CENTER FOR DEMOCRACY & TECHNOLOGY

Digital Decisions

INTRODUCTION

Algorithms play a central role in modern life, determining everything from search engine results and social media content to job and insurance eligibility. Unprecedented amounts of information fuel engines that help us make choices about even mundane things, like what restaurant to visit. Institutions use data to anticipate our interests, determine what opportunities we are afforded, and steer us through digital environments. There are countless automated processes behind each recommendation, prediction, and result — but how are these decisions made?

The designers behind data-driven analytics decide what data to collect and include and what criteria are relevant to the process. Although this sophisticated statistical analysis is a pillar of society in the 21st Century, the technical processes behind the decisions are not transparent to users or regulators. The question of how to regulate algorithmic decision-making and related data practices has implications for all aspects of our lives, including economic opportunity, well-being, and free speech.

CDT is working with stakeholders to develop guidance that ensures the rights of individuals, encourages innovation, and promotes incentives for responsible use of automated technology. Building on principles established by the civil rights community, CDT's guidance for ethical use of automated decision-making technology helps translate principle into action for private industry. We believe that responsible use of data could do more than mitigate unintentional discrimination–it could help address deep-seated cultural bias that contributes to systemic inequality intended to be addressed by civil rights law. This is an opportunity to demystify digital decision-making to create principles for advocates and dynamic and responsive criteria that can be implemented by industry.

I. PAST IS PROLOGUE

In the summer of 2014 Ben Bernanke, former chair of the U.S. Federal Reserve, was denied a mortgage. Since stepping down from his government post earlier in the year, Bernanke had been able to command a reported $250,000 for giving a single speech and had signed a book contract estimated to be in the seven figures. Yet when he and his wife sought a mortgage to refinance their house in the District of Columbia, a house whose value according to tax records was dwarfed by his likely income in the next couple of years, they were turned down. As the New York Times explained, "Ben Bernanke…is as safe a credit risk as one could imagine. But he just changed jobs a few months ago. And in the thoroughly automated world of mortgage finance, having recently changed jobs makes you a steeper credit risk." The numbers were crunched and the decision was made—Mr. Bernanke was denied.

Presumably, the former Fed chair found a banker willing to look more closely at his application and reconsider. But how do less well-resourced individuals fare when decisions about credit and other matters of economic consequence are largely arbitrated by technical systems outside of our control?

The modern world is full of computers crunching numbers on decades of people 'like us' and drawing conclusions about our fate. In a world powered by big data, past is prologue.

II. WHAT YOU NEED TO KNOW

Almost every sector of the economy has been transformed in some way by algorithms. Some of these changes are upgrades, benefiting society by predicting factual outcomes more accurately and efficiently, such as improved weather forecasts. Other algorithms empower tools, such as Internet search engines, that are indispensable in the information age. These advancements are not limited to traditionally computer-powered fields. Algorithms can help doctors read and prioritize X-rays[i], and they are better and faster than humans at detecting credit card fraud. Wall Street fortunes depend on who can write the best trade-executing algorithm. Songwriting algorithms can replicate the styles of legendary composers. Algorithms are everywhere and automation is here to stay.

However, automated decision-making systems powered by algorithms hold equally broad potential for harm. Some of the most crucial determinations affecting our livelihoods—such as whether a person is qualified for a job, is creditworthy, or is eligible for government benefits—are now partly or fully automated. In the worst case scenario, automated systems can deny eligibility without providing an explanation or an opportunity to challenge the decision or the reasoning behind it. This opacity can leave people feeling helpless and discourage them from participating in critical institutions.

Additionally, automated decision-making systems can have disproportionately negative impacts on minority groups by encoding and perpetuating societal biases. A 2014 big data report commissioned by the White House concluded that "big data analytics have the potential to eclipse longstanding civil rights protections in how personal information is used in housing, credit, employment, health, education, and the marketplace." If data miners are not careful, sorting individuals by algorithm might create disproportionately adverse results concentrated within historically disadvantaged groups. Current laws governing fair credit, equal opportunity, and anti-discrimination may not be adequate to address newer ways of ranking and scoring individuals across a range of contexts. For example, the concept and practicality of redress may be meaningless if an individual does not even know she is being assessed—much less what the criteria are.

