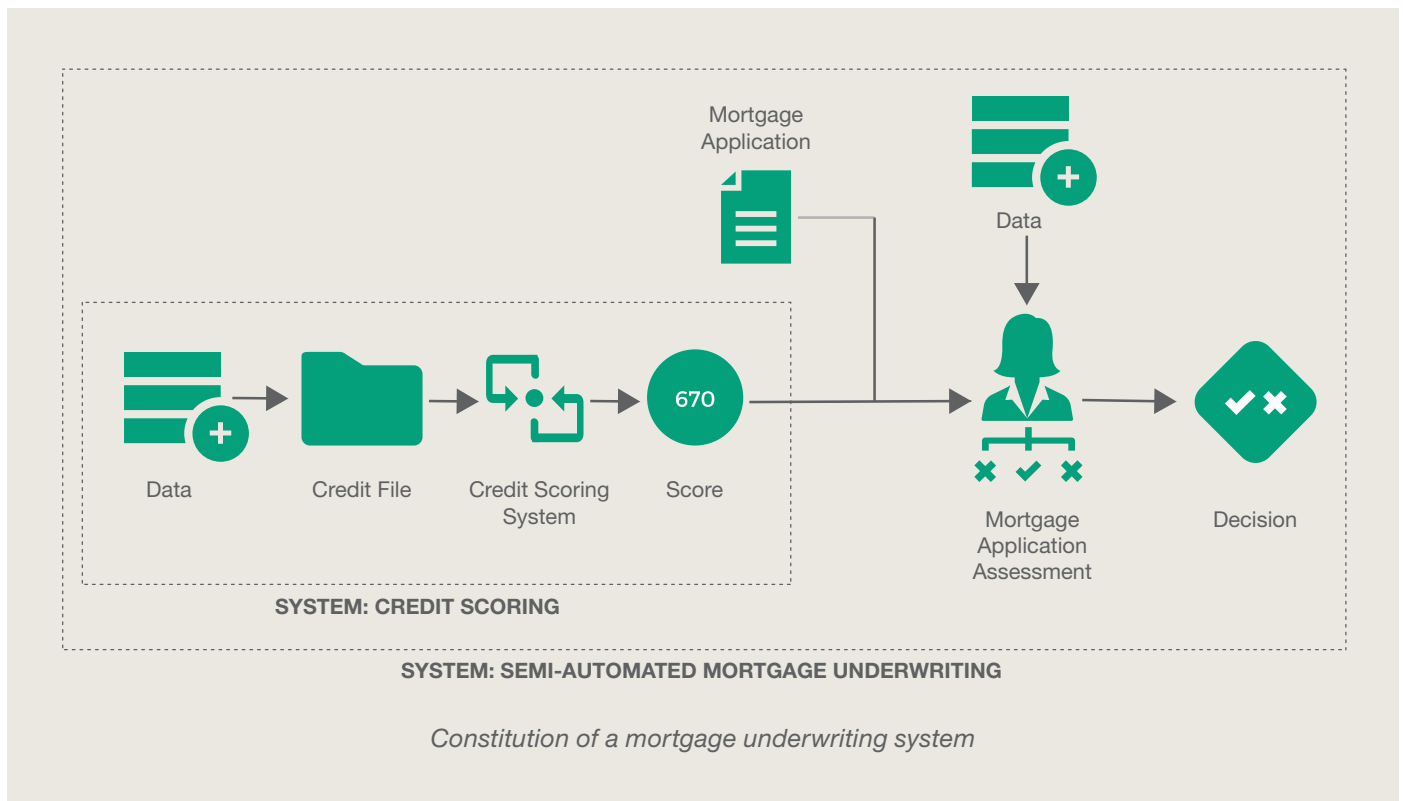
Any analysis of an automated system must begin with knowledge that it **exists**. This basic point is worth emphasizing because automated systems often operate invisibly, behind the scenes. For example, in the United States, certain people were prohibited from boarding commercial aircraft long before the discovery that a “no fly list” — a partly computer-generated list produced with hidden processes and rules — even existed.¹⁵ Likewise, Colombia’s government has announced its intent to leverage big data for unspecified projects, but advocates are struggling to learn what systems are being tested.¹⁶

Purpose

All automated systems are created to serve some **purpose**. Understanding a system’s intended purpose creates the opportunity to debate that system’s role in society, even without more specific details about how it operates. For example, most credit scores are designed to predict the relative likelihood of a negative financial event, such as default on a credit obligation — not to represent a judgment of a person’s general responsibility or character. When employers began using these scores to evaluate job applicants, understanding the actual purpose of the credit scoring system allowed advocates to develop strong arguments against its use in the context of hiring.¹⁷

Constitution

Automated systems can rely on different mixtures of computer software, human discretion, and policies.¹⁸ In practice, “automated” decisions are often only automated partially, to varying degrees.¹⁹ Understanding a system’s **constitution** — the nature of its technical elements, human participation, governing rules, and how they all interact — is critical to guide and inform more detailed inquiry. For example, understanding a social media platform’s content moderation practices requires understanding the role of human and software-driven enforcement, and the policies governing each. Understanding a lending decision requires understanding that credit scores are generated by software programs, and that human analysts review those numbers as part of a final determination.



Constitution is also important because human beings interact with computerized systems in surprising and sometimes unpredictable ways. “Automation bias” is the tendency of people to give too much credence to automated suggestions and to fail to overrule them when necessary.²⁰ But in some situations, people opt to ignore computerized direction, especially to pursue their own self-interest (for example, giving a loan to someone who clearly cannot afford it).²¹ And some research has shown that even when an algorithm performs measurably better at a predictive task than a human does, people are more likely to prefer the human prediction.²² Mapping the constitution of a system can help reveal which, if any, of these competing tendencies are relevant to systems’ outcomes — and determine which other components of a system require closer examination.



Impact

Any automated system of social concern will have some sort of observable consequence, whether for a single person or an entire population. These impacts can be experienced personally, studied anecdotally, or measured quantitatively. For example, an individual might report being turned away at a border crossing, a pre-deployment impact assessment might try to predict broad effects of a system, or a post-hoc investigation could seek to compare how different groups were treated at borders across a jurisdiction. Each instance provides insight, to varying degrees, into how the automated system impacts people in practice, allowing examiners to further focus their scrutiny.

Artifacts



Policies

Even systems that rely heavily on computers can be constrained by policies that govern how both technical and human components of that system should behave. For example, Google and Facebook provide marketers with powerful, highly automated tools with which to target advertisements. But both companies proscribe certain types of harmful advertisements — the products of human judgment — and enforce these prohibitions in different ways, including both human and algorithmic review of ad content and targeting parameters (with varying degrees of success).²³ In another domain, police officers must be *instructed* how to respond to color-coded “threat scores” that are automatically calculated and assigned to locations or residents by predictive policing systems.²⁴ In both cases, policies (and how well those policies are enforced) will play an important role in how the systems affect people in the real world.



Inputs and Outputs

Regardless of their constitution and complexity, all automated systems take some sort of *input* and produce some sort of *output*. For example, most facial recognition systems take a digital image as an input, and produce an output of similar photos from existing data, ranked by probability of a match. Search engines take a user-provided search query as an input, and output search results and advertisements related to that query. Many credit scores take as inputs a range of data about a person’s finances, and produce a number estimating their relative risk of default on a financial obligation.

Sometimes, all a system’s important inputs and outputs are observable and even controllable by outside users (as is the case with many public-facing web-based services). However, in other cases, inputs and outputs are only partially observable. This is particularly true of more complex automated systems with intermediary steps. For example, a job applicant may send her applications materials (an input) to an employer and ultimately receive a rejection (an output). However, the employer could have used intermediary scores and automated analyses that are not at all visible to the applicant or public to reach its conclusion.

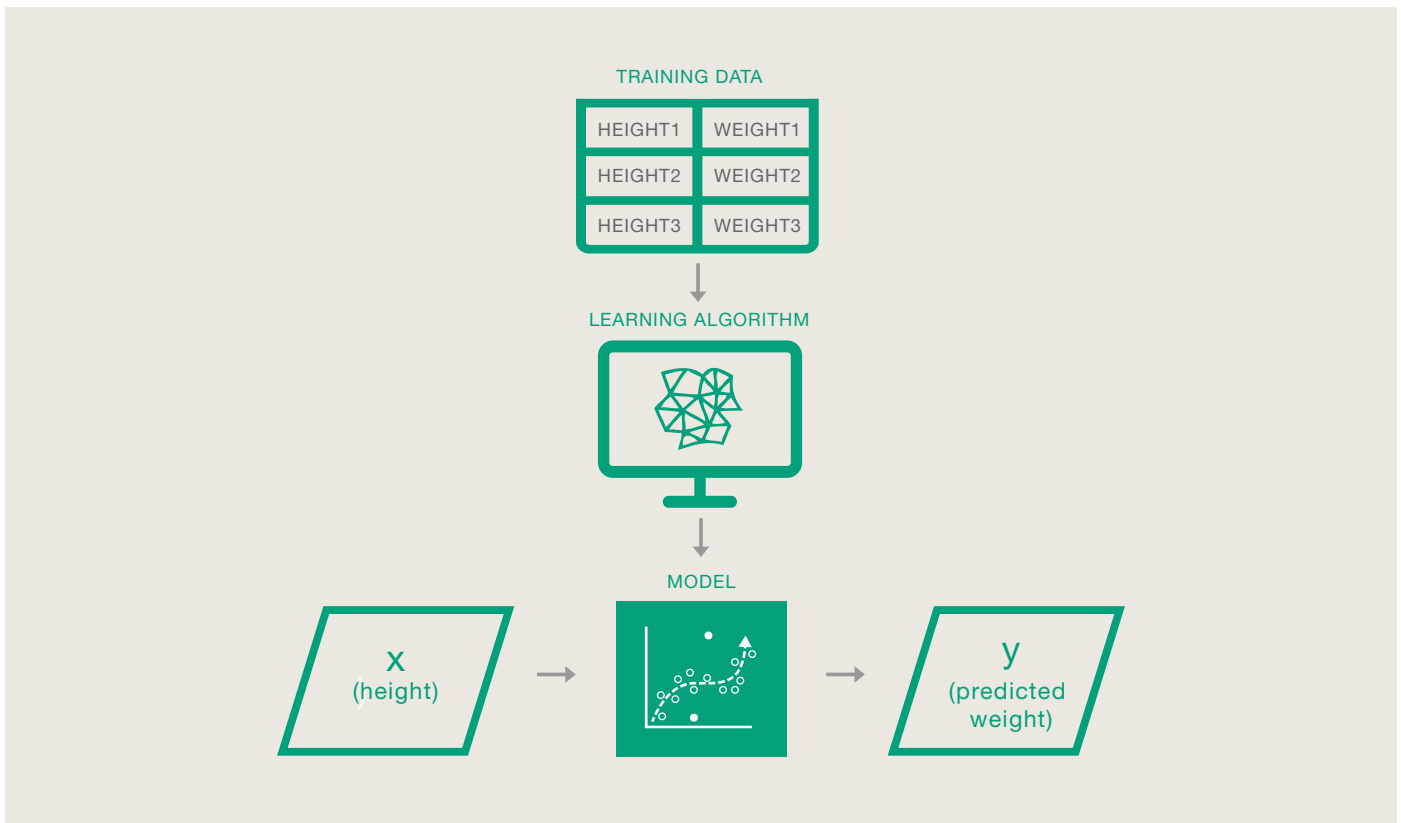
Even systems that rely heavily on computers can be constrained by policies that govern how both technical and human components of that system should behave.



Training Data

Training data refers to a set of historical data used to discover potentially predictive relationships among that data, which end up being represented in a model (commonly referred to as an algorithm). Training data is at the heart of statistical analysis and modern machine learning. Most machine learning algorithms deployed in today’s automated systems are examples of supervised machine learning,²⁵ where the computer learns how to sort example inputs into output categories previously defined by a human being — such as “creditworthy,” or “photo of a puppy.” Such a machine learning algorithm takes a collection of input features (X) and “output” or target variables (Y), and produces a function that maps X to Y (this function is sometimes called a predictive model).

Take, for example, the simple relationship between people’s height and their weight. Shorter people are more likely to weigh less than taller people, and that relationship can be modeled. That model can then be used to predict a person’s weight given their height. The process can be visualized as follows:



A predictive model can only be as good as its training data. The types of data chosen to train a model, and the quality of that data, are some of the most important properties of an automated system that relies on prediction. Inaccurate, incomplete, or irrelevant data will lead to poor results, no matter how sophisticated the mathematical algorithm used to learn from it. Moreover, the machine learning literature is rife with examples of spurious correlations. (For example, Google initially claimed search queries could predict the spread of flu pandemics, but an initial, strong correlation later proved unreliable.²⁶)

Using large amounts of data to make predictions invites some unique hazards, especially in social contexts. When automated pattern-finding is based on historical data, it risks bringing social patterns to the present that were once the norm, but are no longer socially acceptable (e.g., racial discrimination).²⁷ This is a particular concern for predictive policing, courtroom risk assessment, and credit scoring systems, where relevant historical data reflects the biases of society.