There is an ongoing conversation among experts in many different fields who are working to develop techniques and principles to mitigate these risks. But the complexity of the technology and the diversity of contexts in which it is used add up to a very complicated problem. This piece breaks that problem into digestible parts to allow readers to get acquainted with the issue at their own speed. It concludes

with a proposed solution–an interactive tool that prompts data scientists or programmers with questions designed to reveal and mitigate biased design structures.

III. QUICK DEFINITIONS OF KEY CONCEPTS

Algorithms are essentially mathematical equations. However, unlike mathematical equations you may be familiar with from primary school, algorithmic outputs do not necessarily represent a 'right answer,' defined by an objective truth. Imperfect data sets and human value judgements shape automated decisions in intentional and unintentional ways. To understand how this works, it is important to understand some of the basics of algorithms, machine learning, and automation.

Wait, what is an algorithm?

In its most basic form, an algorithm is a set of step-by-step instructions—a recipe—"that leads its user to a particular answer or output based on the information at hand." Applying its recipe, an algorithm can calculate a prediction, a characterization, or an inferred attribute, which can then be used as the basis for a decision. This basic concept can be deployed with varying degrees of sophistication, powered by the huge amounts of data and computing power available in the modern world. Algorithms take large amounts of information and categorize it based on whatever criteria the author has chosen.

What is machine learning?

Computers are able to process very complex algorithms and very large inputs in microseconds, producing what can be opaque and often significant algorithmic decisions. Some algorithms use a process called machine learning. Machine-learning algorithms identify patterns in existing data and use those patterns as rules for analyzing new information. Machine learning can amplify trends that might have been unseen by the researchers. In these cases, researchers "train" the computer using existing datasets, or training data, before setting out to predict results from new data. The process of gleaning insight from unwieldy amounts of information is called data mining. In his presentation to the FTC, Solon Barocas explained that the purpose of data mining is, "to provide a rational basis upon which to distinguish between individuals and to reliably confer to the individual the qualities possessed by those who seem statistically similar."

Some algorithms are designed to predict a future outcome. Designing a predictive algorithm involves coming up with a definition of success and choosing target variables that will bring it about. For example, when designing an algorithm that sifts through job applications to recommend hires, success could mean saving money, hiring a diverse group of employees, or any number of other metrics. The definition of success determines the target variables—the thing the algorithm will actually try to predict. If success means saving money, and employee turnover costs money, then a good hire may be defined as one who is likely to stay at the company for a long time (so the target variable would be longevity).

Target variables get further broken down in a process called feature selection. This is where programmers decide what specific criteria they will prioritize to sort, score, or rank cases. For example, the qualities that determine whether an employee will stay at a company long-term may include the amount of time a person stayed in his or her previous job.

IV. HOW CAN ALGORITHMS BE BIASED?

Rather than unbiased alternatives to human subjectivity, algorithms are imbued with the values of those who create them. Each step in the automated decision-making process creates possibilities for a final result that has a disproportionately adverse impact on protected or vulnerable classes. These steps include designing, building, testing, and refining the model. While the design stage presents multiple opportunities to introduce bias, it also presents opportunities to prevent unintended bias and ensure fairness.

Returning to our example of automating hiring–Suppose the training data shows that employees who live closer to work tend to stay in their jobs for longer than employees who live farther away. A model relying on these patterns could disproportionately reject, and systematically disadvantage, people who live in rural areas. (In fact, analytics company Evolv declined to include this data in their hiring model because of concerns it would have a discriminatory impact.)

Likewise, the selection of the training data can lead to disparate impacts. Machine-learning algorithms identify trends based on statistical correlations in the training data. However, algorithms can only predict the future based on the past—or more specifically on whatever data about past events is on hand. Because of this, the results can unintentionally be discriminatory or exacerbate inequality. For example, if training data for an employment eligibility algorithm consists only of all past hires for a company, no matter how the target variable is defined, the algorithms may reproduce past prejudice, defeating efforts to diversify by race, gender, educational background, skills, or other characteristics.

The characteristics of the data itself can skew predictive models. Data that was collected under biased conditions—or for a purpose unrelated to the goal of the algorithm—may not accurately represent the relevant population. For example, the city of Boston released a smartphone app called StreetBump, which used the app users' GPS data to automatically report potholes to the city, so that they could be repaired. However, since the city was only collecting this pothole data from smartphones, it was under-representative of lower-income neighborhoods, where fewer people own smartphones. The data collection method was unintentionally perpetuating unequal distribution of public services.

Machine-learning algorithms adjust based on new feedback, eventually making decisions based on criteria that may not have been explicitly chosen by human programmers. Absent careful auditing, this process can mask unintended bias. Hard evidence of algorithmic discrimination is somewhat difficult to come by. Discovering algorithmic bias often requires a fortuitous revelation such as Dr. Latanya

Sweeney's accidental discovery that searching her own name prompted ads for an arrest record to be served while searching traditionally white-sounding names did not—or time-consuming forensics.

Many of these issues are not new–discrimination and bias are not modern inventions. However, the scope and scale of automated technology makes the impact of a biased process faster, wider spread, and harder to catch or eliminate. That is why there is an emerging field specifically dedicated to analyzing the role technology is playing in perpetuating or amplifying historically advantaged populations or points of view.

V. HOW CAN DIGITAL DECISIONS BE HARMFUL?

Digital decisions have the potential to increase efficiency, detect and mitigate human bias, and help us better understand our world. But they also have the potential to undermine equality, choice, and other democratic values. Because algorithms work behind the scenes, it can be hard to detect when they make harmful decisions.

Digital decisions may be harmful because (1) they are unreliable, (2) they are low-quality, (3) they cause disparate impacts among different groups or populations, (4) they are unaccountable, or some combination thereof.

Unreliable decisions: violating expectations and betraying trust

Algorithms often make headlines for being better (more accurate, less biased, or more efficient) than humans at making a decision or performing a task. But as with human decision making, there are limits on how much we can or should trust automated decisions. Some decision-making tools overstate or obfuscate their usefulness or accuracy, inducing more trust than they deserve.

Automated decision-making tools can send explicit and implicit signals to users about their reliability. Some tools make explicit promises about their functions and accuracy. For example, Google explains how its algorithms work "to deliver the best results possible," Trivago claims that its "unbiased" searches find "the ideal hotel at the lowest rate," and some weather forecasts come with a confidence level to indicate the strength of the prediction. Implicit signals also tell users how much they should trust a digital decision. For example, users may be more likely to trust a recidivism scoring tool that uses a numerical scale to rate defendants' risk and is marketed for use by judges in sentencing hearings than they are to trust an online quiz called "how likely are you to end up in jail?" even though both may have the same research behind them and deliver similarly accurate results. When a decision-making system's usefulness, accuracy, or underlying logic is misaligned with users' expectations, it can induce misplaced trust resulting in harm.

In his essay, "The Code I'm Still Ashamed Of," developer Bill Sourour wrote about a tool he helped build for a drug company—a health quiz targeting teenage girls. It asked them a series of questions and

then recommended a drug, seemingly based on their answers. But no matter how users answered the question, the tool was designed to always recommend one particular drug. The tool's user interface, the quiz, implied that it was analyzing users' answers, plus some sort of medical knowledge, to make a recommendation. In fact, the underlying algorithm had only one rule: no matter what, recommend the company's drug. The quiz was intentionally deceptive about the nature and usefulness of its decisions. It did not provide the information users needed to assess its value. In other words, it was unreliable.

Reliability is not synonymous with accuracy. We don't always demand or expect a high level of accuracy from automated systems. We don't need perfectly targeted ads or accurate online personality quizzes. When popular medical websites match our symptoms with a long list of diseases, we know we probably don't have most of the conditions on that list. What matters is that automated systems give users enough information to assess how much trust they should place in its digital decisions.

Low-quality decisions

Sometimes, it's not enough for digital decisions to be reliable. Some of the more high-stakes decisions that algorithms make must also be high-quality. High-quality decision-making systems are based on representative and appropriate data and sound logic; are well-suited to the task they are designed to perform; and have high accuracy levels.

The burden of creating high-quality algorithms varies according to the potential consequences of a digital decision. For example, an advertising algorithm that performs poorly—perhaps one that targets tampon ads to males—may not cause as much harm as a flawed hiring algorithm or pre-trial recidivism risk score, which can deprive people of life-changing opportunities. In situations where critical opportunities or rights are at stake, low-quality decisions can cause substantial harm. When companies and agencies use algorithms to administer economic opportunities, access to social services, and criminal justice, they bear a higher burden of ensuring that the algorithms make high-quality decisions.