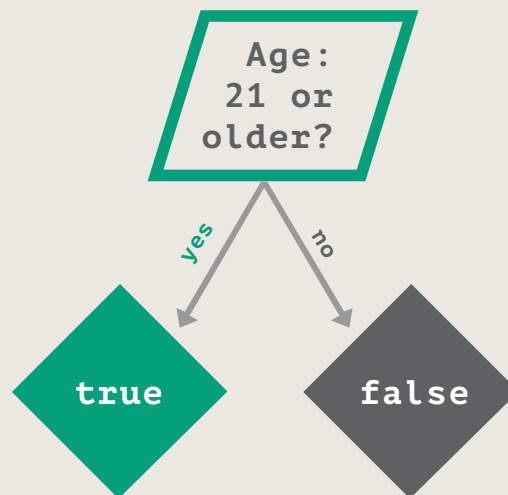
</> Source Code

Source code refers to technical descriptions of software or predictive models within an automated system. In some cases, source code can provide a reliable, objective description of what a computer system is designed to do. But source code can be complex and difficult even for experts to diagnose and understand.

The source code of traditional software will look and behave much differently from the “source code” of predictive models. *Traditional software* refers to computer programs that are structured and defined by human programmers. At a basic level, programmers write procedures — descriptions of rules²⁸ — that react to or manipulate data.²⁹ Conventional software programs are often composed by hand using logical statements (e.g., “if x then y”). For example, the code for a software program designed to check if a person’s age meets a threshold might read as follows:

Source code refers to technical descriptions of software or predictive models within an automated system.

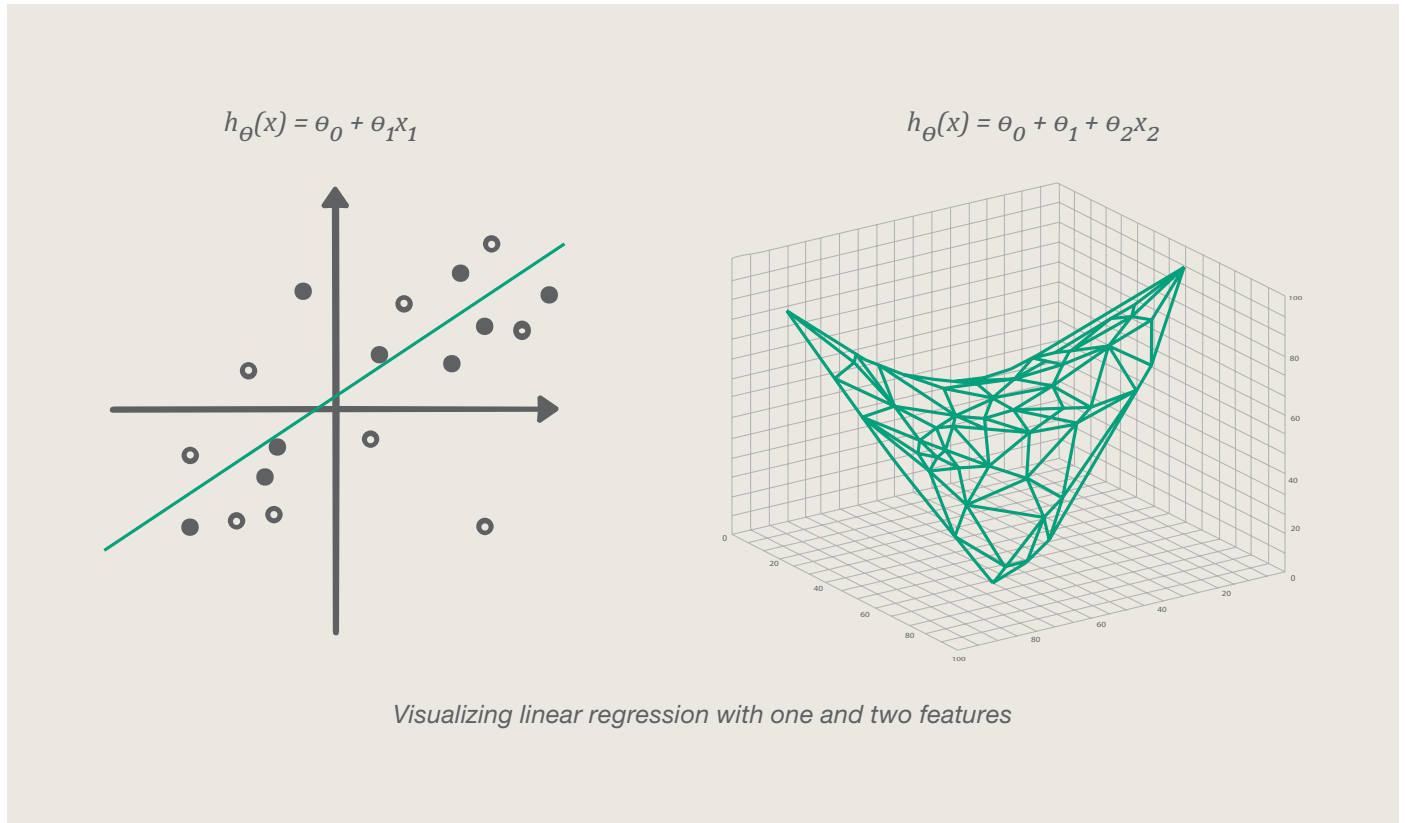
```
if (person.age >= 21) {  
    return true;  
} else {  
    return false;  
}
```



Simple conditional code and a visual representation

Most software programs are much more complex than this example, and thus harder to reason about.³⁰ Real-world programs can involve many thousands of lines of code, collaboratively authored by many different programmers. These programs are likely to rely on a significant number of “dependencies”: other software programs, often written by other people or organizations, that must function in an expected way for the current program to run correctly.³¹ As a result, without care, many software programs become “big balls of mud” that are difficult to understand or modify.³²

Predictive models tend to be different. They don't take the form of declarative steps, but instead express a statistical relationship between different input and output variables. For example, the "code" for a simple predictive model that approximates an output variable as a linear function is likely to be expressed and visualized like this:



The patterns represented in predictive models can quickly become too complex for humans to understand. In practice, it is very hard for most people to think about models with more than two features (visualized in three dimensions above), much less those with thousands of features. Interpreting and constraining machine learning and statistical techniques is a burgeoning area of research, as discussed in more detail later in this report.

This section has described how automated systems can be analyzed piece by piece. They are not a replacement for studying a system's outcomes or impact on society, but can offer a path toward understanding why a consequential system behaves the way it does. In the next section, we describe how these elements can be a useful lens to approach public scrutiny in practice.

Effective public scrutiny of automated systems takes many different forms, and implicates many of the different properties discussed above. Journalists investigate these systems by digging up documents and interviewing inside sources. Researchers probe from the outside, carefully selecting inputs and observing outputs. Data scientists can make inferences by using advanced mathematical techniques, and computer scientists evaluate systems' computer code, looking for errors. This section describes common methods of inquiry and places recent attempts to scrutinize automated systems in the context of the automated system "stack."

Journalism

In recent years, journalists have led the way in highlighting the societal implications of automated decisions.³³ Through interviews, right to information requests, and investigative reporting, journalists can uncover the **existence** of important systems, as well as, most commonly, their **purposes, constitution, and policies**. And journalists' work product — clear, compelling writing — can communicate findings to readers about what can otherwise seem to be dry, technical topics.

Some journalists have begun to use more computationally intensive methods (like "black box testing," discussed below) to augment their reporting. A small but active field of "algorithmic accountability journalism" has grown out of the more established field of "data journalism."³⁴ These approaches are valuable tools for public scrutiny of automated systems, and have been critical in uncovering and drawing attention to systems that affect people's lives.

CASE STUDIES

- The Computational Journalism Lab at the University of Maryland recently began maintaining a list of "potentially newsworthy algorithms used by the U.S. government."³⁵ Contributors scour government websites for a set of relevant terms, compile basic details about systems identified, and evaluate newsworthiness by trying to approximate a system's potential for harm, surprise, and controversy. The project hopes that highlighting the **existence** of these systems will inspire further investigation and research.
- Journalists from UK publication The Guardian obtained and reviewed more than a hundred leaked training manuals, charts, and spreadsheets describing Facebook's content moderation policies.³⁶ These documents shed light on not only the platform's **policies**, but also the **constitution** of Facebook's ranking system, a hybrid of algorithmic sorting and filtering and manual moderation.
- A New York Times investigation revealed that ridesharing company Uber used an algorithm to flag and evade regulators in cities all over the world. Journalists learned about the algorithm's **existence** and **purpose** by speaking with current and former Uber employees and reviewing documents these sources provided.³⁷ The Times' investigation led to broad media coverage and a Department of Justice inquiry into potential criminal behavior by the company.³⁸
- Students in a media law class at the Philip Merrill College of Journalism at the University of Maryland submitted Freedom of Information Act requests to all 50 US states, asking for "documents, mathematical descriptions, data, validation assessments, contracts and source code related to algorithms used in criminal justice."³⁹ Most requests were denied, but the class did receive **source code** for a recidivism prediction algorithm from one state, and information about the **existence, purpose, and constitution** of systems in several others.

Several such journalistic investigations have sparked broad public conversations about the acceptability and performance of key automated systems, suggesting that even exploratory, non-technical scrutiny can be a critical catalyst for further interrogation — while simultaneously driving forward important normative debates.

Qualitative Research

Researchers also use more formal qualitative methodologies, like ethnography, to study automated systems and their use, paying close attention to the **purpose**, **constitution**, and **policies** of the systems they are studying.⁴⁰ These researchers might document the human processes and assumptions related to a particular system, interview people who build and interact with different parts of that system, and observe how people at the receiving end of an automated decision incorporate that guidance into their own decisions and behavior.⁴¹

CASE STUDIES

- Argentinian civil society group Via Libre investigated the Buenos Aires Department of Education’s online enrollment system, which used an automated process to allocate slots in public schools.⁴² The organization used interviews, desk research, and media analysis to describe the **purpose**, **constitution**, and **policies** of the system. Researchers also tried — albeit unsuccessfully — to gain access to the **source code** of the slot allocation algorithm itself.
- Spanish research consulting firm Eticas conducted focus groups with border guards and travelers to better understand technologies used at borders in the European Union — in particular, what aspects of the border crossing process were delegated to algorithms, and how those delegations impacted human decision-making.⁴³ From these interviews, the group was able to map the **constitution** of the system, how automated systems have been deployed differently across European countries, and how people perceive and interact with these systems.
- Stanford sociologist Angèle Christin used ethnographic fieldwork across multiple sites in the US and France to observe how journalists and legal professionals use and interpret algorithms designed to change how these expert professionals make decisions in their work.⁴⁴ The study found that people in these professions use “buffering strategies,” including ignoring or critiquing the output of automated recommendation systems, to minimize the impact of algorithms on their work.

Legal Process

When voluntary methods of inquiry fail, individuals and public actors sometimes turn to legal and regulatory systems to compel companies, government agencies, and others to reveal details about their automated systems. While legal channels are not always successful in revealing desired details about automated systems of concern (particularly **training data** and **source code**, due to barriers like trade secret protections), this method of inquiry can sometimes offer more “teeth” than other tools described in this section.

CASE STUDIES

- Digital rights advocacy group R3D in Mexico is relying on data protection laws and the concept of, and remedies available through, *habeas data*⁴⁵ to challenge before the Data Protection Authorities companies that deny people their right to access data about themselves. By identifying who holds personal data that could be processed automatically, advocates are pursuing a larger strategy to map the **existence** of automated systems of concern.⁴⁶
- The US Federal Trade Commission used its antitrust investigative authority to investigate Google’s prioritization of its own products in search results, demanding the production of internal documents from the company to understand the **purpose**, **constitution**, and **policies** of the search engine’s algorithms.⁴⁷ Even after finding evidence of misbehavior, however, the agency opted not to act against the company.
- US nonprofit the National Consumer Law Center investigated whether credit bureaus have adequate software and human processes to help consumers correct errors in their credit reports. The group’s study, which scrutinized and critiqued the **constitution** and **policies** of automated error dispute systems, relied heavily on discovery documents from relevant court cases.⁴⁸

- Legal scholar Rebecca Wexler has examined legal documents from numerous cases where defense attorneys requested access to **training data** and **source code** of algorithmic systems (like pretrial risk assessment and forensic evidence tools) used in the criminal justice system. These requests are often blocked due to trade secrecy protections.⁴⁹
- A California trucking company appealed the Federal Highway Administration’s denial of a Freedom of Information Act request to disclose the **source code** of its motor carrier safety rating algorithm. The agency claimed the source code of the system was exempt since it related to internal agency practice, and disclosure would risk circumvention of the law.⁵⁰ The judge found that the algorithm was not exempt, and ordered the agency to disclose it.

Black Box Testing

Black box testing is any examination of an automated system without access to its internal structure or code. It covers a range of techniques that can be used to better understand a system’s **inputs and outputs**, and even approximate its underlying **source code** or **models**. Black box testing includes everything from basic input-output observations to more mathematically sophisticated techniques, and is especially useful in cases where the public lacks privileged access to a system, or has not yet identified a clear standard for auditing. While inductive reasoning has its limits — an examiner certainly can’t test every possible hypothesis about how a system works — there is strong evidence that black box testing is a valuable approach.

Simple observation of inputs and outputs

The most basic form of black box testing is to provide a system with different inputs and analyze its outputs. Input-output testing can be simple, with manual testing of a small list of inputs, or complex, using computer systems to automatically generate and observe much larger sets of inputs and outputs. Even in its most basic forms, input-output testing has been a remarkably useful technique for drawing conclusions about the workings of powerful online platforms, including search engines and social networks.