An automated decision-making system may be low-quality because it uses flawed data. A good data set is the essential foundation for high-quality predictions. Data sets that are too small, incomplete, biased, or not suited for the algorithm's intended purpose can create low-quality predictions. Machine-learning algorithms create general rules about how to classify things or people based on the specific examples in the training data. Thus, the quality and representativeness of the data itself can determine the quality of the decisions.

Some algorithms are low-quality because they attempt to predict phenomena or qualities that are very difficult to statistically predict, either because they are influenced by too many factors or because they are rare. For example, algorithms that attempt to predict violent crime have been found to have very low accuracy rates. Because violent crime is a relatively rare event, the probability that any individual will commit a violent crime is very low, and the amount of training data from which algorithms can learn about violent crime represents a tiny percentage of the population. Predicting rare events, such as murders or terrorist attacks, may be tempting, because the ability to prevent such events would benefit

society. However, this endeavor is so fraught with error and potentially meaningless statistical relationships that the hypothetical benefits may not be worth the downsides.

Disparate impacts

Algorithms don't have to be objectively inaccurate to cause harm. In fact, some of the most harmful algorithms are those that rely on statistically significant ways to sort people—often unintentionally—so that historically marginalized groups are denied opportunities or adversely targeted at higher rates.

Disparate impacts can results from algorithms that rely on the past to predict the future. One example of this is hiring algorithms that use data about historically successful job candidates to identify future successful candidates. On the surface, replicating the qualities of past successes, or avoiding hallmarks of failures, seems like a reasonable strategy for a data-driven process. But the patterns in existing data often reflect a history of discrimination and implicit bias.

For example, women and people of color have long been victims of disproportionate employment discrimination and have been historically underrepresented in the workforce. Existing data about successful job candidates, and algorithms designed to recreate that success, are likely to disproportionately favor white males, who enjoy greater workforce representation, particularly in leadership roles.

Eliminating sensitive characteristics, such as race and gender, from an algorithm's equation does not solve the problem of disparate impacts. Predictive models can rely on variables or features that strongly correlate with race, gender, sexual orientation, age, or income. These are known as "proxies." For example, at least one recidivism risk assessment tool asks whether arrestees have a parent who has been imprisoned. However, this factor is also correlated with race, since African-Americans are arrested and imprisoned at higher rates than whites. Since African-Americans are statistically more likely to answer "yes" to this question, an algorithm relying on it could disproportionately rate African-American arrestees as more likely to commit future crimes.

Algorithm designers may unintentionally discriminate when they fail to properly translate their objectives into code. For example, imagine an algorithm for sorting cucumbers to determine which ones go to market and which ones get discarded. Two people with the same objective can come up with completely different formulae because of the values they incorporate into the algorithm and the features they prioritize. For example, the owner of a grocery store chain might want to sort cucumbers in a way that results in the highest sales. If smooth, brightly colored, shiny looking cucumbers sell better than ones with exterior flaws, the store owner's algorithm will prioritize aesthetically pleasing cucumbers. However, an environmental or food sustainability expert might design the algorithm to create as little food waste and feed as many people as possible. This algorithm might only discard cucumbers that are inedible because of mold, insect infestations, or rot. These variations might be OK when it comes to sorting cucumbers, but in a more high-stakes context, such as allocating emergency relief after a natural disaster, the values embedded into an algorithm can have life-or-death impacts.

Discriminatory algorithms are particularly harmful when they arbitrate access to resources, such as credit, insurance, employment, and government programs, and when they determine a person's freedoms. Algorithms with systematic disparate impacts could violate anti-discrimination laws, such as the Fair Housing Act or the Equal Employment Opportunity Act. But technical compliance with the law may not be enough. Those who design and use algorithms to make decisions about others have an ethical obligation to avoid perpetuating harmful biases or marginalizing communities.

Unaccountable decisions

Automated decision-making systems are often referred to as "black boxes," because their existence and logic are unknown to the people whose lives they affect. Opaque decision-making systems can serve predatory loan advertisements or make sentencing decisions without the target ever knowing precisely why. We often don't even know an algorithm was used to make a decision about us. This makes digital decisions difficult to review or challenge—there is often no way to inform an institution or seek redress when we suspect the algorithm got it wrong.

For example, in her book, Weapons of Math Destruction, Cathy O'Neil discusses "e-credit scores," which is a score that credit companies assign to visitors of their websites that is not an actual credit score but is inferred based on things like browsing information. Companies use this "e-credit score" to determine whether to show a user high-interest or low-interest credit card ads. Users who visit a credit card website don't know they're being scored or the criteria or formula behind the score, yet these scores determine their credit opportunities.