- Harvard researcher Latanya Sweeney searched for a predefined set of names (**inputs**) on Google’s search engine and documented the advertisements she observed as **outputs**. She observed that a greater percentage of ads with “arrest” in their text appeared for black-identifying names than for white-identifying names, to an extent that could not plausibly be explained by chance.⁵¹
- Israeli researchers Maayan Perel and Niva Elkin-Koren studied how local online platforms detected and responded to copyrighted materials. They uploaded infringing, non-infringing, and fair use materials (**inputs**) and tracked whether the system correctly classified this content based on whether content was automatically flagged or removed (**outputs**), and found these systems were not reliable judges of infringement.⁵²
- US consumer rights group the Consumer Federation of America obtained rate quotes from major insurance carriers using the same test profile, except for marital status, to test for differences in rates for minimum liability insurance. Holding all other **inputs** equal, they found the quoted rates (**output**) were higher for test subjects who were single, separated, widowed, or divorced than for married people.⁵³

Sophisticated manipulation of inputs and outputs

When an examiner has access to a large historical record of a system’s inputs and outputs, or can automatically generate many different inputs, she can conduct more sophisticated analyses. One family of techniques, called **feature perturbation**, measures the dependence of a system on its inputs by varying the value of one feature while holding the value of other features constant.⁵⁴ Using a related technique, called **model estimation**, an examiner can build her own

simulated model from the inputs and outputs available to her, and then study that model as a stand-in for the original.⁵⁵ These and similar strategies can sometimes allow examiners to draw reliable, sophisticated conclusions about how an automated system functions even without access to the system’s source code.

Both simple and advanced black box testing techniques can be effective in discovering and drawing attention to particular behaviors or automated systems. Basic input-output testing can go a long way in attracting regulatory and public attention, compelling cooperation from the proprietor of a system, or guiding more quantitatively robust inquiries into that system. Advanced methods appear to be most useful when more detail or greater certainty in results is necessary, for example, to guide regulators or preempt rebuttals.

CASE STUDIES

- The US Federal Reserve Board scrutinized the industry-standard practice of credit history scoring by obtaining large sets of relevant data — 3.7 million anonymized consumer credit records enriched with demographic data (**inputs**) and the associated credit scores (**outputs**) — to build its own estimated models in search of evidence that any features served as a proxy for race. The study claimed its analysis showed that the credit characteristics tested “do not serve as substitutes, or proxies, for race, ethnicity, or sex.”⁵⁶
- German researchers examined the “filter bubble” effect of Google search results by defining a set of 16 fixed search **input** terms and capturing the top search results (**outputs**) for those terms across different browser settings and geographies. The researchers used feature perturbation to hold certain inputs constant and measure the weight of others on the estimated model,⁵⁷ and published the dataset so others could analyze it.⁵⁸
- Nonprofit investigative journalism organization ProPublica obtained a set of historical **inputs** and **outputs** for the COMPAS pretrial risk scoring program and, using statistical analysis and model estimation, found it to misclassify black defendants as at high risk of committing a future crime more frequently than it did white defendants.⁵⁹ The resulting article and dataset have catalyzed debates about appropriate calibration of criminal justice algorithms, as well as mathematical research into new ways to measure for fairness in algorithmic models.

Examination of Training Data

When an automated system relies on predictive models, information about the training data used to build those models — including the size of the dataset, its provenance, who and what was included, and how it may have changed over time⁶⁰ — is critical. Such information can allow an examiner to consider the sufficiency and appropriateness of that data for a given **purpose** and whether it matches up with the system’s stated **policies**. Examiners can also think about what data was not included — especially if underrepresentation means the **model** might fail more frequently when applied to new, unfamiliar examples.

CASE STUDIES

- Researchers from the Human Rights and Data Analysis Group analyzed drug crime arrest data from Oakland, California in 2010 — and an approximation of a proprietary model — to test predictive policing systems for the likelihood they would amplify distorted policing and enforcement practices.⁶¹ The researchers pointed out that while systems were intended to predict crime, they used police records as their underlying **training data** — which do not represent all actual crime.
- Street Bump, an app that automatically sent reports to the city of Boston about the condition of its streets to guide maintenance resources and make predictions to inform long-term investment, originally found that there were more potholes in wealthy areas of the city compared to poorer ones. Researchers ultimately found that the **data** collected was not actually representative, since wealthier residents were more likely to own smartphones and use the app.⁶²

Code Review

In many cases, review of source code can be a powerful form of scrutiny. A code audit, also called “white-box testing” in the computer science field, is an analysis of **source code** to discover errors. These audits can also include a review of specific system behavior — logs that record data access, calculations, decision trees, and errors. For some automated systems, “source code” might also refer to the statistical **models** that rank, sort, classify, and score inputs. Models can be more difficult to review than descriptive, deterministic computer code, but in some cases (usually those that rely on a limited number of factors), understanding how different inputs are weighted in an algorithm can be illuminating.⁶³

Reviewing source code has some important practical limitations. In practice, it can be difficult to gain access to source code, which is typically considered sensitive intellectual property. And once source code is obtained, even seasoned experts can miss simple problems buried in complicated code. For even moderately complex programs, it may be necessary to see a program run “in the wild,” with real users and data to truly understand its effects.⁴¹ In practice, code audits are most likely to be useful when there is a clearly defined question about how a software program operates in regulated space, and particular standards against which to measure a system’s behavior or performance.

CASE STUDIES

- US regulators including the Environmental Protection Agency knew that diesel Volkswagen and Audi cars emitted illegal amounts of nitrogen oxides during real-world emissions tests. But they didn’t have a “smoking gun” to show that these results were intentionally engineered. Researchers struggled to piece together the offending **source code**, relying on car hobbyist and auto-performance forums to stitch it together. Finally, researchers were able to examine the code and found the offending section of code that provided strong evidence of purposeful cheating.⁶⁵
- Aware that the move from paper ballots to voting machines could raise election security and integrity issues, Princeton computer science researchers Ariel Feldman, Alex Halderman, and Edward Felten examined the **source code** of the widely used (in the US) Diebold Ballot Station voting machine software. Their code review found that if a malicious actor had access to certain machines, he could “steal votes with little if any risk of detection.”⁶⁶
- Predictive policing company CivicScope released its **model** for outsiders to scrutinize, but critics contended that the code was too generic to evaluate thoroughly in isolation, especially without access to police department data. Moreover, the code could not shed light on how police departments interact with the system — particularly how they act on the outputs that the model provides.⁶⁷

This part has analyzed the many ways that the public can scrutinize automated systems already in use in the world. It has shown that while data and source code transparency can be helpful, there is a range of other ways to scrutinize important systems. Lack of access to the inner workings of an automated system, while an impediment to rapid and thorough analysis, should not necessarily preclude meaningful interrogation of at least some components of that system.

The next part explores emerging opportunities to design automated systems in ways that can make them more accountable to the public.

Designing for Accountability

The methods described above enable mostly post-hoc, external review of automated systems. However, there is a growing range of methods for programmers and insiders to *proactively* design automated systems that are fairer, more interpretable, and more auditable. Most of these methods are still under heavy development in academic and corporate research labs, and so remain largely theoretical today.

Applying Metrics for Fairness

When an automated system relies on patterns derived from training data, steps can be taken to identify and mitigate unfairness and bias at various stages in the model-building process. There is a range of *pre-processing* methods that attempt to remove biases from data before building a model, by suppressing sensitive attributes, changing labels, reweighting attributes, or resampling underlying data.⁶⁸ These methods might involve altering the underlying data or its labels, and sometimes can come at a cost to accuracy. There are also *in-processing* techniques that modify traditional learning algorithms to address discrimination during the training phase, rewarding pre-defined notions of “fairness” when building a model.⁶⁹ Finally, *post-processing* techniques allow the outputs of a model to be reviewed and adjusted to meet particular goals.⁷⁰

Today, the community of data scientists working on these methods remains small, and commercially validated research is limited.⁷¹ However, many advocates hope these types of techniques can one day be proactively applied in areas where automated systems make socially important decisions.

A growing range of methods are available for insiders to proactively design automated systems that are fairer, more interpretable, and more auditable.

RESEARCH EXAMPLES

- Data scientists at Predictive Hire, a company that builds automated job candidate assessment tools, apparently screen their datasets for racial, gender, ethnicity, and age biases before training a predictive model. If they find an imbalance between, for example, males and females on a given performance metric, they turn to pre-processing to mathematically correct for that tendency in the training data.⁷²
- Researchers have described how the US state of Texas’ “Texas Top 10” college admissions program — which guarantees admission to the University of Texas system to the top decile of students — could lead to racial bias in admissions, since “high school attended” data could be a proxy for race. They demonstrate several methods to repair, or pre-process, the data so the program can preserve student ranking information while minimizing or removing disparate impact.⁷³
- Hardt et al. describe several post-processing options in the context of reducing disparate impact of FICO credit scores by adjusting the threshold score(s) for prime-rate loans. For example, a *demographic parity* approach would assign a threshold for each racial group such that the proportion of each group that qualifies for prime loans is the same (e.g., 80 percent of all applicants from each group qualify), while an *equal opportunity* approach would select a different threshold for each group where “the fraction of *non-defaulting* group members that qualify for loans is the same” (e.g., 80 percent of the applicants from each group who never defaulted on a loan qualify).⁷⁴

Interpretability

The “code” of a predictive model built using machine learning — many hundreds or thousands of variables with different associated weights, rather than a logical series of instructions — can be especially difficult to scrutinize. However, there are some emerging techniques to help make predictive models more *interpretable*, or understandable to humans.⁷⁵ These techniques are already common practice when a model will be used in a way that implicates safety (e.g., the medical field) or there are special regulatory requirements (e.g., credit scores). Interpretable models can be easier for both an institution itself to understand, and for outside examiners to test.

There are several well-understood techniques for making interpretable models. Most simply, a modeler can purposely use a smaller number of features. This helps avoid the “curse of dimensionality” — a situation where it is impossible to reason through the relationships among a large number of variables.⁷⁶ However, some promising emerging research claims that thousands of features can be used while maintaining interpretability and accuracy, using new techniques to highlight only the strongest interactions between features for review.⁷⁷ Other researchers imagine “helper algorithms” that could run alongside an automated decision to monitor and interpret its behavior,⁷⁸ or focus on visually mapping or calling out data that is most important in a model (a technique particularly helpful for automated decisions that rely on image analysis or machine vision).⁷⁹

Interpretability continues to receive considerable research attention. While there is a common conception that interpretability almost always comes at the expense of a model’s accuracy, researchers are beginning to challenge this idea.⁸⁰ Many of the professional data scientists we spoke with while preparing this report were optimistic that a significant portion of socially important machine learning tasks could be rendered more amenable to scrutiny — even those relying on hundreds or thousands of variables.

There are some emerging techniques to help make predictive models more interpretable, or understandable to humans.

RESEARCH EXAMPLES

- **Intelligible predictive models can be especially important in domains like medicine, argues Rich Caruana of Microsoft.⁸¹ Famously, in the mid-90s, a model his team trained to predict the probability of death for pneumonia patients learned a surprising rule: that a “history of asthma lowers a patient’s chance of dying from pneumonia.” This rule — although clearly questionable — reflected true patterns in the data, because people with a history of asthma tend to get care faster in real-world hospital settings. The counterintuitive conclusion the model discovered illustrates the importance of understanding what models are learning and why.**
- **Caruana argues that emerging methods for building transparent and editable models, such as “Generalized Additive Models,” which layer a collection of simpler, more interpretable models, are now competing in terms of accuracy with more traditional, opaque kinds of machine learning like neural networks.⁸²**

Procedural Regularity and Audit Trails

Researchers are exploring exciting methods to verify that software systems are behaving as promised, even without having access to their source code. For example, scholars have already shown computer systems can be designed to prove that decisions were made under an announced set of rules consistently applied in each case.⁸³ This property, called *procedural regularity*, is not yet commonly deployed, but could be especially useful when computer systems are entrusted with important decisions.

Other methods, piloted by large data companies, purport to allow third-party auditors full-proven mechanisms to check how data is being processed. Audit logs, when available, can allow an examiner to review important information about a system's use and performance. These logs might reveal when and why data was accessed, and what decisions were made, about whom, and why. Audit logs are familiar in contexts like finance, accounting, and cybersecurity, but their sophisticated use within software systems of social importance is still in the early stages.

RESEARCH EXAMPLES

- An applicant denied a green card through the United States Electronic Diversity Visa Lottery might wish to see proof that she was denied randomly, under a predefined set of constraints — and that the system was not manipulated to favor other individuals or groups. Immigration and security agencies could use cryptography and other mathematical tools to provide proof that the process was random, without revealing the underlying source code.⁸⁴
- Google's DeepMind group is working on systems that log every possible use of data within a system, with the ability “to prove that every data access by every piece of software in the data centre is captured by these logs.”⁸⁵ Participating hospitals and external auditors will be able to review these logs to verify that patient data is only being used for approved purposes.

However, procedural regularity will only be helpful insofar as stakeholders have agreed on what steps and safeguards are to be guaranteed — and on the remedies available when evidence suggests those steps may have been compromised.

The methods described above all have significant potential, but will likely require new laws or policies to spur adoption. The legal and regulatory context around automated decisions, and their scrutiny and design, are discussed in the next section.