Lack of redress for unfair or incorrect automated decisions can lead to frustration, loss of control, and distrust of institutions. The most vulnerable among us may be the most likely to give up on opaque automated decisions that don't work for them and to avoid decision-making institutions. Unreliable or unfair decisions that go unchallenged can contribute to bad feedback loops, which can make algorithms even more likely to marginalize vulnerable populations. When automated systems do not allow for user feedback, they create blind spots that prevent them from learning from their own bad decisions.

VI. SOLUTIONS PART 1: PRINCIPLES

While the problem and harms of algorithmic unfairness has become clearer with more research and attention, solutions are still an emerging field. Academics, journalists, advocates, technologists, industry groups, and even the Obama Administration have raised concerns about the potential harms of unchecked, biased automation. Many of those same groups have also undertaken efforts to identify and rally support behind a set of principles that describe shared values in response to the problem.

This type of groundwork is critical not only because it provides common philosophical ground, but also because it creates a framework for shifting perceptions of technology. Current cultural perceptions of automated decision making technology are out of step with the technical reality. Many present-day

values and beliefs about technology implicitly or explicitly endorse public trust in data-driven decisionmaking as objective or more fair than human decision making. For example, the belief that data captures a neutral or objective view on reality is widespread, and evoked often to explain otherwise improper conclusions like gender playing a role in creditworthiness. And of course data often does capture facts and provide an objective (or at least measurable) insight, but that does not happen by default.

The public interest would be better served by a widespread understanding that data must be used responsibly in order to ensure equitable outcomes. Disrupting beliefs like 'data is objective' requires replacing misconceptions about technology with principles that are ethically and technically sound. The most effective statements recognize the technical pitfalls of relying on big data and automation while drawing from established civil rights and ethical principles to frame their response.

The obscurity of automation reduces fairness, accountability, and transparency. Principles meant to solve these problems need to provide an alternative value structure that is responsive to the nature of current technology. Answering this problem is a work in progress, but common themes have emerged among current proposals. This is true even though principles are written by different communities, with different points of view on the world and the problem. A few core ideas are often repeated, including fairness, explainability, auditability, and accuracy.

Fairness

Fairness, and adjacent concepts like ethics and responsibility, are commonly included among the principles. Including fairness itself as a principle is a direct way to address a core concern around data-driven decisionmaking — that the decisions are unfair. However, conversations around fairness can quickly become fraught as each participant attempts to define fairness based on their particular background and values. While it's difficult to pinpoint what determines whether an algorithm is "fair," the problem of bias in automation requires us to demand fairness based not only on the process, but also on the results.

There are many definitions of fairness in practice. For many technology policy professionals, discussions about fairness are perhaps most familiar in the context of the Fair Information Practice Principles, which is an information governance framework that strives to level the playing field between entities collecting and using personal data and the individuals the data describes. And, while the guidelines define fairness with other, more specific concepts like notice and choice, these principles fundamentally support the idea that all individuals have a right to fair treatment in a process. On the other hand, designers and engineers may be accustomed to seeing the word in the context of statistics where it has a relatively narrow meaning (i.e. a fair die is one that has the same probability of landing on each of its sides). And private companies often argue that fairness is achieved when institutions provide proper and sufficient information to customers so that they can make an informed choice. As a result of these different perspectives, conversations about fairness can sometimes feel circular and unproductive.

In the case of algorithms, we must acknowledge and draw from civil rights traditions to develop an adequate definition of fair. Decisionmaking based on data and algorithms puts those who are already vulnerable or in the statistical minority at greater risk of harm, and so fairness as a principle in this context must mean something greater than equal treatment of individuals. Some civil rights law looks to disparate impact analysis and community-based outcomes to evaluate whether a policy or process is fair.

A group of civil rights advocates created the "Civil Rights Principles for the Era of Big Data" as a response to threats posed to civil rights and equality by automated decision making. The group specifically demanded fairness in automated decisions for all communities, arguing that "computerized decisionmaking … must be judged by its impact on real people, must operate fairly for all communities, and in particular must protect the interests of those that are disadvantaged or that have historically been the subject of discrimination." This principle reflects the reality that data-driven decisionmaking and automation is not equally harmful to all groups or communities.

A fair system or design is demonstrably not biased against individuals or groups of people, but fair outcomes are difficult to prove. Validating that the results of a given equation are fair requires knowing some information about the process and a representative sample of the outcomes it produces.

Explanation

The obscurity of decisionmaking is a longstanding policy issue in the U.S., which is why there are laws requiring explanations; for example, in instances of adverse credit decisions. But technology makes it harder to understand the connections between inputs and outcomes. Developers can't always explain the reasons why their technology gave a particular result, especially when they utilize sophisticated forms of algorithmic decision making that rely on machine learning or neural networks. This is problematic for any principle that proposes a right to explanation for how an automated decision was made about a person — sometimes the question just cannot be answered, and sometimes the answer would be incredibly burdensome to calculate and would require technical expertise on the part of the individual to understand.

Many commercial algorithms are considered proprietary, which also complicates the issue of explanation. While some have proposed algorithmic transparency as a solution to unfairness and explainability, this is not a viable solution. Revealing technical details about how the technology works, even in the service of explaining it to the public, could run the risk of exposing profitable details to competitors. Private companies have a huge financial incentive to resist sharing details that might compromise their products. (It is also not clear what benefit the public would get from access to raw data or equations, which are too obscure for most individuals to interpret and which don't always contain enough information to detect bias when considered out of context.)

While technical explanations can be challenging to provide, the concept of explanation is not limited to a description of mechanics. Many principles statements that include this value are focused on a user understanding the rationale for a decision rather than having access to technical details. For example, the U.S. Association of Computing Machinery asks that "institutions that use algorithmic decision-making …produce explanations regarding both the procedures followed by the algorithm and the specific decisions that are made." Far from a call to provide in-depth technical details, this principle is focused on increasing transparency around two things: the process and the results. This is mirrored in the principles authored primarily by academics. Their statement asks that institutions "ensure that algorithmic decisions as well as any data driving those decisions can be explained to end-users and other stakeholders in non-technical terms." In both frameworks, an institution could fulfill an expectation of explainability with a general sense of what factors are considered in a decision and, if applicable, what qualities of the individual affected were included in the decision.

Companies have found ways to share substantial amounts of insight into the logic of what they are doing with data without compromising the efficacy of their algorithm. This extends to informing people (in plain language) when the data collected or inferred about them will be used as the basis for automated decisions. For example, Fair, Isaac and Company (FICO) provides a general sense of the factors that feed into their credit scoring algorithm as well as the weights given to them. This helps individuals understand how to prove themselves within the system and allows for a cultural dialogue about what it means to be creditworthy as determined by automated systems. Similarly, Google has shared some details about how their PageRank algorithm weights various factors in determining search results. (It's worth noting that Google has additional protection because this technology is patented, rather than protected by trade secrets. Companies might consider the ability to offer increased transparency as one of the benefits of other forms of IP protection.) The decision to share information about how decisions are made can provide some protection from accusations of unfairness since the equation can be questioned and discussed by the public and policymakers.

Auditability

Audits are one method to provide explanations and redress without compromising the intellectual property behind the business model. Designing algorithmic systems that can be easily audited increases accountability and provides a framework to standardize best practices across industries. While explanations can help individuals understand algorithmic decision making, audits are necessary for systemic and long-term detection of unfair outcomes. They also make it possible to fix problems when they arise.

The details of an audit will depend on the nature of the technology and the sophistication of the company. Auditing an algorithm requires deliberate planning and complex record keeping. A decision made at any given moment by an automated system might be unique and difficult to replicate due to the dynamic nature of technology that learns and changes as its running. But being able to identify the source of a problem is necessary to resolving it, as well as providing explanations to the individuals affected. An effective audit requires institutions to maintain internal documentation of the logic or

circumstance behind significant design choices and procedures governing who is responsible for making changes. These systems are best installed as a product develops, rather than retroactively.

Creating a system that can be audited creates accountability and credibility, particularly if the result of an audit can be reviewed externally. The U.S. ACM proposes that creators of automated decisionmaking have a responsibility to record their "models, algorithms, data, and decisions …so that they can be audited in cases where harm is suspected." Some have taken the call for audits further and proposed that systems be accessible by third-parties that can "probe, understand, and review the behavior of the algorithm through disclosure of information that enables monitoring, checking, or criticism." Anticipating and complying with the needs of a large-scale, external audit can help companies detect and mitigate discriminatory impacts of their technology.