Perspectives on the Role of Law

Laws can affect the operation of automated systems in many ways. They can prohibit the use of data deemed sensitive. They can provide the public with information to aid scrutiny and oversight. They can directly constrain the way algorithms are developed. They can govern the outcomes of automated systems, relying on policy frameworks related to nondiscrimination and antitrust. And they can even grant individuals the right to avoid automated decisions entirely.

Below, we offer a more thorough analysis of common approaches to regulating automated decisions, incorporating views from regulators and global digital rights advocates. Some of these approaches are frequently used in practice, while others remain theoretical or largely untested. They are listed roughly in order from most to least frequently mentioned during our outreach.

Restrictions on Data Collection

Limiting the collection of data by institutions can be a powerful way to exercise control over automated decisions. The idea that data should be collected, stored, and used only as far as is reasonably necessary has long been a core privacy principle.⁸⁶ More targeted approaches involve restricting collection of “sensitive” data like race or gender for use in certain kinds of decisions.⁸⁷

Many advocates see limiting data collection as a simple, powerful way to even the playing field between the public that generates data and the powerful institutions that own and depend on it.⁸⁸ This is an especially attractive goal for those in countries with weak legal regimes for data protection and nondiscrimination.⁸⁹ However, there is also widespread recognition that this approach faces strong limitations in practice. Adoption of new technologies and data-driven services is on a meteoric rise, seemingly unperturbed by widespread anxiety about privacy and regulatory concern. Most restrictions on data collection tend not to apply to anonymized or “non-personal” data, even though this data can still be leveraged for important decisions.⁹⁰ And finally, the collection of data — sometimes even sensitive data — can be critical to ensuring that proper decisions are made (e.g., by measuring for bias or properly enumerating different populations for public planning purposes).⁹¹

Many advocates see limiting data collection as a simple, powerful way to even the playing field.

EXAMPLES

- Europe’s General Data Protection Regulation (GDPR) emphasizes data minimization, stating that personal data should be “limited to what is necessary in relation to the purposes for which they are processed” to reduce risks associated with holding or processing that data.⁹² But in practice, there are many avenues for extensive data collection under the Regulation, leaving advocates uncertain about this rule’s effect.⁹³
- To “protect the rights and interests of individuals,” Japan’s privacy regulation bars business operators from collecting sensitive information, including creed or social status without explicit consent except as required by law.⁹⁴ The efficacy of this approach is unclear.
- In a narrower context, advocates in the United States argued that the National Security Agency should not be allowed to collect metadata about phone and Internet activity of US citizens.⁹⁵ Although the government argued that such data can be critical in detecting terrorist activity and that using it still protects privacy and civil liberties,⁹⁶ a federal law eventually ended the agency’s ability to collect this data.⁹⁷

Transparency

Transparency — the ability of the public to examine information and internal deliberations from institutions of interest — has long been celebrated as a tool to hold powerful actors accountable, and is motivated by a range of reasons and theories.

There are a vast number of transparency-related laws and policies. For example, a slight majority of all countries — as well as many smaller jurisdictions — grant public access to details of government activity through open records and freedom of information laws.⁹⁸ “Open government” initiatives have gained traction in recent years, pressing participating government bodies to proactively publish administrative documents and data.⁹⁹ The corporate sector can also be subject to various transparency requirements, such as disclosing risk factors to investors,¹⁰⁰ and also take some voluntary action, like publishing periodic reports about government requests for data.¹⁰¹

Advocates are almost universally supportive of transparency as a policy goal for powerful institutions of all kinds. However, many noted that transparency procedures have been difficult to enforce, and that in practice, requests for information are frequently overridden by other interests. Transparency laws crafted in an analog era, before big data commonly fed predictive and automated systems, may be ill-equipped to handle the realities of today’s automated systems.¹⁰²

EXAMPLES

- **France, through its Digital Republic Bill, is the only country that has explicitly required disclosure of the source code of government-developed algorithms under its open record laws.¹⁰³ The law will require administrations to publish online, in an open standard, key documents, including source code, databases, and any data of public interest. However, the law will take effect gradually, and is currently awaiting implementation decrees.**
- **The US Federal Agency Data Mining Reporting Act of 2007 requires government entities to report on and publicly disclose extensive information about federal use of predictive analytics, as well as the goals and efficacy of related data mining activity. In practice, it is unclear whether this mandate is actually followed.¹⁰⁴**
- **In some legal regimes, operators of automated systems must, in narrow contexts, reveal internal details of those systems upon request. For instance, concerned with preserving airline competition, the US government once legislated that computer reservation systems must share ranking criteria used in sorting algorithms for displayed flights.¹⁰⁵ However, these types of specialized transparency mandates were rarely mentioned by advocates.**

Disclosure of source code and algorithmic formulae may not always serve the needs of the public, either because the requester lacks the technical competence to review the code, or because it allows malicious actors to “game the system” with their intimate knowledge of a system’s behavior. But as the previous sections showed, transparency at different layers of a system can help external scrutinizers understand a system well enough to raise concrete concerns, guide further investigation, and propose potential remedies — a nuance many transparency laws today overlook.

Explanations

The idea that automated decisions should be “explainable” to those affected has recently gained prominence, especially in the EU.¹⁰⁶ Legal approaches to securing the right to explanations range from requiring disclosure of key reasoning behind a particular decision to a meaningful description of the “logic” of an entire system. While many European countries’ data protection laws require some kinds of explanations, the related provisions are often vague or narrow.¹⁰⁷ In the United States, however, consumers routinely receive “reason codes” along with their credit scores — a specific requirement of federal law for this domain.

- The GDPR’s apparent requirement that data controllers provide “meaningful information about the logic involved” to subjects of some automated decisions has commanded significant attention.¹⁰⁸ However, this right only applies to decisions based entirely on an automated process. Several scholars have pointed out that it might not be legally binding.¹⁰⁹ Regulators and data protection authorities have not yet developed guidance for compliance with this requirement, and its effect remains untested.
- India’s Right to Information Act requires public authorities to “provide reasons for its administrative or quasi-judicial decisions to affected persons,” which some speculate may also apply to automated decisions.¹¹⁰ However, a Supreme Court case held that the Act does not require authorities to collect information it does not already have, nor “furnish information which require[s] drawing of inferences and/or making of assumptions”¹¹¹ – an exemption which may have particular significance for automated decisions.
- A more established, narrower approach to explainability can be seen in the Fair Credit Reporting Act (FCRA), which requires consumer reporting agencies to share with consumers up to four key factors that affected their score, listed in order of their effect.¹¹² Regulatory agencies have clarified that consumers must also be notified if they receive less favorable terms for financial products based on low credit scores.¹¹³ These “explanations” are received routinely by US consumers.

Antidiscrimination

The tenet of nondiscrimination is enshrined in the Universal Declaration of Human Rights and in most national laws.¹¹⁴ Many legal regimes aim to secure the values of fairness and equity by prohibiting direct, intentional discrimination of people with certain protected characteristics (e.g., race, gender, religion, or political affiliation). Some also prohibit indirect discrimination, which can be revealed after the fact by comparing treatment of different individuals or measuring adverse impact across protected categories.

Many advocates in the US and the EU spoke about the importance of applying antidiscrimination laws to automated decisions, especially those relying on predictive models built from historical data. However, antidiscrimination laws rarely provide clear quantitative benchmarks, making this a challenging project. In some cases, standards bodies and regulators will need to define key metrics, while in others, normative, society-wide debate and reflection is still necessary.¹¹⁵ Some advocates in Latin America noted that the relative weakness of local nondiscrimination laws will be a barrier to advocacy around discriminatory algorithms.

- In the UK, the Equalities Act of 2010 prohibits direct and indirect discrimination – “less favourable treatment” – in employment, education, and provision of services based on nine protected characteristics (or the perception of having one or several of the characteristics), and requires public authorities to consider equality impact in the design of policies and services.¹¹⁶ No numerical measure of discrimination has been established, and it remains unclear how the UK law might be applied in cases of bias in automated decisions.¹¹⁷
- The US Civil Rights Act of 1964, follow-on legislation, and enforcement activities bar disparate treatment of protected classes in contexts including employment,¹¹⁸ credit,¹¹⁹ and housing.¹²⁰ Jurisprudence and additional laws added the doctrine of disparate impact, and multiple researchers and programmers of automated systems have turned to the Equal Opportunity Employment Commission’s “80 percent” rule to identify adverse impact as a threshold to detect bias in automated decisions.¹²¹

Data Access, Accuracy, and Redress

Many legal regimes give individuals the right to access and correct data about themselves in certain circumstances. These rights, enshrined in a range of Fair Information Practice Principles,¹²² seek to minimize the risk of adverse consequences stemming from inaccurate or missing data, and to allow individuals to know what data about them is held by covered institutions. They are codified differently across the globe: Many Latin American constitutions recognize the right of access and accuracy through the writ of habeas data,¹²³ while in other jurisdictions these rights are anchored in data protection regimes¹²⁴ or consumer protection laws.¹²⁵

When automated decisions are made about individuals, accurate and complete data is critical. For example, rules are especially important in credit and consumer reporting contexts, where in many developed countries, a limited number of entities can hold important data inputs.¹²⁶ However, in some jurisdictions, access rights don't apply to all institutions — and individuals cannot access or correct anonymized data, or change personal data used to train machine learning models that has already been deleted.¹²⁷ Also, there is some evidence that when access rights are provided broadly, data subjects rarely exercise them.¹²⁸

Advocates mentioned these rights only infrequently when discussing legal responses to automated decisions. Some feel that access rights, while welcome, would place an unfair burden on individuals to spot and prevent errors. Others worry that fundamental issues of bias in historical data won't be solved by ensuring the data is accurate. And in some Global South countries, advocates are using access rights as a discovery method to learn more about what data companies hold about people in the first place.

Many legal regimes give individuals the right to access and correct data about themselves in certain circumstances.

EXAMPLES

- In the United States, the FCRA requires certain “consumer reporting agencies” to disclose a consumer’s own information when she requests it, and grants her the right to dispute that information.¹²⁹ Most US consumers have used these rights to review their credit reports.¹³⁰
- In countries that lack robust consumer reporting infrastructure, including many across the Global South, civil society groups depend on access laws to begin to map the existence of datasets that might eventually be used for automated decisions. In Mexico and elsewhere, advocates are only beginning to test the power of habeas data and other data protections that grant access to learn about what data institutions hold.

Opt-outs and Forbearance

Some jurisdictions provide individuals with a right to object to, or “not to be subject to,” certain automated decisions.¹³¹ However, advocates rarely mentioned these policies, perhaps due to their somewhat feeble protections. For example, individual rights tend to be limited to automated decisions that “significantly” affect a subject, and are often subject to broad exceptions for law enforcement and national security-related data processing.¹³² Moreover, placing the onus on individuals to opt out unreasonably requires laypeople to weigh hypothetical, and often intangible, risks.¹³³

In other cases, some entities have opted to formally restrain themselves from developing or deploying automated systems for certain types of decisions, to prevent known or unknown risks.

- Europe’s current Data Privacy Directive directs member states to grant the right not to be subject to “fully automated decisions which produce legal or significant effects,”¹³⁴ and the GDPR includes the right to withdraw consent at any time, as well as to prohibit data processing for marketing purposes.¹³⁵ However, this right does not apply in some cases of government decisions,¹³⁶ particularly those in the realm of criminal justice and national security.¹³⁷
- Some EU member states have opted for even stronger language than what the harmonized law suggests: Estonia, for example, prohibits legally consequential automated assessment of a person’s “character, abilities or other characteristics” without that person’s participation, except in some narrow cases.¹³⁸ Austrian law holds that “nobody shall be subjected to” these sorts of decisions, with an exception for decisions taken on the basis of law.¹³⁹ France prohibits court decisions based on automated processing of personal data.¹⁴⁰
- While no binding national or international laws currently restrict autonomous weapons, in a more extreme example of avoidance, 19 countries have called for a preemptive ban on the technology.¹⁴¹ The US Department of Defense has issued guidance calling for “appropriate levels of human judgment” over uses of force by such systems, for the time being constraining the role of automated decisions in the context of conflict.¹⁴²

Validation and Certification

Premarket validation and certification requirements compel creators of some systems and products to conduct thorough testing — and at times receive external approval from auditors or regulators — before systems are deployed or sold. Restrictive rules are more common in safety-critical contexts where people could be at high risk of grave physical harm, but less binding voluntary certifications and “impact assessments” also fall into this category of ex ante evaluation. Mission-critical algorithms used in flight and nuclear facility software have long been required by regulation to go through verification and validation processes.¹⁴³ Certain medical algorithms that require regulatory approval must also demonstrate that their performance is at least as good as humans.¹⁴⁴

Some in civil society envision voluntary certification of algorithms for fairness and accuracy as a tool to encourage more and earlier testing, as well as a way for companies to demonstrate responsible behavior and legal compliance without disclosing trade secrets.¹⁴⁵ Advocates have proposed that validation requirements be imposed on socially critical algorithms, such as those used in criminal justice contexts.¹⁴⁶ However, a lack of clear consensus around what ought to be measured and how — let alone quantitative benchmarks to indicate success or failure on those measures — means that many of these goals are still aspirational.