Another Proposal: Reliability

The principles above are a step toward fair, accountable, and transparent algorithms. They share a common goal of encouraging designers and users of data-driven systems to commit to deeper investigation into the nature of their data and to deliberate decisionmaking throughout the design process. These principles, and others proposed by stakeholder efforts, work to hold institutions accountable to a standard of demonstrably fair decisionmaking. But the diversity of contexts and consequences of algorithmic decision-making poses a challenge for solutions based only on principles: what happens in the event that we cannot explain, audit, or otherwise prove that technology is fair?

Some argue this renders the technology unusable because the risks are too high. But the expectation that we should only use explainable technology limits its full potential to improve our lives. The complexity of technologies like machine learning, neural networks, and some types of artificial intelligence can make outcomes inscrutable, even to those who are familiar with the technical details. But this does not diminish their importance or usefulness. Instead, advocates and policymakers could shift their expectations from demanding a fully understood technological world, to one that ensures technology is reliable.

Reliability: A system must be able to be trusted to behave as is expected and anticipated by its designers. The ways in which it deviates from these expectations can be monitored and addressed.

Rather than requiring designers to understand the specific reason for an outcome, reliable systems can be trusted to deliver consistent results. Ideally, the results fall within a range of anticipated outcomes the designers have already identified. Expecting that automated systems behave reliably is a flexible method of creating accountability without limiting innovation. Additionally, the description and thought necessary to create reliable systems gives individuals a sense of confidence that they understand the technology they are using.

Designers need to have clear expectations for their data-driven process, including logic that describes the outcomes, and to be able to test if it is meeting them. Holding designers to a "reliability standard"

would force institutions to examine their own technology and establish the expected outcomes, identify the results they can explain, acknowledge the results they can't explain, and monitor the degree of deviation in their system over time. Keeping these metrics provides a basic auditing mechanism by flagging outcomes outside of the expected range for further examination. It also formalizes an otherwise haphazard process, creating a record of decisions by the people involved in creating the system. Once people are asking themselves the right questions, the technology will reflect a higher degree of mindfulness of the risks data-driven technology poses to fair outcomes.

The diversity and scope of automated systems lends itself well to this kind of accountability structure. In fact, similar standards are used to regulate the broad world of technology in other contexts. The Federal Trade Commission's (FTC) Section 5 authority, which empowers the independent agency to regulate an incredibly broad scope of products, is largely premised on ensuring that companies adhere to their own promises. This regulatory structure has plenty of shortcomings. Failing to articulate clear expectations can be a burden for companies and leave consumers exposed to unfair products. But it has incentivized documentation around practices that might otherwise be entirely hidden to the public. Holding algorithms to a reliability standard would encourage higher levels of internal scrutiny and documentation, creating at least some formal structures while advocates, academics, policymakers and journalists develop clear and achievable standards for fairness in algorithms. A flexible standard is a good way to inform the public about the quality of an algorithmic system without chilling innovation.

What's next?

A framework of principles is the first step in developing actionable solutions for the problem of biased automation, but it is far from the last. This foundation, based on fairness, explainability, auditability, and reliability, provides an alternative understanding of data-driven decisions, not as objective measures, but as tools that must be carefully calibrated in order to avoid biased outcomes.

These principles advance important considerations for developers of algorithmic systems, but they are not obviously actionable. To transform these values into practical steps, they have to be aligned with the technology creation process, be accessible to designers who don't necessarily have an intuitive understanding of the stakes, and be responsive to business motivations and innovation.

VII. EMBEDDING PRINCIPLES: FROM IDEA TO DECISION

Principles established by academics, advocates, and policymakers are meant to demonstrate a philosophy that should be embedded throughout automated systems. This puts the burden on designers to understand how to integrate the goals of the principles into the technology itself. While the connection between the harm, the principle, and the remedy may seem obvious to people with the benefit of hindsight, it is hard to anticipate the kind of unfair outcomes that have been documented. Additionally, it is difficult to be specific in providing advice to technology companies who rely on trade

secrets to preserve their economic advantage. The goals of the principles must be connected directly to the process of creating and deploying an algorithm.

RESOURCES:

Link to the CDT Digital Decisions webpage: https://cdt.org/issue/privacy-data/digital-decisions/

Link to the Digital Decisions interactive tool: https://cdt.info/ddtool/