- Illustrating strict certification requirements, the US Food and Drug Administration (FDA) requires validation of software and automated systems used to design or produce food, drugs, and medical devices — including certain diagnostic and predictive software.¹⁴⁷
- The GDPR will require data controllers to conduct “data protection impact assessments” (DPIAs) in cases where data processing is “likely to result in a high risk to the rights and freedoms of natural persons.” Data controllers will need to take steps to mitigate risks identified or consult with regulators, who retain the power to prohibit the high-risk data processing activity, to demonstrate compliance.¹⁴⁸ While potentially constructive, the power of the DPIA process to meaningfully govern automated decisions has yet to be tested, as regulators are still finalizing detailed guidelines.¹⁴⁹

Auditability

The concept of auditability extends beyond transparency, demanding not only access to systems but also that those systems be amenable to meaningful review. As with external inspection of financial records or digital security practices for regulatory compliance, the specter of auditing can encourage actors to self-police while also creating a situation where auditors are able to review systems for procedural integrity and fairness.

General auditability of automated systems is not explicitly required in any major legal regime — indeed, laws enacted to prevent online fraud have created significant barriers to external algorithmic auditing¹⁵⁰ — but some specific legal requirements do exist. A few jurisdictions broadly require or imply a need for audit trails; others require auditability only in narrow contexts. Many advocates have pressed for more external auditing of algorithms, and auditability has been included as a principle in several collections of best practices and codes of conduct for machine learning and artificial intelligence. However, it is often unclear what, exactly, an inspector would be auditing for.

The concept of auditability extends beyond transparency, demanding not only access to systems but also that those systems be amenable to meaningful review.

EXAMPLES

- The Australian government has clarified the importance of comprehensive audit trails to ensure that automated decision systems align with administrative law.¹⁵¹
- In the heavily regulated gaming industry, authorities often retain the right to audit randomization algorithms in gambling machines for unfair behavior.¹⁵² Premarket certification is usually also necessary in these cases — which in practice requires that the programs be intelligible to inspectors.

Competition

Competition law promotes the functioning and benefits of free markets by preventing collusion and market abuses that disadvantage consumers. On reasonable suspicion of unfair practices or abuse of market dominance, antitrust investigative bodies have authority to probe a company's conduct and products, including subpoenas of witnesses and evidence, and to penalize violations of the law.

Competition laws have prompted in-depth investigations of anticompetitive market behavior arising from automated systems. Recently, some of these investigations have involved sorting and filtering in the context of search engines, as well as “algorithmic collusion,” or automatic price fixing.¹⁵³ Regulators in Europe are particularly sensitive to the potential for market abuse by foreign internet companies.¹⁵⁴ But appetite for antitrust enforcement — and style of antitrust inquiry — varies widely across countries and administrations.

EXAMPLES

- The FTC used its antitrust investigative authority to investigate Google's prioritization of its own products in search results — but even after finding evidence of misbehavior opted not to act against the company.¹⁵⁵
- The European Competition Commission began investigating Google for suspected anticompetitive behavior in the company's search results in 2010. A central concern of the Commission was that Google “accord[ed] preferential placement to the results of its own vertical search services to shut out competing services” even though Google characterized search results as “natural” and “algorithmic.”¹⁵⁶ Recently, the Commission found that the company denied “consumers a genuine choice” by using its search engine “to unfairly steer them to its own shopping platform,” resulting in a record \$2.7 billion fine.¹⁵⁷

Product Liability

New product liability regimes could have significant consequences for the design, testing, and deployment of automated systems. Emerging discussions about the applicability of product liability laws for automated decisions have tended to focus on physical systems like robots and autonomous cars.¹⁵⁸ Several academics and policymakers have called for more clarity in currently messy legal frameworks to address the shortcomings of existing liability law for robotic behavior and to provide guidance to industry actors building automated technology.¹⁵⁹ These efforts are in the early stages, but may gain more urgency as more automated physical systems become a reality.

EXAMPLES

- **European parliamentarians have asked the European Commission to draft legislation defining the legal status of autonomous robots and allocating relevant rights and responsibilities to them. The proposal suggests strict liability as a foundation for future rules, and that liability should be proportional to the level of automation in question. Notably, this proposal also says that access to source code should be available to investigate accidents and damage.¹⁶⁰**

Given the vast number of legal approaches discussed above, it can be difficult to draw clear conclusions. When asked about the role of law and policy in scrutinizing automated systems, digital rights advocates from across the globe expressed a range of goals and concerns:

- In the United States, there is no overarching law that governs the collection and sale of data by commercial entities. Many US advocates were particularly concerned about the application of existing antidiscrimination laws to automated decisions. Moreover, the Fair Credit Reporting Act (FCRA), which has regulated key types of data brokers — particularly credit bureaus — with notable success, was frequently highlighted as a model for the future legislation.
- European Union countries enjoy a comprehensive data privacy framework, but the powerful principles of the GDPR are widely viewed as uncertain in their application. Here, almost every advocate we spoke with acknowledged that “algorithms” were a common source of concern, and that “explanations” were a common fixture of the policy discourse. But most were skeptical that a “right to explanation” was defined clearly enough to be useful in practice, or that the new GDPR would provide enough new clarity to significantly alter opportunities for scrutiny.
- Global South countries frequently lack key legal frameworks like data protection and antidiscrimination laws. Many advocates in these countries felt fundamentally unprepared to govern the collection or use of data, or even to know what their governments were using data for in the first place. Although advocates were aware of concerns related to automated decisions, many saw these issues as taking a back seat to more basic policy goals.

Across all regions, there was some recognition that traditional data protection measures, particularly restrictions on data collection, have struggled to handle the realities of data-hungry and complex algorithmic systems. There was also a shared recognition that transparency laws have been critical to uncovering information that has informed both civil society and regulatory conversations and will be crucial moving forward.

Each of these legal and policy approaches applies to different layers of an automated system. When considering enforcement priorities, activism goals, or new policy approaches, the public and policymakers should think about what components of an automated system relate most closely to stated objectives and ensure issues are framed appropriately.

Conclusion and Paths Forward

Automated decisions, increasingly driven by software and vast amounts of data, are the subject of widespread concern. Advocates across the globe have expressed a desire to more closely scrutinize automated systems of social importance. There is a growing consensus that institutions using automated systems should be accountable to the public, as demonstrated by a growing number of high-level principles across the public and corporate spheres. However, a clear agenda for public scrutiny has yet to emerge.

Our research indicates that technical expertise and methods are not the biggest roadblocks to scrutiny today. This might seem surprising, as the rise of software and data-driven systems has created an emphasis on the need for technical approaches to oversight. However, these techniques — many of which require proactive cooperation from institutions of interest — are likely to remain in the lab until civil society makes a clear case for their adoption.

The field needs new ways to obtain knowledge and evidence about how automated systems work in practice, and the domain expertise to wrestle with the difficult normative questions that often lurk just behind the code. Toward this end, we suggest the following paths forward as civil society continues to grapple with the implications of automated systems:

- **Increased investment in exploratory scrutiny, especially by journalists and advocacy organizations.** Today, some of the highest-impact work on automated decisions makes or finds evidence about how important systems work in practice. Many of today's highest-impact efforts have provided just enough insight to spark important debates in domains like criminal justice and credit. To engage wider participation in debates about how automated systems should function, the field needs more work to find evidence about, and clearly explain, how important systems work in practice. This work can help build the case for new policies and technical requirements.
- **Strategic evaluation of right to information laws' ability to facilitate the right kind of transparency for today's automated decisions.** To scrutinize systems used by governments, the public must know they exist, the purposes they have been designed to achieve, and the data they rely upon. Once those basic details are available, other techniques for scrutiny and ensuring integrity will be useful — but all too often, these basic facts remain obscured, and most transparency laws have concerning loopholes that prevent this and other relevant information from being shared with the public. Rather than start with entirely new policy proposals, existing laws should be closely examined with an eye for more consistent enforcement and application.
- **Consideration of policy mandates that certain automated systems be auditable and interpretable.** New technical methods are making it more feasible to design automated systems that are more amenable to scrutiny, but these methods may not be adopted without outside pressures. For public-sector automated systems, this could mean requirements that systems be designed to meet auditability requirements. For private-sector systems, this could mean “interpretability” requirements for important automated systems, such as those used in the context of employment and credit.
- **Further advancement of normative dialogues.** Many new policies and technical proposals presuppose standards and benchmarks that do not yet exist. Policymakers and the public must think more concretely about what “fairness” and “accountability” ought to mean in particular social contexts. For example, what counts as a “fair” way to underwrite an individual for credit? How should society deal with racial disparities when applying risk assessment tools to criminal defendants? How should a social media company organize the information we see? Each of these questions calls for debates about values more than technology, and may require reconciling social and political beliefs of stakeholders that have traditionally failed to reach consensus.

This report explored current and emerging approaches to scrutinizing and governing automated systems. With the rapid profusion of automated decisions, the public and civil society have an important role to play in continuing to investigate automated systems and articulating the social issues and concerns that they raise. It is especially important that the diverse groups working on these issues collaborate and share their insights. At the same time, governments, regulators, and independent bodies must consider whether current governance approaches sufficiently address transparency and accountability challenges presented by all components of automated systems. We hope the tools and insights distilled from this research can inform these efforts.

Appendix: Examples of Public Scrutiny

SYSTEM IN QUESTION	DATE	WHO INVESTIGATED?	WHAT DID THEY ASK?	ELEMENTS EXAMINED	HOW DID THEY INVESTIGATE?	WHAT DID THEY FIND?
US Patriot Missile Targeting System ¹⁶¹	1992	US Government Accountability Office	Why did the Patriot missile defense system in Dhahran, Saudi Arabia fail to track and intercept an incoming Scud missile during Operation Desert Storm, leading to the death of 28 Americans?	Constitution, Inputs and Outputs, Source Code	Interviewed software maintenance officials, analyzed code and system documentation including change logs, performed mathematical calculations and simulations.	A software problem had led to inaccurate tracking calculations that worsened over time when the system was not reset periodically. While the software had been corrected, it had not been updated at the base in question.
US credit scoring ¹⁶²	2007	Federal Reserve Board	Are credit scores accurate, and do they have a negative or differential effect on populations protected under the Equal Credit Opportunity Act?	Purpose, Inputs and Outputs, Training Data	Used 300k+ actual credit records, enriched with demographic data and credit scores, to perform various statistical tests.	The credit scores evaluated were predictive of credit risk for the population as a whole, and for all major demographic groups. Data in credit scoring models do not serve as substitutes, or proxies, for race, ethnicity, or sex.
Staples online price discrimination ¹⁶³	2012	Wall Street Journal	Are major online retailers varying product prices based on type of device or location?	Existence, Inputs and Outputs	Visited the websites from different geographic locations and from different devices; and examined the cookies/scripts associated with those websites.	Major commercial websites were varying prices based on type of device and location.
Google search results ¹⁶⁴	2014	Academics	Are Google's results biased on partisan lines?	Inputs and Outputs	Crowdsourced top 10 Google search results for the names of 16 presidential candidates on a single day, and then coded results for partisanship.	Democrats had more favorable search results than Republicans.
Volkswagen emissions test ¹⁶⁵	2014	Civil society	Are Volkswagen diesel car emissions within legal and advertised ranges?	Existence, Inputs and Outputs, Source Code	Testers drove cars from San Diego to Seattle with portable emission measurement systems attached, and compared results to lab tests by the California Air Resources Board.	The cars perform within limits in lab tests, but emit up to 35 times the legal limit on the road.

Racial bias in Google advertising ¹⁶⁶	2013	Academic	Do people with black-sounding names have more arrest-suggestive ads appear than those with white-sounding names?	Inputs and Outputs	Investigators ran a high volume of searches and noted the content of the personalized ads that were returned for names in each category.	Searches for black names are more likely to trigger arrest-suggestive ads than searches for white names.
Racial bias in COMPAS ¹⁶⁷	2016	ProPublica	Does the COMPAS pre-trial risk assessment system treat blacks and whites fairly?	Existence, Purpose, Constitution, Inputs and Outputs, Training Data	Obtained datasets through FOIA and manual criminal record reconstruction, used statistical models to compare scores across racial groups.	The program was biased against African American defendants on some measures.
Ethics in video surveillance systems ¹⁶⁸	2016	Academic	Was the design of a software system “ethical” and “accountable”?	Purpose, Constitution, Policies	Ethnographers embedded in project team from the start, review of algorithm tests, internal communication, and other documentation.	Accountability was a process distributed across team members and over time, and an ethics board played an important but time- and resource-intensive role.
Censorship on WeChat ¹⁶⁹	2016	Civil society	What is the scale and scope of content filtering, including automated filtering, on WeChat?	Constitution, Inputs and Outputs	Black-box testing: researchers sent messages with various keywords and content across different geographies and measured which messages were received.	Keyword filtering was only enabled for Chinese phone numbers, users were not being told their messages were blocked, and the filtering system changed dynamically.
Facebook profiling ability ¹⁷⁰	2016	Academic	How much information is necessary for Facebook to draw a conclusion about sensitive user characteristics?	Inputs and Outputs	Feature perturbation: examined input and output variables, changing one at a time.	Removing just a few “likes” (~6) significantly reduced Facebook’s inference power.
Flash crash ¹⁷¹	2010-2014	Regulators	What triggered the flash crash of 2010, and did high-frequency trading (HFT) play a role?	Constitution, Policies, Training Data, Source Code	Researchers used audit trail, transaction-level data to identify prevalence and behavior of algorithmic traders, and reviewed code and research containing sensitive information about trade reasoning, training data and proprietary formulas from firms.	High-Frequency Traders (HFTs) did not cause the flash crash, but contributed to it by demanding immediacy ahead of other market participants, leading to a liquidity imbalance.

Predictive chemical toxicity testing ¹⁷²	1997	Academics, private companies, and regulators	Are algorithmic predictive models as accurate as other ways of measuring toxicity?	Inputs and Outputs, Source Code	Compared actual inputs and outputs to predictions by testing model rules on various datasets, information from regulators, public and private data banks, rodent carcinogenicity bioassays, and carcinogenicity databases.	Algorithmic approaches were measurably less accurate than biological approaches.
Gender prediction from photos ¹⁷³	2017	Google UX researcher	What rules did an ML algorithm learn when instructed to classify photos by Male/Female subject?	Inputs and Outputs	Semi-controlled experiment: showed the algorithm different pictures of the same person with different hairstyles and makeup (5 variables / 32 total photos)	Hair length and presence of makeup were determining factors. When a picture did not match the stereotypical norm, it was misclassified.
Stereotypes in Google autocomplete ¹⁷⁴	2013	Academic	Does Google's autocomplete display a racist, sexist, or homophobic bias?	Purpose, Policies, Inputs and Outputs	Interrogated Google searches by entering 2,690 search questions and categorized autocomplete suggestions according to descriptors referenced.	Muslims and Jewish people were linked to questions about aspects of their appearance or behavior, while white people were linked to questions about their sexual attitudes. Gay and black identities appeared to attract higher numbers of questions that reflected negative stereotypes.
Detecting international border personalization in online maps worldwide ¹⁷⁵	2016	Academic	How often and in what circumstances are borders of online maps changed?	Existence, Purpose, Constitution, Inputs and Outputs	Created an automated system to crawl all tiles from a given mapping service from the perspective of every country around the world to identify discrepancies.	Detected the seven instances of border personalization, including two that were not previously documented. (Among them were borders between India and China, between Crimea and Russia, and in the South China Sea.)
Federal Highway Administration Safety Review factors ¹⁷⁶	1992	Private company	How does the agency compute safety ratings (factors and weights) for motor carriers?	Source Code	Submitted an FOIA request for the computer algorithm, appealed on claim of exemption.	Court found the algorithm was not subject to exemption, and ordered the FHWA to turn over documents.

Profiling the unemployed in Poland ¹⁷⁷	2015	Civil society	Is the use of profiling to allocate unemployment benefits accurate and fair?	Existence, Purpose, Constitution, Inputs and Outputs	Looked at list of questions asked during profiling. Collected statistical data on the distribution of active labor market programs across “profiles” at local labor offices, and how representative each “profile” was demographically.	Women, older people, and less educated people are more likely to be categorized as “far” from the labor market / less likely to benefit from services, so were not prioritized as highly as others.
Princeton Review differential pricing ¹⁷⁸	2015	Students / ProPublica	Does The Princeton Review’s pricing system disproportionately assign higher prices based on demographic characteristics?	Existence, Inputs and Outputs	Students in a Harvard data science class found that entering different ZIP codes resulted in different prices. (Their results inspired ProPublica to complete a more robust study.)	Asians were disproportionately represented in ZIP codes that were quoted higher prices. As a result, Asians were 1.8 times as likely to be quoted a higher price than non-Asians. ZIP codes with high median incomes were also more likely to receive higher quotes.
Hacker News ranking system ¹⁷⁹	2013	Blogger	How does Hacker News’ ranking work?	Inputs and Outputs	Crawled several news pages every two minutes and graphed results.	There appeared to be more tweaking of rankings than expected; certain keywords led to penalties in rankings.
Uber Greyball ¹⁸⁰	2017	New York Times	How does Uber evade regulators?	Existence, Purpose, Constitution, Policies, Inputs and Outputs	Interviewed sources	Uber used an algorithm to flag likely regulators based on where they opened the Uber app, credit card information, and other details; regulators were shown a different version of the app.
Facebook content removal policies ¹⁸¹	2017	The Guardian, ProPublica	How does Facebook filter content?	Constitution, Policies	Reviewed leaked internal documents.	Facebook has extensive guidance for human content reviewers.

Endnotes

- 1 See Executive Office of the President, National Science and Technology Council Committee on Technology, [Preparing for the Future of Artificial Intelligence](#), Oct. 2016; also see Executive Office of the President, [Big Data: A Report on Algorithmic Systems](#), Opportunity, and Civil Rights, May 2016.
- 2 Commons Select Committee, [“Algorithms in decision-making inquiry launched.”](#) UK Parliament, Feb. 28, 2017.
- 3 United Nations Office for Disarmament Affairs, [“Background on Lethal Autonomous Weapons Systems.”](#) last accessed on Jun. 26, 2017.
- 4 European Union, General Data Protection Regulation, Article 22.
- 5 Diakopoulos, et al., [“Principles for Accountable Algorithms and Social Impact Statement for Algorithms”](#). This group recently expanded to include related disciplines, now clustered under umbrella group FAT*.
- 6 Association for Computing Machinery, [“Statement on Algorithmic Transparency and Accountability.”](#) Jan. 12, 2017.
- 7 Partnership on AI, [“Tenets”](#).
- 8 [“ITI Unveils First Industry-Wide Artificial Intelligence Policy Principles.”](#) Information Technology Industry Council, Oct. 24, 2017.
- 9 Institute of Electrical and Electronics Engineers, [“The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Automated Systems: General Principles”](#).
- 10 Susumu Hirano [“Policy Issues Toward AI Networking and Guiding Principles for AI Development.”](#) Carnegie Endowment for International Peace.
- 11 See generally Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (2015).
- 12 The Royal Society, [Machine learning: the power and promise of computers that learn from example](#), April 2017; see also Jason Bell, *Machine Learning: Hands-On for Developers and Technical Professionals*, John Wiley & Sons (2014) (“A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.”).
- 13 Algorithms vary widely in their purposes and degrees of complexity, so much so that few useful generalizations can be made about them.
- 14 See, e.g., Cathy O’Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Crown Publishing Group/Penguin Random House (2016), 21 (“[A] key component of every model, whether formal or informal, is its definition of success.”).
- 15 Laura K. Donohue, *The Cost of Counterterrorism: Power, Politics, and Liberty* (Cambridge University Press, 2008), pp. 254–255.
- 16 See, e.g., Fundacion Karisma [“Big Data: un aporte para la discusión de la política pública en Colombia.”](#) Nov. 2, 2016.
- 17 See, e.g., Amy Traub, [Discredited: How Employment Credit Checks Keep Qualified Workers Out of a Job](#) (2012).
- 18 See Danielle Keats Citron, *Technological Due Process*, 85 Wash. U. L. Rev. 1249 (2008) for a seminal treatment of automated decisions in the context of US administrative law. Citron notes that government agencies use systems that “mix automation with human intervention.”
- 19 The term “automation” resists a formal definition. “Few words of recent years have been so twisted to suit a multitude of purposes and phobias as this new word, ‘automation,’” argued James R. Bright, a Harvard business professor in 1958. “Automation simply means something significantly more automatic than previously existed in that plant, industry, or location.” James R. Bright, *Automation and Management* (Cambridge, Mass.: Harvard University, 1958), 4, 6.
- 20 Linda J. Skitka, Kathleen L. Mosier, Mark Burdick, [“Does automation bias decision-making?”](#) Int. J. Human-Computer Studies (1999) 51, 991-1006; R. Parasuraman R and DH Manzey, [“Complacency and bias in human use of automation: an attentional integration.”](#) Hum Factors. 2010 Jun; 52(3): 381-410.
- 21 See, e.g., Henderson et al., [“Credit-score panacea failed to stop U.S. mortgage crisis.”](#) Reuters, May 10, 2007.
- 22 Berkeley Jay Dietvorst, [“Algorithm Aversion.”](#) Publicly Accessible Penn Dissertations, Paper 1686, Jan. 2016, (explaining that “people are especially averse to algorithmic forecasters after seeing them perform, even when they see them outperform a human forecaster. This is because people more quickly lose confidence in algorithmic than human forecasters after seeing them make the same mistake” and finding that “people are considerably more likely to choose to use an imperfect algorithm, and thus perform better, when they can modify its forecasts.”).
- 23 See, e.g., Leadership Conference on Civil and Human Rights, [“Advocates Applaud Google’s Ban on Payday Loan Advertisements.”](#) May 11, 2016.
- 24 David Robinson, [“Buyer Beware: A hard look at police ‘threat scores.’”](#) Jan. 14, 2016, (observing that for police, “information’s accuracy, tone, and context matter greatly in shaping whether it makes officers and members of the public more safe or less so.”).
- 25 Other branches include “unsupervised machine learning,” learning without labels and “reinforcement learning,” which focuses on learning from experience.
- 26 See, e.g., Tim Hartford, [“Big data: are we making a big mistake?”](#) Financial Times, March 28, 2014.
- 27 See, e.g., Solon Barocas and Andrew D. Selbst, *Big Data’s Disparate Impact*, 104 Cal. L. Rev. 671 (2016).
- 28 Traditional software code both describes and causes behavior when it runs. See David A. Patterson & John L. Hennessy, *Computer Organization and Design: The Hardware/Software Interface* (5th ed. 2014).
- 29 See generally, Harold Abelson et al., *Structure and Interpretation of Computer Programs*, Second Edition, MIT Press 1996.
- 30 Code often takes the form of a ‘Big Ball of Mud’: a “haphazardly structured, sprawling, sloppy, duct-tape and bailing wire, spaghetti code jungle.” Joseph Yoder, [“Big Ball of Mud: Is This The Best That Agile Can Do?”](#), Carnegie Mellon University Software Engineering Institute, May 2011.
- 31 Modern computing relies on a mind-boggling array of different procedures interacting with one another. See, e.g., Alex Gaynor, [What happens when ...](#), GitHub/alex, accessed May 14, 2017.
- 32 Brian Foote and Joseph Yoder, *Big Ball of Mud*. Fourth Conference on Patterns Languages of Programs (PLoP ‘97/EuroPLoP ‘97) Monticello, Illinois, September 1997.
- 33 See generally, Appendix.

- 34 Nicholas Diakopoulos, [Algorithmic Accountability: Journalistic investigation of computational power structures](#), 3 Digital Journalism 3 (2015).
- 35 [“Algorithm Tips: Find tips for stories on algorithms”](#).
- 36 Nick Hopkins, [“Revealed: Facebook’s internal rulebook on sex, terrorism and violence.”](#) The Guardian, May 21, 2017.
- 37 Mike Isaac, [“How Uber Deceives the Authorities Worldwide.”](#) The New York Times, Mar. 3, 2017.
- 38 Mike Isaac, [“Justice Department Expands Its Inquiry Into Uber’s Greyball Tool.”](#) The New York Times, Mar. 5, 2017.
- 39 Nicholas Diakopoulos, [“We need to know the algorithms the government uses to make important decisions about us.”](#) The Conversation, May 23, 2017.
- 40 Rob Kitchen, [Thinking critically about and researching algorithms](#), Information Communication & Society, Vol. 20, No. 1, 14-29, 2016.
- 41 Angèle Christin, [“The hidden story of how metrics are being used in courtrooms and newsrooms to make more decisions.”](#) Ethnography Matters, Jun. 20, 2016.
- 42 “Quantified Society: Examining the Consequences of Algorithmic Decision Making for Open Societies — Case Study: Online enrollment for public schools in the City of Buenos Aires, Argentina,” Via Libre Foundation (2015), *on file with authors*.
- 43 Eticas, Big Data at the Border, Section III.1.1, *on file with authors*.
- 44 Angèle Christin, [Algorithms in practice: Comparing web journalism and criminal justice](#), Big Data and Society (2017).
- 45 Explained in greater detail in the following section.
- 46 Interview with R3D.
- 47 [“The FTC Report on Google’s Business Practices”](#).
- 48 Chi Chi Wu, [Automated Injustice: How a Mechanized Dispute System Frustrates Consumers Seeking to Fix Errors in Their Credit Reports](#), National Consumer Law Center, January 2009.
- 49 Rebecca Wexler, [Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System](#), February 20, 2017, available at SSRN.
- 50 [Don Ray Drive-A-Way Co. v. Skinner](#), 785 F. Supp. 198 (D.D.C. 1992).
- 51 Latanya Sweeney, [Discrimination in Online Ad Delivery: Google ads, black names and white names, racial discrimination, and click advertising](#), 11 ACM Queue 3, Apr. 2, 2013.
- 52 Maayan Perel, Niva Elkin-Koren, [Black Box Tinkering: Beyond Disclosure in Algorithmic Enforcement](#), 69 Fla. L. Rev. 181 (2017).
- 53 Consumer Federation of America, [“New Research Shows That Most Major Auto Insurers Vary Prices Considerably Depending on Marital Status.”](#) Press Release, Jul. 27, 2015.
- 54 See, e.g., Julius Adebayo, Lalana Kagal, [Iterative Orthogonal Feature Projection for Diagnosing Bias in Black-Box Models](#), Nov. 15, 2016, Anupam Datta, Shayak Sen, Yair Zick, [Algorithmic Transparency via Quantitative Input Influence](#), 11 Studies in Big Data 71-94, May 10, 2017, Marco Tulio Ribeiro, [“LIME - Local Interpretable Model-Agnostic Explanations.”](#) Apr. 2, 2016.
- 55 See, e.g., Giuseppe De Nicola, Pasquale di Tommaso, Esposito Rosaria, Flammini Francesco, Marmo Pietro & Orazio Antonio, [A grey-box approach to the functional testing of complex automatic train protection systems](#), · EDCC’05 Proceedings of the 5th European conference on Dependable Computing, 305-317 (2005); Julius Adebayo, [“FairML: Auditing Black-Box Predictive Models.”](#) Mar. 9, 2017.
- 56 [The Federal Reserve Board. Report to the Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit](#), Aug. 2007.
- 57 [“Google and the Bundestagswahl: #datadonation.”](#) AlgorithmWatch. See also Tobias D. Krafft, Michael Gamer, Marcel Laessing and Katharina A. Zweig, [“Filterblase geplatzt? Kaum Raum für Personalisierung bei Google-Suchen zur Bundestagswahl 2017.”](#) Algorithm Watch, Sep. 8, 2017.
- 58 [“Daten”](#). See also Cornelius Puschmann, [“How significant is algorithmic personalization in searches for political parties and candidates?”](#) Algorithmed Public Spheres at the Hans-Bredow-Institut for Media Research at the University of Hamburg.
- 59 Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, [“Machine Bias.”](#) ProPublica, May 23, 2016.
- 60 Aaron Plasek, [On the Cruelty of Really Writing a History of Machine Learning](#), IEEE Annals of the History of Computing (2016).
- 61 Kristian Lum, William Isaac, [To predict and serve?](#) 13 Significance 14-19 (2016).
- 62 Elizabeth Good Christopherson, [“Confronting the Data Dilemma.”](#) Rita Allen Foundation, Jul. 25, 2013.
- 63 Researchers are developing new techniques to make more complex models easier to scrutinize through white-box testing. We discuss some of these methods in the following section.
- 64 Christian Sandvig, et al., [Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms](#), May 2015, (“This point is made clear by recent controversies surrounding the publicly disclosed portion of the Reddit ranking algorithm. Even with complete transparency about a particular part of the algorithm, expert programmers have been sharply and publicly divided about what exactly that part of the algorithm does.”).
- 65 Megan Geuss, [“A year of digging through code yields ‘smoking gun’ on VW, Fiat diesel cheats.”](#) Ars Technica, May 28, 2017.
- 66 See, e.g., Ariel J. Feldman, J. Alex Halderman, and Edward W. Felten, [“Security Analysis of the Diebold AccuVote-TS Voting Machine.”](#) Proc. 2007 USENIX/ACCURATE Electronic Voting Technology Workshop (EVT’07), Aug. 2007, at 2.
- 67 Joshua Brustein, [“The Ex-Cop at the Center of Controversy Over Crime Prediction Tech.”](#) Bloomberg (July 10, 2017).
- 68 See generally, Conscientious Classification, 130; Fiasal Kamiran & Toon Calders, [“Data preprocessing techniques for classification without discrimination.”](#) Knowl Inf Syst (2012) 33:1–33.
- 69 See generally, Conscientious Classification, 130.
- 70 Moritz Hardt, [“Equality of Opportunity in Machine Learning.”](#) Google Research Blog, Oct. 7, 2016.

- 71 Conscientious Classification, 124.
- 72 Paul Burley, [“Reducing Bias in Hiring Algorithms – How We Do It At Predictive Hire.”](#) Predictive Hire.
- 73 Michael Feldman, et al., [“Certifying and removing disparate impact.”](#) Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 259-268.
- 74 Moritz Hardt, Eric Price & Nathan Srebro, [“Equality of Opportunity in Supervised Learning.”](#) 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.
- 75 Mike Lee Williams, [“New Research on Interpretability.”](#) Fast Forward Labs Blog, Aug 2. 2017.
- 76 Eamonn Keogh & Abdullah Mueen, “Curse of Dimensionality,” Encyclopedia of Machine Learning (Springer, 2010).
- 77 See, e.g., Yin Lou, Rich Caruana, Johannes Gehrke, & Giles Hooker, [Accurate Intelligible Models with Pairwise Interactions](#); Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin, [“Why Should I Trust You” Explaining the Predictions of Any Classifier](#), KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Aug. 13-17, 2016, (explaining LIME — Local Interpretable Model-agnostic Explanations — a technique that explains the predictions of classifiers by learning an interpretable model locally around the prediction).
- 78 Osbert Bastani, Carolyn Kim & Hamsa Bastani, [“Interpretability via Model Extraction.”](#) 2017 Workshop on Fairness, Accountability, and Transparency in Machine Learning.
- 79 Patrick Hall, Wen Phan, and SriSatish Ambati, [“Ideas on interpreting machine learning.”](#) O'Reilly, Mar. 15, 2017.
- 80 See, e.g., Henrik Brink, Joshua Bloom, [Overcoming the Barriers to Production-Ready Machine Learning Workflows](#), STRATA Feb. 11, 2014; “[W]e show that the same methods scale to large datasets containing hundreds of thousands of patients and thousands of attributes while remaining intelligible and providing accuracy comparable to the best (unintelligible) machine learning methods,” in Rich Caruana et al., [Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission](#), KDD'15.
- 81 Rich Caruana, [“Intelligible Machine Learning for Critical Applications Such As Health Care.”](#) AAAS 2017 Annual Meeting, Feb. 20, 2017.
- 82 Trevor Hastie and Robert Tibshirani, *Generalized additive models* (John Wiley & Sons, Inc.: 1990).
- 83 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 at 634 (2017).
- 84 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 at 634 (2017).
- 85 [Trust, confidence and Verifiable Data Audit.](#)
- 86 See, e.g., Organisation for Economic Co-operation and Development, *Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*, 2013; Asia-Pacific Economic Cooperation, *Privacy Framework*, 2005; The White House, *Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy*, “Consumer Privacy Bill of Rights,” Feb. 2012.
- 87 The Equal Credit Opportunity Act prohibits lenders from using certain categories of data when making loan decisions. See 12 CFR 1002.4.
- 88 See, e.g., Chris Jay Hoofnagle, “The Potemkinism of Privacy Pragmatism,” Slate, Sept. 2, 2014, available at http://www.slate.com/articles/technology/future_tense/2014/09/data_use_regulation_the_libertarian_push_behind_a_new_take_on_privacy.2.html, (Arguing that “[b]ans on data collection are powerful tools to prevent institutions from using certain knowledge in their decision-making,” while “[t]here are deep problems with protecting privacy through regulating use”).
- 89 e.g., Chile, India
- 90 For instance, “smart cities” will collect enormous amounts of de-identified data about traffic, utility use, and other municipal activity; while this does not constitute “personal” data, it will nonetheless inform local government decisions that will affect people’s lives. For a general discussion of interpretations of anonymization, pseudonymization, re-identifiability and profiling as they relate to the legal concept of “personal data” in the EU, see Douwe Korff, [“Comparative Study on Different Approaches to New Privacy Challenges, in Particular in the Light of Technological Developments.”](#) Data protection laws in the EU: The difficulties in meeting the challenges posed by global social and technical developments Working Paper No. 2, European Commission Directorate-General Justice, Freedom and Security, pages 48-58.
- 91 See, e.g., Thibaud Antignac, Daniel Le M’etayer. [Privacy Architectures: Reasoning About Data Minimisation and Integrity](#). Damsgaard Jensen, Christian and Mauw, Sjouke. STM - 10th International Workshop on Security and Trust Management, Sep 2014, Wroclaw, France. Springer, 8743, 2014.
- 92 GDPR Article 5 (1)(c).
- 93 GDPR Article 6 allows processing of data when a subject has given consent, when it’s necessary to fulfill a contractual obligation, or when it “is necessary for the purposes of the legitimate interests pursued by the controller” as long as they are not overridden by the “interests or fundamental rights and freedoms of the data subject ...”
- 94 [The Amended Act on the Protection of Personal Information](#) (Tentative Translation), Ver.1 February 2016.
- 95 See, e.g., EFF and ACLU amicus brief in [Kayman v Obama](#).
- 96 [Opening Statement of General Keith B. Alexander, Director, NSA before the Senate Committee on the Judiciary](#), October 2, 2013.
- 97 Cody M. Poplin, [“NSA Ends Bulk Collection of Telephony Metadata under Section 215.”](#) Lawfare, Nov. 30, 2015.
- 98 111 countries have access to information laws. Access Info Europe, [“Statement by European RTI Community on the world’s First Official Access to Information Day!”](#) Sept. 28, 2016.
- 99 See, e.g., [the Open Government Partnership](#).
- 100 “Where appropriate, provide under the caption ‘Risk Factors’ a discussion of the most significant factors that make the offering speculative or risky.” Regulation S-K, 17 CFR 229 Item 503 (c).
- 101 See Ryan Budish, Liz Woolery and Kevin Bankston, [“The Transparency Reporting Toolkit: Survey & Best Practice Memos for Reporting on U.S. Government Requests for User Information.”](#) New America Open Technology Institute, Mar. 31, 2016.

- 102 See Tal Zarsky, *Transparent Predictions*, 4 Univ. of Ill. L. Rev. 1503 (“[T]he current framework does not provide a sufficient, balanced, and nuanced response to today’s challenges.”).
- 103 The law added the phrase “source code” to the text of the existing Access to Administrative Documents and Re-Use of Public Information law as records that count as administrative documents and are thus subject to disclosure. [More](#).
- 104 U.S. Federal Agency Data Mining Reporting Act. Nicholas Diakopolous notes that Edward Snowden’s documents contradicted the Office of the Director of National Intelligence’s disclosure claiming it did not engage in any form of data mining. Nicholas Diakopolous, [Algorithmic Accountability Reporting: On The Investigation of Black Boxes](#), Tow Center for Digital Journalism, Dec. 2013.
- 105 14 CFR 255.4 (3). The Department of Transportation enforced this law to varying degrees until 2004 at which point it appears all rules sunset.
- 106 Sandra Wachter, Brent Mittelstadt and Luciano Floridi, [“Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation.”](#) International Data Privacy Law, Forthcoming (2017). (“Since approval of the EU General Data Protection Regulation (GDPR) in 2016, it has been widely and repeatedly claimed that the GDPR will legally mandate a ‘right to explanation’ of all decisions made by automated or artificially intelligent algorithmic systems. This right to explanation is viewed as an ideal mechanism to enhance the accountability and transparency of automated decision-making.”).
- 107 Douwe Korff, [“Comparative Study on Different Approaches to New Privacy Challenges, in Particular in the Light of Technological Developments.”](#) Jan. 20, 2010; see also Douwe Korff, Data protection laws in the EU: [The difficulties in meeting the challenges posed by global social and technical developments](#) Working Paper No. 2, European Commission Directorate-General Justice, Freedom and Security, Jan. 20, 2010, page 86.
- 108 GDPR Article 14(2)(g).
- 109 Wachter et al, *supra* note 106.
- 110 See Korff, *supra* note 90.
- 111 [Central Board of Secondary Education & Anr. vs. Aditya Bandopadhyay & Ors](#), Supreme Court of India Civil Appeal No. 6454 of 2011, page 49.
- 112 Fair Credit Reporting Act 15 U.S.C. § 1681g(f)(1).
- 113 12 CFR § 222.72 2010.
- 114 Universal Declaration of Human Rights Article 7 states, “All are equal before the law and are entitled without any discrimination to equal protection of the law. All are entitled to equal protection against any discrimination in violation of this Declaration and against any incitement to such discrimination.”
- 115 Miranda Bogen, Artificial intelligence will force us to confront our values, Equal Future, Sept. 22, 2016 (“As appeals for algorithms and applications of AI to protect our societal values grow in number and volume ... we need to recognize this will be really hard to do when we as a society don’t agree what exactly we want to protect in the first place.”).
- 116 [“Equality Act 2010: Guidance.”](#) UK Government Equalities Office and Equality and Human Rights Commission.
- 117 Written evidence submitted by Dr. Dan Stowell, Dr. Emmanouil Benetos, Dr. Bob Sturm, Dr. Laurissa Tokarchuk, [Centre for Intelligent Sensing to the UK Parliament Science and Technology Committee Algorithms in decision-making inquiry](#) (ALG0036).
- 118 See, e.g., Title VII of the Civil Rights Act and the Equal Credit Opportunity Act.
- 119 “Equal Credit Opportunity Act (ECOA) prohibits discrimination in lending, and Regulation B, which applies ECOA to credit scoring systems. Regulation B sets forth specific data that cannot be used in a credit scoring system, such as: public assistance status, likelihood that any person will bear or rear children, telephone listing, income because of a prohibited basis, inaccurate credit histories, and different standards for married and unmarried persons, race, color, religion, national origin, and sex. 12 C.F.R. § 202.5 (2013).” Danielle Keats Citron & Frank Pasquale, [The Scored Society: Due Process For Automated Predictions](#), 89 Wa. L. Rev 1, 2014.
- 120 Fair Housing Act, 42 U.S.C 3601.
- 121 Michael Feldman, et al, [“Certifying and removing disparate impact.”](#) Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 259-268.
- 122 *Supra* note 61.
- 123 [“Concerning the instruments for the protection of individual rights, the Brazilian Constitution of 1988 granted ‘habeas data,’ an institute to ensure the access to relevant information concerning the petitioner, recorded or stored in the database of governmental institutions or of public nature and also to allow the correction of the information stored, when one does not prefer to correct it via an administrative or judicial case under seal. \(Federal Constitution, article 5, LXXII\). “Diffuse Constitutional Control.”](#) Federal Supreme Court of Brazil; http://www2.stf.jus.br/portalStfInternacional/cms/verConteudo.php?sigla=portalStfSobreCorte_en_us&idConteudo=123036. See generally Andrés Guadamuz, [“Habeas Data: The Latin-American Response to Data Protection.”](#) Journal of Information Law & Technology, 2000 (1); for a discussion of the Indian legal regime for access and redress see Vipul Kharbanda, [“Habeas Data in India.”](#) The Centre for Internet and Society, December 10, 2016.
- 124 European Union law (both the existing Data Privacy Directive and the GDPR) applies to the automated processing of all types of personal data. Similar to the FCRA, albeit across all contexts, European data privacy law requires “data controllers” to give subjects access to any personal data that is being processed, as well as related information such as the purpose of the processing. The law further orders that data subjects have the right to rectify or complete inaccurate personal data. GDPR Article 15(1), 16.
- 125 Fair Credit Reporting Act, 15 USC § 1681 et seq.
- 126 The US has three main consumer credit reporting agencies; European countries have one to three major agencies each. See Association of Consumer Credit Information Suppliers, [“ACCIS Full Members”](#).
- 127 See, e.g., Federal Trade Commission, [“Data Brokers: A Call for Transparency and Accountability.”](#) May 2014.
- 128 Tal Zarsky, *Transparent Predictions*, 4 Univ. of Ill. L. Rev. 1503 at 1524 (2013).
- 129 “... the consumer may, upon providing proper identification, request a free copy of a report and may dispute with the consumer reporting agency the accuracy or completeness of any information in a report.” Fair Credit Reporting Act, 15 USC § 1681b(b)(3)(B)(i)(IV).
- 130 TransUnion, [“One Third of Americans Have Never Checked Their Credit Report Reveals TransUnion Study.”](#) Oct. 29. 2013.

- 131 In some regimes individuals must actively exercise this right, while in others the law prohibits a priori certain fully automated decisions that are based on personal or sensitive data. Cf. Zuiderveen Borgesius, F. J., ["Improving privacy protection in the area of behavioural targeting."](#) UvA Digital Academic Repository (2014).
- 132 ["Proceed With Caution: Flexibilities in the General Data Protection Regulation."](#) EDRi, Jul. 4, 2016.
- 133 See, e.g., Lee Bygrave, ["Minding the machine: Art 15 of the EC Data Protection Directive and automated profiling."](#) Privacy Law and Policy Reporter 67 (2000); Laurens Naudts, ["The Right not to be Subject to Automated Decision-Making: The role of explicit consent."](#) University of Leuven (KU Leuven), Centre for IT and IP, August 2, 2016.
- 134 "Member States shall grant the right to every person not to be subject to a decision which produces legal effects concerning him or significantly affects him and which is based solely on automated processing of data intended to evaluate certain personal aspects relating to him, such as his performance at work, creditworthiness, reliability, conduct, etc." Directive 95/46/EC Article 15.
- 135 GDPR Article 7(3); 21.
- 136 Article 7 of this Framework notes that automated searches require safeguards from relevant governments. "A decision which produces an adverse legal effect for the data subject or significantly affects him and which is based solely on automated processing of data intended to evaluate certain personal aspects relating to the data subject shall be permitted only if authorised by a law which also lays down measures to safeguard the data subject's legitimate interests." Council Framework Decision 200819771JHA, art. 7, 2008 O.J. (L 350)(EC).
- 137 These are governed by a different set of data and EU rules. A recent EU framework addressing this issue is Council Framework Decision 2008/977/JHA.
- 138 Article 17(1) of the Personal Data Protection Act in Estonia.
- 139 Article 49(1) of the Datenschutzgesetz of Austria.
- 140 "No court decision involving the assessment of an individual's behaviour may be based on an automatic processing of personal data intended to assess some aspects of their personality." [Loi Informatique et Libertes Act N°78-17 OF 6 JANUARY 1978 on Information Technology, Data Files and Civil Liberties, Article 10.](#)
- 141 ["Country Views on Killer Robots."](#) Campaign to Stop Killer Robots, December 13, 2016.
- 142 Department of Defense Directive 3000.09 ["Autonomy in Weapon Systems."](#) November 21, 2012.
- 143 DO-178B, Software Considerations in Airborne Systems and Equipment Certification, "provide[s] guidelines for the production of software for airborne systems and equipment that performs its intended function with a level of confidence in safety that complies with airworthiness requirements"; [NASA Space Shuttle Flight Software Verification and Validation Requirements](#), November 21, 1991.
- 144 Bernard Marr, ["First FDA Approval For Clinical Cloud-Based Deep Learning In Healthcare."](#) Forbes, Jan. 20, 2017.
- 145 See, e.g., [Written evidence submitted by the Oxford Internet Institute to the UK Parliament Science and Technology Committee Algorithms in decision-making inquiry](#) (ALG0031).
- 146 See, e.g., [Written evidence submitted by the Information Commissioner's Office to the UK Parliament Science and Technology Committee Algorithms in decision-making inquiry](#) (ALG0036).
- 147 General Principles of Software Validation; [Final Guidance for Industry and FDA Staff](#), January 11, 2002; Transcript: Final Guidance on Medical Device Accessories: [Describing Accessories and Classification Pathway for New Accessory Types, FDA Webinar](#).
- 148 GDPR Article 35 and 36.
- 149 Article 19 Working Party, Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679, Adopted on 4 April 2017.
- 150 "To date, however, efforts to implement algorithm auditing have encountered legal resistance (Farivar, 2016). In the United States, a recent lawsuit brought about by researchers challenges the barriers that the Computer Fraud and Abuse Act (CFAA) has placed against third party 'testing for discrimination on the Internet' (ACLU, 2016). As the plaintiffs note, 'many common website terms of service prohibit . . . [activities] necessary for robust audit testing to uncover discrimination on the Internet', and the CFAA effectively makes such terms of service enforceable as criminal law. Thus one area where the GDPR could have a significant impact on algorithmic discrimination is in providing a legal mechanism for allowing, or even requiring, algorithm auditing." Bryce W. Goodman, ["A Step Towards Accountable Algorithms?: Algorithmic Discrimination and the European Union General Data Protection."](#) 29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.
- 151 [Automated Assistance in Administrative Decision-Making Better Practice Guide, Australian Government](#), February 2007.
- 152 See, e.g., [Jackpot Systems Minimum Technical Requirements \(Casinos\) version 1.0](#), Office of Liquor and Gaming Regulation, Queensland Government; Notice to Licensees: [Pre-Approval Inspection by the Technology Division](#), Nevada Gaming Control Board, April 4, 2017.
- 153 See, e.g., David J Lynch, ["Policing the Digital Cartels."](#) Financial Times, Jan. 8, 2017; Maurice E. Stucke and Ariel Ezrachi, ["How Pricing Bots Could Form Cartels and Make Things More Expensive."](#) Harvard Business Review, Oct. 27, 2017.
- 154 Foo Yun Chee, ["EU's Vestager warns companies against abusing algorithms."](#) Reuters, Mar. 16, 2017.
- 155 Brody Mullins, Rolfe Winkler and Brent Kendall, ["Inside the U.S. Antitrust Probe of Google."](#) The Wall Street Journal, Mar. 19, 2015.
- 156 [Antitrust: Commission probes allegations of antitrust violations by Google, European Commission Press Release Database](#), Nov. 30, 2010.
- 157 Ivana Kottasová, ["EU slaps Google with record \\$2.7 billion fine."](#) CNN, Jun. 27, 2017.
- 158 Ryan Calo, ["When a Robot Kills, Is It Murder or Product Liability?."](#) Slate, Apr. 26, 2016.
- 159 See, e.g., F. Patrick Hubbard, ["Sophisticated Robots": Balancing Liability, Regulation, and Innovation](#), 66 Fla. L. Rev. 1803 (2014), available at [_facpub](#); also see William D. Smart, Cindy M. Grimm, & Woodrow Hartzog, [An Education Theory of Fault for Autonomous Systems](#), (2017).
- 160 [Draft Report with recommendations to the Commission on Civil Law Rules on Robotics](#) (2015/2103(INL)), European Parliament, May 31, 2016.

- 161 [“Report to the Chairman, Subcommittee on Investigations and Oversight, Committee on Science, Space, and Technology, House of Representatives: Patriot Missile Defense - Software Problem Led to a System Failure at Dhahran, Saudi Arabia.”](#) United States General Accounting Office (Feb. 1992).
- 162 [“Report to the Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit.”](#) Board of Governors of the Federal Reserve System (Aug. 2007).
- 163 Jennifer Valentino-DeVries, Jeremy Singer-Vine, and Ashkan Soltani, [“Websites Vary Prices, Deals Based on Users’ Information.”](#) The Wall Street Journal, Dec. 24, 2012.
- 164 Daniel Trielli, Sean Mussenden, and Nicholas Diakopoulos, [“Why Google Search Results Favor Democrats.”](#) Slate (Dec. 7 2015).
- 165 Gregory J. Thompson et al., [“In-Use Emissions Testing of Light-Duty Diesel Vehicles in the United States.”](#) Center for Alternative Fuels, Engines & Emissions West Virginia University, May 15, 2014.
- 166 Latanya Sweeney, [“Discrimination in Online Ad Delivery.”](#) Communications of the ACM, Vol. 56 No. 5, Pages 44-54 (2013).
- 167 Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, [“Machine Bias.”](#) ProPublica, May 23, 2016.
- 168 Daniel Neyland, [“Bearing Account-able Witness to the Ethical Algorithmic System.”](#) Science, Technology, & Human Values Vol. 41(1) 50-76 (2016).
- 169 Lotus Ruan, Jeffrey Knockel, Jason Q. Ng, and Masashi Crete-Nishihata, [“One App, Two Systems: How WeChat uses one censorship policy in China and another internationally.”](#) The Citizen Lab, Nov. 30, 2016.
- 170 Daizhuo Chen, Samuel P. Fraiberger, and Foster Provost, [“Enhancing Transparency and Control when Drawing Data-Driven Inferences about Individuals.”](#) 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016).
- 171 See, e.g., Andrei Kirilenko, Albert S. Kyle, Mehrdad Samadi, and Tugkan Tuzun, [“The Flash Crash: The Impact of High Frequency Trading on an Electronic Market.”](#) US Commodity Futures Trading Commission, Oct. 1, 2014.
- 172 John C. Dearden et al., [“The Development and Validation of Expert Systems for Predicting Toxicity.”](#) ATLA 25, 223-252 (1997).
- 173 Kerry Rodden, [“Is that a boy or a girl? Exploring a neural network’s construction of gender.”](#) Medium (Feb. 24, 2017).
- 174 Nicholas Diakopoulos, [“Sex, Violence, and Autocomplete Algorithms.”](#) Slate (Aug. 2, 2013).
- 175 Gary Soeller, Karrie Karahalios, Christian Sandvig, and Christo Wilson, [“MapWatch: Detecting and Monitoring International Border Personalization on Online Maps.”](#) WWW 2016 (Apr. 2016).
- 176 [Don Ray Drive-A-Way Co. v. Skinner](#), 785 F. Supp. 198 (D.D.C. 1992).
- 177 Jędrzej Niklas, Karolina Sztandar-Sztanderska, and Katarzyna Szymielewicz, [“Profiling the unemployed in Poland: Social and political implications of algorithmic decision making.”](#) Fundacja Panoptykon (2015).
- 178 Keyon Vafa, Christian Haigh, Alvin Leung, and Noah Yonack, [“Price Discrimination in The Princeton Review’s Online SAT Tutoring Service.”](#) JOTS Technology Science, Sep. 1, 2015; Jeff Larson, Surya Mattu and Julia Angwin, [“Unintended Consequences of Geographic Targeting.”](#) ProPublica.
- 179 Ken Shirriff, [“How Hacker News ranking really works: scoring, controversy, and penalties.”](#) Ken Shirriff’s Blog (Nov. 2013).
- 180 Isaac, *supra* note 38.
- 181 Julia Angwin and Hannes Grassegger, [“Facebook’s Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children.”](#) ProPublica, Jun. 28, 2017.

Upturn

Upturn is a team of technology and policy experts based in Washington DC, working toward a world where technology serves the dignity and well-being of all people. We focus on the interests and needs of those at the margins, who are all too often missed when technologies are built, deployed, and governed.

Upturn is a 501(c)(3) non-profit organization.



OMIDYAR NETWORK

Omidyar Network is a philanthropic investment firm dedicated to harnessing the power of markets to create opportunity for people to improve their lives. Established in 2004 by eBay founder Pierre Omidyar and his wife Pam, the organization invests in and helps scale innovative organizations to catalyze economic and social change. Omidyar Network has committed more than \$1 billion to for-profit companies and nonprofit organizations that foster economic advancement and encourage individual participation across multiple initiatives, including Governance & Citizen Engagement, Education, Emerging Tech, Financial Inclusion, and Property Rights.

The Governance & Citizen Engagement Initiative at Omidyar Network works to create open, just, and inclusive societies. Societies where everyone can participate, where people have greater control over the decisions that impact their lives, and where transparency and accountability are the norms.

Our global work is focused on four interrelated areas – Data Governance, Follow the Money, Civic Technology, and Independent Media – where we provide support in the form of nonprofit grants and for-profit investment.

To date we have supported 200 organizations with over \$286M of funding.